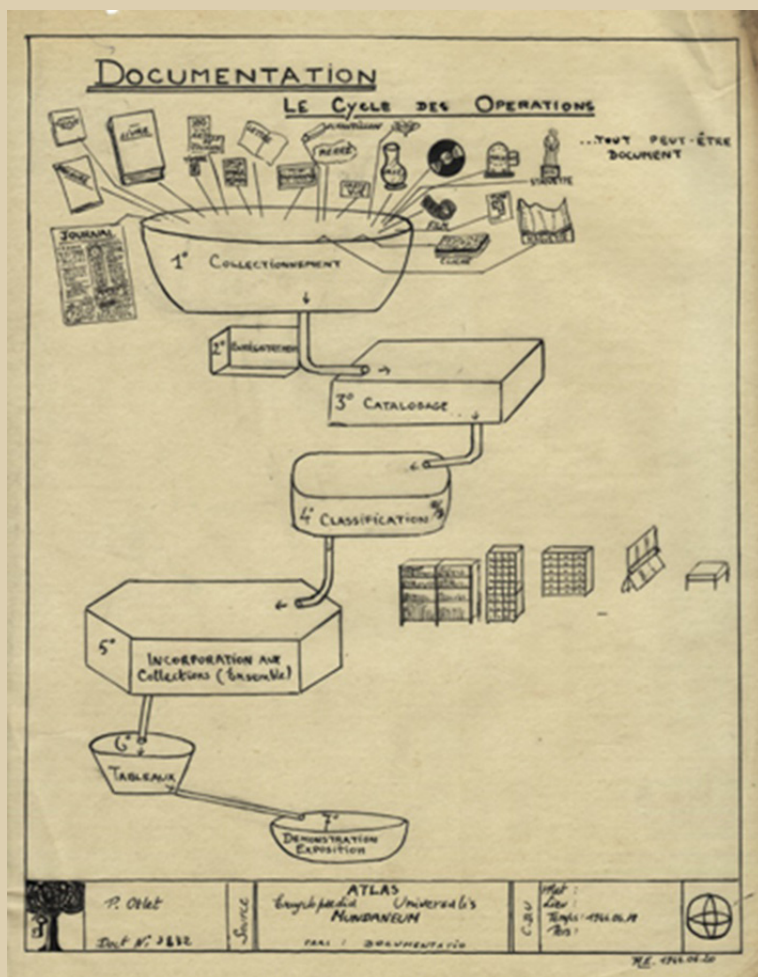


AIDainformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

NUMERO 3-
ANNO 43
LUGLIO-DICEMBRE 2025



AIDAinformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

Fondata nel 1983 da Paolo Bisogno

Proprietario della rivista:

Università della Calabria

Direttore Scientifico:

Roberto Guarasci, *Università della Calabria*

Direttore Responsabile:

Fabrizia Flavia Sernia

Comitato scientifico:

Anna Rovella, *Università della Calabria*;

Maria Guercio, *Sapienza Università di Roma*;

Giovanni Adamo, *Consiglio Nazionale delle Ricerche* †;

Claudio Gnoli, *Università degli Studi di Pavia*;

Ferruccio Diozzi, *Centro Italiano Ricerche Aerospaziali*;

Gino Roncaglia, *Università della Toscana*;

Laurence Favier, *Université Charles-de-Gaulle Lille 3*;

Madjid Ihadjadene, *Université Vincennes-Saint-Denis Paris 8*;

Maria Mirabelli, *Università della Calabria*;

Agustín Vivas Moreno, *Universidad de Extremadura*;

Douglas Tudhope, *University of South Wales*;

Christian Galinski, *International Information Centre for Terminology*;

Béatrice Daille, *Université de Nantes*;

Alexander Murzaku, *College of Saint Elizabeth, USA*;

Federico Valacchi, *Università di Macerata*.

Comitato di redazione:

Antonietta Folino, *Università della Calabria*;

Erika Pasceri, *Università della Calabria*;

Maria Taverniti, *Consiglio Nazionale delle Ricerche*;

Maria Teresa Chiaravallotti, *Consiglio Nazionale delle Ricerche*;

Assunta Caruso, *Università della Calabria*;

Claudia Lanza, *Università della Calabria*.

Segreteria di Redazione:

Valeria Rovella, *Università della Calabria*

Editrice: Cacucci Editore S.a.s.

Via D. Nicolai, 39 – 70122 Bari (BA)

www.cacuccieditore.it

e-mail: riviste@cacuccieditore.it

Telefono 080/5214220

AIDAinformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

«AIDAinformazioni» è una rivista scientifica che pubblica articoli inerenti alle Scienze dell'Informazione, alla Documentazione, all'Archivistica, alla Gestione Documentale e all'Organizzazione della Conoscenza ma amplia i suoi confini in ulteriori campi di ricerca affini quali la Terminologia, la Linguistica Computazionale, la Statistica Testuale, ecc. È stata fondata nel 1983 quale rivista ufficiale dell'Associazione Italiana di Documentazione Avanzata e nel febbraio 2014 è stata acquisita dal Laboratorio di Documentazione dell'Università della Calabria. La rivista si propone di promuovere studi interdisciplinari oltre che la cooperazione e il dialogo tra profili professionali aventi competenze diverse, ma interdipendenti. I contributi pubblicati affrontano questioni teoriche, metodologie adottate e risultati ottenuti in attività di ricerca o progettuali, definizione di approcci metodologici originali e innovativi, analisi dello stato dell'arte, ecc.

«AIDAinformazioni» è riconosciuta dall'ANVUR come rivista di Classe A per l'Area 11 – Gruppo Scientifico Disciplinare 11/HIST-04 – Scienze del libro, del documento e storico-religiose e come rivista scientifica per le Aree 10 – Scienze dell'antichità, filologico-letterarie e storico-artistiche; 11 – Scienze storiche, filosofiche, pedagogiche e psicologiche; 12 – Scienze giuridiche; 14 – Scienze politiche e sociali. È anche annoverata dall'ARES (Agence d'évaluation de la recherche et de l'enseignement supérieur) tra le riviste scientifiche dell'ambito delle Scienze dell'Informazione e della Comunicazione. La rivista è, inoltre, indicizzata in: ACNP – Catalogo Italiano dei Periodici; BASE – Bielefeld Academic Search Engine; ERIH PLUS – European Reference Index for the Humanities and Social Sciences – EZB – Elektronische Zeitschriftenbibliothek – Universitätsbibliothek Regensburg; Gateway Bayern; KVK – Karlsruhe Virtual Catalog; The Library Catalog of Georgetown University; SBN – Italian union catalogue; Ulrich's; Union Catalog of Canada; LIBRIS – Union Catalogue of Swedish Libraries; Worldcat.

I contributi sono valutati seguendo il sistema del *double blind peer review*: gli articoli ricevuti sono inviati in forma anonima a due referee, selezionati sulla base della loro comprovata esperienza nei topics specifici del contributo in valutazione.

AIDainformazioni

Anno 43

N. 3-4 – luglio-dicembre 2025

CACUCCI  EDITORE
BARI

PROPRIETÀ LETTERARIA RISERVATA

© 2025 Cacucci Editore – Bari

Via Nicolai, 39 – 70122 Bari – Tel. 080/5214220

<http://www.cacuccieditore.it> e-mail: info@cacucci.it

Ai sensi della legge sui diritti d'Autore e del codice civile è vietata la riproduzione di questo libro o di parte di esso con qualsiasi mezzo, elettronico, meccanico, per mezzo di fotocopie, microfilms, registrazioni o altro, senza il consenso dell'autore e dell'editore.

Sommario

GIOVANNA ARACRI, ANTONIETTA FOLINO, MARIA TERESA GUAGLIANONE, MARCO MANNA, Identification of thesaurus relationships with the support of ChatGPT and Gemini	9
ALESSANDRA BONSIGNORI, Terminologia giuridica e organizzazione della conoscenza	51
MARIA TERESA CHIARAVALLOTI, MARIA TERESA GUAGLIANONE, ELENA CARDILLO, Metadattazione delle tipologie di documenti clinici: un approccio alla standardizzazione mediante LOINC	73
CARLA BEATRIZ MARQUES FELIPE, CARLA MARIA MARTELOTTE VIOLA, RAIMUNDA FERNANDA DOS SANTOS, Domain Analysis on Data Management	91
CLAUDIO GNOLI, EMANUELA VALENZANO, Che faccetta sei?	111
MANUELA GRILLO, Strumenti di convergenza tra archivi e biblioteche: i thesauri	121
BIANCA LA MANNA, Rappresentare gli atti notarili medievali in Linked Open Data (LOD)	141
CLAUDIA LANZA, L'univocità e la variabilità terminologica a supporto della rappresentatività delle risorse semantiche di organizzazione e rappresentazione della conoscenza specialistica	163
LARISSA MORAES MARTINS, CARLOS EDUARDO DA SILVA CARVALHO, ANA CRISTINA DE ALBUQUERQUE, MARIO BARITÉ, Possibilities for using cultural warrant to complement literary work	181
BIANCA THEREZINHA CARVALHO PANISSET, LUCIA MARIA VELLOSO DE OLIVEIRA, LETICIA SOUZA DA COSTA SAMPAIO, CAMILLA CAMPOI DE SOBRAL, Identification of types of documents and knowledge organization: transparency and access in records management	197
SOYOUNG YOON, CLAUDIO GNOLI, ZIYOUNG PARK, Applying Integrative Levels Classification to Korean KOS Registry	211
EMANUELA NICOLE DONATO, GRAZIA SERRATORE, MARIA TAVERNITI, MARIA TERESA CHIARAVALLOTI, L'evoluzione della struttura documentale nelle cartelle cliniche psichiatriche dell'archivio storico del manicomio di Girifalco (1881-1978)	237
LORENZA TALARICO, Manuscripts in the Digital Age	263
MARIA LÍGIA TRIQUES, ANANDA FERNANDA DE JESUS, MARIO BARITÉ, JOSÉ EDUARDO SANTAREM SEGUNDO, ANA CRISTINA DE ALBUQUERQUE, Vocabularies in the digital context	275

MARCIN TRZMIELEWSKI, Context-Driven Use of Automatic Tools in Medical Ontology Design: A Case Study in Allergology	291
CARLA MARIA MARTELLOTE VIOLA, CARLA BEATRIZ MARQUES FELIPE, Exploring the relationships between the Cross-Cutting Women's Agenda - Multi-Year Plan 2024-2027 and the Knowledge Organization of legislative proposals under processing on women in Brazil	313
ELENA CENCETTI, CAMILLO CARLO PELLIZZARI DI SAN GIROLAMO, ELISABETTA VITI, Modellizzazione delle conoscenze: Thesaurus Nuovo soggettario e Wikidata, due strumenti a confronto	325
PONTISH YERAMYAN, ALEXANDER MURZAKU, NICK UVA, CURT ANDERSON, The transferability of brilliance	353
ANA FLÁVIA DIAS ZAMMATARO, ANA CRISTINA DE ALBUQUERQUE, GIULIA CRIPPA, LUCIA SARDO, CARLO BIANCHINI, MARIA LÍGIA TRIQUES, ISKO Italy in perspective	375

Contributi

Identification of thesaurus relationships with the support of ChatGPT and Gemini

Giovanna Aracri*, Antonietta Folino**,
Maria Teresa Guaglianone***, Marco Manna****

Abstract: The study aims to verify if Large Language Models (LLMs) – ChatGPT and Gemini – can support experts during some critical stages of thesaurus construction: categorization and recognition of semantic relationships between terms/concepts. To this end, domain-independent prompts were defined and tested through a pilot experiment in the sub-domain of archaeological sites. The experimentation required the development of a benchmark consisting of terms and concepts linked by thesaurus' semantic relationships. The results show that LLMs, combined with the experts' skills, can be useful in accomplishing these tasks more efficiently. The interaction between experts and AI services has been mutually beneficial: experts helped AI services fully understand the requests – providing context and improving prompts – while AI services provided reasoning for categorization and semantic network construction, which could provide experts with insights for improving tasks. The study highlights opportunities for improving AI services performances in disambiguating categories and recognizing semantic relationships.

Keywords: Thesaurus Relationships, Categorization, LLMs, ChatGPT, Gemini

1. Introduction

The manual construction of a thesaurus from scratch involves a significant effort in terms of time, resources, and both methodological and domain-speci-

* Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing of the National Research Council (STIIMA-CNR), Milan, Italy. giovanna.aracri@cnr.it. ORCID:0000-0002-8716-3798.

** Department of Culture, Education and Society, University of Calabria. Rende (CS), Italy. antonietta.folino@unical.it. ORCID: 0000-0002-5882-6150.

*** Institute of Informatics and Telematics of the National Research Council (IIT-CNR), Rende (CS), Italy. maria.guaglianone@cnr.it. ORCID:/0000-0002-4070-5828.

**** Department of Mathematics and Computer Science, University of Calabria, Rende (CS), Italy. marco.manna@unical.it. ORCID: 0000-0003-3323-9328.

Author contributions: Conceptualization, All; Methodology, All; Investigation, All; Writing – original draft, G.A. (sections 2.1 - 4.1 - 5), A.F. (sections 1 - 3.1 - 4), M.T.G. (sections 3 - 4.2) and M.M. (sections 2.2 - 3.2); Review and editing, All. Authors are listed alphabetically.

fic expertise. The conceptual effort required for terms and concepts identification and for relationships establishment is very labor-intensive and time-consuming, especially considering that these resources are dynamic and that to guarantee a sufficient representativity of the domain of interest they need to be regularly updated. During the past decades, different attempts have been made to automatize the thesaurus construction process, in particular the identification and establishment of semantic relationships (Tudhope et al. 2001; Yang and Powers 2008; Hazem et al. 2020; Meng et al. 2022; Cardillo et al. 2024), with the aim of speeding and reducing the required work and the employed human resources and of allowing a scalable process.

In the last few years, the widespread use of Artificial Intelligence (AI) systems, which is affecting all research fields, has reached a critical dimension also in the more general field of Text and Document Sciences.

The objective of the research is to observe and analyze the ability of two of the most well-known and widely used AI-based services, namely Chat Generative Pre-trained Transformer (ChatGPT) and Gemini, to classify concepts according to known categories and to establish thesaurus semantic relationships between pairs of terms/concepts according to the ISO standard 25964-1:2011. The study also aims to provide some critical considerations on the possibility that such systems could offer valid support to human experts. The task of identifying semantic relationships between pairs of terms/concepts by using LLMs (Large Language Models) is not entirely a novelty. This topic has been explored in (Aggarwal et al. 2024) with a focus on the development of a pipeline to support automatic ontology generation. In this study a prompt template has been applied to two zero-shot reasoning strategies and tested to six different LLMs. The ontological relationships involved in this process are: is-broader-than, is-narrower-than, is-same-as-than and is-other-than. To evaluate the performance of the prompt template, a set of 1,000 semantic relationships – 250 for each typology – selected from IEEE Thesaurus has been considered.

This research represents a valuable reference for addressing the final aim of the present study, although its objectives and conclusions are different from our work, in which we investigate automatic top-level categorization and thesaural relationships identification by using other LLMs. With this study, we make three main contributions. Firstly, we provide a benchmark dataset in the sub-domain of archaeological sites, consisting of five concept categories, 93 terms assigned to these categories, and 44 exemplary pairs drawn from this set of terms, expressing semantic relationships of three types: equivalence, hierarchical, and associative. The benchmark also includes an excerpt (pp. 66-74) from the document by ICCD (Istituto Centrale per il Catalogo e la Documentazione), *Vocabolario aperto per la definizione dei siti archeologici (applicazione nella scheda SI – Siti archeologici, versione 3.00)* (Ministero per i Beni e le Attività Culturali e per il Turismo. Istituto Centrale per il Catalogo e la Documen-

tazione 2009), which provides appropriate contextual grounding for the terms and categories included in the dataset. Secondly, we formalize a set of rules for identifying semantic relationships, specifying the recommendations to consider to achieve their correct identification and we also define domain-independent LLM prompts, aimed at: (i) assigning each input term to exactly one category from a predefined input set; and (ii) identifying the most appropriate semantic relationship between pairs of already categorized terms/concepts. The prompts were iteratively refined using portions of the benchmark, without any model training or incorporation of benchmark-specific knowledge, and were finalized once a stable formulation was achieved. The resulting prompts remain fully domain-independent. Lastly, a systematic evaluation was carried out in clean sessions to assess the effectiveness of the proposed methodology when applied to two AI-based services. Although the same dataset was used for both prompt refinement and evaluation, the domain-independent design of the prompts allowed us to approximate the best achievable performance without overfitting or introducing benchmark-specific biases (see Sections 3 and 4 for details). The results indicate that the methodology performed well in concept categorization, with accurate assignment of terms to predefined categories. However, some limitations were observed in the recognition of semantic relationships, particularly in the case of equivalence ones, which proved more difficult to detect consistently. By contrast, hierarchical relations were identified with a significantly higher degree of reliability. A broader and more exhaustive evaluation of unseen data is planned as part of future work.

The paper is structured as follows: after presenting the literature concerning the application of AI in the domain of Text and Document Sciences and in the specific field of Knowledge Organization Systems (KOSs) construction, we focus on the adopted methodology by illustrating the main stages, by describing the set of semantic relationships used as a benchmark for the prompt design and the experimentation, then we present and critically comment the obtained results by both quantitative and qualitative point of views. Lastly, strengths and limitations of the adopted approach are outlined and some conclusions are drawn.

2. Related works

2.1. AI and Information Science

In this section, we present a brief review of the literature to highlight the reciprocal relationship between Knowledge Organization (KO) and AI. As will become evident, the cited studies underline their mutual influence and interdependence. A common research question emerging from many of them

is: how AI models can support the representation of a document's conceptual content to enable its description, indexing, classification and retrieval. The gap between KO and AI disciplines has been reduced recently (Qin 2020). Although these disciplines seem to be unrelated, both cover a crucial role in the ongoing digital transformation. To take on the emerging challenges we will have to face, particularly those related to the improvement of knowledge organization and management process, through the adoption of up-to-date and high-performance technologies and enhancement of execution times, a convergence between them is necessary. AI, supported by generative tools like ChatGPT and Gemini, has the capability to automate several tasks including: data discoverability and usability; information extraction and retrieval; creation and diffusion of knowledge (Rezaei 2025). Conversely, KO offers methodologies and tools designed to optimize the quality of data and knowledge modelling (Giunchiglia and Bagchi 2023). In the field of Text and Document Sciences the use of AI systems has gained significant momentum, mainly aimed at automating and making easier routine processes, enhancing accuracy, and fostering novel approaches to KO (Noruzi 2024). This trend has been accelerated by the emergence of LLMs (Shen et al. 2023), effective in analysing, summarizing and creating content, by providing natural language output (Naveed et al. 2024; Raiaan et al. 2024). Several studies highlight the benefits of adopting AI systems in Libraries to enhance both operations and services (Oname and Chinedu Alex-Nmecha 2020). These technologies support experts in carrying out the typical librarian activities: on-site consultation and reading, local and interlibrary loans, access to digital resources (internet, databases, e-books, etc.), assistance in bibliographic research (Santosa 2025). Ahmed et al. (2023) investigate the integration of AI systems in existing tools to improve the automatic indexing of large amounts of documents by using the Annif open-source framework. In Golub et al. (2024) the results obtained by Annif in automated subject indexing and classification are compared to those proposed by ChatGPT, in this way Annif acts as gold standard to evaluate ChatGPT accuracy in document automatic indexing and archiving. Also, the Finto AI service, developed by the National Library of Finland to improve the quality of automated indexing (Suominen et al. 2022) has been conceived starting from Annif. The use of generative AI tools to improve the indexing of short documents is explored in (Bouزيد and Piron 2024), where LLMs are employed to enrich the set of features extracted from short documents, aiming to guarantee a more comprehensive and accurate representation of their content and, thereby enhancing their discoverability and findability. In addition to these more document-centered efforts, there are those which emphasize the use of AI models for the creation of terminological and semantic resources in specific domains. In particular, AI algorithms, by using statistical inferences to analyse relationships between words within a text, can understand linguistic

patterns and generating output by suggesting subject headings, descriptors, and index terms supported by existing KOSs, although the current results are not sufficiently consistent (Noruzi 2024). Extracting relationships by using LLMs – specifically GPT3 – is also the core of the research presented in (Arachchige et al. 2023), whose purpose is to identify and extract ontological relationships from unstructured data to develop a Knowledge Graph which provides a visual representation of the main concepts included in historical documents about the Holocaust domain, in the form of a triple subject-predicate-object sequence.

2.2. Prompt Engineering

LLMs are advanced AI systems trained on massive collections of textual data. They can perform a wide range of language-related tasks, such as summarizing texts, answering questions, and generating human-like content. These models – such as ChatGPT (based on GPT-3.5/GPT-4), developed by OpenAI, and Gemini (formerly Bard), developed by Google – operate by predicting the most likely continuation of a given input sequence, based on patterns learned from vast amounts of text during training (Brown et al. 2020; OpenAI 2023; Zhao et al. 2023).

An LLM works a bit like an advanced autocomplete system: given some words as input, it tries to guess what words are most likely to come next. However, unlike basic autocomplete, it does not just predict one word at a time, but it builds entire sentences and paragraphs that are often coherent, contextually relevant, and grammatically correct. This mechanism allows the model to generate responses that appear fluent and informed. However, it is important to remember that the model does not possess actual human understanding or reasoning abilities. It does not rely on logic or factual knowledge in a structured sense, but rather on statistical associations between words, phrases, and concepts found in its training data (Bender and Koller 2020; Mitchell and Krakauer 2023).

In recent years, LLMs have emerged as a transformative force across both academia and industry, due to their unprecedented ability to process and generate natural language at a level comparable to human experts (Brown et al. 2020; OpenAI 2023; Touvron et al. 2023; Zhao et al. 2023). Thanks to the factual knowledge and reasoning skills acquired during large-scale pre-training, these models demonstrated remarkable performance in tasks involving understanding, inference, summarization, and content creation (Li and Qiu 2023; Liao and Vaughan 2023; Su et al. 2022).

Nevertheless, LLMs present a number of well-documented limitations, since: they may produce hallucinations – statements that are plausible but factually incorrect (Shi et al. 2023; Marvin et al. 2024); they can reflect or

amplify biases inherited from their training data (Zhao et al. 2023); they often display flawed or shallow reasoning, struggling with tasks that require logical consistency, multi-step inference, or grounding in external knowledge (Mitchell and Krakauer 2023); their processing is driven by form rather than meaning, as they do not possess true semantic understanding but rely entirely on statistical regularities. These challenges are particularly critical in domains that demand accuracy, reliability, and semantic rigor.

Prompt engineering has thus emerged as a key technique for effectively interacting with LLMs and maximizing their potential. By carefully crafting natural language instructions – known as *prompts* – users can influence how the model interprets the input and defines the structure, content, or tone of the output. Since LLMs are highly sensitive to input phrasing, the way a request is formulated can strongly influence the result, making prompt design a crucial element in obtaining useful and reliable responses (Liu et al. 2023; White et al. 2023; Shi et al. 2023). Rather than retraining the model or altering its internal parameters, prompt engineering leverages the model’s existing capabilities through well-designed input cues. This allows a single LLM to be adapted to a variety of tasks simply by changing the prompt.

For example, adding instructions like «Answer using only the information provided in the text» or «List the facts step by step» can help reduce hallucinations and improve the model’s reasoning. A similar effect can be observed when asking an LLM to extract all patterns, strings, or entities from a long text that meet specific criteria. In this case, the model might identify only a subset or make mistakes. However, by rephrasing the task and verifying each candidate item one by one, users can guide the model more effectively and obtain more reliable results: breaking down a complex instruction into smaller, focused checks helps the model minimize omissions or errors.

In conclusion, prompt engineering is not only a mechanism for tailoring general-purpose LLMs to specific tasks, but also a strategic method for compensating for their cognitive and reasoning limitations. By shaping how the model processes and outputs information, prompts serve as a bridge between the statistical nature of LLMs and the structured, goal-oriented demands of real-world applications. This makes prompt design a critical component in achieving accurate, trustworthy, and effective interactions with language models.

3. Methodology

The methodological approach outlined an experimental path focused on exploring the potential of generative AI services, supported by domain experts, to provide relevant and accurate results in identifying and establishing proper semantic relationships between terms.

An overview of the methodology is provided in Fig. 1.

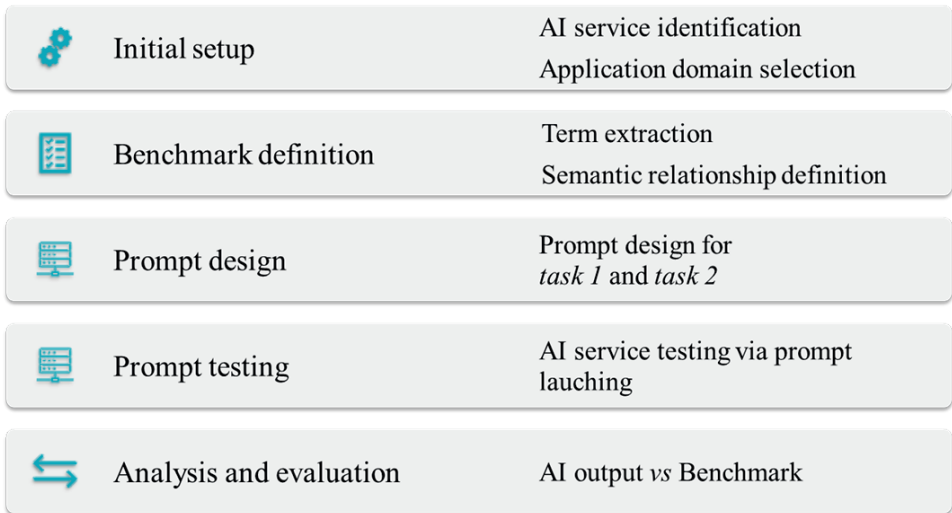


Figure 1. Overview of methodology.

The preliminary phase of the study considered, on the one hand, the identification of specific AI services that could offer a good performance in carrying out this task and, on the other hand, the selection of an application domain which is well defined and documented enough to represent a significant challenge for the considered technologies. Regarding AI services, the GPT-4o version of ChatGPT and the 2.5 flash version of Gemini, have been chosen. Both systems: 1) are freely accessible; 2) are text-based AI models that act as natural multi-language chatbot able to understand and generate responses that convincingly emulate human language; 3) are trained on massive amounts of linguistic data, which have allowed them to acquire a deep understanding of linguistic patterns, contextual semantics and the nuances of natural language; 4) gather and elaborate information from many sources; and 5) represent the latest innovations in generative AI and machine learning. These applications became easily and widely accessible; therefore, their popularity increased rapidly, making them a frequently used resource in people's daily life (Sahin Ozdemir and Ozdemir 2025). The choice of the experimental domain was also a preparatory step in methodological evolution and advancement. It fell upon the Cultural Heritage sector, selected not only for its intrinsic relevance, but above all for the existence of a consolidated and authoritative system of terminological and methodological standards, developed by the ICCD, affiliated with the Italian Ministry of Culture. This system is based on models of detailed descriptive records, terminological tools that promote the use of a shared language, and structured

methodologies for the organization of knowledge relating to Cultural Heritage. Within this vast domain, experimentation focused on a specific set of terms and concepts identifying archaeological sites and, in particular, belonging to the category “place of production activity”, indicating the places where traces of productive activities have been found.

Once preliminary activities – covering AI technology choice and domain selection – and benchmark construction – covering term extraction and semantic relationships definition – were completed, prompt design was performed. Prompt engineering (Bozkurt 2024) provided the principles to carefully formulate the requests. Firstly, AI services were provided with a detailed description of the reference context, elaborated by experts and concerning some of the fundamental notions about thesaurus, relationship typologies, and rules and constraints for their definition. To give instructions to AI services on how to correctly complete the task of building thesaurus relationships, the experts developed a summary of the principles provided by the ISO 25964:1-2011 and tried to make explicit and formalize as much as possible the strategies implemented during the manual execution of the task. All these notions were formalized within a specific instruction called “Context”. After which, a series of requests were developed to be submitted to AI services, containing the tasks to be carried out, namely the categorization and definition of thesaurus relations, respectively task 1 and task 2. Moreover, to support consistent interpretation and domain adherence, it is functional to point out that AI models were also provided for the execution of task 1 – categorization – with the same excerpt from the documentation also used by experts for term extraction and semantic relationships definition during the benchmark construction.

Both systems were asked for the same requests, in the same form and in the same order, to ensure a high degree of standardization and fairness in interaction and to guarantee comparability of the results. The initial requests were established upstream; however, the necessary flexibility was adopted to integrate and adapt these instructions, based on the iterations with the two systems, with the intent of receiving answers in line with the specific needs of the research, and aligned with the reference standards for the construction of controlled vocabularies.

Finally, the observation and comparative analysis of the answers provided by the two AI systems were based on both quantitative and qualitative aspects. This evaluation considered two main dimensions: first, the correspondence of the answers with the manually constructed semantic network, which served as a benchmark; second, the adherence of the answers to the explicit requests of the user, verifying the completeness of the information generated and the coherence and reliability of the reasoning implemented.

This comparison allowed evaluating adequacy and reliability of generative AI systems in correctly identifying and establishing proper semantic relationships between terms, compared to the validated knowledge of experts.

Detailed information about the main phases of the methodological approach are provided in the following subsections.

3.1. Benchmark

In order for the thesaurus relationships to be used as a benchmark for evaluating the AI services' performances, we manually carried out the term extraction phase, the semantic relationships establishment and the final evaluation. Thus, the whole process was based on the exploitation of human experts' conceptual and methodological skills to guarantee compliance with the indications provided by the ISO 25964-1:2011. The first step was oriented towards the identification of relevant terms starting from the above-mentioned document representing the chosen domain. It contains the controlled vocabulary used for the definition of archeological sites, within which terms are structured in two hierarchical levels organized in a tabular form and are furnished with an explicative system oriented to better comprehension of their meanings and use in the cataloguing field. As already stated, we focused on the concept of *Luoghi di attività produttiva* [Places of production activity], that is one of the concepts listed at the first hierarchical level of the cited vocabulary. For some of the corresponding second-level concepts (e.g. *Mulini* [Mills], *Impianti metallurgici* [Metallurgical plants], *Panifici* [Bakeries]) we deeply analyzed the explicative text to manually identify and extract representative terms to be included in our semantic network. It should be noted that most of the concepts and terms were extracted from the ICCD document, but to guarantee the comprehensiveness of the semantic network, this initial set was integrated with terms coming from external sources, namely the *Thesaurus of Nuovo Soggettario* (Biblioteca Nazionale Centrale di Firenze 2025). This was worthwhile in particular for the identification of equivalent terms, whose presence in the ICCD document was considered as not significant for the aim of the experimentation. The integration of such equivalent terms undoubtedly introduced a level of partial equivalence and quasi-synonymousness due to the more general coverage of the *Thesaurus of Nuovo Soggettario*, being conceived for indexing bibliographic resources within the National Library System, so covering all the disciplines. As discussed in the results section, this could have determined a more difficult recognition of inexact equivalence by the AI services.

Terms were subsequently normalized, classified and linked to each other according to the three standard thesaurus relationships: equivalence (USE-UF), hierarchical (BT-NT) and associative (RT). The classification process

was carried out according to the standard set of 13 categories established by the Classification Research Group (Broughton 2023): we identified the 5 most representative ones for our specific subdomain (*Impianti produttivi* [*Production plants*]¹, *Materiali* [*Materials*], *Attività* [*Activities*], *Prodotti* [*Products*] and *Strumenti* [*Tools*])² and we attributed each concept to the most suitable one, having considered these categories as mutually exclusive. This preliminary categorization process allowed us to correctly establish and verify the semantic relationships between concepts, in compliance with ISO 25964-1:2011, such as the possibility to define a hierarchical relationship only between concepts belonging to the same inherent category.

Semantic relationships were established through the TemaTres software (n.d). During the process of semantic network construction and the validation phase, domain experts were involved, in particular to disambiguate terms belonging to the most ambiguous and semantically overlapping categories.

The final benchmark consists of 93 terms, each assigned to exactly one of the five mutually exclusive categories identified for the sub-domain of archaeological production sites. In addition, we selected 44 exemplary pairs from this term set to represent semantic relationships of the three standard thesaurus types. To support consistent interpretation and domain adherence, the benchmark also includes the official excerpt from the above-mentioned document by ICCD, which serves as a foundational reference for term meanings and categorization. The resulting dataset is thus designed not as a full thesaurus, but as a focused, evaluative benchmark suitable for testing semantic reasoning capabilities of AI systems. Thus, it is important to underline that we limited the experimentation to a small set of terms, comparable to a restricted thesaurus branch. Relationships themselves are functional to this set and they could differ if the domain of interest would be extended to other themes or subdomains (e.g. addition of other variants, hierarchical levels, etc.).

The benchmark is available in Appendix A.

¹ *Impianti produttivi* [*Production plants*] corresponds in our subdomain to the standard category “Things”.

² The five categories can be defined as follows (definitions have been worked on from the ICCD document and from *Nuovo Soggettario*): [*Production Plants*] Structures where production cycles or part of them are performed; [*Materials*] Particular substances or set of substances, rough or partially worked, having specific characteristics and having in common their origin or the aim to accomplish particular functions; [*Activities*] The whole of coordinated actions or of accomplished works, in general or in relation with a specific aim or context; [*Products*] The results of a natural process or of an industrial or handcrafted activity, of productive cycles or part of them; [*Tools*] In a technical and scientific meaning, complex of heterogeneous elements (mechanical, electric, etc.) coordinated with each other to make up a system aimed at a specific purpose.

3.2. Prompt definition

The proposed methodology relies on two LLM-based prompts specifically designed to support the following tasks:

1. Assigning each input term to exactly one category from a predefined set.
2. Identifying the most appropriate semantic relationship between pairs of categorized terms.

These prompts were developed through an iterative refinement process using the benchmark dataset created within the archaeological heritage domain (see Section 3.1). The benchmark currently includes the cited excerpt from the official document by ICCD, which provides essential domain-specific context. In the initial versions of the prompts, this contextual information was omitted, resulting in several terms being interpreted according to their most common or generic meanings. This often led to misalignments with their intended usage in the specialized domain. To address this, the ICCD excerpt was later incorporated into the prompts to disambiguate the domain-specific meanings and ensure consistent alignment with expert-defined categories.

To avoid memory interference, task 1 and task 2 were designed to be executed in separate, independent chat sessions. Both prompts share a common introductory section that presents the general framework of KOSs, the ISO 25964-1:2011 definition of a thesaurus, the three types of semantic relationships, equivalence (EQ)³, hierarchical (BT/NT), and associative (RT), and some rules underlying the establishment of these relationships. This shared context was intentionally kept domain-independent to preserve the general applicability of the methodology.

Task 1 is designed to support the categorization of individual terms into a predefined set of mutually exclusive categories. The objective is to assign each input term to exactly one category, relying solely on the category definitions and some additional background knowledge given by the official excerpt from the ICCD document – to help in the categorization process. This task is general in scope and can be instantiated across different domains by adapting the category set and associated definitions. In our case, task 1 was instantiated within the domain of archaeological heritage by applying it to the 93 terms included in the benchmark. These terms were to be classified into one of the above-mentioned five expert-defined categories whose definitions were drawn from authoritative sources such as ICCD and the *Thesaurus of Nuovo Soggettario*.

³ The equivalence relationship is identified by the standard tags USE and UF, respectively introducing preferred and not-preferred terms. However, we decided to use the tag EQ because at this stage of the research we do not ask the AI services to identify which term, among an equivalence group, is the preferred one. Thus, at present stage the use of USE and UF is not significant.

Task 2 is designed to support the identification of semantic relationships between pairs of terms/concepts, with the indication by experts of the category to which they belong. In this task, in fact, we did not include the ICCD document useful to deduce categories, since they were previously assigned and contextually provided within the requests. The task is structured to be reusable across different domains, by supplying appropriate term pairs and contextual grounding. In our instantiation, task 2 was applied to the 44 exemplary pairs included in the archaeological benchmark, with the aim of evaluating the semantic link, if any, between the terms in each pair. For each pair, the model was asked to assess the plausibility of four candidate relationships: hierarchical (in both directions), equivalence, and associative, each scored on a scale from 0 to 1. In early attempts, we explored an alternative, more open-ended formulation of the task, in which the model was asked to identify all possible semantic relationships (BT/NT, NT/BT, EQ, RT) among the 93 terms in the dataset. As expected, this approach produced unsatisfactory results: only a limited subset of relationships was returned, often accompanied by inconsistencies or errors, and the model's decisions were difficult to interpret or verify. These issues highlighted the limitations of an undirected discovery approach and led us to adopt a controlled, pairwise evaluation format. During prompt refinement, it became evident that in some cases all candidate relations were assigned low but non-zero scores, making it ambiguous whether any meaningful connection existed. To clarify such situations, the prompt was extended with a fifth explicit option, allowing the model to indicate that none of the standard relationships apply: «There is no semantic relationship of the type BT-NT, NT-BT, EQ, or RT between t1 and t2». This addition significantly improved the clarity and interpretability of the outputs, especially for borderline or unrelated term pairs.

Throughout the development process, differences in behavior across LLMs were also observed. When evaluating the meaning of terms provided to establish the most appropriate semantic relationship in task 2, the two AI systems behaved differently. ChatGPT consistently respected the categorization specified by the prompt, whereas Gemini occasionally started its own reasoning, proposing alternative categorizations even when instructed not to do so. Specifically, in case of absence of explicit specifications on adherence to the categorization provided, while ChatGPT always adhered to the already identified categories, Gemini, when it disagreed with the proposed categorization, conducted a double reasoning: 1) one, as ChatGPT did, based on the categorization provided in the prompt and 2) the other one autonomous, in which it questioned and reevaluated the expert category assignment and proposed its own categorization, on which it subsequently based the choice of the semantic relationship to establish. Gemini adopted this dual behaviour despite the explicit request in the prompt to strictly adhere to the provided categories

and to not consider alternative categorization other than those proposed in the instructions. In some instances, Gemini justified hierarchical relationships based on its own reinterpretation of category membership, conflicting with expert knowledge. This is the case of the pairs $t1 = \textit{Gualchiere}$ [Fulling mills] <Strumenti [Tools]> and $t2 = \textit{Mulini}$ [Mills] <Impianti Produttivi [Production Plants]>⁴. Based on expert knowledge, the two considered terms belong to the same production site, but are assigned to different categories, hence they are not in the same hierarchy and semantically linked by an associated relationship.

This behavior highlighted the need for prompt customization and refinement that enforces strict adherence to the provided framework and explicitly discourages autonomous reinterpretation. During prompt tuning, the formulation of instructions was progressively refined to reduce ambiguity and prevent models from misinterpreting or omitting parts of the expected output. For instance, although the shared context already stated that equivalence is a symmetric relation, we found it beneficial to restate this explicitly within task 2: «The EQ relation is symmetric, so $t1 \text{ (EQ)} \rightarrow t2$ if and only if $t2 \text{ (EQ)} \rightarrow t1$ ». Similar clarifications were added for hierarchical relationships to improve output completeness and consistency across systems.

The final prompt templates used for both tasks are available in Appendix B.

4. Results and discussion

This section reports on the experimental evaluation about the efficiency degree of two state-of-the-art AI services, ChatGPT and Gemini, performed when guided by domain-independent LLM prompts described in the previous sections. The objective was also to test whether prompts designed to be domain-independent can support tasks that require domain-specific knowledge, such as categorizing concepts and identifying semantic relationships, specifically within the archaeological heritage sector. Both prompts for task 1 and task 2 were applied to the same benchmark used during prompt refinement. Although this might suggest a risk of overfitting, no learning occurred on the benchmark: the refinement phase involved only adjusting the wording of the prompts, without providing the LLMs with domain knowledge or examples reusable during evaluation. The final prompts include no references to benchmark terms, categories, or relationships, and thus cannot encode benchmar-

⁴ Differently from the categorization experts indicated, Gemini argued that «A *Gualchiera* is a type of *Mulino*. This is a clear genus-species hierarchy. The categories given (Strumenti vs Impianti Produttivi) describe *what they are* in a broader sense, but the relationship *between them* is fundamentally hierarchical», since «A “gualchiera” is a type of “mulino”. A “mulino” is an “Impianto Produttivo”. A “gualchiera” is also an “Impianto Produttivo” (specifically, a fulling mill) which also can be described as a “Strumento”».

k-specific patterns. Importantly, the results do not indicate artificially inflated performance. Even after prompt refinement, the models did not reach perfect accuracy on the benchmark, suggesting that the outcomes reflect their intrinsic capabilities rather than dataset-specific adaptation. Testing the prompts on unseen datasets would therefore mainly serve to assess the stability and scalability of the results and is planned as future work. For each task, the AI outputs were compared to the benchmark, enabling the analysis of both performance accuracy and reasoning strategies. It is important to distinguish the AI services' performances while running the two tasks. In task 1 they demonstrated a coherent alignment with expectations since each term/concept has been assigned to only one category and in most cases the corresponding category was successfully identified. On the contrary, task 2 regarding identification of semantic relationships remains more challenging. Whereas, the analysis of semantic relations highlighted some difficulties in the identification of the relationship of equivalence, which is often determined by nuances in meaning. Hierarchical and associative links, by contrast, were generally more consistently recognized. In the following subsections we present the quantitative results, we report and discuss the most representative examples, and we outline some conclusive observations on the main insights emerged from the interaction with the AI services.

4.1. Task 1

In this section we discuss the results produced by ChatGPT and Gemini in running task 1. Overall, as shown in Table 1 the findings suggest a positive outcome for both the systems. Indeed, the models have allocated all terms in a specific category, and out of a total of 93 terms, only seven terms in ChatGPT and six terms in Gemini are not assigned to the corresponding category, in line with the benchmark. This means that the models have respectively reached 92% and 94% of the performance, providing an output coherent with what requested in the prompt. Considering the performance improvement due to the prompt refinement phase, we can deduce that the reliability of the results is due to the accurate and detailed information provided in input to the AI model (knowledge domain of reference, context, and criteria to accomplish the categorization task).

	ChatGPT	Gemini
Number of terms assigned to the benchmark's corresponding categories	86/93	87/93
Number of terms assigned to the benchmark's non-corresponding categories	7/93	6/93

Table 1. ChatGPT and Gemini results for Task 1.

The absence of correspondence concerns, on the one hand, the <Materials> and <Products> categories and, on the other hand, the <Production plants> and <Tools> categories. This may be due to the different meanings that some concepts can have according to different contexts of use, resulting in a certain degree of semantic overlap. The following examples clarify this phenomenon:

- *Forge* [*Forges*] <Strumenti [Tools]> and *Fucine* [*Forges or Smithies*] <Strumenti [Tools]>⁵ are intended by ChatGPT and Gemini as an independent physical space and not as an integrated part of a more complex structure;
- *Gualchiere* [*Fulling mill*] is considered by ChatGPT and Gemini as belonging to <Production plants> although it can be intended as a <Tool> used for a specific purpose or work.

With regard to the correlation between the categories <Materials> and <Products> the semantic overlapping is influenced by the point of view adopted for the classification of the concept and by the specific context in which the study is collocated. For example, <Metalli [Metals]> may be considered as a raw material, input of a transformation process for the production of marketable goods, or as finished products, ready for use and/or sale.

To reduce the level of ambiguity and understand the most appropriate category, as already stated, in task 1 we provided AI services with the ICCD document used for benchmark construction. However, a certain level of uncertainty persists and causes some cases of non-correspondence.

4.2. Task 2

In this section we describe the results concerning task 2, aimed at identifying thesaurus relationships between pairs of terms/concepts. As we progressively analyse the results, we will discuss the factors that could have affected the degree of relationship appropriateness with respect to those manually identified in the benchmark. To test the capability of ChatGPT and Gemini models in interpreting the request and recognizing the appropriate thesaurus relationship, we provided both pairs of terms/concepts linked by the three thesaurus relationship types and pairs of terms/concepts not related by any kind of semantic relationship.

Table 2 shows the total number of correspondences between the responses given by the AI services and the relationships manually identified in the 44 pairs of terms/concepts used as test set to evaluate the AI models' performance.

⁵ Manually assigned to the <Tools> category and automatically assigned to the <Production plants> category by ChatGPT and Gemini. As demonstrated by the different English translations of *Fucine* the meaning this concept can assume is different according to the context (*Tool* or *Workshop*).

In case where AI services return partial score within the range 0-1, the relationship with the higher value has been considered the correct one.

Detailed results are illustrated through a contingency table in Appendix C.

Metric	ChatGPT	Gemini
Total number of identified relationships corresponding with the benchmark	37/44 (84%)	37/44 (84%)
Number of identified relationships of equivalence corresponding with the benchmark	5/11 (45%)	8/11 (73%)
Number of identified hierarchical relationships corresponding with the benchmark	14/14 (100%)	13/14 (93%)
Number of identified associative relationships corresponding with the benchmark	10/10 (100%)	10/10 (100%)
Number of identified null relationships corresponding with the benchmark	8/9 (89%)	6/9 (67%)

Table 2. ChatGPT and Gemini results for Task 2.

As it concerns equivalence relationship, orthographical variants, such as *Mulini* and *Molini* [*Mills*], are generally correctly recognized, except for *Calcare* (*luoghi di produzione*) and *Calchère* in ChatGPT results. In fact, they are recognized as hierarchically connected, although the definitions provided by the ICCD document qualify them as local variants: «t1: calcare (*luoghi di produzione*) (category: Impianti Produttivi) refers to lime production sites or kilns used for producing lime (*calce*), a broader industrial category. t2: calchère (category: Impianti Produttivi) A dialectal or regional term (particularly northern Italy) for lime kilns – essentially synonymous with lime production sites». Whereas, in Gemini the equivalence relationship for the same pair of terms is correctly identified, even though with the partial score 0.8. This means that the equivalence degree between the two terms is not fully recognized. Analysing the reasoning provided by Gemini, we can deduce that in this specific example the model does not have the capability to take advantage of the qualifier provided in round brackets, which should have helped Gemini since it further specified the term meaning.

On the contrary, the quasi-synonyms *Pelli* [*Skins/Hides*] and *Pellami* [*Hides*]⁶ are considered by ChatGPT respectively as BT and NT (with score 1), because, as it explains, *Pelli* «generally refers to raw or untreated animal skins» and *Pellami* «typically refers to processed or tanned leather materials»⁷. Gemini, in this case, proposes two relationships with different degrees of appropriateness and correctly it prefers the equivalence relationship to the associative one with a score of 0.8, understands the nuances of meaning and specifies that these two terms cannot be considered synonyms in all contexts, but in reference to the archaeological sites they can be seen as *quasi-synonyms* or as *terminological variants*.

⁶ These two terms are considered equivalents in the *Nuovo Soggettario*.

⁷ «This would mean that pelli is a broader term than pellami. Since pellami are derived from or processed forms of pelli, this fits a genus-species or raw-material → processed-material relationship».

The hierarchical relationships have been correctly identified by both AI services. In particular – also thanks to the instructions provided to them – they have observed the rule according to which two concepts can be related by a hierarchical relationship only if belonging to the same category. A valid example is represented by the pair of concepts *Pietre* [Stones] and *Granito* [Granite], which in the benchmark are linked by a hierarchical relationship: ChatGPT and Gemini recognized the same relationship (score 1) specifying that «They belong to the same category (Materiali) and have a clear genus-species relationship, where “granito” is a type of “pietra”».

The associative relationship is also well recognized by both the AI services. The only exception in ChatGPT is represented by the pair *Presse a vite* [Screw presses] and *Stiratura* [Ironing], between which it does not identify any semantic connection, although the benchmark indicates an RT one («There is no strong associative relationship in the archaeological classification context: although tools and activities can be related in practice, here we are constrained by strict category adherence»). Gemini, on the contrary, correctly identifies the associative relationship (score 0.8), highlighting that the link between the two concepts is based on the historical use of screw presses to press textiles or other materials, even though this does not refer to ironing as it is currently intended.

The opposite situation concerns the pair *Vasi* [Vases] and *Farina* [Flour]. Gemini recognized an associative relationship with a score 1, although such a relationship has not been manually expected. Gemini favors this association considering the historical function of vessels, as an object used to contain and store food. In the benchmark the concept of [Vases] is intended as a result of a transforming process applied to raw material within a specific <Production plant>. There is not a meaningful divergence between the choice of human experts and reasoning followed by Gemini since the associative relationship between the two concepts can be contemplated. However, the intention, in this specific case, has been to highlight the object's production process perspective rather than its function. ChatGPT, on the contrary, does not identify any relationship between them.

It is also worthwhile providing some examples related to the identification of more than one relationship, each one with a different score, as this shows a certain level of ambiguity. An example is provided in Gemini by the pair of concepts *Manufatti ceramici* [Ceramic artifacts] and *Pelli* [Skins/Hides]. The benchmark established a null relationship, since between the two concepts the RT relationship is considered weak or non-existent. Gemini attributed a score of 0.5 to the associative relationship but also the same score to the null one. On the one hand, the associative relationship is considered plausible but not primary and direct. On the other hand, given the weak appropriateness of the RT relationship – as motivated by Gemini itself – the null relationship has a

moderate degree of appropriateness, reflecting the ambiguous nature of their conceptual link.

Thus, as can be noticed, equivalence seems to be the most difficult relationship to recognize. In general, this can be due to the intrinsic nature of this semantic link: different levels of equivalence can exist between two or more terms, which can be exactly equivalents or partially/inexactly equivalents (synonyms or quasi-synonyms), depending on specific contexts of use or on the representativeness of the corresponding concepts in regard to the domain covered by a thesaurus. Furthermore, as already stated, in this specific case, equivalents have also been found in external sources, not necessarily related to the domain of archeological sites, so different levels of equivalence have been introduced.

Generally, the two AI systems perform quite similarly in identifying and defining semantic relationships between the provided pairs of terms/concepts: both evaluate the possible definable relationships and provide reasoning and score regarding the degree of appropriateness for each type of relationship evaluated, in line with what requested by experts. The reasoning is based not only on the degree of closeness in meaning between terms, but also on those correlations that are considered admissible by the types of relationships provided for thesauri, contextually trying to give proper motivations accordingly with the constraints coming from the thesauri's construction rules.

The AI services correctly start the analysis by evaluating, in the order indicated in the prompt of task 2, the possible hierarchical, equivalence, associative and null relationships. In fact, when evaluating the degree of appropriateness of the existence of an associative relation between two terms, the systems correctly consider the priority of the hierarchical relation in the semantic analysis and, where this has been identified as the most appropriate relation, they subsequently discourage the combination with the associative relation since the reasoning would be inconsistent.

As mentioned, AI systems in most cases act coherently with rules and constraints provided by experts. Thus, for example, while evaluating the possibility to establish a hierarchical correlation, even if they recognize the belonging of two terms to the same semantic area (e.g. *Calzature* [Footwear] and *Calzaturifici* [Shoe Factories]), they are able to affirm that a hierarchical relationship is not admissible since the two terms do not belong to the same category, a condition required by the ISO standard for a hierarchical relationship between two terms to be admissible.

Furthermore, AI systems, in accordance with what is described and prescribed in the context, are often able to distinguish and provide, even if not required as task output, the mutually exclusive subtypes of hierarchical relationship, such as the genus-species subtype, and additionally to apply the *all-and-some test*. This is the case of *Cave* [Quarries] and *Cave in galleria* [Tunnel quarries],

for which both ChatGPT and Gemini give the following subtype of hierarchical relationship and motivations:

- ChatGPT: «*Cave* is a broader term than *Cave in galleria* [*Underground/mined quarries*]. Since *Cave in galleria* is a specific type of *Cave*, this fits the genus-species subtype of hierarchical relationship»;
- Gemini: «*Cave (Quarries)* is a broader term that encompasses various types of *Quarries*. *Cave in galleria* (*Gallery quarries* or *Tunnel quarries*) is a specific type of *Quarry*. Both terms belong to the *Impianti Produttivi* category. This fits the “Genus-species” type of hierarchical relationship, where «all *Cave in galleria* are *Cave*, and some *Cave* are *Cave in galleria*».

Similarly, both systems provide motivations for the choice of the correct hierarchical relationship direction, in line with recommendations for the consistent definition of thesaurus relationships. For the inverse relationship, for which they gave score 0 as degree of appropriateness, they correctly specify that:

- ChatGPT: «This would imply that *Cave in galleria* is broader than *Cave*, which contradicts the correct hierarchical direction identified above. Because the BT-NT relation is antisymmetric, this must be ruled out»;
- Gemini: «This is the inverse of the correct hierarchical relationship. A specific type cannot be the broader term for the general category».

Finally, both also provide term definitions when deemed useful to give conceptual clarification and to better explain their reasoning⁸.

5. Conclusion and future perspectives

This study presents a pilot experiment carried out to explore if AI services can support human experts to accomplish top level categorization and to correctly identify and recognize semantic relationships between pairs of terms/concepts, considering that currently they are time and resource consuming tasks performed by experts. As previously clarified, the manual construction of a valid benchmark and consequently the definition of efficient prompts able to achieve optimal results for both task 1 and task 2 represented fundamental activities and required continuous improvement and testing stages. The definition of a well-structured benchmark and of domain-independent LLM

⁸ Below are, as examples, the definitions of the terms *Macinazione* [*Milling*] and *Riduzione in polvere* [*Grinding into powder*] provided by ChatGPT and Gemini respectively: 1) «*Macinazione*: Grinding, milling. This refers to the process of reducing something to smaller particles, often but not exclusively to a powder»; «*Riduzione in polvere*: Reduction to powder, pulverization. This specifically means making something into a powder»; 2) «*Macinazione* generally refers to the process of grinding or milling materials into smaller particles»; «*Riduzione in polvere*: specifically refers to pulverizing something until it becomes powder – a finer and more specific outcome than general grinding».

prompts – resulting from significant prompt tuning activity, which required a considerable amount of time and expertise spent in explaining all rules and constraints that normally regulate expert activities – were essential to outline a methodology that is efficient and replicable across different contexts, by providing helpful contributions to those approaching similar tasks of categorization and definition of thesaurus relationships.

The results shown in Section 4 are encouraging but not completely satisfactory. In fact, in some cases, the answers reveal the lack of correspondence between the outcome and the expected results. Furthermore, in other circumstances, the AI services demonstrated a certain degree of uncertainty, especially indicating as appropriated multiple relationships that have almost equivalent scores. By providing seemingly plausible reasoning, it could give expert insights to improve conceptualization but also generate confusion on specific semantic choices. On the contrary, without the expert supervision in evaluating the responses, an inexperienced user, lacking the specialized skills, could be misled and choose the inappropriate relationship for the target domain among those reported.

As a conclusion, we can affirm that AI services are useful for the accomplishment of the final objectives, but they cannot completely replace the human expertise required for high conceptual tasks such as categorization and semantic relationships identification. These activities are strongly dependent on context, final users and thesaurus aim, which introduce a level of necessary balance between flexibility and robustness in semantic network construction, which at present are difficult to be automatically reached.

The results of the present study can be considered encouraging enough to justify hypotheses of future proposals, including further developments of the outlined methodology, with the adjunction for AI services of new tasks to accomplish. It would be worthwhile to design a specific prompt oriented to the recognition of the preferred term – in the presence of the relationship of equivalence – among an equivalence group of terms including synonyms or quasi-synonyms. In addition, it could be interesting to explore the support given by LLMs also to refine associative relationships defined between related terms by asking the specification of a set of RT subtypes. The aim is to avoid that intrinsic vagueness and heterogeneity of this kind of thesaurus relationship affect the objective of a semantic network to realize terminological control within itself.

These two future perspectives would require the modification of the initial context about thesaurus relationships definition provided by including the explanation by expert of further rules and constraints to be used by AI services for executing the decision-making process and for giving additional information about correlation between terms, that are closer to ontological relations, so as to complete the methodological framework outlined so far.

Acknowledgments

The authors gratefully acknowledge Assunta Caruso – Researcher at the University of Calabria – for her precious support in assessing the English translation of the context and the prompts provided by AI services, and Carlo Fanciulli – Researcher at the Institute of Condensed Matter Chemistry and Technologies for Energy of the Italian National Research Council (ICMA-TE-CNR) – for his valuable help in refining and validating the semantic network included in the benchmark. Marco Manna was partially supported by NextGenerationEU through project FAIR (CUP H23C22000860006), Italian Ministry of University and Research through project PRODE (CUP H53D23003420006), Italian Ministry of Enterprises and Made in Italy through projects ASVIN (CUP B29J24000200005) and PreDiCI (CUP B29J24000300005), and Calabrian Region through project NextGenGuides (CUP J39I24002040005).

References

- Aggarwal, Tanay, Angelo Salatino, Francesco Osborne, and Enrico Motta. 2024. “Identifying Semantic Relationships Between Research Topics Using Large Language Models in a Zero-Shot Learning Setting.” In *Sci-K 2024: 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment 11/12, November 2024 – Baltimore, MD, USA*. CEUR Workshop.
- Ahmed, Mustak, Mondrita Mukhopadhyay, and Parthasarathi Mukhopadhyay. 2023. “Automated Knowledge Organisation: AI/ML-based Subject Indexing System for Libraries.” *DESIDOC Journal of Library & Information Technology* 43 (1): 45-54. <http://dx.doi.org/10.14429/djlit.43.01.18619>.
- Arachchige, Isuri Anuradha Nanomi, Le An Ha, Ruslan Mitkov, and Vinitar Nahar. 2023. “Evaluating Large Language Models in Relationship Extraction from Unstructured Data: Empirical Study from Holocaust Testimonies.” In *Proceedings of Recent Advances in Natural Language Processing, Varna, Sep 4–6, 2023*, 117-23.
- Bender, Emily M., and Alexander Koller. 2020. “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185-98.
- Biblioteca Nazionale Centrale di Firenze (BNCF). 2025. *Nuovo Soggettario – Thesaurus*. <https://thes.bncf.firenze.sbn.it/ricerca.php>.

- Bouazid, Sara, and Loïs Piron. 2024. "Leveraging Generative AI in Short Document Indexing." *Electronics* 13(17):3563. <http://dx.doi.org/10.3390/electronics13173563>.
- Bozkurt, Aras. "Tell Me Your Prompts and I Will Make Them True: The Alchemy of Prompt Engineering and Generative AI." *Open Praxis*.
- Broughton, Vanda. 2023. "Facet Analysis: The Evolution of an Idea." *Cataloging & Classification Quarterly*, 61(5-6), 411-38. <https://doi.org/10.1080/01639374.2023.2196291>.
- Brown, Tom B., Benjamin Mann, Nick Ryder, et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems (NeurIPS)* 33.
- Cardillo, Elena, Alessio Portaro, Maria Taverniti, Claudia Lanza, and Raffaele Guarasci. "Towards the automated population of Thesauri using BERT: a use case on the Cybersecurity domain." In *Advances in Internet, Data & Web Technologies. EIDWT 2024. Lecture Notes on Data Engineering and Communications Technologies* 193, edited by L. Barolli. Springer, Cham. https://doi.org/10.1007/978-3-031-53555-0_10.
- Giunchiglia, Fausto, and Mayukh Bagchi. 2024. "From Knowledge Representation to Knowledge Organization and Back." In *International Conference on Information (iConference) 2024, Wisdom, Well-being, Win-win*. Springer LNCS, Springer Cham Switzerland. <https://arxiv.org/abs/2312.07302>.
- Golub, Koraljka, Jue Wang, and Johannes Widegren. 2024. "Using ChatGPT for (semi-) automatic subject indexing of different document types." In *Digital Humanities in the Nordic and Baltic Countries, 8th Conference Reykjavik 27-31 May 2024*.
- Hazem, Amir, Beatrice Daille, and Claudia Lanza. 2020. "Towards Automatic Thesaurus Construction and Enrichment." In *Proceedings of the 6th International Workshop on Computational Terminology, May 2020, Marseille, France*. 62-71.
- ISO 25964-1:2011. Information and documentation – Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval.
- Li Xiaonan, and Xipeng Qiu. 2023. "Finding Supporting Examples for In-Context Learning." <https://doi.org/10.48550/arXiv.2302.13539>.
- Liao, Q. Vera, and Jennifer Wortman Vaughan. 2023. "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap." <https://doi.org/10.48550/arXiv.2306.01941>.

- Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Comput. Surv.* 55(9): 195:1-195:35.
- Mancinelli, Maria Letizia. 2018. "Gli standard catalografici dell'Istituto Centrale per il Catalogo e la Documentazione." In *Le voci, le opere e le cose. La catalogazione dei beni culturali demotnoantropologici*, a cura di Roberta Tucci. 279-302. Roma: Istituto centrale per il catalogo e la documentazione – Ministero dei beni e delle attività culturali e del turismo.
- Marvin, Ggaliwango, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2024. "Prompt Engineering in Large Language Models." In *Proceedings of Data Intelligence and Cognitive Informatics – ICDICI 2023. Algorithms for Intelligent Systems*. Singapore: Springer.
- Meng, Fan, Kaile Zhou, Yi Bu, et al. 2022. "Keywords Extraction and Thesaurus Construction for Domain News." *Procedia Computer Science* 214: 837-44. <https://doi.org/10.1016/j.procs.2022.11.249>.
- Ministero per i Beni e le Attività Culturali e per il Turismo. Istituto Centrale per il Catalogo e la Documentazione. 2009. *Strumenti terminologici. Vocabolario aperto per la definizione dei siti archeologici (applicazione nella scheda SI – Siti archeologici, versione 3.00)*. Ultimo aggiornamento 2023.
- Mitchell, Melanie, and David C. Krakauer. 2024. "The debate over understanding in AI's large language models." *Proc. Natl. Acad. Sci. U.S.A.* 120 (13), <https://doi.org/10.1073/pnas.221590712>.
- Naveed, Humza, Asad Ullah Khan, Shi Qiu, et al. 2025. "A Comprehensive Overview of Large Language Models." *ACM Transactions on Intelligent Systems and Technology*. <https://doi.org/10.1145/3744746>.
- Noruzi, Alireza. 2024. "The use of artificial intelligence in knowledge organization and subject indexing." *Informology*, 3(1): 1-8.
- Oname, Isaiah, and Juliet Chinedu Alex-Nmecha. 2020. "Artificial Intelligence in Libraries." In *Managing and adapting Library Information Services for future users*. IGI Global. <https://doi.org/10.4018/978-1-7998-1116-9.ch008>.
- OpenAI, 2023. *GPT-4 Technical Report*. CoRR abs/2303.08774.
- Qin, Jian. 2020. "Knowledge Organization and Representation under the AI Lens." *Journal of Data and Information Science* 5(1):3-17. <https://doi.org/10.2478/jdis-2020-0002>.

- Raiaan, Mohaimenul Azam Khan, Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, and Most Marufatul Jannat Mim. 2024. "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges." *IEEE Access* 12: 26839-874. <https://doi.org/10.1109/ACCESS.2024.3365742>.
- Rezaei, Mojtaba. 2025. "Artificial intelligence in knowledge management: Identifying and addressing the key implementation challenges." *Technological Forecasting & Social Change* 217. <https://doi.org/10.1016/j.techfore.2025.124183>.
- Sahin Özdemir, Meryem, and Yusuf Emre Ozdemir. 2024. "Comparison of the Performances between ChatGPT and Gemini in Answering Questions on Viral Hepatitis." *Scientific Reports* 15 (1). <https://doi.org/10.1038/s41598-024-83575-1>.
- Santosa, Faizhal Arif. 2025. "Artificial Intelligence in Library Studies: A Textual Analysis." *JLIS.It* 16 (1):61-71. <https://doi.org/10.36253/jlis.it-626>.
- Shen, Yiqiu, Laura Heacock, Jonathan Elias, et al. 2023. "ChatGPT and Other Large Language Models Are Double-edged Swords." *Radiology* 307(2). <https://doi.org/10.1148/radiol.230163>.
- Shi, Freda, Xinyun Chen, Kanishka Misra, et al. 2023. "Large Language Models Can Be Easily Distracted by Irrelevant Context." In *Proceedings of the International Conference on Machine Learning (ICML 2023)*. 31210-27. <https://doi.org/10.48550/arXiv.2302.00093>.
- Su, Hongjin, Jungo Kasai, Chen Henry Wu, et al. 2023. "Selective Annotation Makes Language Models Better Few-Shot Learners." In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*. <https://doi.org/10.48550/arXiv.2209.01975>.
- Suominen, Osmo, Juho Inkinen, and Mona Lehtinen. 2022. "Annif and Finto AI: Developing and Implementing Automated Subject Indexing." *JLIS.It* 13 (1): 265-82. <https://doi.org/10.4403/jlis.it-12740>.
- Tematres. n.d. Last accessed July 14, 2025. <https://vocabularyserver.com/web/>.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." <https://doi.org/10.48550/arXiv.2302.13971>.
- Tudhope, Douglas, Harith Alani, and Christopher Jones. 2001. "Augmenting thesaurus relationships: possibilities for retrieval." *Journal of Digital Information* 1 (8).
- White, Jules, Quchen Fu, Sam Hays, et al. 2023. "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT." <https://doi.org/10.48550/arXiv.2302.11382>.

- Yang, Dongqiang, and David M. V. Powers. 2010. "Using Grammatical Relations to Automate Thesaurus Construction." *Journal of Research and Practice in IT* 42(2): 129-46. <https://search.informit.org/doi/10.3316/informit.448235738201280>.
- Zhao, Wayne Xin, Kun Zhou, Junyi Li, et al. 2023. "A Survey of Large Language Models." <https://doi.org/10.48550/arXiv.2303.18223>.

Appendix A – Benchmark

Terms (n. 93) assigned to categories (n. 5)

N.	Terms	Manually assigned category
1	battitura a caldo	attività
2	calcare (luoghi di produzione)	impianti produttivi
3	calcare (roccia calcarea)	materiali
4	calce	materiali
5	calce viva	materiali
6	calchère	impianti produttivi
7	calzature	prodotti
8	calzaturifici	impianti produttivi
9	carbone	materiali
10	cardatura	attività
11	cave	impianti produttivi
12	cave a cielo aperto	impianti produttivi
13	cave in galleria	impianti produttivi
14	ceramica	materiali
15	combustibili	materiali
16	combustione	attività
17	concerie	impianti produttivi
18	concia	attività
19	cottura	attività
20	crogioli	strumenti
21	crogiuoli	strumenti
22	cuoio	Prodotti
23	depurazione	attività
24	estrazione	attività
25	fabbriche di feltro	impianti produttivi
26	falegnamerie	impianti produttivi
27	farina	prodotti
28	feltro	Prodotti
29	filatura	attività
30	folloni	Strumenti
31	fonderie	impianti produttivi

33	forge	strumenti
33	fornaci	Impianti produttivi
34	fucine	strumenti
35	fusione	attività
36	gioielli	prodotti
37	granito	materiali
38	grano	materiali
39	gualchiere	Strumenti
40	impianti ceramici	impianti produttivi
41	impianti metallurgici	impianti produttivi
42	industrie della lana	impianti produttivi
43	lana	materiali
44	laterizi	Prodotti
45	lavaggio	attività
46	lavanderie	impianti produttivi
47	lavatura	attività
48	legna	materiali
49	legno	materiali
50	macerazione	attività
51	macinazione	attività
52	macine	Strumenti
53	manufatti ceramici	prodotti
54	marmo	materiali
55	mattonifici	impianti produttivi
56	metalli	materiali
57	metalli preziosi	materiali
58	minerali	materiali
59	minerali metalliferi	materiali
60	minerali non metalliferi	materiali
61	mole	Strumenti
62	mole (attrezzi agricoli)	Strumenti
63	molini	impianti produttivi
64	mortai	Strumenti
65	mulini	impianti produttivi
66	mulini a vento	impianti produttivi
67	mulini ad acqua	impianti produttivi
68	officine	Impianti produttivi
69	officine di laterizi	impianti produttivi

70	officine figuline	impianti produttivi
71	oreficerie	impianti produttivi
72	ossidiana	materiali
73	ossido di calcio	materiali
74	pane	prodotti
75	panifici	impianti produttivi
76	pellami	materiali
77	pelli	materiali
78	pettinatura	attività
79	pietre	materiali
80	presse a vite	Strumenti
81	pulitura	attività
82	raschiatura	attività
83	riduzione in polvere	attività
84	riscaldamento	attività
85	selce	materiali
86	smacchiatura	attività
87	stiratura	attività
88	stoffe	Prodotti
89	tanno	materiali
90	terrecotte	prodotti
91	vasi	prodotti
92	vetrerie	impianti produttivi
93	vetro	materiali

Pairs of terms/concepts (n. 44) randomly sorted

1	Consider the terms “calzature” (category: “Prodotti”) and “calzaturifici” (category: “Impianti Produttivi”).	RT
2	Consider the terms “cave” (category: “Impianti Produttivi”) and “cave in galleria” (category: “Impianti Produttivi”).	BT/ NT
3	Consider the terms “pelli” (category: “Materiali”) and “pellami” (category: “Materiali”).	EQ
4	Consider the terms “macinazione” (category: “Attività”) and “riduzione in polvere” (category: “Attività”).	EQ
5	Consider the terms “minerali” (category: “Materiali”) and “minerali metalliferi” (category: “Materiali”).	BT/ NT

6	Consider the terms “vasi” (category: “Prodotti”) and “farina” (category: “Materiali”).	Nulla
7	Consider the terms “pietre” (category: “Materiali”) and “granito” (category: “Materiali”).	BT/ NT
8	Consider the terms “cave” (category: “Impianti Produttivi”) and “cave a cielo aperto” (category: “Impianti Produttivi”).	BT/ NT
9	Consider the terms “pane” (category: “Prodotti”) and “panifici” (category: “Impianti Produttivi”).	RT
10	Consider the terms “mole (attrezzi agricoli)” (category: “Strumenti”) and “macine” (category: “Strumenti”).	EQ
11	Consider the terms “concia” (category: “Attività”) and “impianti di piscicoltura” (category: “Impianti Produttivi”).	Nulla
12	Consider the terms “minerali metalliferi” (category: “Materiali”) and “metalli preziosi” (category: “Materiali”).	BT/ NT
13	Consider the terms “minerali” (category: “Materiali”) and “minerali non metalliferi” (category: “Materiali”).	BT/ NT
14	Consider the terms “pietre” (category: “Materiali”) and “marmo” (category: “Materiali”).	BT/ NT
15	Consider the terms “officine di laterizi” (category: “Impianti Produttivi”) and “mattonifici” (category: “Impianti Produttivi”).	EQ
16	Consider the terms “combustibili” (category: “Materiali”) and “carbone” (category: “Materiali”).	BT/ NT
17	Consider the terms “crogioli” (category: “Strumenti”) and “crogiuoli” (category: “Strumenti”).	EQ
18	Consider the terms “lavaggio” (category: “Attività”) and “lavatura” (category: “Attività”).	EQ
19	Consider the terms “legno” (category: “Materiali”) and “falegnamerie” (category: “Impianti Produttivi”).	RT
20	Consider the terms “forge” (category: “Strumenti”) and “fucine” (category: “Strumenti”).	EQ
21	Consider the terms “vetrerie” (category: “Impianti Produttivi”) and “pettinatura” (category: “Attività”).	Nulla
22	Consider the terms “cuoio” (category: “Prodotti”) and “conceerie” (category: “Impianti Produttivi”).	RT
23	Consider the terms “minerali non metalliferi” (category: “Materiali”) and “calce” (category: “Materiali”).	BT/ NT

24	Consider the terms “minerali non metalliferi” (category: “Materiali”) and “pietre” (category: “Materiali”).	BT/ NT
25	Consider the terms “gualchiere” (category: “Strumenti”) and “mulini” (category: “Impianti Produttivi”).	RT
26	Consider the terms “depurazione” (category: “Attività”) and “laterizi” (category: “Prodotti”).	Nulla
27	Consider the terms “cardatura” (category: “Attività”) and “legna” (category: “Materiali”).	Nulla
28	Consider the terms “manufatti ceramici” (category: “Prodotti”) and “pelli” (category: “Materiali”).	Nulla
29	Consider the terms “terrecotte” (category: “Prodotti”) and “panifici” (category: “Impianti Produttivi”).	Nulla
30	Consider the terms “presse a vite” (category: “Strumenti”) and “stiratura” (category: “Attività”).	RT
31	Consider the terms “calce viva” (category: “Materiali”) and “ossido di calcio” (category: “Materiali”).	EQ
32	Consider the terms “tanno” (category: “Materiali”) and “gioielli” (category: “Prodotti”).	Nulla
33	Consider the terms “combustibili” (category: “Materiali”) and “legna” (category: “Materiali”).	BT/ NT
34	Consider the terms “metalli” (category: “Materiali”) and “battitura a caldo” (category: “Attività”).	RT
35	Consider the terms “fusione” (category: “Attività”) and “fonderie” (category: “Impianti Produttivi”).	RT
36	Consider the terms “mulini” (category: “Impianti Produttivi”) and “mulini a vento” (category: “Impianti Produttivi”).	BT/ NT
37	Consider the terms “mulini” (category: “Impianti Produttivi”) and “molini” (category: “Impianti Produttivi”).	EQ
38	Consider the terms “cardatura” (category: “Attività”) and “lana” (category: “Materiali”).	RT
39	Consider the terms “pietre” (category: “Materiali”) and “selce” (category: “Materiali”).	BT/ NT
40	Consider the terms “pietre” (category: “Materiali”) and “ossidiana” (category: “Materiali”).	BT/ NT
41	Consider the terms “mulini” (category: “Impianti Produttivi”) and “mulini ad acqua” (category: “Impianti Produttivi”).	BT/ NT
42	Consider the terms “gualchiere” (category: “Strumenti”) and “mulini” (category: “Impianti Produttivi”).	Nulla

43	Consider the terms “folloni” (category: “Strumenti”) and “gualchiere” (category: “Strumenti”).	EQ
44	Consider the terms “gioielli” (category: “Prodotti”) and “oreficerie” (category: “Impianti Produttivi”).	RT
45	Consider the terms “crogioli” (category: “Strumenti”) and “pane” (category: “Prodotti”).	Nulla
46	Consider the terms “calcare (luoghi di produzione)” (category: “Impianti Produttivi”) and “calchère” (category: “Impianti Produttivi”).	EQ

Appendix B – Prompts

Context 1

According to the website of the International Society for Knowledge Organization, a Knowledge Organization System (KOS) is a structured set of concepts and their semantic relationships, designed to facilitate the management and retrieval of information. Examples of KOS include classifications, subject headings, thesauri, and ontologies.

A thesaurus, according to ISO 25964-1:2011, is a knowledge organization tool that provides a list of interconnected terms organized in a hierarchical structure, with the aim of representing concepts and their relationships. It facilitates information retrieval and the integration of documentary resources through synonyms, terminological variants, and relationships between concepts, such as equivalence, hierarchy, and association.

According to ISO 25964-1:2011, a thesaurus is based on three types of strong semantic relationships that link terms to one another. The semantic relationships are as follows:

- **Equivalence Relationship (EQ)** – Indicates that two terms are synonyms, lexical or orthographic variants, alternative forms, abbreviations or acronyms, and quasi-synonyms. Example: *Automobile (EQ) → Car*; *Car (EQ) → Automobile*; *PC (EQ) → Computer*; *Computer (EQ) → PC*.
Preferred Terms (USE – Use) – In a group of terms connected by an equivalence relationship (synonyms, lexical and orthographic variants, alternative forms, abbreviations or acronyms, and quasi-synonyms), a single preferred term must be identified. The preferred term should be selected based on the following criteria: relevance to the context of use, frequency of use, and unambiguity of meaning.
 For example, if we have:
“Automobile (EQ) → Car” and *“Car (EQ) → Vehicle”* and *“Machine (EQ) → Automobile”*,
 we identify *“Automobile”* as the preferred term and therefore assert that:
“Automobile (USE) → Car”, *“Automobile (USE) → Vehicle”*, and *“Automobile (USE) → Machine”*.
- **Hierarchical Relationship (BT-NT – Broader Term/Narrower Term)** – Indicates an inclusion relationship between broader terms (BT) and more specific terms (NT), which must belong to the same category. The hierarchical relationship is divided into the following mutually exclusive types:
 - *Genus-species*.
 Example: *Mammals (BT) → Canids (NT)*, where all Canids are Mammals and some Mammals are Canids.
 - *Part-whole*, which can only be established between terms belonging to one of the following categories:
 - § bodily systems and organs. Example: *Cardiovascular System (BT) → Arteries (NT)*
 - § geographical locations. Example: *Canada (BT) → Toronto (NT)*; *North America (BT) → Canada (NT)*
 - § disciplines or fields of discourse. Example: *Biology (BT) → Botany (NT)*

- § hierarchical social structures. Example: *Army (BT) → Regiment (NT)*
 - *Class-instance (instance dependency)*. Example: *Painters (BT) → Claude Monet (NT)*
- **Associative Relationship (RT – Related Term)** – Indicates a conceptual link between two terms that are related but not hierarchically connected or linked by equivalence. This relationship is symmetrical, meaning that if A is related to B, then B is related to A.
Example: *Pharmacy (RT) ↔ Drugs*, or similarly, *Drugs (RT) ↔ Pharmacy*.

Context 2

As a second context, consider the text included in the attached pdf file.

Instruction

We will indicate below a series of tasks to be carried out based on the aforementioned context and the information at your disposal.

Task 1

We are operating within the domain/context/field of [DOMAIN_OR_CONTEXT_OR_FIELD_NAME]. Consider the following set of [NUMBER_OF_CATEGORIES] mutually exclusive categories, along with their respective definitions:
[CATEGORY_NAME_1]: [CATEGORY_DEFINITION_1]
[CATEGORY_NAME_2]: [CATEGORY_DEFINITION_2]
...
[CATEGORY_NAME_N]: [CATEGORY_DEFINITION_N]

Consider the following set of [NUMBER_OF_TERMS] terms, where terms followed by a word in parentheses indicate the specific meaning to be considered:

{[TERM_NAME_1], [TERM_NAME_2], ..., [TERM_NAME_N]}

Determine, for each of the [NUMBER_OF_TERMS] terms provided, which of the [NUMBER_OF_CATEGORIES] aforementioned categories is most appropriate/ relevant. When a term is followed by its specific meaning in parentheses, take this into account to determine the correct category.

Provide the output in the form:

{([TERM_NAME], [CATEGORY_NAME]), ([TERM_NAME], [CATEGORY_NAME]), ...}, without translating them, and ensuring that all [NUMBER_OF_TERMS] terms are included in the output.

Task 2

We are still operating within the domain/context/field of [DOMAIN_OR_CONTEXT_OR_FIELD_NAME].

Consider the terms “[TERM_1]” (category: “[CATEGORY_NAME_1]”) and “[TERM_2]” (category: “[CATEGORY_NAME_2]”). Denote the first term as $t1$ and the second as $t2$.

Determine, on a scale from 0 to 1 (where 0 = not at all and 1 = fully), the degree of appropriateness of the following assertions:

- $t1$ (BT) $t2$ (NT), where BT = Broader Term and NT = Narrower Term
- $t2$ (BT) $\rightarrow$ $t1$ (NT), where BT = Broader Term and NT = Narrower Term
- $t1$ (EQ) $\rightarrow$ $t2$, where EQ = Equivalent
- $t1$ (RT) $\rightarrow$ $t2$, where RT = Related Term
- There is no semantic relationship of the type BT-NT, EQ, or RT between $t1$ and $t2$

Importantly, take into account the following:

- At most, only one of “ $t1$ (BT) $\rightarrow$ $t2$ (NT)” and “ $t2$ (BT) $\rightarrow$ $t1$ (NT)” can have a value greater than 0, since binary hierarchical relations like BT-NT are by nature antisymmetric.
- The EQ relation is symmetric, so “ $t1$ (EQ) $\rightarrow$ $t2$ ” if and only if “ $t2$ (EQ) $\rightarrow$ $t1$ ”
- The RT relation is symmetric, so “ $t1$ (RT) $\rightarrow$ $t2$ ” if and only if “ $t2$ (RT) $\rightarrow$ $t1$ ”
- Strictly adhere to the provided categories; do not consider alternative categorizations or real-world functional relationships that go beyond how $t1$ and $t2$ have been classified here.

Appendix C – Detailed results of task 2

Contingency table showing scores and results about ChatGPT and Gemini performing task 2

t1 (cat_t1)	t2 (cat_t2)	Semantic Relationship	Benchmark	ChatGPT		Gemini	
				Score	Result	Score	Result
Calzature (Prodotti)	Calzatur- ifici (Impianti Produttivi)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	1		1	
		None		0		0	
Cave (Impianti produttivi)	Cave in galleria (Impianti produttivi)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None		0		0	
Pelli (Ma- teriali)	Pellami (Materiali)	t1 (BT) → t2 (NT)		1	x	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	0		0.8	
		t1 (RT) → t2		0		0.2	
		None		0		0	
Maci- nazione (Attività)	Riduzione in polvere (Attività)	t1 (BT) → t2 (NT)		1	x	1	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	0		0	
		t1 (RT) → t2		0		0	
		None		0		0	
Minerali (Materiali)	Minerali metalliferi (Materiali)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None		0		0	

Vasi (Prodotti)	Farina (Materiali)	t1 (BT) → t2 (NT)		0	√	0	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		1	
		None	1	1		0	
Pietre (Materiali)	Granito (Materiali)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None		0		0	
Cave (Impianti Produttivi)	Cave a cielo aperto (Impianti Produttivi)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None		0		0	
Pane (Prodotti)	Panifici (Impianti Produttivi)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	1		1	
		None		0		0	
Mole <attrezzi agricoli> (Strumenti)	Macine (Strumenti)	t1 (BT) → t2 (NT)		1	x	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	0		1	
		t1 (RT) → t2		0		0	
		None		0		0	
Concia (Attività)	Impianti di piscicoltura (Impianti Produttivi)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None	1	1		1	
Minerali metalliferi (Materiali)	Metalli preziosi (Materiali)	t1 (BT) → t2 (NT)	1	0.7	√	0	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0.3		1	
		None		0		0	

Minerali (Materiali)	Mine- rali non metalliferi (Materiali)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None		0		0	
Pietre (Materiali)	Marmo (Materiali)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None		0		0	
Officine di laterizi (Impianti Produttivi)	Matto- nifici (Impianti Produttivi)	t1 (BT) → t2 (NT)		1	√	0.8	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	0		0.7	
		t1 (RT) → t2		0.2		0	
		None		0		0	
Com- bustibili (Materiali)	Carbone (Materiali)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0.2		0	
		None		0		0	
Crogioli (Strumen- ti)	Crogiuoli (Strumen- ti)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	1		1	
		t1 (RT) → t2		0		0	
		None		0		0	
Lavaggio (Attività)	Lavatura (Attività)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	1		1	
		t1 (RT) → t2		0		0	
		None		0		0	
Legno (Materiali)	Falegna- merie (Impianti produttivi)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	1		1	
		None		0		0	

Forge (Strumen- ti)	Fucine (Strumen- ti)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	1		1	
		t1 (RT) → t2		0		0	
		None		0		0	
Vetriere (Impianti produttivi)	Pettinatura (Attività)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0.1		0	
		None	1	0.9		1	
Cuoio (Prodotti)	Concerie (Impianti produttivi)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	1		1	
		None		0		0	
Mine- rali non metalliferi (Materiali)	Calce (Materiali)	t1 (BT) → t2 (NT)	1	1	√	0	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0.2		1	
		None		0		0	
Gualchiere (Strumen- ti)	Mulini (Impianti produttivi)	t1 (BT) → t2 (NT)		0.6	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	0.7		1	
		None		0		0	
Depu- razione (Attività)	Laterizi (Prodotti)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None	1	1		1	
Cardatura (Attività)	Legna (Materiali)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None	1	1		1	

Manufatti ceramici (Prodotti)	Pelli (Ma- teriali)	t1 (BT) → t2 (NT)		0	√	0	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0.5	
		None	1	1		0.5	
Terrecotte (Prodotti)	Panifici (Impianti Produttivi)	t1 (BT) → t2 (NT)		0	√	0	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		1	
		None	1	1		0	
Presse a vite (Stru- menti)	Stiratura (Attività)	t1 (BT) → t2 (NT)		0	x	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	0		0.8	
		None		1		2	
Calce viva (Materiali)	Ossido di calcio (Materiali)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	1		1	
		t1 (RT) → t2		0		0	
		None		0		0	
Tanno (Materiali)	Gioielli (Prodotti)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None	1	1		1	
Com- bustibili (Materiali)	Legna (Materiali)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0.2		0	
		None		0		0	
Metalli (Materiali)	Battitura a caldo (Attività)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	0.8		1	
		None		0		0	
Fusione (Attività)	Fonderie (Impianti Produttivi)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	1		1	
		None		0		0	

Mulini (Impianti Produttivi)	Mulini a vento (Impianti Produttivi)	$t1 (BT) \rightarrow t2 (NT)$	1	1	$\sqrt{}$	1	$\sqrt{}$
		$t2 (BT) \rightarrow t1 (NT)$		0		0	
		$t1 (EQ) \rightarrow t2$		0		0	
		$t1 (RT) \rightarrow t2$		0.2		0	
		None		0		0	
Mulini (Impianti Produttivi)	Molini (Impianti Produttivi)	$t1 (BT) \rightarrow t2 (NT)$		0	$\sqrt{}$	0	$\sqrt{}$
		$t2 (BT) \rightarrow t1 (NT)$		0		0	
		$t1 (EQ) \rightarrow t2$	1	1		1	
		$t1 (RT) \rightarrow t2$		0		0	
		None		0		0	
Cardatura (Attività)	Lana (Ma- teriali)	$t1 (BT) \rightarrow t2 (NT)$		0	$\sqrt{}$	0	$\sqrt{}$
		$t2 (BT) \rightarrow t1 (NT)$		0		0	
		$t1 (EQ) \rightarrow t2$		0		0	
		$t1 (RT) \rightarrow t2$	1	1		1	
		None		0		0	
Pietre (Materiali)	Selce (Ma- teriali)	$t1 (BT) \rightarrow t2 (NT)$	1	1	$\sqrt{}$	1	$\sqrt{}$
		$t2 (BT) \rightarrow t1 (NT)$		0		0	
		$t1 (EQ) \rightarrow t2$		0		0	
		$t1 (RT) \rightarrow t2$		0.3		0	
		None		0		0	
Pietre (Materiali)	Ossidiana (Materiali)	$t1 (BT) \rightarrow t2 (NT)$	1	1	$\sqrt{}$	1	$\sqrt{}$
		$t2 (BT) \rightarrow t1 (NT)$		0		0	
		$t1 (EQ) \rightarrow t2$		0		0	
		$t1 (RT) \rightarrow t2$		0.2		0	
		None		0		0	
Mulini (Impianti Produttivi)	Mulini ad acqua (Impianti Produttivi)	$t1 (BT) \rightarrow t2 (NT)$	1	1	$\sqrt{}$	1	$\sqrt{}$
		$t2 (BT) \rightarrow t1 (NT)$		0		0	
		$t1 (EQ) \rightarrow t2$		0		0	
		$t1 (RT) \rightarrow t2$		0.2		0	
		None		0		0	
Folloni (Strumen- ti)	Gualchiere (Strumen- ti)	$t1 (BT) \rightarrow t2 (NT)$		0.9	x	1	x
		$t2 (BT) \rightarrow t1 (NT)$		0		0	
		$t1 (EQ) \rightarrow t2$	1	0		0	
		$t1 (RT) \rightarrow t2$		0.6		0	
		None		0		0	

Gioielli (Prodotti)	Oreficerie (Impianti produttivi)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2	1	0.8		1	
		None		0		0	
Crogioli (Strumen- ti)	Pane (Pro- dotti)	t1 (BT) → t2 (NT)		0	√	0	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0		0	
		None	1	1		1	
Calcare <Luoghi di pro- duzione> (Impianti produttivi)	Calchère (Impianti produttivi)	t1 (BT) → t2 (NT)		0.8	x	0.2	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	0.2		0.8	
		t1 (RT) → t2		0.3		0	
		None		0		0	

Terminologia giuridica e organizzazione della conoscenza

Metodo per la ricostruzione della struttura concettuale del diritto penale in ottica contrastiva

Alessandra Bonsignori*

Abstract: This paper deals with the issue of comparing legal concepts of different legal systems in the field of criminal law, and aims to show how the preliminary reconstruction of conceptual relations within the systems under analysis is crucial to comparison, mainly in establishing the degree of conceptual equivalence – an aspect that is relevant to both legal professionals and translators of legal texts. Due to that, the study focuses on the area of criminal offences against property, as defined in two different legal systems: Italian and Czech. Conceptual relations are implicitly present in texts, which are concrete evidence of the way in which professionals organize their knowledge. Maintaining a knowledge driven approach, the study proposes a comparison-oriented method of reconstruction of conceptual relationships.

Keywords: Criminal law, Terminology, Concepts, Italian, Czech.

1. Premessa

Lo studio si inserisce nella più ampia ricerca dell'autore concernente la redazione di un database terminologico bilingue relativo ai domini del diritto penale sostanziale e processuale, condotta mantenendo un approccio che parte dal presupposto per cui: – il concetto sia un elemento extralinguistico, ma accessibile esclusivamente tramite la verbalizzazione, ossia l'uso del linguaggio; – i testi, nel nostro caso scritti, costituiscano verbalizzazioni dei singoli concetti, dunque possono essere considerati materiale linguistico che rende accessibile il contenuto extralinguistico costituito dalla conoscenza di settore. Il prodotto della ricerca, nelle intenzioni di chi scrive, si figura come potenziali destinatari e fruitori non soltanto i traduttori, ma anche i giuristi ed i professionisti del diritto e si pone come obiettivo quello di coniugare i contenuti concettuali della disciplina con le metodologie di ricerca terminologiche.

Sulla base di tali presupposti, verrà preso in considerazione uno degli aspetti principali nell'analisi terminologica – contrastiva dei domini di riferimento:

* Università della Boemia Meridionale di České Budějovice. ORCID: 0009-0005-1906-9889.
bonsia00@ff.jcu.cz.

la ricostruzione delle relazioni concettuali. A parere di chi scrive, infatti, tale lavoro preliminare consente di portare alla luce quello che è il procedimento di acquisizione dei concetti che gli stessi professionisti del diritto, all'inizio della loro formazione, attivano per acquisire la conoscenza della materia. Partendo dunque dal materiale testuale – prodotto tangibile della concettualizzazione – si ricostruiranno le modalità di organizzazione del pensiero giuridico, con specifico riferimento al diritto penale. A tal fine, a titolo esemplificativo della metodologia di lavoro applicata, è stata presa in considerazione una porzione del dominio del diritto penale dotata di sufficiente autonomia in quanto rientrante in una delle categorie di reato classificate dal legislatore, ossia quella dei reati contro il patrimonio.

Il materiale utilizzato ai fini della raccolta dati è stato in particolar modo costituito dai codici penali in vigore nei due ordinamenti di riferimento, nonché dalla manualistica accademica. Come è infatti noto, in ambito giuridico sono in generale individuabili tre categorie di materiale testuale, ognuna avente le proprie caratteristiche e peculiarità e nessuna da sola sufficiente ai fini della ricostruzione della concettualizzazione: 1. Fonti normative: si tratta dei testi di legge veri e propri, scritti con la finalità di regolare e disciplinare gli aspetti della convivenza sociale dei cittadini, nel caso del diritto penale sostanziale, in particolare definendo i comportamenti umani illeciti e dunque passibili di sanzione penale¹; 2. Dottrina: ossia testi accademici finalizzati al commento dei testi normativi, volti all'apprendimento del contenuto concettuale della disciplina; 3. Giurisprudenza: si tratta delle decisioni di casi singoli, contenute in sentenze, a cura delle autorità del potere giudiziario, compiute applicando a casi concreti le fonti normative.

Per procedere nella ricostruzione delle relazioni concettuali nella porzione di dominio considerata, si è ritenuto più efficace partire dal testo normativo, procedendo poi, nella fase di definizione dei concetti ricavati, ad integrare quanto ricavato dalle previsioni di legge con le indicazioni più dettagliate ricavabili dalla letteratura di settore. Si è scelto inoltre di privilegiare la fonte normativa in quanto in diritto penale vale il cd. principio di tassatività, ossia il principio per cui le norme incriminatrici debbano avere confini ben individuati e non siano suscettibili di estensione in via analogica. Le condotte illecite, dunque, sono esclusivamente quelle previste come tali dal legislatore e non possono essere create *ex novo* facendo ricorso ad interpretazione analogica estensiva della portata di altre norme (Padovani 2023, Jelínek et al. 2017).

¹ Nell'ambito della ricerca l'autore si è basato sui codici penali, ben consapevole tuttavia del fatto che le previsioni normative in materia penale non si esauriscono nel solo codice, bensì alcuni settori, anche molto ampi, costituiscono oggetto di leggi speciali o di norme penali contenute in testi normativi di per sé non aventi carattere penale *tout court*. Al fine di circoscrivere l'ambito di analisi, in questa fase il lavoro si è dunque limitato ai codici penali. Non si escludono in futuro ulteriori integrazioni con la normativa penale *extra codicem*.

La giurisprudenza, in generale nella ricerca portata avanti dall'autore, di cui il presente studio costituisce un esempio, è stata presa in considerazione in via residuale, con lo scopo di chiarire eventuali interpretazioni discordanti date a livello accademico, in modo che, in sede di comparazione e confronto tra i due ordinamenti, si possa avere un maggiore livello di chiarezza.

Il trattamento dei testi è stato portato avanti senza l'ausilio di particolari tecnologie e l'estrazione dei concetti è stata effettuata manualmente, facendo riferimento diretto ai testi presi in considerazione e senza la mediazione di supporti tecnologici. Si è ritenuto infatti fondamentale, ai fini della ricostruzione della concettualizzazione, l'approccio diretto al materiale testuale, basato sullo studio e sulla previa acquisizione sistematica del contenuto concettuale nel suo complesso, rimandando alla fase di inserimento delle reti concettuali nella banca dati predisposta l'eventuale uso di tecnologie per rendere visibili le relazioni tra concetti all'utente finale.

Dati i presupposti della ricerca, le sue finalità e le modalità di lavoro, nonché la natura ibrida del lavoro stesso, il quale parte dal pregresso studio del diritto e della disciplina in questione e lo unisce alle metodologie di analisi proprie della terminologia, si ritiene che la comparazione non sia pienamente assimilabile al diritto comparato classico, ma sia compenetrata da elementi più strettamente linguistici quali quelli semantici e definitori.

2. Inquadramento teorico – metodologico

È generalmente riconosciuto e ampiamente studiato lo stretto rapporto esistente tra il diritto e la lingua, elemento che rende il diritto una scienza alquanto peculiare dato che il suo oggetto ha carattere convenzionale (le leggi sono fatte dal Parlamento, il quale altro non è che l'insieme dei cittadini eletti a rappresentare la collettività) e materialmente costituito da testi, insiemi di parole a carattere prescrittivo, aventi però ripercussioni concrete sulla vita dei singoli individui, tenuti a confrontarsi e conformarsi a tali prescrizioni (si veda, sul linguaggio giuridico e sulle sue caratteristiche in particolare Parise 2022, Cornu 2005). Come si legge inoltre in Sacco (2010, 22): «La legge è, in modo evidente, una dichiarazione. Essa esprime un imperativo [...] e promana da un'autorità».

Per quanto riguarda nello specifico il diritto penale, oggetto della presente ricostruzione, esso viene definito come «il complesso di norme giuridiche che, nel prevedere l'irrogazione di sanzioni afflittive a carico di chi tenga determinati comportamenti offensivi, pone dei confini alla libertà di agire dei singoli, al fine di garantire le condizioni essenziali della convivenza sociale» (Garofoli 2022, 3).

Considerata tale funzione specifica della disciplina, è possibile sostenere che, all'atto pratico, tale branca del diritto compie delle operazioni di classificazione e categorizzazione delle condotte umane, definendone le caratteri-

stiche che le fanno risultare contrarie ai principi dell'ordinamento e dunque sanzionabili. Alla luce di tali caratteristiche peculiari, si comprende come l'organizzazione del dominio parta da tale esigenza di classificazione e segua dei criteri che possono essere ricostruiti attraverso i testi specialistici.

Ogni sapere scientifico ha l'esigenza di organizzare ed ordinare il proprio contenuto di conoscenza ricavato dall'osservazione dei fenomeni che studia e a tal fine stabilisce dei criteri appositi secondo i quali procede a tale operazione (a tal proposito, per un approfondimento sul meccanismo di categorizzazione studiato dalla psicologia, si veda in particolare lo studio di Halford et al. 2014). Nel diritto tale operazione è fatta a monte dal legislatore e i pratici della disciplina, poi, sulla base delle categorie fornite appunto dallo stesso legislatore, osservano la realtà loro circostante, costantemente posta a confronto col modello predefinito costituito dalla norma. Per questo motivo è possibile considerare l'operazione compiuta dal legislatore nella redazione di norme come una operazione di costruzione di categorie (Mauri 2021).

Pare opportuno operare preliminarmente una distinzione fondamentale circa i concetti di classificazione e categorizzazione, così come utilizzati negli studi di psicologia relativi alle performance cognitive umane. Anche se a prima vista parrebbe possibile utilizzare come sinonimi i due termini, in realtà sussistono delle differenze sostanziali di significato e, in particolare, deve intendersi per classificazione la divisione in sottogruppi di un insieme di elementi, mentre per categorizzazione si intende il fatto che una determinata categoria a cui un elemento appartenga abbia delle caratteristiche che si riverberano sull'elemento stesso e pertanto sia utile a fornire indicazioni circa la natura dell'elemento in questione (Estes 1994).

Applicando tale assunto al dominio di riferimento, si vede come il diritto penale sia il frutto dell'attuazione di tali procedimenti: i reati, oggetto della disciplina, sono ripartiti in gruppi in base ai principi che vedremo in seguito nel dettaglio analizzando la struttura del codice penale e i gruppi in questione – categorie, per l'appunto – hanno determinate caratteristiche comuni che raccolgono tutti gli elementi che ne fanno parte, come vedremo nell'analisi della categoria specifica dei reati contro il patrimonio.

Come sopra detto, tali operazioni cognitive, alla base dell'umana conoscenza, sono poste in essere anche a livello normativo dal legislatore in sede di redazione delle norme e ciò vale per tutti i settori del diritto. Successivamente, chi si accinge allo studio della materia dovrà poi far propri tali schemi e interiorizzarli per applicarli nell'osservazione del reale e dei fenomeni che di volta in volta si presenteranno alla sua valutazione nello svolgimento della propria professione.

In una prospettiva comparatistica che abbia l'obiettivo di rendere accessibile il contenuto del dominio del diritto penale di due ordinamenti diversi ad operatori che siano non solo giuristi, ma anche linguisti o traduttori che lavo-

rino nel settore giuridico, pare imprescindibile esplicitare tale processo cognitivo al fine di agevolare la comprensione concettuale non fine a se stessa, bensì alle finalità professionali del singolo operatore. Come sopra indicato, infatti, la comparazione non può essere intesa come un'operazione legata al solo diritto comparato, bensì coinvolge anche il livello semantico e definitorio, il modo in cui i singoli legislatori degli ordinamenti presi in considerazione abbiano scelto di verbalizzare il contenuto concettuale, tenendo comunque ben fermo, in ogni caso, l'approccio onomasiologico basato sul concetto senza cadere nella mera sovrapposizione dei termini usati nelle due lingue studiate.

Nell'esplicitare i processi cognitivi dei professionisti del settore, a fini contrastivi, un ruolo chiave è rivestito dalle strutture concettuali, ossia dalle relazioni esistenti tra i concetti facenti parte del dominio di riferimento. Proprio dalle fonti testuali – in particolare da quelle normative, per il rilievo che esse ricoprono nel dominio giuridico – parte il lavoro di ricostruzione delle relazioni concettuali di cui il contributo fornisce un esempio.

La ricostruzione delle relazioni concettuali ricavate dal materiale testuale che le esplicita e le verbalizza costituisce il primo passo dell'analisi terminografica ed è considerato fondamentale per comprendere le modalità di organizzazione del dominio e per riscontrare analogie e differenze, di cui si forniranno esempi pratici nel prosieguo. Una volta proceduto in tal senso, si procederà ad entrare più dettagliatamente nel dominio e nei singoli concetti in esso evidenziati.

Considerando poi gli aspetti più strettamente connessi alla terminologia e al lavoro terminografico, l'autore aderisce ad un approccio cd. *Knowledge driven* (L'Homme 2020) in cui il punto di partenza dell'analisi è il concetto – considerato come un elemento sì extralinguistico, ma percepibile e comprensibile solo attraverso il linguaggio, di cui i testi – in particolare i codici di riferimento – forniscono prova tangibile.

3. Relazioni tassonomiche

Il primo livello di struttura del dominio può essere riscontrato direttamente dalla fonte normativa, costituita, nel nostro caso, dal codice penale italiano e da quello ceco (Tab. 1).

CODICE PENALE (Testo coordinato ed aggiornato del Regio Decreto 19 ottobre 1930, n. 1398)	Zákon č. 40/2009 Sb. trestní zákoník
Libro primo Dei reati in generale	Část první – Obecná část (§ 1 - § 139)
Libro Secondo dei delitti in particolare	Část druhá - Zvláštní část (§ 140 - § 418)
Libro Terzo delle contravvenzioni in particolare	Část třetí - Přechodná a závěrečná ustanovení (§ 419 - § 421)

Tabella 1. Sezioni dei codici penali di riferimento.

Anche solo osservando la tabella sopra riportata, può riscontrarsi, in entrambi gli ordinamenti, una tendenza del legislatore ad organizzare il contenuto del codice procedendo dal generale al particolare, e ciò vale per entrambi gli ordinamenti. Nel diritto ceco, infatti, si riscontra una tripartizione delle sezioni del codice seguendo una terminologia che nell'ordinamento italiano è diversa da quella del legislatore ma è presente in ambito accademico. Il legislatore ceco, infatti, parla di parte generale (*Obecná část*) e parte speciale (*Zvláštní část*), dedicando la sezione finale del codice alle norme finali e transitorie (*Přechodná a závěrečná ustanovení*).

Ricavato questo primo livello di struttura, occorre adesso procedere ad un ulteriore livello di specificità e andare a scomporre ulteriormente la struttura in questione, verificando quale siano i contenuti dei singoli libri del codice, per entrambe le lingue. Di seguito si riportano due tabelle, una per ciascuno degli ordinamenti considerati, che riproducono tale ripartizione rendendola immediatamente percepibile a livello visivo (Tab. 2 e 3):

Trestný čin	• Trestné činy proti životu a zdraví • Trestné činy proti svobodě a právům na ochranu osobnosti, soukromí a listovního tajemství • Trestné činy proti lidské důstojnosti v sexuální oblasti • Trestné činy proti rodině a dětem • Trestné činy proti majetku • Trestné činy hospodářské • Trestné činy proti životnímu prostředí • Trestné činy proti České republice, cizímu státu a mezinárodní organizaci • Trestné činy proti pořádku ve věcech veřejných • Trestné činy proti branné povinnosti • Trestné činy vojenské
-------------	--

Tabella 2. Ripartizione della macro-categoria Reato riscontrabile dal codice penale ceco.

Reato	Delitto	- o Dei delitti contro la personalità dello Stato - o Dei delitti contro la Pubblica amministrazione - o Dei delitti contro l'amministrazione della giustizia - o Dei delitti contro il sentimento religioso - o Dei delitti contro l'ordine pubblico - o Dei delitti contro l'incolumità pubblica - o Dei delitti contro l'ambiente - o Dei delitti contro la fede pubblica - o Dei delitti contro l'economia pubblica - o Dei delitti contro il patrimonio culturale - o Dei delitti contro la moralità pubblica e il buon costume - o Dei delitti contro gli animali - o Dei delitti contro l'integrità e la sanità della stirpe - o Dei delitti contro la famiglia - o Dei delitti contro la persona - o Dei delitti contro il patrimonio
	Contravvenzione	- o Delle contravvenzioni di polizia - o Delle contravvenzioni concernenti l'attività sociale della pubblica amministrazione - o Delle contravvenzioni concernenti la tutela della riservatezza

Tabella 3. Ripartizione della categoria Reato riscontrabile nel codice penale italiano.

Le tabelle aiutano dunque ad evidenziare un primo tipo di relazione concettuale che viene seguendo l'approccio *knowledge driven*: la tassonomia (L'Homme 2020). Si tratta della relazione esplicitabile in termini di "è un"/"è un tipo di", riscontrabile a vari livelli all'interno del dominio. Le relazioni tassonomiche si contraddistinguono per la transitività: ciò significa che il concetto più specifico condividerà con il concetto generico le sue caratteristiche, ma si differenzierà per aggiunta di caratteristiche ulteriori (L'Homme 2020) e pertanto la relazione tassonomica esistente tra il concetto Reato/*Trestný čin* e i concetti gerarchicamente subordinati (ad esempio, delitti contro la persona/*trestné činy proti životu a zdraví*) farà sì che l'elemento più specifico condivida con quello più ampio (reato/delitto) alcune delle sue proprietà, come vedremo meglio nel prosieguo.

Rimanendo per il momento su tale livello di organizzazione del contenuto specialistico ed osservando le due tabelle, relative ai codici penali in esame, è possibile notare che, apparentemente, nella concettualizzazione dell'italiano esiste un maggiore livello di specificità, perché il concetto di reato è più ampio di quello di delitto (come si nota dalla ripartizione della categoria in ulteriori subcategorie). In realtà, anche l'ordinamento ceco conosce una ripartizione intermedia ulteriore della categoria *Trestný čin*, come si evince dall'art. 14 del codice penale ceco, il quale parla di *Přečiny* e *zločiny* come di ripartizioni ulteriori

della categoria di reato². Specularmente, si ritrova la definizione normativa che distingue tra delitti e contravvenzioni anche nel c.p. italiano, all'art. 39³.

Nella classificazione delle tipologie di reato adottata dal codice ceco, si nota come la ripartizione in sotto-categorie avvenga a partire dalla categoria più ampia di *Trestný čin* e non mantenendo la ripartizione intermedia (conosciuta tuttavia dal codice). La suddivisione ricavata dall'art. 14 sopra citato, dunque, può essere considerata rilevante a livello formale – cioè se adottassimo come criterio distintivo solo la pena inflitta – ma a livello sostanziale – ossia considerando come criterio le caratteristiche materiali delle singole condotte – parrebbe che le sottocategorie siano enucleate partendo dalla categoria più ampia di reato.

L'ordinamento italiano, invece, mantiene l'ulteriore ripartizione intermedia della categoria, ossia quella di delitti e contravvenzioni – avente carattere quantitativo in quanto basata sulla durata della pena, anche nell'enucleare le ulteriori ripartizioni delle condotte illecite.

Tale prima osservazione consente di evidenziare come, nel mantenere un approccio *knowledge driven* e nella ricostruzione delle relazioni concettuali, sia di fondamentale importanza la conoscenza sistematica ed integrale del dominio. Se è vero infatti che il codice e la sua struttura rappresentano di per sé un insieme di relazioni concettuali che possono essere ricostruite esclusivamente dalla struttura del codice, è vero anche che il codice rappresenta un primo livello fondamentale di contenuto concettuale, ma che tale contenuto è soltanto una parte – seppur essenziale e fondamentale – rispetto ad un contenuto più articolato, come ricordato in premessa circa le tipologie di fonti.

Altro aspetto da tener presente è costituito dal fatto che l'operazione di concettualizzazione della categoria reato da parte del legislatore, operata in sede di redazione del codice, non rappresenta l'unica modalità di classificazione: la dottrina, infatti, opera ulteriori classificazioni dei singoli reati effettuate sulla base di diversi criteri distintivi diversi da quelli adottati dal legislatore, i quali, in entrambi gli ordinamenti presi in considerazione, possono essere riconducibili al criterio dell'interesse leso dalla condotta, come si evince dalla formulazione «delitti contro.../ *Trestné činy proti...*». Non ci soffermeremo in questa sede sulle diverse classificazioni dei reati e sui criteri adottati dalla

² Si riporta di seguito l'articolo del codice penale citato, in lingua originale: (1) *Trestné činy se dělí na přečiny a zločiny.* (2) *Přečiny jsou všechny nedbalostní trestné činy a ty úmyslné trestné činy, na něž trestní zákon stanoví trest odnětí svobody s horní hranicí trestní sazby do pěti let.* (3) *Zločiny jsou všechny trestné činy, které nejsou podle trestního zákona přečiny; zvlášť závažnými zločiny jsou ty úmyslné trestné činy, na něž trestní zákon stanoví trest odnětí svobody s horní hranicí trestní sazby nejméně deset let.*

³ «I reati si distinguono in delitti e contravvenzioni, secondo la diversa specie delle pene per essi rispettivamente stabilite da questo codice».

dottrina per operarle, ritenendo prevalente il testo normativo e non essendovi sempre univocità nella definizione di tali categorie⁴.

4. Concettualizzazione e traduzione

Sebbene la ricerca dell'autore non sia finalizzata strettamente alla traduzione e alle modalità di traduzione, bensì alla ricostruzione e mappatura del contenuto concettuale del diritto penale nei due ordinamenti di riferimento e alla successiva comparazione concettuale legata agli aspetti giuridici, la comparazione è comunque strettamente legata anche alla traduzione e alle modalità di resa dei singoli concetti per soddisfare appunto lo scopo di rendere possibile la comprensione del contenuto ricostruito e pertanto pare opportuno accennare brevemente agli orientamenti principali che si sono susseguiti in materia di traduzione giuridica. Uno dei problemi con cui il traduttore si trova a dover fare i conti è infatti costituito dal problema della incongruenza concettuale (Gémar 2014), problema a cui si cerca, almeno in parte, di offrire una risposta mediante il presente contributo e, più in generale, mediante il progetto di ricerca nel suo complesso.

Come ben sottolineato in Šarčević (1997), le specificità della traduzione giuridica sono state discusse ed affrontate sia in ambito linguistico che in ambito giuridico e ognuno di questi ambiti ha attribuito un ruolo diverso al testo e alla terminologia. Se, infatti, la linguistica e le teorie sulla traduzione generale hanno attribuito maggiore importanza al testo, trascurandone in parte la funzione specifica, in ambito giuridico la terminologia specialistica ha assunto una posizione centrale a scapito degli aspetti testuali. Šarčević (1997), evidenzia inoltre come lo scopo primario della traduzione giuridica sia quello di veicolare un contenuto per l'appunto giuridico e come il maggiore ostacolo sia costituito, nell'ambito di tale operazione, dalle differenze concettuali esistenti tra i vari ordinamenti⁵.

L'esperienza canadese e gli studi nella cd. linguistica giuridica (*jurislinguistics*) mostrano come il problema della differenza concettuale sia stato affrontato in ottica pragmatica data l'esistenza di una esigenza concreta di redigere testi normativi e giurisprudenziali in due lingue ufficiali, ai sensi dell'*Official Languages Act* del 1969. Tale previsione normativa ha portato infatti al coinvolgimento, negli anni Settanta, di giuristi francesi ed inglesi chiamati a collaborare col legislatore alla stesura delle leggi (Gémar 2014). Il coinvolgimento diretto dei giuristi in sede di redazione delle norme, eludendo, in un certo senso, la traduzione, è un esempio lampante di come rivesta un'importanza fonamen-

⁴ Per approfondire la questione, si considerino Garofoli 2022 e Jelínek et al. 2017 rispettivamente per l'ordinamento italiano e per quello ceco.

⁵ Per un approfondimento sugli studi relativi alla traduzione giuridica e ai diversi approcci teorici si veda inoltre Biel 2022.

tale la comparazione non tanto a livello linguistico, quanto piuttosto a livello concettuale. Per questo, secondo G  mar, ogni atto di traduzione giuridica rappresenta un'operazione di diritto comparato, tesi che ci sentiamo di condividere pienamente, ma chiaramente l'autore    ben consapevole del fatto che anche gli elementi testuali, in particolare per quanto riguarda lo stile e la modalit   di redazione, hanno una loro rilevanza e sono portatori della cultura giuridica di una determinata nazione, espressa attraverso il linguaggio (G  mar 2015).

A dimostrazione di quanto detto sopra, gli esempi che saranno analizzati nel prosieguo, costituiti in particolare dai reati di furto e rapina, serviranno a mostrare da un lato, nel caso del furto, come la concettualizzazione differente della fattispecie nei due ordinamenti porti a far s   che, nonostante l'equivalenza apparente, permangano degli elementi non esattamente sovrapponibili, dall'altro, nel caso della rapina, si mostrer   come in presenza di elementi analoghi all'interno delle fattispecie presenti nei due ordinamenti, la collocazione sistematica sia diversa e dunque tale elemento dovr   poi essere valutato in sede di comparazione, come vedremo in seguito.

5. Reati contro il patrimonio

Nella cornice sopra delineata, prenderemo ad esempio il concetto ampio di reati contro il patrimonio e provveremo a riscontrare le relazioni tassonomiche esistenti in questa sezione del dominio. Ricordiamo che, in virt   della transitivit   riscontrabile nelle relazioni tassonomiche, tutti i reati contro il patrimonio rappresentano, come qualsiasi reato, comportamenti umani che la legge definisce illeciti e dunque sanzionabili con la tipologia di sanzioni prevista dagli ordinamenti. L'elemento di specificit   che unisce tali reati, differenziandoli rispetto agli altri,    costituito dall'offesa che la condotta reca al patrimonio, inteso come «insieme di beni economicamente valutabili appartenenti, di diritto o di fatto, ad un soggetto» (Piergallini 2022, 13)⁶.

Andiamo dunque a scomporre ulteriormente il concetto, andando a verificare quali fattispecie di reato i due ordinamenti raccolgano all'interno dei reati contro il patrimonio (Tab. 4).

⁶ Si noti come il concetto di patrimonio non rilevi soltanto a fini penalistici, bens   anche in altri settori del diritto ed in particolare nel diritto civile, tanto da far parlare di «stretta contiguit   con fondamentali concetti del diritto civile» (Piergallini 2022, 12). Tale circostanza depone ancor pi   a favore di una conoscenza sistematica del dominio giuridico, stavolta inteso in senso lato come Diritto in generale, in ogni sua branca e, a livello pi   strettamente linguistico, si nota come le modalit   di concettualizzazione siano trasversali e non circoscrivibili ad un settore esclusivo, ma operino attraverso rimandi ad altri settori non solo giuridici. Anche tale aspetto costituisce una delle caratteristiche fondamentali del lessico giuridico (si veda Cornu 2005 e Serianni 2003).

Delitti contro il patrimonio	Trestné činy proti majetku
Delitti contro il patrimonio mediante violenza alle cose o alle persone	§ 205 Krádež § 206 Zpronevěra
Art. 624 – Furto	§ 207 Neoprávněné užívání cizí věci
Art. 624 bis – Furto in abitazione e furto con strappo	§ 208 Neoprávněný zásah do práva k domu, bytu nebo k nebytovému prostoru
Art. 626 – Furti minori	§ 209 Podvod
Art. 628 – Rapina	§ 210 Pojistný podvod
Art. 629 – Estorsione	§ 211 Úvěrový podvod
Art. 630 – Sequestro di persona a scopo di estorsione	§ 212 Dotační podvod § 213 Provozování nepoctivých her a sázek
Art. 631 – Usurpazione	§ 216 Legalizace výnosů z trestné činnosti
Art. 632 – Deviazione di acque e modificazione dello stato dei luoghi	§ 217 Legalizace výnosů z trestné činnosti z nedbalosti
Art. 633 – Invasione di terreni o edifici	§ 218 Lichva
Art. 633 bis – Invasione di terreni o edifici con pericolo per la salute pubblica o l'incolumità pubblica	§ 219 Zatajení věci
Art. 634 – Turbativa violenta del possesso di cose immobili	§ 220 Porušení povinnosti při správě cizího majetku
Art. 634 bis – Occupazione arbitraria di immobile destinato a domicilio altrui	§ 221 Porušení povinnosti při správě cizího majetku z nedbalosti
Art. 635 – Danneggiamento	§ 222 Poškození věřitele
Art. 635 ter – Danneggiamento di informazioni, dati e programmi informatici pubblici o di interesse pubblico	§ 223 Zvýhodnění věřitele
Art. 635 quater – Danneggiamento di sistemi informatici o telematici	§ 224 Způsobení úpadku
Art. 635 quater 1 – Detenzione, diffusione e installazione abusiva di apparecchiature, dispositivi o programmi informatici diretti a danneggiare o interrompere un sistema informatico o telematico Art. 637 – Ingresso abusivo nel fondo altrui	§ 225 Porušení povinnosti v insolvenčním řízení
Art. 635 quinquies – Danneggiamento di sistemi informatici o telematici di pubblico interesse	§ 226 Pletichy v insolvenčním řízení
Art. 636 – Introduzione o abbandono di animali nel fondo altrui e pascolo abusivo	§ 227 Porušení povinnosti učinit pravdivé prohlášení o majetku

Art. 637 Ingresso abusivo nel fondo altrui	§ 228 Poškození cizí věci
Art. 638 – Uccisione o danneggiamento di animali altrui	§ 229 Zneužívání vlastnictví
Art. 639 – Deturpamento e imbrattamento di cose altrui	§ 230 Neoprávněný přístup k počítačovému systému a neoprávněný zásah do počítačového systému nebo nosiče informací
Delitti contro il patrimonio mediante frode	§ 231 Opatření a přechovávání přístupového zařízení a hesla k počítačovému systému a jiných takových dat
	§ 232 Neoprávněný zásah do počítačového systému nebo nosiče informací z nedbalosti
Art. 640 – Truffa	
Art. 640 bis – Truffa aggravata per il conseguimento di erogazioni pubbliche	
Art. 640 ter – Frode informatica	
Art. 640 quinquies – Frode informatica del soggetto che presta servizi di certificazione di firma elettronica	
Art. 641 – Insolvenza fraudolenta	
Art. 642 – Fraudolento danneggiamento dei beni assicurati e mutilazione fraudolenta della propria persona	
Art. 643 – Circonvenzione di persone incapaci	
Art. 644 – Usura	
Art. 645 – Frode in emigrazione	
Art. 646 – Appropriazione indebita	
Art. 648 – Ricettazione	
Art. 648 bis – Riciclaggio	
Art. 648 ter – Impiego di denaro, beni o utilità di provenienza illecita	
Art. 648 ter 1 – Autoriciclaggio	
Art. 648 quater – Confisca	

Tabella 4. Reati contro il patrimonio – ordinamento italiano e ordinamento ceco.

La tabella 4 riporta l'elenco dei reati contro il patrimonio elencati nei codici penali di riferimento, dal quale sono state espunte le norme riguardanti esclusivamente il trattamento sanzionatorio e le norme abrogate. Anche in questo caso, è possibile evidenziare come l'ordinamento italiano mantenga una ripartizione intermedia e suddivida ulteriormente la categoria dei reati contro il

patrimonio in due sottocategorie, in base alle modalità con cui l'offesa viene posta in essere (mediante frode/mediante violenza alle cose o alle persone). Le fattispecie elencate in tabella, inoltre, seguono l'ordine in cui compaiono all'interno dei rispettivi codici e, in questa fase, data la finalità di partenza del lavoro, ossia la ricostruzione delle relazioni tassonomiche nei singoli ordinamenti, non è stata ancora effettuata una comparazione tra singole fattispecie per riscontrare il livello di corrispondenza.

Riepilogando, dunque, abbiamo tratto dal testo normativo una struttura del dominio nella quale si riscontrano, a due diversi livelli di specificità, delle relazioni tassonomiche del tipo «è un/ è un tipo di». Come prevedibile, i due ordinamenti concettualizzano in maniera differente i reati contro il patrimonio e, anche solo a livello numerico, si vede come l'elenco del codice italiano preveda più fattispecie rispetto all'ordinamento ceco.

Per dar conto del livello successivo di analisi, quello di comparazione tra le singole fattispecie appartenenti ai due domini, di seguito analizzeremo il reato di furto, presente in entrambi gli ordinamenti, a titolo di esempio della problematicità della comparazione tra ordinamenti diversi, mostrando come, ancora una volta, la conoscenza sistematica dell'intero dominio – in una lingua e nell'altra – venga in soccorso in sede di comparazione.

6. Il reato di furto: un'applicazione del metodo proposto in fase di comparazione tra norme

Prima di procedere alla comparazione delle norme vigenti nei due ordinamenti, pare opportuno citare Padovani (2023, 128-129):

Quando [...] si tratta di considerare il reato come fenomeno normativo, che deve essere riferito ad un determinato accadimento, occorre verificare molteplici profili, ciascuno dei quali assume una sua peculiare rilevanza: in primo luogo è necessario verificare la corrispondenza del fatto storico, nella sua dimensione obiettiva, alla fattispecie incriminatrice; in secondo luogo bisogna verificare se esso sia stato o meno commesso in presenza di una causa di giustificazione; in terzo luogo accertare con quale atteggiamento psicologico il fatto sia stato commesso e se sussistano particolari circostanze che scusino il soggetto. La considerazione del reato non può dunque essere che analitica.

Il passaggio è significativo in quanto esplicita la modalità con cui concettualmente si procede all'analisi del reato ed individua i passaggi salienti dell'analisi in questione. Tali passaggi sono poi fondamentali per scomporre il reato, ossia la previsione incriminatrice del legislatore, nelle sue componenti essenziali.

Sul punto, storicamente si sono susseguiti vari orientamenti circa quelli che in diritto penale sono definiti gli elementi costitutivi del reato, riconducibili

alla cd. teoria bipartita, nelle sue varie declinazioni, o alla cd. teoria tripartita. Si rimanda a Padovani (2023) per un resoconto più dettagliato delle teorie circa gli elementi costitutivi del reato, ma il motivo per cui in questa sede si richiamano tali teorie è che, all'interno del dominio di riferimento, si parla di elementi costitutivi della fattispecie, pertanto, in sede di comparazione tra ordinamenti diversi, proprio da tali elementi costitutivi sarà necessario partire per poter evidenziare nel dettaglio le caratteristiche delle norme e dunque, successivamente, i livelli di equivalenza.

In tale prospettiva, alla luce del fatto che si considera il dominio come frutto di operazioni di concettualizzazione, nonché del fatto che, all'interno del dominio stesso, si definisca il reato in termini di elementi costitutivi, dal punto di vista più strettamente linguistico pare opportuno procedere all'analisi delle fattispecie proprio scomponendole nei loro elementi costitutivi, definibili secondo quella che, nella parte generale del diritto penale, viene definita struttura del reato.

Si potrebbe obiettare che quanto affermato citando la dottrina italiana non valga per l'ordinamento ceco. Tale obiezione è tuttavia infondata, in quanto, in primo luogo, entrambi i sistemi appartengono alla cd. *civil law* e, soprattutto, perché si riscontra anche nella manualistica accademica ceca la stessa concezione analitica del reato e una adesione alla cd. teoria tripartita.

Si veda, a tal proposito, Šámal et al. (2022, 100-103):

Podle platného českého trestního zákoníku se vyznačuje každý trestný čin (každé provinění) těmito znaky [...] a) Protiprávnost činu (§ 13 odst. 1), kterou posuzujeme z pohledu celého právního řádu, neboť trestným činem může být pouze jednání právním řádem zakázané a nedovolené. [...]. Naplnění skutkové podstaty některého trestného činu sice indikuje, že posuzovaný čin může být protiprávní a trestný, avšak výjimečně může být dána některá z okolností vylučujících protiprávnost (např. nutná obrana, krajní nouze, svolení poškozeného). [...] b) Typové znaky tvořící tzv. skutkovou podstatu trestného činu, které charakterizují individuální objekt trestného činu (tj. chráněnou právní hodnotu), jeho objektivní stránku (zejména jednání, následek a kauzální vztah mezi nimi), pachatele (subjekt) (speciální či konkrétní subjekt) a subjektivní stránku (zejména formu zavinění – úmysl či nedbalost). [...] c) Obecné požadavky na subjekt (pachatele).

È possibile riscontrare, nei passaggi citati e nella descrizione degli elementi del reato, una forte analogia con le metodologie di ricostruzione del significato sviluppatesi tra gli anni Cinquanta e Sessanta del Novecento in America ed in Europa, definite come analisi componenziale (Taylor 2004, Geeraerts 2010).

In particolare, l'analisi componenziale si basa sull'assunto per il quale il significato di una determinata parola possa essere ricostruito sulla base di elementi costitutivi, detti caratteristiche o componenti semantici. In ambito americano ed europeo il presupposto delle teorie ha come antecedente l'analisi

fonologica strutturalista basata sui tratti distintivi, ma nelle due aree geografiche in questione il metodo dell'analisi componenziale ha avuto sviluppi e sorti diverse. Se in Europa l'analisi componenziale si sviluppa dalla teoria dei cd. campi semantici di Trier, grazie a esponenti come Pottier, Coseriu e Greimas, in ambito americano essa trova terreno fertile nel campo della linguistica antropologica, la quale non prende in considerazione le connessioni con la teoria strutturalista dei campi semantici (Geeraerts 2010).

Tornando al dominio di riferimento e alla concezione analitica del reato, riteniamo che, al fine di portare avanti una analisi comparatistica delle norme, sia opportuno concentrarsi su quell'elemento costitutivo del reato che nella letteratura specialistica è definito come fatto tipico (*skutková podstata trestného činu*). Se, infatti, il professionista deve verificare la corrispondenza tra fatto storico e descrizione astratta della norma incriminatrice per verificare se è stato integrato un reato, prendendo in considerazione la condotta, il soggetto che la pone in essere, e via dicendo (Padovani 2023, Šámal et al. 2022), così in sede di comparazione tra ordinamenti andremo a comparare quali elementi le relative norme prevedano quali termini di paragone con la realtà materiale per stabilire se si sia o meno in presenza di un reato. Gli altri elementi, infatti, possono dirsi non rilevanti in sede di comparazione tra norme in quanto esterni rispetto alla condotta considerata illecita dall'ordinamento: l'antigiuridicità è infatti valutata alla luce dell'assenza o meno di cause che escludano l'integrazione del reato, mentre la colpevolezza riguarda l'attribuibilità del fatto commesso al reo, e dunque non incide sulla definizione e sulla classificazione del fatto come illecito.

Partiamo dunque dalla fattispecie di furto e procediamo confrontando il testo della norma italiana con quello della norma ceca, estrapolando quelli che i relativi ordinamenti considerano come elementi costitutivi del fatto tipico (Tab. 5):

§ 205 Krádež	Art. 624 Furto
(1) Kdo si přisvojí cizí věc tím, že se jí zmocní, a a) způsobí tak na cizím majetku škodu nikoliv nepatrnou, b) čin spáchá vloupáním, c) bezprostředně po činu se pokusí uchovat si věc násilím nebo pohrůžkou bezprostředního násilí, d) čin spáchá na věci, kterou má jiný na sobě nebo při sobě, nebo e) čin spáchá na území, na němž je prováděna nebo byla provedena evakuace osob, bude potrestán odnětím svobody až na dvě léta, zákazem činnosti nebo propadnutím věci.	Chiunque si impossessa della cosa mobile altrui, sottraendola a chi la detiene, al fine di trarne profitto per sé o per altri, è punito con la reclusione da sei mesi a tre anni e con la multa da euro 154 a euro 51. Agli effetti della legge penale, si considera cosa mobile anche l'energia elettrica e ogni altra energia che abbia un valore economico. [...]
	Art. 624 bis Furto in abitazione e furto con strappo Chiunque si impossessa della cosa mobile altrui, sottraendola a chi la detiene, al fine di trarne profitto per sé o per altri, mediante introduzione in un edificio o in altro luogo destinato in tutto o in parte a privata dimora o nelle pertinenze di essa, è punito con la reclusione da quattro a sette anni e con la multa da euro 927 a euro 1.500. Alla stessa pena di cui al primo comma soggiace chi si impossessa della cosa mobile altrui, sottraendola a chi la detiene, al fine di trarne profitto per sé o per altri, strappandola di mano o di dosso alla persona [...]

Tabella 5. La fattispecie di furto – ordinamento ceco e norme italiane.

Scomponiamo adesso le norme prese in considerazione per verificare quali siano gli elementi del fatto tipico che devono essere integrati per poter affermare di essere in presenza di un furto, elementi ricavati dalla letteratura specialistica. Dato che «il contenuto minimo indefettibile del fatto tipico è rappresentato da una condotta umana obiettivamente rilevabile [...]» (Padovani 2023, 134), ci aspetteremo, in linea di massima, che le norme indichino: - un soggetto che compie l'azione penalmente rilevante; - una determinata azione espressa da un verbo; - un eventuale soggetto passivo che subisce l'azione; - lo scopo per cui tale azione viene posta in essere, qualora rilevi ai fini dell'applicazione della norma⁷. I componenti sopra menzionati, all'interno del dominio,

⁷ Non tutti gli elementi sono sempre presenti in tutte le fattispecie di reato previste dal legislatore, per questo nella dottrina sono operate ulteriori classificazioni in base alla presenza/ all'assenza o alle caratte-

sono identificati mediante termini specifici, riportati nella tabella sottostante (Tab. 6):

	Trestní zákoník § 205	Codice penale (artt. 624 e 624 bis)
Soggetto attivo	Kdo	Chiunque
Condotta	Kdo si přisvojí cizí věc tím, že se jí zmocní, a a) způsobí tak na cizím majetku škodu nikoliv nepatrnou b) čin spáchá vloupáním, c) bezprostředně po činu se pokusí uchovat si věc násilím nebo pohrůzkou bezprostředního násilí, d) čin spáchá na věci, kterou má jiný na sobě nebo při sobě, nebo e) čin spáchá na území, na němž je prováděna nebo byla provedena evakuace osob, bude potrestán odnětím svobody až na dvě léta, zákazem činnosti nebo propadnutím věci.	Chiunque si impossessa della cosa mobile altrui, sottraendola a chi la detiene
		mediante introduzione in un edificio o in altro luogo destinato in tutto o in parte a privata dimora o nelle pertinenze di essa, strappando [la cosa mobile altrui] di mano o di dosso alla persona
Elemento soggettivo – dolo o colpa		al fine di trarne profitto per sé o per altri

Tabella 6. Identificazione dei termini specifici.

Il confronto mediante gli elementi costitutivi del reato di furto consente di osservare innanzitutto che ad una norma dell'ordinamento ceco corrispondono due diverse fattispecie nell'ordinamento italiano, diverse per la modalità con cui la condotta di appropriazione e spossessamento propria del furto viene attuata: con ingresso in un edificio privato o con strappo. Šámal et al. 2022 affermano tuttavia che anche la norma ceca fa riferimento a fattispecie diverse, distinte sulla base delle differenze della condotta riportate alle lettere da b) a e).

Ulteriore elemento da tener presente è che il codice italiano prevede circostanze aggravanti, cioè elementi esterni al reato che rilevano per il solo trattamento sanzionatorio e non rientrano negli elementi costitutivi del reato, ma che, a livello terminologico, danno vita a tutta in giurisprudenza ad una

ristiche di uno degli elementi considerati (ad es. classificazione dei reati in reati comune e reati propri, basata sul fatto che il reato comune può essere commesso da chiunque, mentre quello proprio solo da particolari categorie di soggetti specifici, o la distinzione tra reati commissivi e reati omissivi, laddove la condotta, nel secondo caso, è rappresentata da un non fare una determinata azione per impedire il prodursi delle conseguenze dannose del reato). Si rimanda a Padovani 2023, nonché a Šámal et al. 2022 per ulteriori approfondimenti.

specifica terminologia (ad esempio, art. 625, n. 2: se il colpevole usa violenza sulle cose – furto con scasso), a conferma della necessità di una conoscenza sistematica dell'intero dominio e di ogni sua componente.

L'esempio mostrato conferma da un lato la caratteristica della transitività esistente nelle relazioni tassonomiche, per cui gli elementi subordinati e specifici conservano una parte delle caratteristiche del concetto sovraordinato, dall'altro è possibile identificare un'altra tipologia di relazione, di tipo partitivo, tra le singole componenti delle fattispecie, confermata dall'approccio mantenuto nelle fonti.

Per quanto riguarda gli aspetti connessi al livello di equivalenza, l'analisi degli elementi costitutivi delle fattispecie nei due ordinamenti consente di evidenziare come possa essere riscontrato un nucleo concettuale analogo connesso alla condotta di impossessamento e sottrazione di cosa altrui che consente di associare i termini *Krádež* e furto. Per le ulteriori precisazioni concettuali, collegate alle diverse modalità di attuazione delle condotte appropriative, alle quali spesso corrisponde un diverso concetto verbalizzato con termini di fonte giurisprudenziale o dottrinale, occorrerà poi procedere all'integrazione della fonte normativa con le suddette fonti ulteriori qui citate, a riprova del fatto che ciascuna, singolarmente considerata, si rivela inadatta ad una mappatura completa del dominio. Occorrerà inoltre tener presente il fatto che alcuni concetti a cui rimanda la norma non sono strettamente giuridici, ma si riferiscono alla comune esperienza o, ancora, pur essendo giuridici non sono di esclusiva pertinenza del diritto penale, per un ordinamento e per l'altro.

7. Un esempio di differenze nella categorizzazione – il reato di rapina

L'esempio qui considerato mostra come, nonostante le fattispecie presentino i medesimi elementi costitutivi, gli ordinamenti di appartenenza scelgano, a propria discrezione, di classificare le condotte in una categoria diversa, a seconda della maggiore o minore rilevanza attribuita agli elementi costitutivi delle condotte stesse.

Trestni zakonik - Zvláštní část (část druhá) > trestné činy proti svobodě a právům na ochranu osobnosti, soukromí a listovního tajemství (hlava ii) > trestné činy proti svobodě (díl 1) > Loupež (§ 173)	Codice Penale > LIBRO SECONDO - Dei delitti in particolare > Titolo XIII - Dei delitti contro il patrimonio > Capo I - Dei delitti contro il patrimonio mediante violenza alle cose o alle persone > Rapina (art. 628)
--	--

Tabella 7. Il reato di rapina: collocazione all'interno del codice penale ceco ed italiano.

	Trestní zákoník § 173	Codice penale art. 628
Soggetto attivo	Kdo	Chiunque
Condotta	Kdo proti jinému užije násilí nebo pohrůžky bezprostředního násilí	mediante violenza alla persona o minaccia, s'impadronisce della cosa mobile altrui, sottraendola a chi la detiene
Elemento soggettivo – dolo o colpa	v úmyslu zmocnit se cizí věci	per procurare a sé o ad altri un ingiusto profitto

Tabella 7. Elementi costitutivi del reato di rapina nell'ordinamento ceco ed in quello italiano.

Anche se, a giudicare dalla norma, sia l'elemento soggettivo che quello oggettivo paiono sovrapponibili nei due ordinamenti, l'ordinamento italiano colloca la rapina tra i delitti contro il patrimonio, quello ceco in quelli contro la libertà personale, dando dunque maggior rilievo all'elemento della violenza e della minaccia a terzi rispetto a quello della sottrazione patrimoniale.

Ecco dunque che, in tal caso, nella ricostruzione della concettualizzazione e delle modalità classificatorie, occorrerà segnalare questo elemento di differenza dovuto anche a ragioni storico – culturali, dato che il codice ceco è entrato in vigore nel 2009, mentre quello italiano è il codice Rocco del 1930. Questo dato di carattere storico mostra, a livello macroscopico, come la normativa sia il prodotto di due epoche profondamente diverse, con valori e politiche criminali radicalmente opposte.

Conclusioni

Il presente contributo vuol presentare un metodo di lavoro volto alla ricostruzione delle modalità di concettualizzazione sottese al dominio del diritto

penale. Il presupposto alla base del metodo adottato è che, in ottica contrattiva, non sia necessario comparare esclusivamente elementi linguistici ma sia necessario procedere alla comparazione concettuale. Per poter comparare concettualmente due realtà, però, occorre penetrare nelle modalità con cui tale concettualizzazione viene messa in atto nei singoli ordinamenti di riferimento, per poi incrociare i risultati ottenuti dalla ricostruzione. A tale scopo, si è ritenuto essenziale e preliminare procedere ad evidenziare quali siano le relazioni concettuali esistenti tra gli elementi presenti nei domini del diritto penale italiano e ceco, procedendo dal generale al particolare e prendendo come riferimento i codici penali.

L'approccio alle fonti, costituenti testi e dunque composte da materiale linguistico, è stato di tipo *knowledge - driven*, basato su una conoscenza pregressa dell'intero dominio a livello concettuale.

Il primo livello di analisi è stato quello di ricostruire la struttura generale del dominio del diritto penale partendo dal codice ed evidenziando le relazioni tassonomiche riscontrabili dal codice stesso. Si è poi selezionata una parte specifica del dominio dotata di una sua autonomia, ossia quella dei reati contro il patrimonio, mostrando il criterio di ripartizione di tali fattispecie in categorie nei due ordinamenti.

Il secondo livello di analisi, di cui si è dato un esempio mostrando l'analisi dettagliata della fattispecie di furto, è quello che concerne l'individuazione delle componenti essenziali delle norme e la successiva comparazione sulla base degli elementi riscontrati nelle fattispecie prese in considerazione. Nel procedere in tal senso, è parso di riscontrare forti analogie con la teoria semantica dell'analisi componenziale, basata in linea di massima sul fatto che il significato di determinate parole fosse ricostruibile sulla base di elementi costitutivi essenziali. Si è ritenuto di poter procedere ad un recupero di tale teoria in quanto il dominio di riferimento adotta un metodo analitico e sistematico nella concettualizzazione, come dimostrato attraverso la letteratura specialistica citata. Si parla di elementi costitutivi del reato, le singole fattispecie specifiche sono legate al concetto generale di reato da relazioni tassonomiche e dunque vige il principio di transitività per cui i concetti più specifici, costituiti dai singoli reati, condividono alcuni elementi del concetto più ampio. Tali elementi condivisi sono quelli presi in considerazione nell'analisi della fattispecie di furto, ossia gli elementi che, secondo la dottrina, sono parti essenziali del fatto tipico: soggetto attivo, eventualmente soggetto passivo, condotta, elemento soggettivo (dolo o colpa). L'esempio della fattispecie di rapina, inoltre, mostra come anche fattispecie aventi elementi costitutivi molto vicini e quasi sovrapponibili, possano essere oggetto di riconduzione ad una diversa categoria. Nel caso in esame, infatti, l'ordinamento ceco colloca la rapina tra i reati contro la libertà personale e non nei reati contro il patrimonio, circostanza dovuta

anche a ragioni storiche, vista la differenza tra le date di entrata in vigore dei due codici nei due ordinamenti (1930 in Italia, 2009 in Repubblica Ceca).

In ragione del fatto che le categorie di riferimento per il dominio sono create artificialmente dal legislatore al fine di rispondere all'esigenza di classificazione dei comportamenti umani per tracciare un confine tra lecito e illecito, si ritiene che l'attribuzione di un elemento ad una o ad un'altra categoria avvenga conformemente a caratteristiche comuni (per tornare ai reati contro il patrimonio, ciò che essi hanno in comune è l'offesa a beni economicamente valutabili nella disponibilità di un determinato soggetto).

In fase di comparazione tra due distinti ordinamenti, dove l'esigenza è quella di stabilire il livello di equivalenza tra norme di due diversi sistemi giuridici non per finalità applicative a casi singoli, bensì a scopo di traduzione, si ritiene di dover tener ferme le categorie previste dalla legislazione dei singoli ordinamenti e cercare, per quanto possibile, le somiglianze, tenendo ben presenti – ed evidenziando esplicitamente – le differenze.

Bibliografia

- Biel, Lucja. 2022. "Research into legal translation: an overview of the 2010 trends from the perspective of translation studies." In *Comparative Perspective on Law and Language*, edited by A. Parise e O. Moréteau. Eleven.
- Cornu, Gérard. 2005. *Linguistique juridique – Le vocabulaire juridique*. 3a ed. Montchrestien.
- Estes, William K. 1994. *Classification and cognition*. Oxford University Press.
- Garofoli, Roberto. 2022. *Compendio di diritto penale, parte generale e speciale*. Nel Diritto Editore.
- Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford University Press.
- Gémar, Jean-Claude. 2014. "Catching the spirit of Law: from translation to co-drafting." In *Comparative Law – engaging translation*, edited by S. Glarent. Routledge.
- Gémar, Jean-Claude. 2015. "De la traduction juridique à la jurilinguistique." *Meta* 60 (3): 476–93.
- Halford, Graeme S., William H. Wilson, Glenda Andrews, and Steven Phillips. 2014. *Categorizing Cognition: Toward Conceptual Coherence in the Foundations of Psychology*. MIT Press.
- Jelínek, Jiří, Zdeněk Sovák, Antonín Drašík, et al. 2017. *Trestní právo hmotné – obecná část, zvláštní část*. 6a ed. Nakladatelství Leges.
- L'Homme, Marie-Claude. 2020. *Lexical semantics for terminology – An introduction*. John Benjamins Publishing Company.

- Mauri, Caterina. 2021. "Ad hoc categorization in linguistic interaction." In *Building Categories in Interaction: Linguistic Resources at Work*, edited by C. Mauri, et al. John Benjamins Publishing Company.
- Padovani, Tullio. 2023. *Diritto penale*. 10a ed. Giuffrè Editore.
- Parise, Augustín. 2022. "Law and Language: a Plea for Experts Sans Frontières." In *Comparative Perspective on Law and Language*, edited by A. Parise e O. Moréteau. Eleven.
- Piergallini, Carlo. 2022. *Reati contro il patrimonio*, Giappichelli editore.
- Sacco, Rodolfo. 2010. "Azione, pensiero, parola nella creazione del diritto." In *Lingua e diritto. Livelli di analisi*, a cura di J. Visconti. LED.
- Šámal, Pavel, Tomáš Gřivna, Lukáš Bohuslav, et al. 2022. *Trestní právo Hmotné*. 9a ed. Wolters Kluwer.
- Šarčević, Susan. 1997. *New approach to legal translation*. Kluwer Law International.
- Taylor, John R. 2004. *Linguistic Categorization*. Oxford University Press.

Metadattazione delle tipologie di documenti clinici: un approccio alla standardizzazione mediante LOINC

Maria Teresa Chiaravalloti, Maria Teresa Guaglianone, Elena Cardillo*

Abstract: Achieving semantic interoperability in the healthcare information ecosystem is crucial for efficient data exchange, yet challenging due to the heterogeneous naming conventions used for clinical documents. Although there are systems such as LOINC® to provide unique identifiers, the heterogeneity of document names creates ambiguity and hinders their exchange and long-term preservation, especially in federated electronic health record systems. This paper describes the methodology adopted to standardize document metadata for naming document types within a national digital preservation system by mapping local codes to the LOINC® standard. The results demonstrate the advantages of using LOINC as a tool for standardizing document type names and achieving interoperability. Furthermore, the study's qualitative reflections contribute to the debate among experts about knowledge organization, digital preservation and categorization of clinical documents.

Keywords: Metadata, Clinical document, Standardization, Document types, LOINC.

1. Introduzione

L'interoperabilità tecnica e semantica fra sistemi informativi sanitari favorisce lo scambio di dati clinici e l'integrazione fra applicativi differenti. Affinché essa si realizzi, fondamentale è l'uso di sistemi standardizzati di codifica dei dati clinici. In essi i codici svolgono il ruolo di identificativo univoco di un concetto, il quale può essere, però, in realtà rappresentato mediante denominazioni differenti. Essendo declinazioni linguistiche, questi nomi hanno un

* Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Rende (CS), Italia. maria.chiaravalloti@iit.cnr.it. ORCID: 0000-0003-4695-2026; maria.guaglianone@iit.cnr.it. ORCID: 0000-0002-4070-5828; elena.cardillo@iit.cnr.it. ORCID: 0000-0001-5003-205X. Le autrici hanno lavorato collegialmente alla definizione concettuale e metodologica del contributo, coordinate da E. Cardillo in qualità di responsabile dell'Accordo di collaborazione Medas-CNR. Tuttavia, per la stesura e la revisione dei paragrafi le responsabilità sono così suddivise: M.T. Chiaravalloti, in qualità di prima autrice, ha curato la struttura generale dell'articolo e nello specifico la redazione dei paragrafi 1, 2.1, 3 e 4; M.T. Guaglianone ha curato la redazione dei paragrafi 2.2, 4 e 5; E. Cardillo ha curato la redazione dei paragrafi 1, 3 e 5.

certo margine di ambiguità interpretativa. L'eterogeneità nella denominazione dei documenti clinici è, difatti, un problema molto dibattuto in letteratura (Zuo et al. 2024; Dugas et al. 2009; Hyun et al. 2009), in quanto ostacolo primario all'interoperabilità semantica in un contesto quale, ad esempio, quello di un sistema di cartelle cliniche federate. La standardizzazione dei nomi dei documenti clinici è essenziale per l'uso, la gestione e lo scambio ottimale di documenti all'interno e tra istituzioni (Chen et al. 2010). Sistemi di questo tipo non sono solo Knowledge Organization Systems (KOSs) nel senso stretto del termine, ma di fatto organizzano e disciplinano nella pratica il flusso operativo del sistema informativo sanitario cui sono applicati. La tipologia di un documento clinico determina, infatti, chi può produrlo e consultarlo, ma anche i tempi e i modi relativi alla sua conservazione a lungo termine. Proprio nel contesto della conservazione dei documenti informatici (Agenzia per l'Italia Digitale 2020), siano essi clinici o amministrativi, gli standard di metadatozione permettono l'identificazione univoca del tipo di documenti posti in conservazione (ad esempio, cartella clinica, fattura, verbale di pronto soccorso, referto di laboratorio, ecc.), al fine di favorire una loro corretta archiviazione e recupero, qualora ne venga richiesto l'accesso. Non essendoci una indicazione normativa univoca in tal senso, generalmente i privati che offrono servizi di conservazione a lungo termine in outsourcing gestiscono il metadato DocumentType secondo il contesto di riferimento e le esigenze dei clienti (strutture sanitarie, enti, ecc.) che versano in conservazione i documenti. Questo ha portato, quindi, a un proliferare di sistemi locali di denominazione e codifica, i quali rendono di conseguenza i dati e i metadati, e in generale il master coding index utilizzato nel sistema di conservazione, non interoperabili, ostacolando, di fatto, la possibilità di migrazione ad altro conservatore prevista dalle Linee Guida AgID (Agenzia per l'Italia Digitale 2020). Il confronto con uno dei Conservatori qualificati nel panorama nazionale ha, inoltre, permesso di rilevare un'ulteriore problematica dal punto di vista semantico, attinente alla difficoltà di identificare la tipologia del documento sulla sola base della sua denominazione e alla correlata resistenza dei clienti ad inquadrare le proprie tipologie documentali in etichette predeterminate e non omonime rispetto ai nomi dei propri documenti.

Il fine del presente studio è quello di mappare verso LOINC®, acronimo di Logical Observation Identifiers Names and Codes, il sistema di codifica aziendale del Conservatore per l'identificazione dei documenti clinici posti in conservazione, al fine di renderne l'organizzazione più coerente e semanticamente interoperabile e gestire la variabilità nelle denominazioni dei documenti clinici provenienti da diversi soggetti produttori. Inoltre, l'utilizzo dello standard LOINC dà la possibilità di ereditare direttamente dai metadati di indicizzazione nel FSE il codice relativo alla tipologia documentale, in quanto, come previsto dall'Affinity Domain Italia (AD) (Fascicolo Sanitario Elettroni-

co n.d.), il metadato *XDSDocumentEntry.typeCode* prevede l'uso di un value set LOINC, attualmente costituito da 33 codici. Questo numero rappresenta una minima parte delle molteplici tipologie di documenti clinici gestiti dalle strutture sanitarie. Pertanto, questo lavoro apre la strada alla definizione di potenziali nuovi codici LOINC per l'identificazione di nuove tipologie di documenti clinici nel value set del citato metadato di AD.

Il presente contributo ha l'obiettivo di descrivere la metodologia adottata e i risultati ottenuti nell'ambito dell'attività di collaborazione scientifica con il suddetto Conservatore, una delle più note aziende del settore privato che realizza servizi di conservazione in outsourcing, con specializzazione nell'ambito sanitario. Il lavoro svolto ha riguardato la valorizzazione del metadato *DocumentType* all'interno del sistema di conservazione aziendale mediante lo standard internazionale LOINC con il fine ultimo di standardizzare il proprio catalogo a beneficio del recupero dei documenti e dell'interoperabilità semantica del proprio sistema¹. Questa attività ha permesso di formulare una serie di riflessioni teoriche e metodologiche circa l'organizzazione e la rappresentazione della conoscenza nel dominio dei nomi assegnati ai documenti clinici, che vengono presentate in questo lavoro al fine di stimolare la discussione proattiva fra gli esperti di settore: clinici, conservatori, documentalisti, archivisti. Una prima mappatura del metadato *DocumentType* verso lo standard LOINC era stata realizzata nel 2021 in un lavoro di tesi e a seguito di un periodo di tirocinio presso il suddetto Conservatore (Mezzatesta et al. 2021). Questo lavoro è stato recentemente rivisto ed evoluto, in considerazione delle release che aggiornano semestralmente lo standard LOINC e delle molteplici novità tecniche e normative introdotte in Italia in merito all'indicizzazione dei documenti nel Fascicolo Sanitario Elettronico (FSE) (Pasceri e Chiaravallotti 2025). Al contempo, l'Istituto di Informatica e Telematica del Consiglio nazionale delle Ricerche (IIT-CNR) ha avuto modo di formalizzare un accordo di collaborazione scientifica con il suddetto Conservatore, al fine di contribuire con le proprie competenze specialistiche, anche in qualità di partner italiano ufficiale di LOINC, a formalizzare e finalizzare l'implementazione dello standard. Questa attività ha permesso anche di formulare una serie di proposte metodologiche di soluzioni condivise per l'univoca identificazione dei documenti clinici sia in fase di conservazione, che di formazione e gestione degli stessi, ribadendo così la complementarità di questi tre momenti della vita del documento informatico. Il seguito del contributo è organizzato secondo la seguente suddivisione in paragrafi: nel paragrafo 2 sono discussi i principali riferimenti di letteratura; il paragrafo 3 descrive i materiali e la metodologia di

¹ L'attività rientra in un processo di revisione e standardizzazione più ampio che include anche l'analisi e mappatura ad altri metadati (sia di tipo clinico che amministrativo) utilizzati nel sistema di conservazione documentale aziendale, ma nel presente lavoro ci si focalizza sui risultati del processo di mappatura al LOINC.

lavoro; il paragrafo 4 riporta i principali risultati ottenuti ed apre ad una serie di riflessioni teoriche e metodologiche, che sfociano in conclusioni e prospettive future, presentate nel paragrafo 5.

2. Letteratura di riferimento

2.1. Contesto nazionale ed internazionale sulla standardizzazione delle tipologie documentali in ambito sanitario

In Italia, l'unico riferimento ad una standardizzazione delle tipologie di documentali in ambito sanitario è presente all'interno del documento di AD, il quale definisce gli standard di codifica da adottare per garantire l'interoperabilità tra i sistemi regionali di FSE, con particolare attenzione ai sistemi di codifica clinica previsti dal DPCM sul FSE (178/2015). Esso include inoltre specifiche codifiche amministrative necessarie per valorizzare correttamente i metadati nelle transazioni Integrating the Healthcare Enterprise (IHE), impiegate nello scambio di messaggi tra i sistemi FSE regionali. Le tipologie di documenti indicizzabili nel FSE sono univocamente identificate mediante codici LOINC all'interno del metadato DocumentType, al quale è possibile collegare mediante relazione gerarchica di dipendenza il metadato EventCodeList, che permette di specificare ulteriormente il tipo di documento indicizzato, i.e. la tipologia di prestazione specialistica refertata. I valori associabili a questo metadato appartengono a diversi standard di codifica, fra cui anche LOINC. Questi metadati, opportunamente combinati con gli altri indicati nel documento, consentono al personale sanitario autorizzato, per il tramite dell'Infrastruttura Nazionale di Interoperabilità (INI), la ricerca ed il recupero dei documenti indicizzati nei sistemi di FSE regionali, nel rispetto della normativa sulla privacy e con il consenso dell'interessato. Similmente, i metadati di indicizzazione dei documenti nel FSE sono anche il perno attorno a cui ruota l'interoperabilità transfrontaliera tra gli Stati membri dell'Unione Europea. Al fine di facilitare la mobilità dei pazienti e la continuità delle cure in Europa sono, difatti, molte le iniziative progettuali nate per coordinare gli sforzi dei singoli verso una meta comune, i.e. eHealth Network, MyHealth@EU, eHDSI, X-eHealth e Joint Action TEHDAS.

La stessa interoperabilità semantica non si riscontra, tuttavia, quando si scende a livello locale, ossia della singola struttura sanitaria. Seppur esiste, infatti, una semantica condivisa, derivante da normativa e consuetudini linguistiche di utilizzo, per denominare i documenti clinici, di fatto poi nella realtà esistono sfumature lessicali nelle denominazioni di documenti identici per contenuto, ma anche differenze semantiche nei contenuti di documenti denominati allo stesso modo. Questa variabilità potrebbe generare *misclassification* di documenti, tanto in fase di indicizzazione nel FSE quanto in fase di

metadattazione per la conservazione a lungo termine. Da queste motivazioni e con questi presupposti nasce il presente studio.

Numerose sono le pubblicazioni che hanno affrontato il tema della rappresentatività dei codici LOINC rispetto a subset documentali di ambito clinico più o meno ampi, con metodi differenti e giungendo talvolta a proposte di miglioramento concrete anche nei confronti dello standard (Dugas et al. 2009; Hyun et al. 2009; Chen et al. 2010). In considerazione della loro rilevanza per via dell'applicazione di metodologie più innovative, sono di seguito argomentati solo gli studi più recenti. In uno studio sulla mappatura automatica di documenti clinici, provenienti da cinque differenti istituzioni, alla LOINC Document Ontology (DO) è stato sviluppato un sistema basato su modelli linguistici che, basandosi solo sulla denominazione e senza accedere al contenuto dei documenti, ha raggiunto un'accuratezza del 90% (Zuo et al. 2024). Similmente al presente contributo, questa analisi è stata finalizzata a standardizzare le denominazioni dei documenti clinici per rendere più puntuale la loro condivisione ed ha evidenziato la copertura attuale della LOINC DO e le potenziali aree di estensione. Maggiormente innovativo uno studio che propone un framework per la categorizzazione automatica dei documenti clinici, mappandoli sulla LOINC DO tramite un approccio *Bag of Words* e metadati strutturati. Dopo una fase iniziale di training con mappatura manuale, il sistema ha raggiunto una copertura del 73,4% classificando 132 milioni di documenti in meno di 2 ore, e dimostrandosi quindi più efficiente rispetto ai metodi basati su Natural Language Processing (Khan et al. 2024). In Australia la piattaforma "My Health Record" (Pearce and Bainbridge 2014), che, garantisce a tutti gli australiani di accedere a un summary delle proprie informazioni sanitarie online e in modo sicuro, utilizza i metadati document class e document types come elementi critici per migliorare la ricerca e l'individuazione di documenti clinici, applicando LOINC e il National Clinical Terminology Service (NCTS), basato sullo standard Fast Healthcare Interoperability Resources (FHIR), ai fini della loro valorizzazione e codifica.

2.2. Il metadato DocumentType come mediatore nella standardizzazione della denominazione dei documenti sanitari

Il metadato DocumentType rappresenta un elemento fondamentale per promuovere la comunicazione e l'interoperabilità tra ecosistemi informativi sanitari frammentati. Grazie al suo ruolo di facilitatore esso richiama i concetti di *boundary object* (Star 1989) e *boundary factors* (Bjørn et al. 2009), i quali si configurano in questo contesto come validi e pertinenti strumenti analitici per comprendere la sua natura e la sua funzione.

Il concetto di *boundary object* – utilizzato dalla sociologia e dagli studi scientifici e tecnologici – descrive un artefatto che è sufficientemente plastico da adattarsi alle esigenze e ai vincoli locali di comunità eterogenee pur mantenendo la stabilità e la robustezza necessarie per conservare un'identità comune, così da rimanere riconoscibile e utile in contesti diversificati. Il DocumentType può essere interpretato come un *boundary object* nell'informatica medica, poiché, dato il carattere notoriamente frammentato dell'assistenza sanitaria, l'implementazione del metadato DocumentType mira al superamento di una logica operativa a “compartimenti stagni” tra sistemi eterogenei, attraverso un nucleo condiviso, nel quale le differenze locali possono essere ricondotte.

La standardizzazione del DocumentType non si configura come un'operazione meramente tecnica, bensì come un fenomeno complesso e multidimensionale, le cui implicazioni derivano da pratiche quotidiane, quadro normativo, dinamiche organizzative e socioculturali. Essendo lo stesso LOINC uno standard tecnico, la dimensione immediatamente più evidente è in questo caso quella tecnica, che include la definizione dei codici, la granularità degli stessi e delle loro definizioni, le regole di implementazione e la compatibilità tra diverse piattaforme tecnologiche. La dimensione legislativa si concretizza nell'insieme di leggi e regolamenti nazionali e internazionali che rivestono un ruolo primario nella standardizzazione dei dati sanitari per la realizzazione di requisiti legali inerenti all'archiviazione, all'accesso e alla conservazione. La dimensione prasseologica è incarnata dalla pratica clinica, riflettendo le esigenze operative dei professionisti sanitari che quotidianamente creano, utilizzano e interpretano i documenti, categorizzandoli per finalità di assistenza al paziente, comunicazione interprofessionale e organizzazione dei flussi di lavoro. L'integrazione della standardizzazione del documentType nella realtà organizzativa sanitaria introduce inevitabilmente anche una dimensione organizzativa, poiché ogni struttura sanitaria possiede propri flussi di lavoro, politiche di gestione dei documenti e priorità strategiche ed economiche. Infine, la dimensione socioculturale, sebbene più sottilmente percepibile, è tuttavia pervasiva, poiché l'adozione di un nuovo metodo per denominare le tipologie documentali può scontrarsi con routine consolidate o con la percezione che l'informatizzazione e la standardizzazione limitino l'autonomia professionale.

Ciascuna di queste dimensioni, dunque, contribuisce alla configurazione finale del processo di standardizzazione e adeguamento del metadato documentType, fungendo sia da contesto sia, talvolta, da vincolo. La loro coesistenza e interazione generano tensioni intrinseche nella realizzazione della standardizzazione della categorizzazione dei documenti sanitari. La tensione più immediatamente evidente si manifesta tra la dimensione prasseologica e quella tecnica. Gli operatori sanitari potrebbero aver bisogno di documentare sfumature che uno standard tecnico, per sua natura, potrebbe non rappresentare con il livello di dettaglio atteso, portando a una percezione di perdita di

granularità o di significato a livello locale. Questa dinamica illustra la sfida posta dai *boundary factors* e dalle *contextual contingencies* che intervengono nel processo di standardizzazione mirando a individuare, rispettivamente, gli elementi che devono essere stabili e condivisi e quelli che, invece, possono essere localizzati e riconfigurabili (Bjørn et al. 2009). Questi concetti evidenziano come la standardizzazione di un metadato come il *documentType*, sebbene tecnicamente ben definito, possa fallire qualora non tenga conto di fattori contestuali e vincolanti. A titolo esemplificativo, nell'ambito della conservazione a lungo termine, si osserva come, per uno stesso conservatore, le implementazioni della valorizzazione del *documentType* possano variare significativamente a seconda del settore dei clienti serviti (ad es. sanità vs. pubblica amministrazione) e, in particolare, delle specifiche esigenze contrattuali. A tal fine, la soluzione adottata è spesso quella di fornire *valueset* personalizzati per il cliente, provvedendo alla mappatura a standard quando necessario. Esiste poi la tensione tra la dimensione legislativa e le dimensioni prasseologica e organizzativa. Le normative, infatti, possono imporre requisiti non perfettamente allineati con le capacità organizzative esistenti.

Infine, anche la tensione tra la dimensione socioculturale e tutte le altre dimensioni citate può rappresentare un forte ostacolo all'adozione di un sistema standard. La resistenza al cambiamento, la preferenza per terminologie locali consolidate, la sfiducia nell'imposizione di standard o la percezione che l'informatizzazione e la standardizzazione burocratizzino la pratica clinica potrebbero minare l'efficacia di uno standard, per quanto esso sia intrinsecamente valido.

In sintesi, la standardizzazione del campo *documentType* rappresenta un microcosmo in cui coesistono dinamicamente diverse dimensioni e nel quale si manifestano le principali complessità della digitalizzazione in ambito sanitario. Il successo delle politiche di standardizzazione risiede nelle capacità intrinseche di un *boundary object*, nel caso attuale LOINC e in particolare il metadato *documentType*, di gestire e mediare le tensioni esistenti tra le diverse dimensioni, offrendo una soluzione universalmente riconoscibile e localmente adattabile. È auspicabile, dunque, che la standardizzazione si configuri come un processo di negoziazione, cooperazione e co-costruzione in relazione ai contesti di utilizzo.

3. Materiali e metodologia

LOINC è uno standard internazionale di codifica di osservazioni di laboratorio, osservazioni cliniche e tipologie di documenti clinici. Tralasciando le prime due finalità di codifica, non rilevanti in questa sede, la LOINC DO è uno strumento che permette la standardizzazione della denominazione dei documenti clinici, utilizzando un approccio multiasse, e funge da framework per

una denominazione dei documenti coerente e interoperabile tra sistemi diversi (Li et al. 2011). Nello specifico, la LOINC DO si compone di identificativi univoci frutto della combinazione di 5 assi chiave (si veda un esempio in Fig. 1): il tipo di documento (referto, lettera, ecc.), il tipo di servizio (consulto, valutazione, ecc.), la branca specialistica di afferenza (cardiologia, chirurgia, ecc.), il ruolo dell'autore del documento (medico, infermiere, ecc.) e il contesto clinico di produzione (ricovero, emergenza, ecc.).

Subject Matter Domain		Role	Setting	Type of Service	Kind of Document
Part Number	Part Name				
LP263512-8	Addiction medicine				
LP172912-8	Addiction psychiatry				
LP172915-3	Aerospace medicine				
LP345851-9	Allergy				
LP172918-7	Anesthesiology				
LP172981-5	Pain medicine				
LP248728-8	Sleep medicine				
LP172919-5	Audiology				
LP439983-8	Child life				
LP172923-7	Chiropractic medicine				
LP172933-6	Critical care medicine				
LP172934-4	Dentistry				
LP172935-1	Dermatology				
LP172987-2	Pediatric dermatology				
LP172948-1	Emergency medicine				

Figura 1. Estratto dell’asse Subject Matter Domain della LOINC DO, ossia la branca specialistica di riferimento.

Un sesto asse relativo al metodo viene specificato solo se dirimente rispetto ad altri nella produzione della tipologia di documento. Lo standard LOINC è implementato in Italia da diversi anni (il suo utilizzo è stato introdotto dal citato DPCM sul FSE (178/2015), grazie anche al Gruppo di lavoro LOINC Italia (LOINC Italia n.d.), afferente all’IIT-CNR, che ne cura la versione ufficiale in lingua italiana e fornisce servizi di training e validazione delle map-pature allo standard su tutto il territorio nazionale. La versione corrente dello standard è la 2.80, rilasciata a febbraio 2025, contenente un totale di 104.672 codici. Ai fini del presente contributo è stata utilizzata la versione 2.72 (com-

prendente 98.268 LOINC terms), la più recente disponibile al momento dello svolgimento del lavoro di mappatura. Tuttavia, questo non rappresenta una criticità poiché ad ogni release LOINC è rilasciato il file “MapTo”, che traccia gli aggiornamenti ai codici esistenti e, quindi, consente di mantenere valide le mappature effettuate.

L’approccio metodologico adottato ha, pertanto, previsto dapprima un’accurata analisi della documentazione inerente al sistema di conservazione proprio del Conservatore, con una specifica attenzione alla strutturazione del suo sistema di metadattazione e alle modalità di costruzione delle relative tabelle di codifica. Successivamente, è stata effettuata una valutazione dei precedenti lavori di mappatura citati in premessa, al fine di valutare il loro grado di maturità e stato di attuazione, anche in relazione agli aggiornamenti dello standard ed al rinnovato contesto normativo. Lo sforzo è stato quindi diretto a colmare le lacune di mappatura riscontrate, soprattutto mediante continui confronti con il Conservatore.

Per ridurre l’eterogeneità delle denominazioni dei documenti provenienti da soggetti produttori differenti, il Conservatore ha realizzato un proprio sistema di transcodifica così da ricondurre a univocità documenti che, seppur denominati in modo differente, hanno lo stesso contenuto semantico.

Esso si basa sulla distinzione tra *genus* e *species*, dove *genus* rappresenta il *core* della tipologia documentale (ad esempio, referto), mentre *species* la sua declinazione specialistica (ad esempio, cardiologia). Essi corrispondono, quindi, a due dei sei assi della LOINC DO sopra descritti: tipo di documento e branca specialistica di appartenenza. Sia *genus* che *species* sono rappresentati da una notazione numerica a tre cifre e, quando combinati, sono separati da un punto fermo. Tuttavia, il sistema notazionale non è univoco, in quanto si può riscontrare la stessa sequenza numerica in oggetti diversi del sistema di codifica. Per finalizzare l’implementazione di LOINC per la gestione del metadato documentType sono state quindi considerate, come base di partenza, le tabelle di corrispondenza già predisposte nell’ambito del suddetto lavoro di tesi e rese disponibili dal Conservatore. Si è deciso di revisionare anche le mappature dei codici precedentemente identificati come matching esatto, in considerazione delle evoluzioni dello standard LOINC nel tempo intercorso fra il lavoro di tesi e l’ultima release disponibile, così da avere infine un insieme di mappature più coerente ed esteso. Quindi è stata creata una tabella di corrispondenze (di cui si può vedere un estratto esemplificativo in Fig. 2), contenente i seguenti campi:

- Local Code: codice associato alla tipologia documentale;
- Local Label: label associata alla tipologia documentale;
- Local Description: descrizione della casistica di utilizzo e/o del contesto di produzione della tipologia documentale;

- commenti riconciliazione: eventuali annotazioni circa la mappatura/ assenza di mappatura tra un local code e un codice LOINC, nonché suggerimenti di aggiunta, eliminazione, adeguamento di code e/o label, ecc.;
- commenti riconciliazione normalizzati: uniformazione dei commenti per motivi computazionali;
- LOINC Code: codice LOINC verso cui è stata effettuata la mappatura;
- LOINC Long Common Name: denominazione associata al codice LOINC mappato.

D	E	F	G	H	I	J	K	L
concat servizio	code	label	description	Commenti riconciliazione	not	Commenti riconciliazione NORMALIZZATI	LOINC CODE	LOINC COMMON NAME
020.170 - Certificato di ricovero	020.170	Certificato di ricovero	Ricovero	ok 86532-9 Admission notification note		ok	86532-9	Admission notification note
020.180 - Certificato Assistenza al Parto (CedAP)	020.180	Certificato Assistenza al Parto (CedAP)	Assistenza al parto	che info contiene questo doc?		che info contiene questo doc?		
020.190 - Attestazione parto	020.190	Attestazione parto	Attestazione di parto	che info contiene questo doc?		che info contiene questo doc?		

Figura 2. Esempio di mappatura dei document type locali verso i codici LOINC, con sistema di colori per la validazione della correttezza delle associazioni effettuate.

Il catalogo locale dei document type codes consta di 479 record. Tuttavia, rispetto al totale, 46 codici sono stati identificati come “not document type”, in quanto identificativi di parametri, misurazioni e relazioni. Nel riconsiderare le mappature precedentemente eseguite, sia quelle esatte, che quelle verso codici LOINC di significato più ampio o più specifico, ci si è tuttavia resi conto della presenza di document type simili nelle denominazioni, ma di fatto assimilabili nella semantica, ad esempio, *Attestazione parto* e *Attestazione di avvenuta nascita*. Da un confronto con il Conservatore è emerso che questa sovrapposizione può essere dovuta a successive richieste di aggiunta da parte dei clienti di etichette di tipologie documentali maggiormente rispondenti alle proprie. Questo ha fatto ipotizzare un loro utilizzo limitato nella pratica e per tal motivo è stato di concerto deciso di effettuare una valutazione sulla base di un criterio di necessità, determinato dalle reali occorrenze, ovvero dal numero di documenti conservati nei diversi impianti del Conservatore, descritti con un determinato document type. Si è ritenuto, quindi, di impostare a 10 occorrenze documentali (su oltre 200 milioni di documenti conservati) la soglia di utilizzo di ciascun document type code da considerare per la mappatura. Questo ha portato ad escludere in questa fase 278 record. Il lavoro si è pertanto focalizzato sull’analisi di 186 codici di document type, per i quali sono state

realizzate le mappature verso codici LOINC con diversi gradi di corrispondenza semantica o revisionate le mappature precedentemente realizzate. Nello specifico, rispettando i principi di mapping e in continuità con i molteplici studi relativi alla valutazione della copertura semantica della LOINC DO, i mapping sono stati qualificati nel seguente modo:

- mapping esatto (equivalenza) ossia con corrispondenza 1:1 e semanticamente completa fra il concetto di partenza ed il concetto di destinazione;
- mapping parziale ossia con corrispondenza 1:n o n:1 e semanticamente non completa fra il concetto di partenza ed il concetto di destinazione, in quanto quest'ultimo è troppo generico o troppo specifico rispetto al primo;
- nessun mapping ossia mancata corrispondenza fra il concetto di partenza ed il concetto di destinazione.

I document type che non è stato possibile mappare e quelli per cui è stato identificato un mapping parziale sono stati sottoposti a riesame per determinare se: i) il document type possa essere chiarito o rinominato in modo che esso possa essere nuovamente mappato; ii) nel caso di un termine del document type concatenato o eccessivamente complesso, sia possibile scomporlo senza perdere il contesto; iii) il document type abbia davvero esigenza di essere mappato e il perché o quali possano essere gli effetti collaterali nel caso in cui resti non mappato (oltre all'interoperabilità); iv) prima di effettuare una *new term submission request* alla LOINC Committee, sia possibile creare un caso d'uso da parte del Conservatore che supporti tale richiesta. Laddove, al termine di questo riesame, i document type risultino ancora non mappati, si avvia un processo di formalizzazione secondo le regole di denominazione LOINC, per essere sottomessi come richieste di aggiunta di nuovi codici allo standard. È, quindi, prevista ad ogni nuova release dello standard una fase di manutenzione che verifichi l'aggiornamento delle mappature mancanti o parziali mediante i nuovi codici creati.

4. Risultati e discussione

La collaborazione scientifica fra esperti di gestione e conservazione documentale, di organizzazione della conoscenza e referenti di LOINC Italia ha consentito di condurre lo studio presentato in maniera interdisciplinare, ottenendo risultati interessanti sia in termini quantitativi che qualitativi. Essi consentono di trarre alcune considerazioni metodologiche riguardo l'organizzazione e la strutturazione delle conoscenze relative alla definizione dei tipi di documenti clinici attualmente prodotti, gestiti e conservati nelle strutture sanitarie italiane.

L'analisi di tipo quantitativo ha permesso di discriminare tra: document type code mappati verso un codice LOINC, secondo tre diversi gradi di copertura semantica; candidati per *new term submission request* alla LOINC Committee; document type code non considerati in questa fase, ma soggetti a una rivalutazione da parte del Conservatore perché ambigui o attestantisi sotto la soglia delle 10 occorrenze documentali ad essi ricollegate; codici locali da non considerare perché non identificanti document type ma parametri osservabili o relazioni fra gli oggetti del sistema di codifica del Conservatore (Tab. 1).

Metadata Medas	N. Record originali	N. Mapping effettuati	N. record per new LOINC submission requests	N. record not considered document type	N. Record non considerati	N. record da riconsiderare
documentType	479	82	64	46	278	9

Tabella 1. Dati statistici relativi ai mapping dei DocumentType code del Conservatore verso LOINC.

Dei 479 record del file fornito dal Conservatore: 82 sono stati mappati verso un codice dello standard LOINC; 64 sono stati identificati come candidati per *new term submission request* alla LOINC Committee in quanto mancante la corrispondenza semantica dei concetti o, se presente, a volte più generica a volte più specifica rispetto al concetto espresso dal documentType del Conservatore; 9 sono i record da riconsiderare insieme al Conservatore per approfondire il contenuto semantico dei documenti e quindi se eventuali mappature proposte possono o meno essere corrette; 46 quelli che nell'ottica dello standard sono da considerarsi parametri osservabili o misurabili o relazioni fra gli oggetti del sistema di codifica piuttosto che vere e proprie tipologie documentali; 278 i codici di document type esclusi per numero di occorrenze documentali ad esso associate inferiore a 10. Sono di seguito riportati degli esempi significativi per ciascuna delle categorie sopra elencate, con le relative argomentazioni del caso:

- mapping esatto (1:1): *Referto di ECG* è stato mappato verso il codice LOINC 11524-6 *EKG study*;
- mapping parziale:
 - corrispondenza 1:n ossia document type più specifico rispetto al document type LOINC: *Referto di endocrinologia e malattie del ricambio* è più specifico rispetto al codice LOINC 34898-7 *Endocrinology Note*, per cui è necessario richiedere un nuovo codice che esprima anche il concetto “malattie del ricambio”;
 - corrispondenza n:1 ossia document type più generico rispetto al document type LOINC: *Referto di chirurgia toracica* è più generico rispetto al codice LOINC 5436-4 *Cardiothoracic surgery Note*, per

cui è necessario richiedere un nuovo codice che esprima solo il concetto “chirurgia toracica”.

- nessun mapping: *Referto di neuropsichiatria infantile* non trova alcuna corrispondenza in LOINC, in quanto manca la definizione della branca specialistica nell’asse deputato;
- record identificati come “da riconsiderare”:
 - documentType code del Conservatore per i quali è necessario valutare un esempio di documento clinico opportunamente anonimizzato, onde definirne l’area semantica di riferimento in base al contenuto, al fine di ricondurli ad un eventuale codice LOINC esistente o di includerli fra quelli candidati per la richiesta di aggiunta di un nuovo codice alla LOINC Committee;
 - documentType code riferibili a schede di valutazione, che secondo lo standard LOINC non sono documenti clinici in sé e per sé, ma sono considerati osservazioni in forma di panel, ossia multiple di test di laboratorio/misurazioni cliniche/domande di questionari standardizzati a livello internazionale e/o riconosciuti dalla comunità scientifica di riferimento. Pertanto, non si ritengono considerabili per il mapping di questa tipologia di metadato.

Molti dei document type identificati come nuovi codici da richiedere sono stati sottomessi. Di questi, alcuni sono stati creati come new LOINC terms, ad esempio il document type *Lettera di fine trattamento* ha dato luogo al codice LOINC 101136-0 *Radiation oncology End of treatment letter*, mentre altri sono in fase di valutazione e/o lavorazione da parte della LOINC Committee.

L’analisi quantitativa della copertura semantica dello standard rispetto alle tipologie documentali da codificare presenta percentuali di successo inferiori rispetto agli studi citati nel paragrafo 2.1. Difatti, è stato possibile mappare verso LOINC solo il 53% dei document type di partenza, il 6% sono mappature dubbie, mentre il 41% di essi rappresentano tipologie di documenti per le quali è richiesta la modellazione di nuovi codici specifici.

L’analisi qualitativa dei risultati ottenuti ha permesso di formulare una serie di considerazioni metodologiche migliorative del sistema di classificazione delle tipologie documentali in uso presso il Conservatore ed alcune riflessioni riguardo all’organizzazione delle informazioni in settori in rapida evoluzione, come quello sanitario soprattutto dopo l’avvento del digitale e gli apporti scientifici derivanti da periodi di crisi quale quello pandemico.

Una delle principali problematiche riscontrate durante la mappatura, presente anche per altri tipi di metadati analizzati nel corso del progetto, è l’uso di uno schema notazionale di base non univoco, in quanto può essere che lo stesso codice sia utilizzato senza distinzione all’interno di metadati differenti per indicare valori associati a concetti diversi. Ciò contrasta nettamente con uno dei *desiderata* fondamentali di un sistema di codifica (Cimino 1998), ossia

che la notazione sia univoca ed il suo valore sia pienamente espresso dal codice in sé e non dipendente dal metadato cui è associata. Nel caso del metadato *documentType* del Conservatore, la notazione esprime nella prima parte la tipologia di documento clinico e nella seconda parte la branca specialistica cui essa fa riferimento. Tuttavia, la seconda parte della notazione non è sempre univocamente riferita alla stessa classe di specialità.

Questa scelta non comporta problemi nel momento in cui il codice è sempre accompagnato dalla sua etichetta di denominazione; tuttavia, limita la potenzialità espressiva del sistema notazionale. Sono state, quindi, formalizzate e fornite al Conservatore delle indicazioni che suggeriscono di rendere, innanzitutto, le notazioni univoche in tutto il sistema di conservazione e non solamente nell'ambito del singolo metadato, prevedendo quindi la definizione di un master coding index coerente e ben strutturato. Secondariamente, sono stati evidenziati i vantaggi derivanti dalla costruzione di un albero classificatorio che renda il codice notazionale parlante. Tuttavia, anche in questo caso, devono essere soppesati i vantaggi derivanti dalla classificazione gerarchica con il rischio di un potenziale raggiungimento di un limite espressivo nell'espansione dei rami foglia. Ciò è vero soprattutto nel settore sanitario, che, grazie all'avvento delle moderne tecnologie, sta attraversando un periodo di grandi cambiamenti e innovazioni. È necessario, inoltre, considerare che il processo di mapping richiede un importante sforzo iniziale da parte del Conservatore, ma minimo per assicurare poi una costante manutenzione della correttezza e della validità delle mappature. A tal fine, è importante prevedere la figura di un responsabile che abbia il compito di preservare l'integrità dei dati mappati, fornendo al contempo aggiornamenti tempestivi e completi. Inoltre, si potrebbe automatizzare il processo di aggiornamento prevedendo il collegamento ad un servizio terminologico integrato di dominio.

Dai risultati descritti risulta evidente che non è sempre possibile realizzare una completa progettazione a priori di una terminologia o catalogo standard prima del suo utilizzo nello specifico contesto di destinazione. Questa problematica potrebbe essere minimizzata adottando una visione incentrata sui citati *boundary object* e valutando la necessaria e vantaggiosa distinzione tra *boundary factors* e *contextual contingencies*. Il conflitto tra standardizzazione e personalizzazione è, dunque, un processo di negoziazione tra gli stakeholders coinvolti, che hanno interesse a raggiungere un accordo su uno specifico artefatto. Non è essenziale che si concordi su ogni dettaglio, piuttosto è necessario che il raggiungimento della soluzione sia percepito come un processo di negoziazione in cui il nucleo del sistema (terminologico, in questo caso) sia concordato e riconosciuto come stabile, mentre le specificità locali siano identificate e riconfigurate a livello locale (Bjørn et al. 2009). La stessa standardizzazione dei nomi e dei codici delle tipologie documentali attraverso il processo di map-

patura a LOINC è possibile solo se realizzata in stretta sinergia con il Conservatore e si configura come un processo dinamico e in continua evoluzione.

5. Conclusioni e prospettive future

Il presente contributo descrive un caso di studio sulla standardizzazione del metadato documentType nel contesto della conservazione digitale a lungo termine dei documenti sanitari e pone l'attenzione sugli aspetti che rendono fondamentale realizzare un processo di standardizzazione in determinati ambiti e su quanto sia importante che questo avvenga in completa sinergia con i destinatari ed utilizzatori dei sistemi stessi. I risultati ottenuti hanno messo in evidenza quanto sia essenziale un processo di mappatura dei codici localmente utilizzati verso codici universalmente riconosciuti, al fine di garantire un efficiente iter di descrizione, organizzazione, indicizzazione, conservazione, ricerca e recupero dei documenti da repository eterogenei, anche nel lungo periodo.

Il sistema di codifica LOINC si conferma il più idoneo a rappresentare questa tipologia di dati, nonostante le lacune o le imprecisioni semantiche riscontrate e descritte nei precedenti paragrafi, in quanto è lo standard di riferimento di settore; quindi, agevola l'interoperabilità semantica e la portabilità dei dati, ed è aggiornabile su richiesta degli utenti con un processo semplice e rapido, se comparato ad altri sistemi di codifica dei dati clinici (ICD-10/11, ICPC-3, SNOMED CT). I vantaggi della mappatura verso LOINC risultano evidenti e comprendono: la riduzione dell'ambiguità attraverso l'identificazione univoca e coerente dei documenti in termini di contenuto informativo e finalità; la realizzazione dell'interoperabilità, poiché document type standardizzati favoriscono la condivisione di contenuti all'interno dell'ecosistema digitale dei dati e dei documenti sanitari; il recupero dei dati in modo efficiente; il supporto all'analisi dei dati e alla produzione della reportistica in diversi contesti sanitari.

Simili interventi sono migliorativi dell'efficienza del processo di metadattazione dei documenti clinici nella conservazione a lungo termine, con evidenti ricadute anche sulle strutture sanitarie che usufruiscono dei servizi di conservazione. Il cambiamento di approccio del Conservatore coinvolto, finalizzato alla standardizzazione del metadato documentType, supportato anche dalle best practices in uso per la realizzazione dell'interoperabilità semantica, rappresenta un valido test case che ha condotto a soddisfacenti risultati e che si configura come facilmente replicabile da parte altri enti Conservatori, al fine di favorire la standardizzazione a livello nazionale dei document type.

La principale prospettiva di ricerca del presente studio riguarda la possibilità di evolvere la metodologia, che ha previsto finora l'uso di un approccio manuale, verso un approccio automatico di mapping, sfruttando tool specifi-

catamente basati sia su regole di codifica specifiche per LOINC che su Large Language Models (LLMs).

Ringraziamenti

L'attività è stata finanziata dall'Accordo di Collaborazione Scientifica Medas – IIT-CNR, prot. IIT CNR n. 0004358/2021 del 29/06/2021 (CUP B25F21000800007). Un doveroso ringraziamento va a Umberto Ferri e Carlo Alzati di Medas Srl per aver supportato lo studio ed aver condiviso i dati utilizzati ai fini della mappatura. Un ulteriore ringraziamento va a Matteo Savoldi, in quanto referente scientifico iniziale dell'Accordo, per il supporto fornito nell'ambito della prima fase dell'approccio. Si precisa che i dettagli sul sistema di conservazione di Medas e le tabelle di codifica non sono mostrati in questo studio a causa dei diritti di proprietà; pertanto, vengono utilizzati solo analisi aggregate ed esempi significativi per evidenziare i risultati dello studio. Si ringraziano anche Maria Mezzatesta ed Erika Pasceri, autrici di studi che hanno preceduto e ispirato questa ricerca.

Riferimenti bibliografici

- Agenzia per l'Italia Digitale. 2020. *Linee guida sulla formazione, gestione e conservazione dei documenti informatici*. Determinazione dirigenziale n. 407/2020, 10 settembre 2020.
- Bjørn, Pernille, Stacey Burgoyne, Veronica Crompton, Teri MacDonald, Barbe Pickering, and Sue Munro. 2009. "Boundary Factors and Contextual Contingencies: Configuring Electronic Templates for Healthcare Professionals." *European Journal of Information Systems*, 1-14.
- Chen, Elizabeth S., Genevieve B. Melton, Mark E. Engelstad, and Indra Neil Sarkar. 2010. "Standardizing Clinical Document Names Using the HL7/LOINC Document Ontology and LOINC Codes." *AMIA Annual Symposium Proceedings* 2010: 101-05. PMID: 21346949; PMCID: PMC3041326.
- Cimino, James J. 1998. "Desiderata for Controlled Medical Vocabularies in the Twenty-First Century." *Methods of Information in Medicine* 37 (4-5): 394-403. PMID: 9865037; PMCID: PMC3415631.
- Dugas, Martin, Sylvia Thun, Thomas Frankewitsch, and Kai U. Heitmann. 2009. "LOINC Codes for Hospital Information Systems Documents: A Case Study." *Journal of the American Medical Informatics Association* 16 (3): 400-03. <https://doi.org/10.1197/jamia.M2882>. PMID: 19261942; PMCID: PMC2732230.

- Fascicolo Sanitario Elettronico. n.d. "Affinity Domain." Fascicolo Sanitario Elettronico – Portale nazionale." Ultima consultazione 17 Novembre 2025. <https://www.fascicolosanitario.gov.it/portale/affinity-domain>.
- Hyun, Sookyung, Jason S. Shapiro, Genevieve Melton, et al. 2009. "Iterative Evaluation of the HL7-LOINC Document Ontology for Representing Clinical Document Names: A Case Report." *Journal of the American Medical Informatics Association* 16 (3): 395-99.
- Khan, Huzaifa, Abu Saleh M. Mosa, Vyshnavi Paka, et al. 2024. "Mapping Clinical Documents to the Logical Observation Identifiers, Names and Codes (LOINC) Document Ontology Using Electronic Health Record Systems Structured Metadata." *AMIA Annual Symposium Proceedings* 2023: 1017-26.
- Li, Li, C. Paul Morrey, and David Baorto. 2011. "Cross-Mapping Clinical Notes between Hospitals: An Application of the LOINC Document Ontology." *AMIA Annual Symposium Proceedings* 2011: 777-83. PMID: 22195135; PMCID: PMC3243240.
- LOINC Italia. n.d. "LOINC Italia." Ultima consultazione 17 novembre 2025. <https://www.loinc.it>.
- Mezzatesta, Maria, Maria Teresa Chiaravalloti, and Erika Pasceri. 2021. "Mapping LOINC Document Type Codes: An Experience in Managing Clinical Documents for the Long-Term Preservation." *LOINC Conference*, Virtual, March 9-12, 2021.
- Pasceri, Erika, and Maria Teresa Chiaravalloti. 2025. "Indexing Clinical Documents for Long-Term Preservation through an International Coding System." *JLIS.It* 16 (2): 32-45. <https://doi.org/10.36253/jlis.it-632>.
- Pearce, Christopher, and Michael Bainbridge. 2014. "A Personally Controlled Electronic Health Record for Australia." *Journal of the American Medical Informatics Association* 21 (4): 707-13. <https://doi.org/10.1136/amia-jnl-2013-002068>.
- Star, Susan Leigh. 1989. "The Structure of Ill-Structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving." In *Distributed Artificial Intelligence* 2: 37-54, edited by Les Gasser and Michael N. Huhns. San Mateo, CA: Morgan Kaufmann.
- Zuo, Xuo, Yujia Zhou, Jon Duke, et al. 2024. "Standardizing Multi-Site Clinical Note Titles to LOINC Document Ontology: A Transformer-Based Approach." *AMIA Annual Symposium Proceedings* 2023: 834-43.

Domain Analysis on Data Management

A View from Brazilian Information Science

Carla Beatriz Marques Felipe*, Carla Maria Martellote Viola**,
Raimunda Fernanda dos Santos***

Abstract: The general objective of this work is to describe what constitutes the domain of Data Management in Brazilian Information Science. To this end, it develops a domain analysis on the subject. Two approaches were selected for development: bibliometric study and epistemology. The methodology used is bibliographic and documentary, with the analysis of Brazilian Postgraduate Programs in Information Science that offer disciplines on Data Management and bibliographic analysis on the subject. As a result, it is observed that what is researched in the Postgraduate Programs is reflected in the bibliographic production. In addition, it is noted that Brazilian Information Science develops more research-on-Research Data Management.

Keywords: Data Management, Domain Analysis, Brazilian Information Science.

1. Introduction

Due to e-Science and the Open Science Movement – which advocates for making information accessible to all – the sharing of research data has gained prominence in the academic context. In this case, data are considered products generated throughout scientific research, in various formats (texts, images, sounds, spreadsheets, videos, etc.) and across multiple fields of knowledge.

Sharing and reuse by researchers will only be possible if efficient data management is in place. Therefore, data management must be considered throughout the entire research process and should be developed through actions related to the organization of data information and the sharing infrastructure, always aiming for the data to remain usable even after the research has concluded.

* Federal University of Rio de Janeiro, Course in Library Science and Information Unit Management; Rio de Janeiro, Brasil. felipecarla12@gmail.com. <https://orcid.org/0000-0001-5277-9165>.

** Instituto Brasileiro de Informação em Ciência e Tecnologia; Brasília -DF, Brasil. viola.carla@gmail.com. <https://orcid.org/0000-0002-0895-8163>.

*** Federal University of Rio Grande do Norte, Library Science Course; Natal, Brasil. raimunda.fernanda@ufrn.br. <https://orcid.org/0000-0002-7750-3269>.

According to Alonso-Arévalo (2019, 82), data management operates along two lines: «long-term preservation of datasets in storage and sharing systems, and reuse of datasets for research purposes». As actions related to data preservation are implemented, sharing and subsequent reuse are facilitated.

The BRIET¹ research group has, as one of its goals, the development of studies on the concept of research data and the processes involved in transforming data into information and scientific knowledge. In addition, it conducts research on Knowledge Organization.

Knowledge Organization (KO) emerges as the discipline within Information Science aimed at studying and developing systems for organizing units of knowledge (concepts) and for information retrieval when represented by these concepts.

In this discipline, studies focus on the thematic and descriptive treatment of information. The former deals with the subject matter addressed by an object, while the latter concerns the characteristics of the object. For the discipline to function, it develops tools that support knowledge representation, known as Knowledge Organization Systems (KOS), which are intended to assist information professionals in the representation of information.

Among the research efforts that contribute to the development of KOS. According to Hjørland (2017, sec. 2.1) «a domain, on the other hand, is a specialization in the division of cognitive labor that is theoretically coherent or socially institutionalized. [...] domains are not ready-made divisions of the world but are dynamic, developing, and theory dependent».

One of the proposed research objectives within the group is the development of a Thesaurus on Research Data Management. To construct the thesaurus, it is essential to have in-depth knowledge of the subject area it will cover, as only after this stage can the selection of terms that will comprise the thesaurus be defined. In this context, domain analysis proves to be an important tool, as it provides an overview of how a field of knowledge or discipline develops.

As Freitas and Albuquerque (2017, 2) explain, «the theoretical-methodological approach of Domain Analysis – specifically in studies related to Information Science – has among its goals to support the research process and the construction of tools for knowledge organization». In this context, domain analysis serves to understand what is being researched and discussed regarding a particular concept.

Thus, the central question of this study is: How can the domain analysis approach, stemming from KO, be used to map the field of Data Management, gaining substantial knowledge to support the development of a thesaurus on the topic?

¹ Acronym of *Biblioteconomia, Representação, Interoperabilidade, E-science e Tecnologia*, translated Librarianship, Representation, Interoperability, E-science and Technology.

The general objective is to describe what constitutes the domain of Data Management in Brazilian Information Science. The specific objectives include investigating the scientific output in the field and identifying postgraduate-level courses and disciplines related to the topic.

Data management is a multidisciplinary field, involving various areas beyond Information Science, such as computer science, management, and applied fields like healthcare. However, the aim of this work is precisely to analyze how the topic of data management has been addressed specifically within the scope of Information Science, considering its theoretical frameworks, terminologies, and specific approaches.

We acknowledge, however, that this delimitation may result in a partial perspective, especially regarding the construction of a controlled vocabulary such as a thesaurus. In view of this, the text will emphasize the justification for this delimitation, indicating that it is a conscious methodological choice aimed at understanding the specificity of the Information Science domain within a broader field.

In this way, it is believed that research can contribute to studies on KO and information representation in data management, as well as support the development of a thesaurus on the topic.

2. Data Management

According to Borgman (2012), data is the driving force behind research in any field of knowledge. Their format will depend on their origin, purpose, and academic discipline. For example, a rock may serve as data for both History and Geology. A photograph can be considered research data in Anthropology, History, Medicine, Biodiversity, and other fields.

In this sense, research data are not generated exclusively in digital formats; notebooks, analog photographs, drawings, handwritten calculations, among others, are also considered research data, as long as they are produced within the context of a research activity.

Today, there are various ways to share data; however, data sharing alone does not ensure data reuse. Effective data management is essential. Researchers seek to ensure long-term access to their data (Shah et al. 2025).

Data management encompasses aspects related to acquisition, organization, storage, use and reuse, and preservation of data. All of this must be supported by appropriate policies and infrastructures. According to Majid, Zhang and Foo (2018, 2) «Research data management entails all activities and processes which are undertaken or done to ensure that research data is properly documented, organised, stored, archived and curated so that it is available for access, use and reuse whenever the need arises after the research has been done and reported».

It can be stated that research data management encompasses the set of activities and mechanisms that span from data capture to its reuse. Processes that ensure data organization—such as description and storage—are essential. «Data management covers all aspects related to handling, organizing, documenting, and adding value to data, and it plays a crucial role as a facilitator in data sharing processes, ensuring long-term data sustainability and accessibility» (Sayao and Sales, 2015, 5).

In turn, Kotronoulas et al. (2023, 2) offer a more simplified definition: «Crucially, the starting point is to carefully prepare a data set ready for analysis. Numbers (also known as raw data) must be put in a form that makes them suitable for analysis. This is called data management». That is, data management enables data to be used in a simplified and efficient manner.

Examples of data management activities include addressing the data lifecycle (from capture to reuse), access policies, sharing and reuse policies, data preservation actions, infrastructure selection, and data representation.

Data do not speak for themselves; therefore, KOS assist in their description within data repositories, data papers, and data journals. In this context, KO can play a key role in supporting effective data management.

3. Organization of Knowledge

Knowledge is a fundamental part of human life and is intrinsically linked to the functioning of society. Therefore, aspects such as science, religion, and politics can only be developed through shared knowledge—and for this to happen, it must be recorded. However, recorded knowledge alone does not guarantee its dissemination. It must also be organized and made accessible to all. According to Guimarães (2014, 14): «in order for this socially produced knowledge to have social use, a mediating process of organization is necessary, in which “substitutes of knowledge” are established, so that the contexts of production and use can be brought into dialogue».

In this context, KO emerges as a discipline whose aim is to ensure that knowledge is retrievable, playing a fundamental role within the fields of Information Science and Librarianship. It can be inferred that the development of these fields is grounded in the construction of information retrieval techniques, based on theories of knowledge organization. On this matter, Smiraglia (2012, 225) states: «Knowledge organization (also well-known by its acronym KO) is the domain in which the order of knowledge is both the primary paradigm for scientific investigation and the primary application in the development of systems».

Thus, based on the theories that underpin discipline, tools are created to support the representation and retrieval of information, such as KOS. For Mazzocchi (2024, sec. 1.0) «Knowledge organization system (KOS) is a ge-

neric term used for referring to a wide range of items (e.g. subject headings, thesauri, classification schemes and ontologies), which have been conceived with respect to different purposes, in distinct historical moments».

Knowledge Organization Systems have among their functions the task of assisting in thematic representation when a document needs to be described using keywords. Documents are written in natural language (the language of the document), whereas KOS provide standardized terms that can be used to represent the record consistently.

Those systems can take various forms, including thesauri, subject heading lists, glossaries, controlled vocabularies, ontologies, taxonomies, among others. Their structures vary according to their intended purpose, which may include improving information retrieval, enhancing navigation, facilitating communication among specialists, mapping knowledge domains, supporting intelligent machine processing, and more. In this way, KOS have specific characteristics that distinguish them from one another.

Therefore, KOS can support the retrieval of information within research data. For Given and Olso (2003, 160) «the principal concepts of knowledge organization and their interrelations, when construed broadly, offer a model for the intellectual work involved in preparing data for analysis».

It is not possible to understand a dataset without its exhaustive description. For the reuse of datasets made available in sharing environments to occur, a proper description is necessary for comprehension. According to Meschini and Francelin (2022, 4), «Knowledge Organization has the potential to act as a mediator between issues related to the ease of data acquisition and its complex interpretation, contributing to the semantic expansion of this data phenomenon, especially concerning scientific data».

Moreover, KOS, such as a thesaurus, can assist researchers not only during the sharing and reuse of data but also in knowledge representation when research outputs like abstracts and articles are developed.

As mentioned in the introduction of this study, the current research aims to support the BRIET research group in the development of a thesaurus on data management, with the goal of helping Brazilian researchers better represent their texts according to the topic.

To develop a thesaurus, it is necessary to understand the domain to which it will belong. Thus, the objective of this research was to carry out a domain analysis of data management within the context of Brazilian Information Science. The step-by-step process of how the research was conducted will be presented in the next section.

4. Methodological Procedures

To meet the objectives, a Domain Analysis study was conducted, developed according to Hjørland (2002), as it is an approach widely used in the field of Information Science research, specifically within the scope of Knowledge Organization. According to Smiraglia (2011, 1):

Domain analysis is the study of the evolution of discourse around theoretical poles represented through the formal products, usually the literature, of a research community. Domain analysis is one way of generating new knowledge about the interaction of communities of scholars with information. Domain analysis of international research communities brings the promise of new comprehension of how people interact with information in different places.

For its development, several methodological steps must be undertaken. In line with the objectives proposed for the present research, the approaches suggested by Hjørland (2002) were considered. Specifically, Hjørland (2002, 422) proposes 11 approaches, which are:

producing literature guides and subject gateways; producing special classifications and thesauri; research on indexing and retrieving specialties; empirical user studies; bibliometrical studies; historical studies; document and genre studies; epistemological and critical studies; terminological studies, LSP (languages for special purposes), discourse studies; studies of structures and institutions in scientific communication; and domain analysis in professional cognition and artificial intelligence.

For the development of the present research, two approaches were selected: bibliometric study and epistemology. The bibliometric study aims to demonstrate how scientific production occurs within a given domain and/or subject. In turn, according to Hjørland (2002), epistemological and critical studies make it possible to understand the theories, methodologies, and applications of a domain. This approach may seem more basic, but it is necessary. According to the author, when this approach is not employed, domain analysis tends to be more superficial.

For the epistemological study, academic Graduate Programs in Information Science were selected. The justification for choosing these programs is twofold: first, there are no undergraduate degrees in Information Science in Brazil; second, these programs are advanced courses that focus on the study and practice of information management, organization, retrieval, and dissemination. These programs typically address topics such as librarianship, archival science, information technologies, data analysis, and information policy. Therefore, they deal directly with information and are also focused on discussing new demands for information organization, such as the topic of data management.

To identify and select the Graduate Programs, a search was conducted on the Sucupira Platform. The Sucupira Platform is a digital platform that centralizes information from the National Graduate System. Graduate Programs fill out the platform annually with data that the government can use for planning and decision-making. The tool enables the management of this level of Brazilian education.

The search was conducted using the filters provided by the platform, which are: Major Area of Knowledge, Evaluation Area, and Field of Knowledge. For this research, the selected categories were, respectively: Major Area of Knowledge – Applied Social Sciences; Evaluation Area – Communication and Information; and Field of Knowledge – Information Science.

A total of 28 programs were retrieved, of which 3 were not yet operational, being still under development. Therefore, 25 programs were considered valid for the research. After identifying the programs, their websites were examined to determine whether they offered courses that address issues related to Data Management. The search was carried out in March 2025. From the 25 programs retrieved, 11 programs were selected that offered courses related to the theme. As part of the data analysis, the course syllabi were reviewed.

One program was excluded because it was not possible to access or read the course syllabi.

To analyze data related to postgraduate programs, thematic coding was selected. According to Medelyan (2025) coding is the process of labeling and organizing your qualitative data to identify different themes and the relationships between them. That is, thematic coding allows the researcher to structure the data in a way that facilitates its subsequent interpretation. In other words, coding is not just a technical act, but an interpretative and organizational step.

Still according to Medelyan (2025), when coding customer feedback, you assign labels to words or phrases that represent important (and recurring) themes in each response. These labels can be words, phrases, or numbers; we recommend using words or short phrases, as they're easier to remember, skim, and organize.

For this research, the first step was to read the course syllabi to gain a general understanding of the topics. Following this, the topics were divided based on the course syllabi. Then, the topics were grouped by theme, and finally, the areas covered by the courses were identified. Thus, thematic coding was performed using the course titles, following Mendelay's (2025) guidelines. The results of the final coding, with the grouping of courses and graduate programs, can be seen in the next chapter.

For the bibliometric study, the selected database was the *Base de Dados em Ciência da Informação (BRAPCI)*, a Brazilian digital infrastructure designed for the collection, preservation, and access to scientific literature in the field of Information Science. It includes a wide range of publications, such as journal

articles, conference papers, books, and book chapters, primarily from Brazilian and Latin American sources (BRAPCI n.d.). The database also includes some journals from Portugal.

Currently, the database includes over 100 journals in the field and publishes proceedings from 8 national events. It contains articles from the year 1962 to the present. The platform allows searches by fields such as author, abstract, keywords, title, and a combined search across all five fields. Filters can be applied to narrow results to Brazilian journals, foreign journals, conferences, or books and book chapters. The search also supports Boolean operators, as shown in the Figure 2.



Figure 2. Brapci home screen.

To encompass current research, a time frame from 2010 to 2024 was applied. This range allows for an updated overview of the topic. The search was conducted using the titles of the works. The choice of the title field was made because the title is considered an important element in the development of academic works. A well-structured title summarizes the main topic and the purpose of the research. It is also believed that a clearly described title improves discoverability in databases and digital libraries, as it is the first element users encounter in publications.

Two search terms were used in the database: *Data Management* and *Research Data Management*, aiming for a more precise retrieval. A total of 128 works were retrieved: 96 articles with the term *Data Management* and 32 with *Research Data Management*. After removing duplicates, 97 articles were analyzed.

For the article analysis, the title, abstract, keywords, and introduction were read. As the focus of the research is to understand the themes and approaches related to data management, other information related to bibliometric studies, such as quantitative data, authorship analysis, and collaboration networks, were not discussed. The articles were categorized by thematic areas for analysis and will be presented in the results chapter.

5. Results

The first analysis carried out was of the Graduate Programs. Of the 10 programs selected for the final stage of the research, 1 program focuses on Information Management and the remaining 9 are centered on Information Science.

Even though it is not part of the thematic coding, it is considered necessary to present a description of the research lines of the graduate programs investigated, as they offer courses that aim to contribute to their scope.

The *Information Management Program* at UDESC (*Santa Catarina State University*) focuses on the discussion of topics related to the management of organizations that deal with information, aiming to propose solutions to concrete problems. It is committed to ethics and to the economic, social, and educational role of information in citizenship, preparing professionals capable of solving complex problems in information management while generating and disseminating technical and scientific knowledge for innovation.

In turn, the *Graduate Program in Information Science* at UFPE (*Federal University of Pernambuco*) is centered on Information as its main theme, with Memory and Technology as its theoretical framework. Even so, it offers a course that covers the scope of Data Management.

The *Information Science Program* at the *Federal University of Santa Catarina (UFSC)* has “Information and Society” as its area of concentration. According to the program’s official webpage, it emphasizes the study of phenomena related to core processes such as management, selection, processing, organization, retrieval, use, and dissemination of information, aiming primarily to address, explore, understand, and explain these phenomena.

The *Information Science Program* at *São Paulo State University (UNESP)* is focused on “Information, Technology, and Knowledge”. It develops research that encompasses aspects of information and technology, the production and organization of information, and the management, mediation, and use of information.

Meanwhile, the *Information Science Program* at the *Federal University of Alagoas (UFAL)* focuses on “Information, Technology, and Innovation”. Research in this program is directed toward the production, mediation, communication, and management of information and technological processes.

The *Information Science Program* at the *Federal University of Rio Grande do Sul (UFRGS)* concentrates on “Information, Science, and Society”. Its goal is to develop research on the relationships between information and science, as well as in different social contexts, involving the discussion and analysis of contemporary problems to improve information flows and their management in various environments.

The *Information Science Program* at the *Federal University of Rio Grande do Norte (UFRN)* focuses on “Information in Contemporary Society”. This concentration area aims to address trends in theories, methodologies, and practices in Information Science. It considers issues specific to post-industrial society related to the production, organization, mediation, management, dissemination, access, and use of information and knowledge.

The *Information Science Program* at the *Federal University of Ceará (UFC)* conducts research focused on the *Representation and Mediation of Information and Knowledge*. To this end, its focus is divided into studies on information, organization, communication and technology, and information, society, and culture.

The *Information Science Program* at the *Brazilian Institute for Information in Science and Technology (IBICT)* has its area of concentration in *Information and Social and Technological Mediations for Knowledge*. Therefore, researchers in the program focus on aspects of information and its social and technological mediations for knowledge.

The *Information Science Program* at the *University of São Paulo (USP)* is centered on *Information and Culture*. It deals with the relationships that characterize the processes of construction and/or reconstruction of meaning and/or cultural products when information is transformed into knowledge and the cultural product into a cultural asset. It proposes the observation of the necessary actions – within the context of cultural institutions – for information to be preserved and socially circulated (collection, selection, organization, access), and the analysis of the cultural contexts in which these processes occur and acquire social meaning.

After selecting the programs, a total of 18 courses related to the topic were identified. These courses address aspects related both to general *Data Management* and to more specific topics such as *Research Data Management*. Below is a description of what was recovered in the first reading of the course syllabi.

It is evident that the courses offered are intrinsically linked to the field of Information Science, and each program provides courses aligned with its area of concentration or scope. From an epistemological perspective of Information Science, the courses found reflect the objectives set forth in the programs’ scopes, contributing to issues such as the organization, description, preservation, visualization, and use of data for information generation.

The *Information Management Program* at *Santa Catarina State University (UDESC)* offers two courses focused specifically on data management: *Data Governance and Research Data Management*. Regarding the *Data Governance* course, its objective is to understand governance concepts and their application in the context of scientific data, including Open Science and Citizen Science contexts, as well as the creation of a Scientific Data Management Plan. The *Research Data Management* course aims to discuss e-Science practices, focusing on Research Data: concepts, types, and structure. It also covers platforms for Data Management, including Research Data Repositories.

Meanwhile, the *Federal University of Pernambuco (UFPE)*, through its Graduate Program in Information Science with a focus on Memory and Technology, offers the course *Digital Curation*. The Digital Curation course provides a methodological framework for issues related to the preservation and management of data and digital information, involving the maintenance and enhancement of research data throughout its lifecycle. This concept aims to ensure the accessibility and use of research data in the future, as well as to promote its discovery, interpretation, and reuse.

The *Federal University of Santa Catarina (UFSC)* offers three courses related to the theme of data management: *Big Data and Semantic Web*; *Web Data Management for Information Science*; and *LGPD (General Data Protection Law) and the New Legal Framework for Innovation: Interfaces between Law and Information Science*.

The *Big Data and Semantic Web* course presents concepts and discusses aspects related to the Semantic Web and supporting technologies. It also covers the conceptual foundations of ontologies applied to information systems, practices of Linked Data, semantic information retrieval, Big Data, and associated information technologies.

The *Web Data Management for Information Science* course focuses on the characteristics of Web data, modeling of semi-structured data, access to Web data, as well as the integration, processing, analysis, and visualization of data.

The *New Legal Framework for Innovation: Interfaces between Law and Information Science* course discusses aspects related to the General Data Protection Law (LGPD), which is a law designed to establish rules for the processing of personal data, whether in physical or digital form, by individuals or legal entities.

In the Graduate Program at *São Paulo State University (UNESP)*, three courses were identified: *Information Science and the Data Lifecycle: Research Frontiers*; *Open Science and Research Data Management: Theory and Practice*; and *Data Science for Information Science*.

The course *Information Science and the Data Lifecycle: Research Frontiers* discusses the role of Information Science in the collection, processing, storage, dissemination, and preservation of data.

The course *Open Science and Research Data Management: Theory and Practice* aims to provide a theoretical deep dive into the concepts of Open Science and Research Data Management, as well as the points of convergence between these two movements. Topics discussed include the FAIR Principles, Data Librarianship, Data Management Plans, Data Stewardship, and other aspects related to research data management.

The course titled *Data Science for Information Science* covers discussions on Data Science applied to Information Science, enabling students to understand and apply the principles, methods, and techniques of Data Science within the field of Information Science. This involves studying aspects related to data collection, organization, analysis, and interpretation, allowing the extraction of relevant knowledge that contributes to efficient information management.

In the Graduate Program in Information Science at the *Federal University of Alagoas (UFAL)*, two courses were identified: *Open Science and Digital Scientific Marketing and Open Data and E-Governance*.

The *Open Science and Digital Scientific Marketing* course discusses collaborative and open scientific initiatives and practices, such as e-science and research data. It also presents concepts of digital marketing applied to scientific products.

The *Open Data and E-Governance* course covers concepts related to open government data and electronic governance.

In the Graduate Program in Information Science at the *Federal University of Rio Grande do Sul (UFRGS)*, three courses were identified: *Open Science and Scientific Reproducibility; Scientific Communication; and Scientific Data Management*.

The course *Open Science and Scientific Reproducibility* studies aspects of Open Science and data openness, data publication, FAIR (Findable, Accessible, Interoperable, Reusable) Principles, and other aspects of research data management.

The *Scientific Communication* course addresses issues related to open science and open access, as well as topics concerning open research data.

The *Scientific Data Management* course aims to investigate the conceptual and practical relationship between the research process and data handling across all knowledge areas. It includes the study of the main international organizations involved in developing policies and tools for scientific data management, types of scientific data, and the development of data management plans as well as preservation and reuse processes in repositories. The course also covers the scientific data lifecycle in relation to Open Science and the mastery of fundamental aspects necessary for the digital preservation of scientific data, in addition to studying the role of information professionals.

The Graduate Program in Information Science at the *University of São Paulo (USP)* offers a course on the topic titled *Data Visualization and Analytical Cul-*

ture. This course focuses on the introductory exploration of software and tools for creating dashboards based on usage scenarios, as well as theoretical reflection on these informational objects within an analytical culture that emerges from the mediation of Information and Communication Technologies.

In the Graduate Program in Information Science at the *Federal University of Rio Grande do Norte (UFRN)*, the course *Scientific Communication and Open Science* was identified. This course discusses Open Science, presenting concepts, schools of thought, principles, and collaborative and open scientific practices. It also covers open data, research data management, and Open Science policies.

The *Federal University of Ceará (UFC)*, in its Graduate Program, also offers a course titled *Scientific Communication*. This course presents discussions about Open Access and Open Data.

In the Graduate Program of the *Brazilian Institute of Information for Science and Technology (IBICT)*, the course also called *Scientific Communication* addresses issues related to open data and the publication of such data.

In summary, as part of the thematic coding, the following is a description of the disciplines with the most prominent emphasis:

- Data Governance;
- Digital Curation;
- Open Science;
- Data Visualization;
- Personal Data Protection (LGPD);
- Scientific Communication;
- Repositories and Preservation;
- Semantic Web and Big Data;
- Data Lifecycle.

The initial codes are based on the first reading and thematic grouping. These are:

Excerpt	Code
“understand Governance concepts and their application in the context of scientific data”	Data governance
“Open Science and Citizen Science”	Open Science
“Creation of the Scientific Data Management Plan”	Data Management Plan (DMP)
“Digital curation” / “maintenance and valorization of research data throughout the lifecycle”	Digital Curation / Data Lifecycle
“Big Data and Semantic Web”	Big Data / Web Semântica
“FAIR principles”	FAIR principles

“Data Stewardship, Data Librarianship”	Data librarianship
“Data preservation and reuse in repositories”	Data reuse
“General Data Protection Law”	Personal data protection (LGPD)
“Data Visualization and Dashboards”	Data visualization

Table 1. Initial codes, extracted directly from the data.

After the analysis carried out in the coding, it is possible to perceive the central themes and their groupings, together with the postgraduate programs.

Universidade	Disciplinas	Main Themes (Thematic Coding)
UDESC (Santa Catarina State University)	- Data Governance - Research Data Management	- Data Governance - Open Science - Citizen Science - Data Management Plan - Data Platforms and Repositories
UFPE (Federal University of Pernambuco)	- Curadoria Digital	- Digital curation - Data preservation and lifecycle - Data accessibility and reuse
UFSC (Federal University of Santa Catarina)	- Big Data and Semantic Web - Web Data Management for Information Science - LGPD and the New Legal Framework for Innovation	- Big Data - Semantic Web - Ontologies - Web Data and Analytics - General Data Protection Law (LGPD)
UNESP (São Paulo State University)	- Information Science and the Data Lifecycle - Open Science and Research Data Management - Data Science for Information Science	- Data Lifecycle - Open Science - FAIR Principles - Data Librarianship and Data Stewardship - Data Science - Data-Based Information Management

Universidade	Disciplinas	Main Themes (Thematic Coding)
UFAL (Federal University of Alagoas)	- Open Science and Digital Scientific Marketing - Open Data and E-Governance	- Open Science - Digital Science Marketing - Open Government Data - E-Governance
UFRGS (Federal University of Rio Grande do Sul)	- Open Science and Scientific Reproducibility - Scientific Communication - Scientific Data Management	- Open Science - Scientific Communication - FAIR Principles - Repositories and Preservation - Information Professionals in Data Management
USP (University of São Paulo)	- Data Visualization and Analytical Culture	- Data visualization - Analytical culture - Tools and dashboards
UFRN (Federal University of Rio Grande do Norte)	- Scientific Communication and Open Science	- Open Science - Open Data - Open Science Policies
UFC (Federal University of Ceará)	- Scientific Communication	- Scientific communication - Open access - Open data
IBICT (Brazilian Institute of Information for Science and Technology)	- Scientific Communication	- Scientific communication - Publication and access to open data

Table 2. Subjects and Topics Related to Data Management in Postgraduate Programs in Information Science – Brazil (2025).

The disciplines demonstrate how the field of Information Science is integrating:

- Emerging epistemologies (open science, FAIR data);
- Digital infrastructures (repositories, platforms, big data);
- Legal and social demands (LGPD, reproducibility, innovation);
- Interdisciplinary practices (curation, data science, law).

This integration reveals an evolution of the field toward a systemic management of scientific information, strengthening its social, scientific, and political relevance.

It is noticeable that, when analyzing the courses offered by these programs, such perspectives promote discussion on epistemological innovations in Information Science. It is possible to observe the evolution of the field alongside the advancement in knowledge availability, as seen in data management. Moreover, the field has been utilizing its theories and studies involving issues of organization, processing, archiving, preservation, and use to support theoretical questions about this new domain, which is Data Management.

The second approach developed in the present research was the analysis of bibliographic production, which was conducted using one of the specific databases for Brazilian Information Science, *Brapci*.

When analyzing the scientific production of the domain, it becomes evident that it reflects what is being studied in the Graduate Programs. The topics retrieved in this part of the research present study results strongly linked to the themes of the courses, such as Open Science, research data, data visualization, the General Data Protection Law, FAIR Principles, digital curation, as well as the role of the information professional.

It can be observed that the main topics developed are closely connected to research data management, including Open Science, e-Science, the data lifecycle, data repositories, FAIR Principles, and data management plans. These aspects are recurrent in both the scholarly works analyzed and the courses offered by the Graduate Programs.

Within the theme of data curation, several studies address digital preservation, especially within data repositories, and refer to the data lifecycle. Still within the scope of curation, works were also found that deal with the use of metadata for data curation and for the description of data in repositories.

Regarding the theme of the FAIR Principles, these are directly linked to research focused on data repositories, as they serve as guiding principles for such technologies. Additionally, there are studies on FAIR objects and FAIR data management plans.

Additional studies were also found that focus on the *data librarian* and the role of the librarian in data management—an aspect that was not identified in the analysis of the Graduate Programs. Since *Librarianship and Information Science* are closely linked, it is believed that the presence of bibliographic production on this topic reflects what is being studied in undergraduate Librarianship programs in Brazil.

After reading the articles, one can find aspects that reflect the epistemology of Information Science. Such as the idea that research data are not merely technical records, but fundamental elements for the construction, validation, and circulation of scientific knowledge. There is a premise that knowledge needs to be organized, described, and accessible to be legitimized.

The bibliographic production also includes articles discussing *data visualization*, *services related to data management*, *data quality*, *Big Data*, and the

General Data Protection Law (LGPD). This Brazilian law, approved in 2018, regulates privacy and the use/processing of personal data in Brazil.

Based on the results presented, the following Figure 2 attempts to synthesize the findings by combining epistemological study and the bibliometric study.

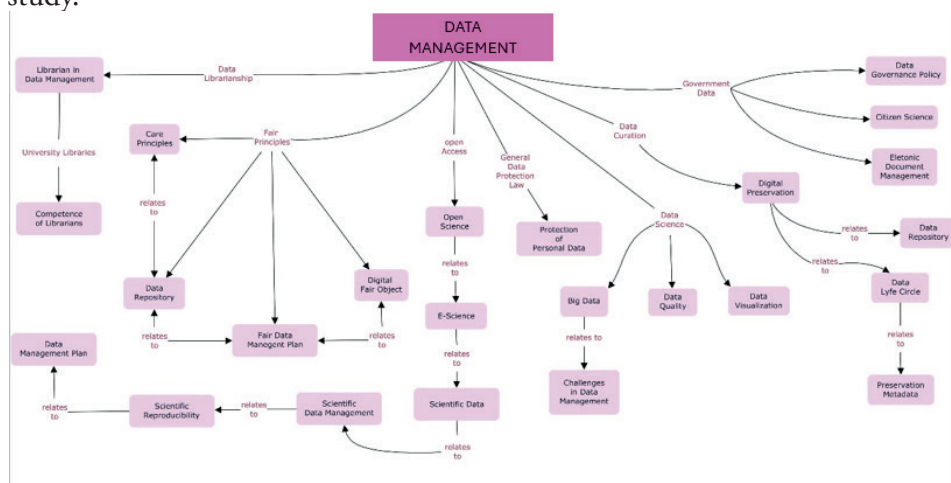


Figure 2. Conceptual map about Data Management.

The choice to develop a *concept map* stem from its ability to provide an initial visualization of the field, which can contribute to the creation of the *Briet Group's thesaurus*, since a thesaurus is composed of terms organized in a hierarchical structure (Grupo Briet n.d).

When observing the map, the central field is *Data Management*, which branches into six subfields. These, in turn, are formed by secondary topics that constitute the research landscape. Thus, it can be inferred that the field of Data Management in Brazilian Information Science is composed of seven main themes: *Data Librarianship*, *FAIR Principles*, *Open Access*, *Data Science*, *General Data Protection Law*, *Data Curation*, and *Government Data*. These are supported by secondary, yet significant, topics that enrich discussions both within Graduate Programs and in the construction of the bibliographic field.

This demonstrates that Brazilian Information Science has been evolving in its *epistemological aspects*, embracing new nuances in the creation, organization, and dissemination of knowledge.

6. Final Considerations

Domain analysis proves to be an important tool for the understanding and development of fields such as *Data Management*. It is hoped that this work

will effectively contribute to the progress proposed by the *Briet research group*. Based on the analysis of the results, it is evident that, in the investigated context, *Brazilian Information Science* is primarily focused on *Research Data Management*.

It was observed that both the courses and the articles analyzed have strong *epistemological connections* with Information Science, as it is a discipline that deals with various aspects of information. This demonstrates that Information Science is in tune with the new demands of *knowledge production and organization*.

It is acknowledged that choosing the field of Information Science as the corpus to support the research and, subsequently, to serve as a basis for the construction of a thesaurus entails certain methodological and conceptual challenges. These implications can be observed in several aspects, such as:

- *Incomplete domain representation*: Considering that the field of Data Management has a multidisciplinary nature, there is a risk of excluding relevant terms, which would compromise the conceptual scope of the thesaurus.
- *Interoperability issues*: Thesauri and controlled vocabularies are often used in information systems that require interoperability. The absence of standardization or alignment between systems may hinder communication among them.
- *Low usefulness for users*: An effective thesaurus must comprehensively reflect users' information retrieval needs. Otherwise, its practical applicability may be significantly reduced.
- *Weaknesses in evolution and updating*: There is a risk that the thesaurus may become obsolete if it remains restricted to a specific thematic approach, making it difficult to adapt to new informational demands.

To mitigate these risks, it is recommended that, before finalizing the thesaurus, the team responsible for its development carefully consider these implications. It is essential to adopt strategies that enable the inclusion of the various subfields that make up the Data Management domain, thereby enriching the thesaurus's base corpus and reducing the occurrence of conceptual gaps.

Additionally, it is suggested that the thesaurus be subject to periodic reviews and updates to ensure its relevance, adaptability, and alignment with the evolution of the thematic domain.

References

- Alonso-Arévalo, Julio. 2019. "La gestión de datos de investigación en el horizonte de las bibliotecas universitarias y de investigación." *Cuadernos de Documentación Multimedia* 30: 75-88. <https://doi.org/10.5209/CDMU.62806>.

- Borgman, Christine L. 2012. "The conundrum of sharing research data." *Journal of the American Society for Information Science and Technology* 63, no. 6 (June):1059-78. <https://doi.org/10.1002/asi.22634>.
- BRAPCI (Base de Dados em Ciência da Informação). n.d. "Sobre a Brapci!" Last Accessed November 2, 2024. <https://brapci.inf.br/about/brapci>.
- Freitas, Lidiane Marques, and Ana Cristina de Albuquerque. 2017. "As abordagens da análise de domínio como aporte metodológico para a classificação arquivística." *Tendências da Pesquisa Brasileira em Ciência da Informação* 10(2): 1-19. <https://revistas.ancib.org/tpbci/article/view/424>.
- Given, Lisa M., and Hope A. Olson. 2003. "Knowledge organization in research: A conceptual model for organizing data." *Library & Information Science Research* 25, no. 2 (Summer): 157-76. [https://doi.org/10.1016/S0740-8188\(03\)00005-7](https://doi.org/10.1016/S0740-8188(03)00005-7).
- Grupo Briet. n.d. "Briet: Biblioteconomia, Representação, Interoperabilidade, E-science e Tecnologia." Last Accessed December 16, 2024. <https://grupobriet.com/>.
- Guimarães, José Augusto Chaves. 2014. "Análise de domínio como perspectiva metodológica em organização da informação." *Ciência da Informação* 43, no. 1 (jan.-abr.):13-21. <https://doi.org/10.18225/ci.inf.v43i1.1415>.
- Hjørland, Birger. 2002. "Domain analysis in information science: eleven approaches traditional as well as innovative." *Journal of Documentation* 58, no. 4 (August):422-62. <http://dx.doi.org/10.1108/00220410210431136>.
- Hjørland, Birger. 2017. "Domain analysis." *Knowledge Organization* 44 (6):436-64. <https://doi.org/10.5771/0943-7444-2017-6-436>. Also available in ISKO Encyclopedia of Knowledge Organization, edited by Birger Hjørland and Claudio Gnoli, https://www.isko.org/cyclo/domain_analysis.
- Kotronoulas, Grigorios, Suzana Miguel, Maura Dowling, et al. 2023. "An Overview of the Fundamentals of Data Management, Analysis, and Interpretation in Quantitative Research." *Seminars in oncology Nursing* 39, no. 2 (April):e161398. <https://doi.org/10.1016/j.soncn.2023.151398>.
- Majid, Shaheen, Shubert Foo, and Xue Zhag. 2018. "Research Data Management by Academics and Researchers: Perceptions, Knowledge and Practices". In *Maturity and Innovation in Digital Libraries. ICADL 2018*, edited by Milena Dobrev, Annika Hinze, and Maja Žumer. *Lectures Notes in Computer Science* 11279. Cham: Springer. https://link.springer.com/chapter/10.1007/978-3-030-04257-8_16.
- Mazzocchi, Fulvio. 2024. "Knowledge Organization System (KOS)." *ISKO Encyclopedia of Knowledge Organization*, November 11. <https://www.isko.org/cyclo/kos>.

- Medelyan, Alyona. 2025. "Coding Qualitative Data: How to Guide." Last updated February 24, 2025. <https://getthematic.com/insights/coding-qualitative-data>.
- Meschini, Fabio Orsi, and Marivalde Moacir Francelin. 2022. "Organização do Conhecimento e suas contribuições em um contexto Big Data". *Transinformação*, Campinas, 34:e210075. <https://doi.org/10.1590/2318-0889202234e210075>.
- Plataforma Sucupira. n.d. "Informações e serviços." Last Accessed March 1, 2025. <https://sucupira.capes.gov.br/>.
- Sayão, Luis Fernando, and Luana Farias Sales. 2015. *Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores*. CNEN/IEN. <https://www.aben.com.br/Arquivos/420/420.pdf>.
- Shah, Naimat Ullah, Salman Bin Naeem and Robina Bhatti. 2025. "Digital data sets management in university libraries: challenges and opportunities". *Global Knowledge, Memory and Communication* 74(1-2):446-62. <https://doi.org/10.1108/GKMC-06-2022-0150>.
- Smiraglia, Richard P. 2011. "Domain coherence within Knowledge Organization: People, Interacting Theoretically, Across Geopolitical and Cultural Boundaries." In *Proceedings of the 39th Annual Conference of CAIS*, edited by Pam Mckenzie, Catherine Johnson, and Sarah Stevenson. University of New Brunswick. <https://doi.org/10.29173/cais601>.
- Smiraglia, Richard P. 2012. "Knowledge organization: some trends in an emergent domain." *El profesional de la información* 21, no. 3 (mayo-junio):225-27. <https://doi.org/10.3145/epi.2012.may.01>.

Che faccetta sei?

Organizzazione della conoscenza a scuola col metodo Ma.Vi.

Claudio Gnoli*, Emanuela Valenzano**

Abstract: Such authors as Comenius, Bliss, Soergel and Szostak find that education is an important field of application for knowledge organization. In particular, in language acquisition learners should be able to fit new words and concepts into their existing knowledge networks. We introduce the Ma.Vi. (Marcatori Visivi, visual markers) method, by which movements and coloured marks are used to facilitate learning of phonology, morphology, syntax and textuality thus improving reading and writing skills. Experimentation is carried out to extend the Ma.Vi. method to semantics by applying fundamental categories of relationships from the bibliographic tradition of facet analysis, particularly those of the Integrative Levels Classification (ILC): quality, quantity, part, property, change, opposition, agent, space, time and perspective. In the context of primary school, categories are expressed by standard colours and by basic questions, which can be used in various linguistic games. Preliminary results and further planned research are discussed.

Keywords: Education, Facet analysis, Integrative Levels Classification, Language acquisition, Ma.Vi. method

1. KO per l'educazione

I principi dell'organizzazione della conoscenza (*knowledge organization*, *KO*), oltre ad attingere più o meno esplicitamente ad elementi della filosofia e della linguistica, si sono consolidati specialmente con le loro applicazioni in biblioteconomia e scienze dell'informazione, e più recentemente in informatica. Si tratta quindi di un dominio interdisciplinare.

In aggiunta alle discipline citate, esiste un altro campo di primaria importanza che necessita di organizzare i propri contenuti conoscitivi: quello dell'educazione. Già il pedagogista seicentesco Comenius, infatti, sosteneva una presentazione unitaria dell'insieme delle conoscenze e un suo insegnamento

* Università di Pavia, Biblioteca della Scienza e della Tecnica Pavia, Italia. claudio.gnoli@unipv.it, ORCID: 0000-0002-4721-7448.

** Centro Ma.Vi. / Ma.Vi. Centre; Asti, Italia. valenzanolinguista@gmail.com.

universale (*pansofia*) fondato appunto sull'organizzazione delle idee; invece della consueta memorizzazione meccanica, esso avrebbe dovuto utilizzare ausili visivi e connessioni logiche con gli elementi già acquisiti: «si inizi dalle cose più note e il processo verso quelle ignote avvenga con lentezza e gradualmente, così come, in una catena, ogni anello sostiene e trascina l'altro anello» (Comenius 1639, cit. in Rossi 2000, 209).

H.E. Bliss dedica un paragrafo della sua autorevole opera *The organization of knowledge and the system of the sciences* a "The organization of knowledge in education": citando svariati psicologi e pedagogisti, anch'egli sottolinea l'importanza dei rapporti di "correlazione" fra i temi che vengono insegnati, in modo da costruire progressivamente un sistema coerente; l'educazione stessa non sarebbe altro che «un'organizzazione dell'esperienza e della conoscenza» (Bliss 1930, 81-88). L'"ordine pedagogico" in cui conviene disporre le materie si affianca all'ordine logico delle scienze – scienze astratte, fisiche, biologiche, psicologiche, sociali – e all'ordine del loro sviluppo storico, ai quali è connesso seppur non identico (226-228); dai loro ordinamenti paralleli, presentati in tavole sinottiche (232-235), deriva anche l'ordine delle discipline nella classificazione bibliografica.

Nella letteratura contemporanea sulla KO, le applicazioni educative emergono purtroppo raramente. Fa eccezione la visione di Dagobert Soergel che, in particolare in un articolo di rassegna, rilancia la concezione di Bliss e l'idea che una presentazione organizzata dei materiali didattici, anche attraverso schemi e tabelle (Jonassen et al. 1993), sia un potente facilitatore dell'apprendimento dei contenuti:

La maggioranza delle teorie sull'apprendimento e la creazione di senso concordano sull'idea che le memorie delle persone consistano di strutture che tengono insieme molti pezzi di informazione e che i discenti debbano assimilare le nuove informazioni in relazione a tali strutture di conoscenza preesistenti nelle loro menti, collocare fatti nuovi nelle strutture esistenti, aggiornare tali strutture, oppure creare strutture del tutto nuove nel caso che l'informazione non vi si adatti (Soergel 2013, 233)¹.

Un'idea, quest'ultima, in accordo con quella di "coerenza esplicativa" che secondo Thagard (1993) caratterizza le tassonomie scientifiche e il loro cambiamento: nuove teorie, come la tettonica delle placche, implicano modifiche delle tassonomie preesistenti.

In generale, un fattore favorevole all'apprendimento è evidentemente un'organizzazione chiara e coerente dei testi (Dickson e Kameenui 1995). Tra

¹ «Most learning and sensemaking theories agree that peoples' memories consist of structures that tie many pieces of information and knowledge together and that learners must assimilate new information in relation to these knowledge structures existing in their minds, fitting new facts into existing structures, updating these structures, or creating entirely new structures if the new information does not fit».

gli strumenti didattici suggeriti da Soergel (2013) ci sono «schemi di tipi di relazione molto generali» come quelli sperimentati da McCagg e Dansereau (2010); l'analisi a faccette, «uno strumento potente per comprendere la struttura concettuale di un dominio»; e le mappe concettuali messe a punto da Novak (2012), sia presentate dagli insegnanti che, meglio ancora, sviluppate dagli allievi stessi. La loro efficacia può essere valutata con test di comprensione della lettura (Morsy et al. 2010).

Di recente, anche Rick Szostak (2024) ha incluso l'educazione fra tre «scopi aggiuntivi» della KO, in quanto essa «consente agli studenti (e altri) di cogliere il fatto che ogni idea che incontrano si situa entro un complesso più ampio» e «dovrebbe stimolare il loro interesse e la loro capacità di cercare informazioni su qualsiasi tema»: a questo scopo egli ha suggerito che gli studenti vengano messi in contatto con un sistema di organizzazione della conoscenza (KOS) offrendo così loro una struttura di riferimento per quanto viene man mano appreso.

2. L'apprendimento di nuove parole e relazioni di significato

Nell'ambito educativo, un ruolo fondamentale è notoriamente svolto dall'insegnamento della lingua. Uno dei suoi obiettivi è quello della riflessione progressiva sul significato delle parole, sia in termini di aumento quantitativo delle parole conosciute sia in termini di relazione tra significati, come raccomandato nelle *Indicazioni nazionali per il curriculum della scuola dell'infanzia e del primo ciclo di istruzione* (MIUR 2012).

In particolare, nel lavoro educativo con bambini e ragazzi con bisogni linguistici speciali (bambini stranieri, sordi, con ritardo di linguaggio, neurodivergenti ecc.) l'ampliamento del vocabolario e un migliore utilizzo dello stesso è un obiettivo fondamentale, al fine di sviluppare competenze linguistico-comunicative carenti e, di conseguenza, migliorare la qualità di vita (Sebastiani 2014; Sebastiani e Valenzano 2023; Diadori et al. 2015).

Nella prassi educativa l'arricchimento del vocabolario dei bambini viene spesso promosso spiegando la parola sconosciuta con la sua definizione, solo verbalmente o al più elencandola in una rubrica. Dall'esperienza emerge tuttavia che ciò non basta: la memoria di queste parole resta evanescente, poco significativa, legata al «qui e ora» del contesto linguistico. Un apprendimento di questo tipo si può considerare «meccanico». L'apprendimento «significativo», invece, «si verifica quando chi apprende decide di mettere in relazione le nuove informazioni con le conoscenze che già possiede» (Novak 2012, 47).

È quindi evidente come sia necessario trovare nuovi percorsi teorici e didattici per riflettere sull'aspetto semantico.

Poiché «nella linguistica odierna è ormai assodato che le parole non sono unità isolate, bensì sono connesse da una fitta rete di relazioni associative» (Ca-

sadei e Basile 2019), riteniamo che sia possibile incoraggiare la comprensione del significato delle parole con l'uso di una *rete semantica*, cioè spingendo il discente non solo ad apprendere la definizione o il concetto, ma a riflettere sulle relazioni che sussistono tra parole e tra concetti, attraverso il collegamento tra parole nuove e parole già acquisite.

In generale, esistono diverse possibili rappresentazioni di reti semantiche. Nel caso più semplice esse sono formate da alberi gerarchici di parole, nelle quali i termini posti in basso sono iponimi di quelli più in alto (forma che corrisponderebbe a quella di un tesauro costituito esclusivamente da relazioni BT/NT). All'estremo opposto troviamo reti molto estese, come le mappe concettuali di Novak, nelle quali le parole possono essere collegate da relazioni di qualsiasi tipo, in modo diverso a seconda del contesto di apprendimento. Mentre le prime ci sembrano limitanti della varietà di relazioni tra termini, le seconde ci sembrano poco chiare e legate a una proliferazione contestuale di relazioni.

È per noi necessario creare invece una struttura visiva a rete regolata da principi chiari, riproducibile e riconoscibile. Essa non solo dovrà essere una guida pratica per l'elaborazione delle informazioni semantiche, ma pensiamo possa fornire un modello operativo, uno «schema per organizzare conoscenze, una struttura in cui incorporare e attraverso la quale interpretare nuove informazioni, e punto di partenza da cui lanciare azioni future» (Borkowsky e Muthukrishna 2011, 16). Si può così adottare un'ottica metacognitiva, nella quale non solo si apprende, ma si *impara ad apprendere*: una rete semantica dovrà riprodurre nel piccolo la rete dei significati presenti nella mente umana, fornendo un modello visivo del suo funzionamento e accrescimento.

Utilizzando le parole di Mara Vendrame (in Sebastiani 2014, 120), pensate per i discenti sordi ma estendibili a tutti, «le rappresentazioni visuo-spaziali possono così facilitare la consapevolezza di aver creato nuovi schemi di conoscenza, attraverso tre principali processi»: accrescimento, creazione e aggiustamento di collegamenti in una rete semantica.

3. Il metodo Ma.Vi.

A partire dal 2003 Maria Michela Sebastiani, Emanuela Valenzano e collaboratori hanno sviluppato il metodo *Ma.Vi.* (*marcatori visivi*) inizialmente destinato all'educazione linguistica di bambini e ragazzi sordi, ma in seguito esteso a qualsiasi difficoltà di apprendimento della lettura e della scrittura e allo stesso insegnamento quotidiano a scuola nelle ore di italiano (Sebastiani e Valenzano 2023).

I Ma.Vi., il cui scopo è migliorare la comprensione e la produzione di testi, si articolano su diversi livelli linguistici – fonologia, morfologia, sintassi, testualità – che accompagnano l'apprendimento della lettura e della scrittura e

sono accomunati dall'utilizzo di movimenti e segni grafici colorati, che aiutano ad aumentare la consapevolezza delle relazioni esistenti fra le diverse parole e parti di un testo.

L'incontro con l'organizzazione della conoscenza stimola l'estensione del metodo alla componente semantica: i termini incontrati all'interno di un testo possono essere messi in relazione con altri termini conosciuti, spostandosi così dal testo al vocabolario, considerato come un vero e proprio KOS.

Il contesto linguistico da cui si parte è dunque il testo (contesto sintagmatico), nel quale la parola ha un preciso significato in relazione a quanto si sta dicendo. Identificando e selezionando una parola il cui significato è sconosciuto, o sulla quale si vuole lavorare, essa viene estratta dal contesto testuale e posta al centro della rete semantica. A questo punto la riflessione riguarda il suo significato a prescindere dal suo uso nel testo iniziale: il contesto diventa paradigmatico.

Ad esempio, in un testo sugli uomini primitivi che costruivano tende, la parola *tenda* viene estratta e analizzata nei suoi significati a prescindere da quello particolare del testo: ad esempio con la relazione "di cosa è fatta?" si potranno indicare diversi materiali, compresa la plastica che non era ovviamente considerata nel contesto sintagmatico. Questa riflessione a prescindere dal testo di partenza permetterà un migliore ampliamento del vocabolario e una permanenza in esso più lunga, rispetto alla semplice definizione contestuale.

Certo, l'obiettivo è che quel termine, acquisito o su cui è migliorata la consapevolezza del significato, sia poi utilizzato dal discente all'interno di precisi contesti sintagmatici (ad es. parlando delle tende per il campeggio, delle tende degli indiani ecc.). In altre parole, attraverso una visualizzazione delle relazioni tra significati si ottiene un apprendimento significativo.

4. Le categorie dell'ILC

L'organizzazione della conoscenza offre qui un contributo determinante nell'aver identificato degli insiemi di poche relazioni chiave alle quali si può ricondurre qualsiasi collegamento tra concetti. La tradizione dell'analisi a faccette, iniziata da S.R. Ranganathan e sviluppata dal Classification Research Group, ha infatti definito certe *categorie fondamentali* che in ciascun dominio di conoscenza si realizzano nelle particolari *faccette* che lo caratterizzano (Gnoli 2004).

Da questa tradizione prende le mosse anche il progetto della Integrative Levels Classification (ILC), che anziché discipline ("zoologia") elenca fenomeni del mondo reale ("animali") ed appare perciò più vicina al vocabolario comune che si incontra durante l'apprendimento.

Il sistema di categorie fondamentali della ILC è composto, oltre che dalle sottoclassi ovvero tipi, da dieci relazioni rappresentate da cifre, che ne control-

lano meccanicamente l'ordine di presentazione in un elenco di voci o sugli scaffali di una biblioteca. Nelle tavole dello schema (ISKO Italia 2023) ciascuna di esse è visualizzata con un proprio colore, scelto in modo da assegnare colori più scuri alle faccette con valore ordinale maggiore e da avere colori adiacenti sufficientemente diversi.

Ogni categoria esprime un tipo di relazione attraverso cui qualsiasi fenomeno si può rapportare ad altri fenomeni: di esso è così possibile specificare le dimensioni, le parti e i materiali componenti, le proprietà, gli aspetti dinamici, i conflitti interni o esterni, i fattori che agiscono su di esso, il contesto spaziale e quello temporale; la categoria finale della prospettiva permette di esprimere una generica relazione di un fenomeno con qualsiasi altro o dal punto di vista di qualsiasi altro, analogamente al simbolo dei due punti nella Classificazione Decimale Universale.

Ai fini delle esperienze nella scuola primaria, i nomi delle categorie e i loro codici numerici sono stati sostituiti da delle semplici domande alle quali la categoria risponde:

<i>a-y</i>	tipo	“che cos'è?”	nero
9	qualità	“quale, di/per cosa è?”	marrone
8	quantità	“quanto è?”	viola
7	parte	“con/di cos'è fatto?”	verde scuro
6	proprietà	“com'è?”	rosso
5	cambiamento	“cosa gli succede?”	arancione
4	opposizione	“a cosa si oppone?”	indaco
3	agente	“da cosa è cambiato/fatto?”	rosa
2	luogo	“dov'è?”	verde chiaro
1	tempo	“quando è?”	azzurro
0	prospettiva	“a cos'altro lo colleghiamo?”	giallo

5. Le reti semantiche: una prima sperimentazione

La sperimentazione avviata consiste nell'applicazione delle categorie della ILC al lavoro semantico, volto a far prendere coscienza ai discenti della diversa natura delle relazioni fra parole.

La parola target è posta al centro e da essa partono i collegamenti colorati verso altre parole. Il risultato è una rete semantica a raffigurazione radiale che permette di visualizzare i legami tra le parole.

Le attività didattiche possono essere diverse: posizionare in modo corretto le parole, scoprire i legami tra parole, costruire il significato di una parola nuova attraverso i suoi collegamenti ecc.

Le prime esperienze ci hanno condotto ad affinare alcuni aspetti del lavoro. Abbiamo innanzitutto migliorato la forma delle domande in modo da evitare che risultassero fuorvianti rispetto alla logica delle categorie; ad esempio, per la

categoria del *cambiamento* la domanda è stata corretta dall'iniziale "cosa fa?" a "cosa gli succede?" in modo da suggerire cambiamenti intransitivi o passivi, il cui agente sia eventualmente esprimibile nella relativa categoria, come previsto nell'analisi a faccette che pone al centro dell'interesse (ovvero in prima posizione nell'ordine di citazione) l'entità che subisce il cambiamento.

Altro aspetto importante è stato la necessità di utilizzare un colore neutro (nero o grigio o legno) per le parole, riservando la marcatura colorata alle sole relazioni fra parole: infatti ogni parola presente nella parte periferica, legata a quella centrale da un certo colore, potrebbe essere a sua volta al centro di nuovi legami di colori diversi.

Rispetto al sistema di dieci categorie, abbiamo inoltre ritenuto che le prime esperienze nella scuola primaria debbano limitarsi a un insieme più ristretto, comprendente ad esempio soltanto tipo, parte, proprietà e cambiamento, con l'aggiunta successiva di quantità, agente, luogo e tempo ed eventualmente un uso della categoria di opposizione quando utile. Un uso dell'intero insieme di categorie ci sembra ipotizzabile dalle classi successive, ampliando le riflessioni con i ragazzi della scuola secondaria; alcune categorie di base (tipo, parte, proprietà) potrebbero essere applicate anche nella scuola dell'infanzia.

Il protocollo Ma.Vi. Semantica, da noi sviluppato insieme alle insegnanti di scuola primaria Lisa Borgogno e Rossella Buemi, dovrebbe dunque comprendere le seguenti fasi:

- introdurre l'idea di rete semantica;
- trovare parole collegate a una parola di riferimento con una relazione data;
- identificare le corrette relazioni tra parole date;
- aumentare il numero di relazioni considerate;
- apprendere parole nuove mediante la rete semantica;
- comprendere l'asimmetria delle relazioni;
- collegare diverse reti.

Dopo una breve introduzione mediante spiegazione, i bambini di una classe che adotta il protocollo vengono coinvolti in giochi guidati dall'insegnante che realizzano le fasi successive. Si può lavorare su diversi supporti (lavagna, LIM, fogli, cartelloni...) e con materiali diversi.

In particolare, ci sembra significativa la possibilità di proporre giochi di movimento, nei quali visualizzare le relazioni tra parole: i bambini si identificano con le parole, si posizionano all'interno di cerchi e si collegano tra loro tramite fettucce colorate dei colori standard delle categorie ILC. Tale metodologia permette di riflettere sulle relazioni attraverso un'esperienza motoria, più significativa rispetto alla semplice attività "a tavolino" (sull'apprendimento dell'italiano attraverso il movimento vedere Marchetti e Valenzano 2025).

Una prima sperimentazione è stata svolta nel corso dell'a.s. 2024-2025 da insegnanti Esperte Ma.Vi. con classi pilota nelle scuole primarie di Buttigliera d'Asti (Istituto comprensivo di Castelnuovo Don Bosco), di Tigliole (I.C. San Damiano d'Asti) e "Bernard Damiano" di Cuneo. I primi risultati indicano ovunque:

- una comprensione quasi immediata delle categorie da parte degli scolari;
- una loro buona capacità di assegnare le parole a categorie corrette;
- segnali di miglioramento nella comprensione dei nuovi significati.

6. Discussione e sviluppi

Nell'a.s. 2025-2026 si prevede un'ulteriore sperimentazione, nella forma di studio di caso, finalizzata a raccogliere dati quantitativi sull'applicazione del protocollo Ma.Vi. semantica e, in particolare, a comprendere se il protocollo migliori l'acquisizione di nuovi significati, aiuti l'elaborazione di significati già conosciuti e favorisca la riflessione sul tipo di connessione tra due significati.

L'integrazione tra metodi per l'educazione linguistica e metodi sviluppati originariamente per scopi bibliografici, quale l'analisi a faccette, ci appare fin d'ora promettente. Emerge la necessità di armonizzare le terminologie delle diverse tradizioni e di adattare alcuni principi nati per funzioni diverse, quali la notazione o la costruzione passiva su cui è impostata l'analisi a faccette. Rimarrebbe peraltro da esplorare l'utilità cognitiva e didattica delle funzioni di ordinamento sequenziale fra le diverse componenti semantiche (ordine di citazione delle faccette, ordine fra i loro possibili valori), messa a punto nella tradizione classificatoria.

L'intuitiva consonanza di senso riscontrata tra applicazioni educative e applicazioni bibliografiche è probabilmente un effetto della comunanza di certi principi fondamentali che risiedono alla base dell'organizzazione della conoscenza umana: ibridazioni come quella qui presentata possono aiutare ad evidenziare meglio tali principi in termini generalizzati e meno gergali.

A maggior ragione, un rilancio delle applicazioni della KO al settore educativo, eccessivamente trascurato dalla letteratura recente, sembra quindi rappresentare una fruttifera direzione di ricerca.

Riferimenti bibliografici

- Bliss, Henry Evelyn. 1930. "The organization of knowledge and the system of the sciences." *Nature* 126 (199). <https://doi.org/10.1038/126199b0>.
- Borkowsky, John G., e Nithi Muthukrishna. 2011. *Didattica metacognitiva. Come insegnare strategie efficaci di apprendimento*. Erickson.

- Casadei, Federica, e Grazia Basile. 2019. *Lessico ed educazione linguistica*. Carocci.
- Comenius, Iohannes Amos. 1639. *Pansophiae prodromus*.
- Diadori, Pierangela, Massimo Palermo, e Donatella Troncarelli. 2015. *Insegnare l'italiano come seconda lingua*. Carocci.
- Dickson, Shirley V., Deborah C. Simmons, and Edward J. Kameenui. 1995. *Text organization and its relation to reading comprehension: a synthesis of the research*. Technical Report 17, National Center To Improve the Tools of Educators. <https://eric.ed.gov/?id=ED386864>.
- Gnoli, Claudio. 2004. *Classificazione a faccette*. Enciclopedia tascabile 26. Associazione Italiana Biblioteche.
- ISKO Italia (International Society for Knowledge Organization Italia). 2023. "ILC developing version." Consultato il 5 novembre 2025. <https://www.iskoi.org/ilc/ilc.php>.
- Jonassen, David H., Katherine Beissner, and Michael Yacci. 1993. *Structural knowledge: Techniques for representing, conveying and acquiring structural knowledge*. Lawrence Erlbaum Associates.
- Marchetti, Rosalba, e Emanuela Valenzano. 2025. *Joy of moving: gioco e leggo, muovo e scrivo con i Ma. Vi.*, con la supervisione di Caterina Pesce. Calzetti & Mariucci.
- Metodo Ma.Vi. (Il) Consultato il 5 novembre 2025. <https://ilmetodomavi.org/>.
- Mazzoni, Giuliana. 2000. *L'apprendimento: comportamento, processi cognitivi, neurobiologia*. Carocci.
- McCagg, Edward C., and Dansereau, Donald F. 2010. "A convergent paradigm for examining knowledge mapping as a learning strategy." *Journal of Educational Research* 84: 317-24. <https://doi.org/10.1080/00220671.1991.9941812>.
- MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca). 2012. *Indicazioni nazionali per il curriculum della scuola dell'infanzia e del primo ciclo di istruzione*. https://www.mim.gov.it/documents/20182/51310/DM+254_2012.pdf.
- Morsy, Leila, Michael Kieffer, and Catherine Snow. 2010. *Measure for measure: a critical consumer's guide to reading comprehension assessments for adolescents*. Final report for Carnegie Corporation of New York's Council on Advancing Adolescent Literacy. <https://eric.ed.gov/?id=ED535299>.
- Novak, Joseph D. 2012. *Costruire mappe concettuali: strategie e metodi per utilizzarle nella didattica*. Erickson.

- Rossi, Paolo. 2000. *Clavis universalis: arti della memoria e combinatoria da Lullo a Leibniz*, 3a ed. Il Mulino.
- Sebastiani, Maria Michela, e Emanuela Valenzano. 2023. *Educare alla lettura con il metodo Ma.Vi.*, 2a ed. Amazon Italia.
- Sebastiani, Maria Michela. 2014. *Processi pedagogici e sordità: ipotesi di interventi autenticamente educativi*. Aracne.
- Soergel, Dagobert. 2013. "Knowledge organization for learning: Conjectures and methods of study." extended and enlarged version of the invited presentation at ISKO 2014, *Ciência da informação*, 42(2): 232-54.
- Szostak, Rick. 2024. "Additional purposes of knowledge organization." In *Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities: Proceedings of the Eighteenth International ISKO Conference 20-22 March 2024 Wuhan, China*, edited by Wei Lu, Thiago Henrique Bragato Barros, Lu An and Daniel Martínez-Ávila. Advances in knowledge organization 20. Würzburg: Ergon, Baden-Baden, 209-18.
- Thagard, Paul. 1993. *Conceptual revolutions*. Princeton University Press.

Strumenti di convergenza tra archivi e biblioteche: i thesauri

Manuela Grillo*

Abstract: The aim of the article is to reflect on the use of thesauri in an archival context: as systems of authority control, thesauri have contributed to building “bridges” between cultural institutions such as libraries, archives and museums. Thesauri have contributed to the pooling of cultural resources between cultural institutions. The contribution proposes a reflection on some of the most significant national and international experiences, examining in detail methods and characteristics of a specialised thesaurus created and used for the semantic indexing of ancient archival documents, developed as a knowledge organisation tool for the database of the European Research Council research project “Practicing, Producing and Protecting the Arts (1650-1750). Towards a Shared History of Performing Arts in Rome in Family Archives”. The main objective of the project is the study of the performing arts in the milieu formed by aristocratic families in 17th and 18th century Rome.

Keywords: GLAM, Knowledge Organisation, Performart Thesaurus, Semantic Indexing, Thesauri.

1. Introduzione

Il presente contributo ha per oggetto i thesauri come strumenti di concreta convergenza tra archivi, biblioteche e altre istituzioni di conservazione della memoria.

Grazie alle loro caratteristiche di standardizzazione, i thesauri hanno progressivamente acquisito nuovi ruoli, nuove funzioni, oltre al tradizionale e consolidato impiego per l'indicizzazione semantica dei contenuti, e una importanza sempre maggiore nell'ambiente dei linked data del web semantico (Guerrini e Possemato 2015; Marinelli 2018; Daquino 2021), per il loro contributo alla modellizzazione della conoscenza, fissando la semantica che le ontologie – descrizioni formali esplicite dei concetti di un dato dominio – arricchiscono con

* Sapienza Università di Roma (Biblioteca centrale della Facoltà di Ingegneria civile e industriale “G. Boaga”, Roma, Italia). *Sapienza University of Rome (Library “G. Boaga”, Faculty of Civil and Industrial Engineering; Rome, Italy).* manuela.grillo@uniroma1.it. ORCID: 0000-0002-1179-127X.

modelli concettuali e teorie logiche per consentire il ragionamento da parte degli agenti software (Tomasì 2022).

I thesauri si dimostrano strumenti dinamici che supportano le potenzialità del machine learning e dell'intelligenza artificiale, nonché dei motori di ricerca, perché con i thesauri è possibile risolvere quei problemi connessi all'uso del linguaggio naturale che – con polisemie, sinonimie, omonimie, ambiguità – ne inficiano la piena funzionalità.

Essi sono componenti essenziali per l'integrazione dei dati sul web, in particolare per la mappatura e l'assistenza all'interoperabilità tra risorse eterogenee: con l'adozione dei formati del web semantico, come RDF/SKOS (Resource Description Framework/Simple Knowledge Organization System), e seguendo gli standard, i thesauri si sono evoluti permettendo il riutilizzo tra diversi framework; consentono di aumentare il multilinguismo e le equivalenze concettuali, collegando informazioni e metadati prodotti da istituzioni di diversi Paesi.

Un fattore importante nella spinta ad usare strumenti comuni di descrizione nell'ambito di contesti documentari diversi è stata la costruzione di portali che danno accesso a risorse informative di varia natura. Tuttavia, relativamente alla forma dei nomi e alla ricchezza delle informazioni che li qualificano, nonché all'analisi semantica, persiste un significativo divario fra i mondi di biblioteche, musei ed archivi.

In particolare, il contributo intende riflettere sull'uso di thesauri – strumenti di Knowledge Organization (Gnoli et al. 2006) tipicamente sviluppati per l'indicizzazione semantica in ambito bibliotecario – in contesto archivistico. In quanto sistemi di controllo dell'autorità, negli ultimi anni i thesauri hanno contribuito a costruire “ponti” tra istituzioni culturali e di conservazione della memoria come biblioteche, archivi e musei, contribuendo alla messa a fattore comune delle risorse culturali (Lucarelli 2022a). Di qui l'interesse per una riflessione a partire da alcune delle più significative esperienze di indicizzazione semantica di documenti archivistici condotte in ambito nazionale ed internazionale.

In un contesto archivistico tradizionale le logiche di ricerca e di recupero delle informazioni in un archivio storico sarebbero diverse e legate al contesto di produzione dei documenti. Ma la varietà di fonti e la granularità di corpora documentari “creati” dall'assemblaggio dei ricercatori, in cui le singole unità documentarie sono decontestualizzate (Valacchi 2015), rendono l'indicizzazione semantica e l'uso di un thesaurus molto utili, perché al contenuto informativo delle risorse documentarie inserite in un archivio digitale, frutto delle attività di ricerca, vengono infatti a mancare quegli elementi di contesto che ne qualificano la piena portata informativa (Valacchi 2023).

L'uso di tecnologie legate al semantic web ha peraltro contribuito a ridefinire modi e sistemi di valorizzazione del potere informativo espresso dalle

descrizioni archivistiche, favorendo considerazioni sulla natura stessa dei dati archivistici e sul loro riuso nell'ottica di un dialogo con altre fonti di dati, al fine di estendere le capacità semantiche ed espressive del patrimonio informativo negli archivi custodito (Tomasi e Daquino 2015).

2. Indicizzazione semantica in contesto archivistico: alcune esperienze

Nel panorama nazionale, rilevanti esperienze sono state realizzate da un gruppo di lavoro afferente a Musei Archivi Biblioteche (MAB) Toscana, comitato territoriale di MAB Italia, coordinamento permanente per esplorare le prospettive di convergenza tra i mestieri e gli istituti in cui operano i professionisti degli archivi, delle biblioteche, dei musei¹. Il Gruppo linguaggi di MAB Toscana ha dato conto dell'esperienza in un incontro pubblico – realizzato presso la Soprintendenza archivistica per la Toscana nell'ottobre 2015 – dal titolo “Il nome delle cose, il linguaggio come punto d'incontro tra archivi, biblioteche e musei” (Capetta 2015), con l'intento di presentare il lavoro del gruppo, i risultati raggiunti e le progettualità future, nella speranza di incuriosire e stimolare altri a lavorare sulla scia dell'integrazione. Il Gruppo ha individuato come macro temi di riflessione da un lato l'integrazione e l'uniformazione degli elementi descrittivi prodotti da istituzioni di natura diversa e dall'altro le modalità di codifica e architettura dei dati catalografici prodotti. Esaminati diversi sistemi italiani e stranieri per evidenziarne le caratteristiche tecniche, le modalità di integrazione e le funzionalità di ricerca, il Gruppo ha analizzato alcuni aspetti organizzativi dei progetti individuati come casi di studio: modello di collaborazione, risorse utilizzate e natura dei finanziamenti. Nell'esame dei sistemi, l'attenzione si è focalizzata sulle questioni relative agli standard adottati, alle soluzioni per l'interoperabilità, volte a garantire la ricerca simultanea su più fonti, e alla navigabilità fra le informazioni (Bruni et al. 2016).

Il gruppo ha sostenuto iniziative di indicizzazione per soggetto di risorse archivistiche attraverso metodologie e strumenti nati in ambito bibliotecario, in particolare con l'utilizzo del Thesaurus generale e multidisciplinare Nuovo Soggettario, realizzato dalla Biblioteca Nazionale Centrale di Firenze (BNCF) e impiegato da biblioteche italiane e altre istituzioni nell'indicizzazione per soggetto di risorse di varia natura. Nuovo Soggettario è caratterizzato da formati aperti e i protocolli standard che facilitano progetti di integrazione fra data base di diverse istituzioni della memoria: i riferimenti normativi sono lo standard ISO 5963:1985 (rivisto e confermato nel 2020, ha per oggetto i metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei ter-

¹ MAB è l'acronimo con cui Associazione Italiana Biblioteche (AIB), Associazione Nazionale Archivistica Italiana (ANAI) e International Council of Museum – Comitato Nazionale Italiano (ICOM Italia) hanno dato vita al coordinamento permanente.

mini di indicizzazione), gli standard ISO 25964-1:2011 e ISO 25964-2:2013 (rivisti e confermati rispettivamente nel 2017 e 2018, hanno per oggetto la costruzione e lo sviluppo di thesauri per il recupero dell'informazione nella Parte 1 e l'interoperabilità fra differenti thesauri e con altri tipi di sistemi di organizzazione della conoscenza nella Parte 2), i modelli concettuali della famiglia FR (Functional Requirements) sviluppati nel corso degli anni dall'International Federation of Library Associations and Institutions (IFLA), nati dall'esigenza di comprendere, descrivere e rappresentare formalmente le informazioni bibliografiche (Biblioteca Nazionale Centrale di Firenze 2021, Lucarelli 2022b).

Proprio allo scopo di incrementare la fruibilità e l'integrazione delle risorse degli istituti culturali, i termini del Thesaurus Nuovo Soggettario sono collegati a voci corrispondenti in altri vocabolari o altri strumenti impiegati per l'indicizzazione, la descrizione o denotazione delle proprie risorse in archivi o musei. In particolare, il Thesaurus è collegato al catalogo della BNCF e a quello del Servizio Bibliotecario Nazionale (SBN), mediante il menù Notizie bibliografiche che si trova all'interno del record di ciascun termine preferito. Per entrambi i cataloghi, il link Opere indirizza l'utente alle notizie bibliografiche contenenti la parola cercata all'interno delle stringhe di soggetto. Nel campo Risorse di Archivi e Musei, uno o più link permettono di accedere alle risorse presenti in banche dati di archivi e/o cataloghi online di archivi e musei (al momento, le Gallerie degli Uffizi, il Museo della Fondazione scienza e tecnica, il Museo Galileo e l'Archivio Ricordi), indicizzate o descritte con lo stesso termine del Thesaurus. Questa integrazione di risorse è un esempio di come a partire da un termine di thesaurus si possa migliorare ed aumentare anche l'esperienza utente attraverso il disvelamento di legami semantici fra le fonti (Fig. 1).

Figura 1. Collegamento a risorse di banche dati o cataloghi online a partire dai termini del Nuovo Soggettario.

Inoltre da anni è garantita l'interoperabilità con gli Archivi di personalità del Sistema Informativo Unificato per le Soprintendenze Archivistiche (SIUSA). Questo aspetto è stato alla base dell'esperienza di indicizzazione delle carte di Giovanni Pascoli, digitalizzate a cura della Soprintendenza archivistica e bibliografica della Toscana (Capetta 2017), così come dell'indicizzazione semantica dell'archivio di persona appartenuto ad Ernesto Rossi, conservato a Firenze, presso gli Archivi storici dell'Unione Europea² che dal 1983 custodiscono e rendono disponibili archivi di istituzioni europee, documentazione di singoli e di organizzazioni internazionali legate al processo di integrazione europea.

L'archivio Rossi, depositato nel 1999, completamente digitalizzato e consultabile liberamente online, contiene una ricchissima quantità di materiali sull'attività del politico antifascista, federalista europeo, economista, giornalista e scrittore. Il valore culturale di questo archivio (peraltro sottoposto in passato a una parziale soggettazione in francese), le sue caratteristiche, l'accuratezza del suo trattamento nel data base che lo ospita, lo hanno reso particolarmente interessante per sperimentarne l'indicizzazione.

Organizzato sulla base di criteri tematici e cronologici, il Fondo Ernesto Rossi è diviso in sottofondi, serie, sottoserie (tutti corredati di abstract), unità

² Gli Archivi storici dell'Unione europea (ASUE) sono il centro di deposito ufficiale dei documenti d'interesse storico prodotti dalle istituzioni dell'Unione europea. Gli archivi sono parte integrante dell'Istituto Universitario Europeo (IUE) di Fiesole (Firenze).

archivistiche (fascicoli) e documenti. L'indicizzazione, che è stata attuata per alcuni dei suoi vari livelli, in molti casi fino a quello più specifico dell'unità documentaria, oltre a dare l'occasione di integrare il vocabolario di terminologia non ancora presente nel Thesaurus, ha avuto l'obiettivo di testare l'applicabilità del Nuovo soggetto in contesti extra-bibliotecari e di attivare – tramite lo “snodo” dei dati aperti del Thesaurus – l'interoperabilità fra cataloghi di biblioteche, in particolare quello della BNCF collegato al Thesaurus, e i data base degli Archivi storici dell'Unione europea (Becherucci et al. 2016).

La pratica di indicizzazione della semantica dei documenti archivistici è attestata anche a livello internazionale: ad esempio, la Library of Congress soggetta anche i documenti archivistici usando le intestazioni di soggetto Library of Congress Subject Headings (LCSH), senza limite numerico, con la conseguenza che talvolta una risorsa documentaria ha decine di voci di soggetto. Questo aspetto, attinente a questioni di scelte indicizzatorie, merita una riflessione – anche in relazione al PerformArt Thesaurus (PATH) di cui si tratterà di seguito – legata alla differenza tipologica delle risorse archivistiche rispetto a quelle bibliografiche.

Si tratta peraltro, ad avviso di chi scrive, di uno degli aspetti più delicati ed interessanti dell'indicizzazione semantica di materiale d'archivio, sia esso moderno o antico.

3. Un thesaurus per l'indicizzazione semantica di documenti antichi: il caso di studio del PATH

Su questa scia si inserisce l'esperienza che ho condotto e curato con la creazione e l'uso di un thesaurus concepito per l'indicizzazione semantica di documenti d'archivio antichi.

Il presente capitolo si propone quindi di illustrare metodi e caratteristiche di un thesaurus specializzato sviluppato ed usato come strumento di organizzazione della conoscenza per il database del progetto di ricerca European Research Council (ERC) intitolato *Practicing, Producing and Protecting the Arts (1650-1750). Towards a Shared History of Performing Arts in Rome in Family Archives*³.

PerformArt ha studiato la storia delle arti dello spettacolo a Roma a partire dagli archivi di famiglia dell'alta aristocrazia tra 1644 e 1740, allo scopo di

³ Il progetto *Promoting, Patronising and Practising the Arts in Roman Aristocratic Families (1644-1740). The Contribution of Roman Families' Archives to the History of Performing Arts*, finanziato da European Research Council (ERC) nell'ambito del programma di innovazione e ricerca dell'Unione europea Horizon 2020 (*grant agreement* n° 68415), si è svolto dal settembre 2016 all'agosto 2022, sotto la direzione di Anne-Madeleine Goulet (Centre d'Etudes Supérieures de la Renaissance du Centre National de la Recherche Scientifique – CESR-CNRS) e il coordinamento delle attività di ricerca di Michela Berti.

approfondire il ruolo delle arti nella vita quotidiana delle élite, di chiarire le condizioni e le motivazioni del loro mecenatismo e di valutare l'importanza delle arti nel processo di costruzione identitaria di queste grandi famiglie. Gli obiettivi erano dichiaratamente studiare il posto delle arti sceniche nel sistema di magnificenza aristocratica che prevaleva a Roma e proporre un modello per l'analisi storica degli spettacoli, basato sul concetto di performance.

L'équipe di ricerca – costituita da trenta studiosi, tra cui storici, musicologi, storici del teatro, storici della danza, storici dell'economia, provenienti da otto Paesi europei – ha effettuato una descrizione sistematica dei documenti presenti in undici fondi d'archivio familiari conservati principalmente a Roma, Frascati e Subiaco, per raccogliere il materiale necessario a rinnovare lo studio delle arti dello spettacolo nell'ambito delle famiglie aristocratiche romane: si tratta degli archivi Aldobrandini, Borghese, Caetani, Chigi, Colonna, Lante della Rovere, Orsini, Ottoboni, Pamphilj, Ruspoli e Vaini; l'investigazione è stata inoltre completata da ricerche svolte nel fondo del Collegio Nazareno, in quello dell'Accademia dell'Arcadia e nell'Archivio del Sovrano Ordine di Malta.

Lo scopo della ricerca era studiare un patrimonio culturale immateriale – le arti performative: musica, danza, teatro, eventi transitori e unici – attraverso l'utilizzo e l'analisi di patrimonio culturale materiale: documenti archivistici (in particolar modo di archivi familiari) ma anche partiture, iconografie, mappe antiche, oggetti, materiali e abiti di scena, etc.: le ricerche hanno implementato un repository digitale attraverso il quale la connessione tra fonti e dati di diversa natura è stata esplicitata, ha permesso di rendere la complessità del sistema mecenatistico e ha consentito la ricostruzione degli eventi performativi, patrimonio culturale definito dall'immaterialità per antonomasia⁴.

La creazione di un thesaurus, pur prevista dalla Principal Investigator (PI) nelle specifiche iniziali del progetto (Goulet 2020), non era stata elaborata e dettagliata nella sua struttura e nel piano di sviluppo. Tuttavia, la grande quantità di dati raccolti ha reso necessario sviluppare uno strumento per il recupero dell'informazione all'interno del corpus documentario per garantire un efficace recupero dell'informazione.

⁴ Il tema del rapporto tra patrimonio culturale materiale e patrimonio culturale immateriale è stato approfondito in una relazione, condivisa con Michela Berti (Conservatorio F. Morlacchi di Perugia), dal titolo *PerformArt Thesaurus (PATH): uno strumento per la semantica delle arti dello spettacolo* per il XXXIII Convegno dell'Associazione Italiana per la Terminologia (Ass.I.Term) *Terminologia del patrimonio culturale materiale e immateriale: analisi e approcci di studio*, Università degli Studi di Napoli L'Orientale, 7 novembre 2023, i cui atti sono in corso di pubblicazione

3.1. Struttura del database

L'équipe di ricerca ha implementato un repository digitale, utilizzato come strumento di lavoro durante gli anni del progetto e divenuto al suo termine una fonte di informazioni per la comunità scientifica (PerformArt 2023).

Il DB è formato da diverse categorie, divise in tre grandi macro aree (Fig. 2):

- macro area delle Fonti per Documenti (documenti archivistici, ma anche partiture e libretti), Realia (descrizioni di oggetti come strumenti, abiti per commedie, scarpe da ballo, scenografie, etc.), Iconografia (descrizioni di documenti iconografici di ogni tipo come mappe, ritratti, incisioni di scenografie, etc.);
- macro area del Contesto, che si articola nelle categorie Persone (database prosopografico comprendente un rimando al Virtual International Authority File – VIAF), Collettività (gruppi di persone, sia istituzionalizzati che non istituzionali, come accademie, collegi, confraternite, cappelle musicali e le famiglie stesse), Opera (ogni tipo di produzione artistica), Luoghi (tutti i luoghi utili alla ricostruzione degli eventi performativi, ma anche alla biografia delle persone, ai luoghi di riunione delle collettività, etc.), Eventi (eventi storici, eventi relativi alla storia delle famiglie a Roma ed eventi culturali);
- macro area di Autorità di dati esterni al database, che si articola nelle categorie Bibliografia (sia monografie e articoli scientifici, che fonti stampate dell'epoca), Autori (autori delle pubblicazioni inserite nella categoria Bibliografia) e Thesaurus.



Figura 2. Struttura del database PATH.

Per questa architettura di dati, vista la quantità delle informazioni raccolte (circa 6.000 schede Documento e 8.000 schede Persona, oltre 200 schede Realia, Iconografia e Collettività, più di 1.600 schede Opera, oltre 1.200 schede Luoghi, circa 2.500 schede Evento), si è resa necessaria la costruzione e l'implementazione di uno specifico thesaurus per indicizzare il contenuto delle schede, attraverso l'attribuzione di parole-chiave scelte non dalla lingua naturale ma da un serbatoio di termini controllati, garantendo così la possibilità di recuperare le informazioni contenute nei records.

3.2. Caratteristiche e sviluppo

PATH è stato sviluppato in collaborazione con il progetto Nuovo soggettario. Il Thesaurus della Biblioteca Nazionale Centrale di Firenze è stato la base di sviluppo di PATH, sia relativamente alla macro-struttura che per molti termini generici, che sono stati importati. Allo stesso tempo, Nuovo Soggettario è stato arricchito dalle proposte di PATH, laddove la terminologia relativa alle arti performative e alla storia dell'economia era carente.

La scelta di sviluppare PATH a partire da una base thesaurale preesistente e multidisciplinare è derivata principalmente da valutazioni concernenti la sostenibilità del lavoro e da una attenta valutazione delle caratteristiche dello strumento.

Come già detto, Nuovo soggettario è un sistema di indicizzazione per soggetto di risorse di varia natura, realizzato in conformità ai principi stabiliti dall'IFLA e alle indicazioni degli standard internazionali, nonché disponibile in SKOS/RDF⁵ e presente fra i dataset disponibili nella “nuvola” dei Linked Open Data cloud (LOD cloud) sulla piattaforma dati.cultura.gov.it del Ministero della Cultura (MiC). Il Thesaurus Nuovo Soggettario, in continua evoluzione e crescita⁶, è parte di un sistema informativo finalizzato a descrivere le risorse, a mettere a disposizione dell'utente strumenti e modalità di ricerca efficaci che gli consentano di trovare e selezionare le più appropriate, di esplorare quelle che hanno fra loro relazioni significative, di scoprire nuove risorse e, infine, di accedere a quelle selezionate. Queste caratteristiche hanno portato alla firma del protocollo di collaborazione tra BNCF e il progetto di ricerca a partire dalla primavera 2018.

La struttura di PATH – mutuata dal thesaurus Nuovo Soggettario – comprende quattro macrocategorie (Agenti, Azioni, Cose e Tempo), caratterizzate da tredici categorie che coincidono con altrettanti *top terms*, ovvero i termini apicali delle categorie, corrispondenti a concetti generali. Si fornisce di seguito l'elenco delle tredici categorie e una breve nota di contenuto:

- Attività: l'insieme delle azioni coordinate o dei lavori compiuti, in generale o in rapporto a una finalità o ad un contesto particolare;
- Discipline: le azioni che, in particolari ambiti, prevedono apprendimento o insegnamento, ricerca e diffusione di conoscenze, strutture organizzative – come università, centri di studio, etc. – che fanno loro da supporto;

⁵ Famiglia di linguaggi formali progettati per rappresentare glossari, classificazioni, tassonomie e qualsiasi tipo di vocabolario strutturato, che formalizza in modo standard le relazioni semantiche tra i termini.

⁶ Alla data dell'ultimo aggiornamento (marzo 2025) contiene più di 74.490 (di cui 34.470 termini preferiti).

- **Forme:** rappresentazioni, aspetto esteriore, manifestazioni di entità, oggettuali o astratte o frutto del pensiero e della creatività, comprese le forme della comunicazione – ad es. letterarie, musicali, etc. – che possono avere anche una componente documentale;
- **Materia:** ciò che costituisce tutti i corpi, ossia la sostanza fisica, che, essendo dotata di massa che assume forme diverse nello spazio, può essere oggetto di esperienza sensibile ed è, in generale, concepita come esistente indipendentemente dalla coscienza individuale;
- **Oggetti:** ogni cosa, specialmente solida, che può essere percepita dai sensi e in particolare mediante la vista o il tatto;
- **Organismi:** l'insieme degli organismi umani, animali, vegetali, etc.;
- **Organizzazioni:** i complessi organizzati di risorse, di persone e di regole per il conseguimento di uno scopo;
- **Persone:** l'insieme delle varie categorie e tipologie di esseri umani, compresi i loro vari raggruppamenti e aggregati;
- **Processi:** gli insiemi di fenomeni in successione che presentino una certa unità o si svolgano in modo omogeneo e regolare, e in genere tutti gli aspetti della realtà in quanto siano espressione di un divenire;
- **Spazio:** entità illimitata, indefinita, o variamente limitata nella quale sono situati corpi o oggetti della realtà dotati di dimensioni ed eventualmente capaci di spostamenti;
- **Strumenti:** entità, elementi, cose, situazioni o strumenti astratti e concettuali che costituiscono mezzi o modalità per realizzare o caratterizzare qualcosa;
- **Strutture:** manufatti che occupano uno spazio in maniera stabile e che non sono mobili;
- **Tempo:** svolgimenti cronologici come elementi principali o esclusivi dei termini che appartengono alla faccetta.

PATh consta di circa 2.300 termini, utilizzati per indicizzare tutte le schede Eventi e parte delle schede Documenti, attraverso l'attribuzione di una o più parole chiave scelte tra i termini del thesaurus.

Per gli aspetti metodologici della redazione del thesaurus – rilevanti nella costruzione dell'albero thesaurale, quali la poligerarchia (condizione di un termine di avere due termini sovraordinati) e la disambiguazione (tecnica impiegata per distinguere i differenti significati di uno stesso termine) – si rinvia ad altro contributo, che entra nel vivo della costruzione del thesaurus, con un corredo di termini esemplificativi (Grillo 2021).

3.3. Le criticità legate alla diacronia

L'applicazione di un linguaggio di indicizzazione a materiale archivistico antico è uno degli aspetti che ha presentato più criticità ed è, pertanto, uno dei

più interessanti. In terminologia la dimensione diacronica è stata a lungo un terreno poco esplorato (Piccini et al. 2019) e in ambito biblioteconomico solo di recente è stata oggetto di specifica riflessione metodologica (Grillo 2015, Lucarelli 2019). Al tema è stata dedicata attenzione nell'ambito della riflessione biblioteconomica dell'Istituto Centrale del Catalogo Unico, in particolare nelle *Linee guida per la catalogazione per soggetto e la classificazione nel Servizio Bibliotecario Nazionale (SBN)* (Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche 2023): nel capitolo 3. Politiche di indicizzazione di particolari tipologie di risorse, paragrafo 3.4. Selezione cronologica, si dà conto di come tradizionalmente le opere vengano classificate a prescindere dalla data in cui sono state composte o editate, ma nella maggior parte delle biblioteche italiane opere del passato siano state escluse dalla indicizzazione per soggetto pur in presenza di un soggetto individuabile e di possibile interesse da parte degli utenti. Le linee guida sottolineano, al contrario, l'auspicabilità dell'indicizzazione per soggetto delle opere antiche, nei casi in cui ciò sia ritenuto utile per gli utenti (e/o possibile rispetto alle forze in organico nelle istituzioni bibliotecarie) e secondo le indicazioni fornite dalla seconda edizione della Guida (Biblioteca nazionale centrale di Firenze 2021)⁷.

Le domande fondamentali da porsi rispetto all'indicizzazione semantica di risorse antiche sono quindi relative alla fattibilità e all'utilità dell'operazione. Per ciò che concerne la fattibilità, essa attiene esclusivamente a scelte di opportunità dei singoli istituti e non ai principi dell'indicizzazione: né lo standard ISO 5963:1985 sull'analisi concettuale dei documenti né lo standard ISO 25964-1:2011 sui thesauri prevedono alcun criterio di selezione cronologica. IFLA Library Reference Model (LRM) fa riferimento alla "selezione cronologica", ma non come esclusione dalla soggettazione delle opere anteriori a una certa data, piuttosto come possibilità per l'utente di selezionare, in fase di ricerca, tramite la data di creazione dell'opera, le risorse recuperate in una ricerca per soggetto, potendo poi scegliere opere recenti per un'informazione corrente o potendo scegliere opere scritte prima di una certa data o in un determinato periodo, se l'utente interessato a conoscere il modo in cui un tema è stato inteso e spiegato in un determinato tempo.

Quindi, se gli stessi principi valgono sia per le opere moderne sia per quelle antiche, ciò che le differenzia ai fini dell'indicizzazione per soggetto sono alcune implicazioni, derivanti dalla dimensione temporale delle opere antiche, riguardanti sia l'analisi concettuale dei documenti che il controllo terminologico. Certamente in queste due operazioni ci sono delle complessità pratiche che richiedono attenzione, ma sulla fattibilità possiamo riassumere con le parole di Anna Lucarelli, scelte per il titolo di un suo contributo: "un'opera è un'opera" (Lucarelli 2019).

⁷ In particolare ai paragrafi C.2 e 3.6.2.5.1.

Per ciò che concerne l'utilità, il tema – ciò di cui un'opera tratta o *aboutness* – è l'oggetto centrale dell'interesse dell'utente e, se sarebbe assurdo pensare che tutte le opere del passato si debbano soggettare e/o sia utile fare questo, è pur vero che è opportuno indicizzare la semantica di un'opera che trasmette contenuti informativi interessanti per gli storici di una certa disciplina o per una qualunque categoria di utenti.

Rispetto alle criticità, sia nel caso di risorse bibliografiche che nel caso di risorse archivistiche, esse attengono in primo luogo all'analisi concettuale dei documenti.

Nel caso delle risorse bibliografiche, l'individuazione di temi pertinenti e rilevanti può essere resa più complessa dalla natura spesso composta, eterogenea del libro antico, che può rendere opportuna la segnalazione di più aspetti di uno stesso soggetto o di più temi specifici, siano essi trattati come temi indipendenti o facenti parte di un tema più ampio.

Nel caso delle risorse archivistiche, la molteplicità di argomenti spesso trattati all'interno di un singolo documento – si pensi ad esempio di un conto per l'allestimento di un evento performativi, contenente molteplici concetti a cui è potenzialmente possibile e utile dedicare un accesso semantico – fa sì che non si possa procedere a sommarizzazione⁸, ma che sia necessario stabilire a che livello di approfondimento di analisi procedere⁹.

In generale, nell'analisi concettuale delle risorse del passato la valutazione dei fattori contestuali e di quelli testuali può richiedere una procedura più complessa e articolata rispetto alle risorse moderne.

Altro tipo di criticità possono essere relative al piano terminologico, cioè alle operazioni per scegliere i termini appropriati per rappresentare i concetti, ove sono necessarie uniformità e univocità dei termini, in relazione all'evoluzione lessicale.

Occorre chiaramente evitare gli anacronismi derivanti dalla scelta di un termine associato oggi a un significato diverso da quello del passato, così come è da evitare l'uso di un termine che rappresenta un concetto non pertinente rispetto al contesto storico e culturale dell'epoca.

Se i termini di indicizzazione scelti per rappresentare un concetto sono normalmente quelli di uso corrente, è opportuno creare punti di accesso allo

⁸ In base al numero e alla tipologia di concetti identificati, selezionati e rappresentati nell'indicizzazione, si distinguono due procedimenti, noti come sommarizzazione e indicizzazione spinta: con il primo s'intende un livello di analisi finalizzato prevalentemente all'indicizzazione del tema di base di un'opera, con il secondo anche di temi particolari e di altre caratteristiche o aspetti del suo contenuto valutati come meritevoli di essere segnalati (Biblioteca Nazionale Centrale di Firenze 2021). La sommarizzazione è quindi una delle operazioni che caratterizzano le fasi dell'analisi concettuale delle risorse bibliografiche, volte ad individuare il tema di base della pubblicazione da esprimere con un'unica stringa di soggetto.

⁹ La questione verrà argomentata nel paragrafo 3.3 Politica di indicizzazione.

stesso concetto anche da termini impiegati in epoche passate e non più in uso, purché siano effettivamente attestati nelle opere e nelle fonti lessicografiche.

In fase di strutturazione del PATH, in ragione dell'altezza cronologica del materiale trattato, la diacronia ha spesso comportato delle criticità, non tanto per le attestazioni terminologiche dei concetti nei repertori, quanto per la gestione dei termini descrittivi all'interno delle gerarchie thesaurali generaliste e relative a concetti contemporanei derivate da Nuovo Soggettario, ma la struttura modulare ha permesso la gestione di termini rappresentanti concetti moderni assieme a termini che rappresentano concetti antichi.

Talvolta è stato necessario discostarci dalle gerarchie di Nuovo Soggettario, in particolar modo per quei termini rappresentanti concetti che appartengono tanto al mondo seicentesco quanto al mondo attuale e che non risentono quindi di problemi diacronici, ma che nel Nuovo Soggettario sono stati assunti in una gerarchia connotata dall'attributo della modernità, inaccettabile in un progetto di ricerca in chiave diacronica, inaccettabile: l'organizzazione dei concetti – e dei termini che li rappresentano – all'interno della struttura del PATH deve inevitabilmente riflettere il contesto e la temporalità dei concetti espressi, proponendo una categorizzazione e dei principi di raggruppamento/divisione dei concetti secondo caratteristiche comuni tipiche del dominio e del tempo. La questione ha riguardato tipicamente i mestieri: portiamo ad esemplificazione il caso del termine Fabbri (Fig. 3 e 4).

Fabbri

GERARCHIA

Macrocategoria: **Categoria** Agenti: Persone e gruppi

Usato per [Fabbri ferrai](#), [Magnani](#)

Termine apicale [Persone](#)

Termine più generale [Lavoratori del ferro](#)

Termine associato [Artigiani](#)

[Ferro](#)

[Fucina](#)

[Fucine](#)

[Maniscalchi](#)

Ha come Variante storica [Fabbri ferrai](#)

Nota storica Precedentemente: Fabbri ferrai (BNI 1956-1985)

Fonti BNI 1956-1985: Fabbri ferrai; Treccani.it; [DeM](#); PT; VT; WebDewey(IT); [Wikipedia\(IT\)](#)

DDC (WebDewey) 682.092

Equiv. in altri strumenti di indicizzazione

LCSH: [Blacksmiths](#)

RAMEAU: [Forgerons](#)

Notizie bibliografiche

- [Catalogo della BNCF](#)
 - Opere
 - Stringhe di soggetto
- [Catalogo SBN](#)
 - Opere

Suggerimenti sul termine

Figura 3. Termine Fabbri nel Thesaurus Nuovo Soggettario.



Figura 4. Gerarchia del termine Fabbri nel Thesaurus Nuovo Soggettario.

Assumere questa gerarchia all'interno di uno strumento di organizzazione della conoscenza dedicato all'antico era possibile solo saltando l'etichetta di nodo¹⁰ [Personae nelle attività del settore secondario]. Del resto, come è noto, le etichette di nodo non sono termini di indicizzazione e, anche nelle stesse gerarchie Nuovo Soggettario, possono cambiare in relazione al modo di classificare i termini, per cui saltarla – ed evitare l'anacronismo del concetto relativo “al settore secondario” in un thesaurus per documenti seicenteschi – ci è sembrato un ottimo compromesso (Fig. 5).

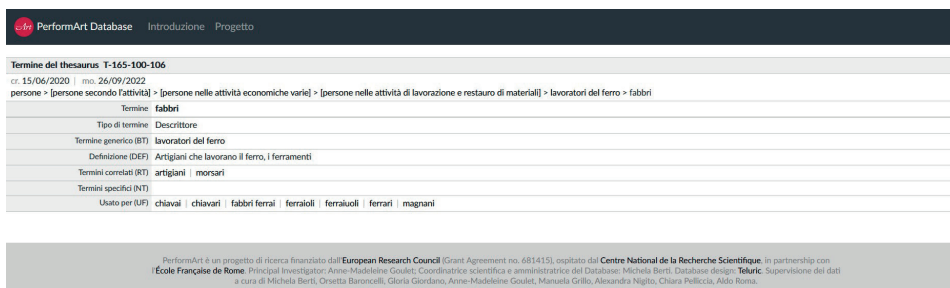


Figura 5. Termine Fabbri nel PATH.

La stessa logica è stata seguita per Parrucchieri che non potevano avere nel PATH un termine di gerarchia come [Personae nelle attività del settore terziario], saltata nella gerarchia del termine (Fig. 6).

¹⁰ Nella pratica dell'analisi a faccette, le etichette di nodo sono degli elementi di raggruppamento dei termini thesaurali secondo una certa caratteristica.

PerformArt Database		Introduzione	Progetto
Termine del thesaurus T-161-030-180			
cr. 02/06/2020 mo. 02/06/2020			
persone > [persone secondo l'attività] > [persone nelle attività economiche varie] > parrucchieri			
Termine	parrucchieri		
Tipo di termine	Descrittore		
Termine generico (BT)	[persone nelle attività economiche varie]		
Definizione (DEF)			
Termini correlati (RT)			
Termini specifici (NT)			
Usato per (UF)			

Figura 6. Termine Parrucchieri nel PATH.

Ovviamente il problema si è posto anche per tutti quei mestieri che appartengono al passato, come i Nevaroli (ovvero coloro che, in passato, raccoglievano, conservavano e vendevano la neve ghiacciata per conservare alimenti o realizzare gelati), che in Nuovo Soggettario vengono considerati un termine gerarchicamente sotto-ordinato di Commercianti, ovvero persone nelle attività del settore terziario, mentre in PATH sono stati raggruppati assieme ad altri mestieri del passato sotto l'etichetta di nodo dedicata [Persone nelle attività economiche varie nel passato], assieme ad Acquaroli, Bottiglieri, Depositari, Maestri di casa, Mercanti, Morsari, Nevaroli, Oglierari, Palafrenieri, Scopatori, Speciali, Trinaroli (Fig. 7).

PerformArt Database		Introduzione	Progetto
Termine del thesaurus T-052-520-160			
cr. 14/05/2018 mo. 27/08/2018			
persone > [persone secondo l'attività] > [persone nelle attività economiche varie] > [persone nelle attività economiche varie nel passato]			
Termine	[persone nelle attività economiche varie nel passato]		
Tipo di termine	Etichetta di nodo		
Termine generico (BT)	[persone nelle attività economiche varie]		
Definizione (DEF)			
Termini specifici (NT)	acquaroli bottiglieri depositari maestri di casa mercanti morsari nevaroli oglierari palafrenieri scopatori speciali trinaroli		

PerformArt è un progetto di ricerca finanziato dall'European Research Council (Grant Agreement no. 661415), ospitato dal Centre National de la Recherche Scientifique, in partnership con l'École Française de Rome. Principal Investigator: Anne-Madeleine Goulet. Coordinatrice scientifica e amministrativa del Database: Michela Berti. Database design: Teluric. Supervisione dei dati a cura di Michela Berti, Ornella Baroncelli, Gloria Giordano, Anne-Madeleine Goulet, Manuela Grillo, Alexandra Nigito, Chiara Pelliccia, Aldo Roma.

PerformArt Database è concesso in licenza CC BY-NC-ND 4.0

Figura 7. Termini specifici (NT) dell'etichetta di nodo [Persone nelle attività economiche varie nel passato] nel PATH.

Sembra importante mettere in evidenza questi possibili casi di cortocircuito logico che di per sé non sono comunque stati un ostacolo alla reciproca collaborazione¹¹. Anzi, la sempre maggiore specializzazione – che si può introdurre

¹¹ Il tema della collaborazione tra thesauri generali e thesauri specialistici e di ricerca è stato esplorato in una relazione, condivisa con Elisabetta Viti (Biblioteca nazionale centrale di Firenze), dal titolo

implementando interi rami gerarchici, come ad esempio quello sulle professioni antiche, in un thesaurus generalista come Nuovo Soggettario – contribuisce a fornire uno strumento sempre più ricco e raffinato a coloro che indicizzano le risorse negli istituti culturali, a beneficio delle comunità di ricerca che basano le loro attività proprio sul patrimonio documentale conservato nei nostri istituti culturali. Detto in altre parole: chi fa ricerca si rivolge a biblioteche, archivi ed istituti di conservazione e se questi possono usare per l'indicizzazione strumenti ricchi, sempre più raffinati grazie anche alla collaborazione con i progetti di ricerca, sono gli utenti – spesso essi stessi ricercatori – a trarne benefici in termini di recupero dell'informazione.

3.4. Politica di indicizzazione

Le tecniche di indicizzazione sottendono ad un'operazione quanto mai complessa: è l'attività di valutazione del messaggio, ovvero del contenuto concettuale di un testo (Guarasci e Guerrini 2022), che si configura come mediazione. La capacità di conoscere, interpretare e restituire in forma accessibile contenuti per loro natura ostici o comunque difficili da trovare, capire e usare è un valore aggiunto (Valacchi 2022) che rende particolarmente delicate, nel caso di indicizzazione semantica di materiale documentario, le scelte di “politica di indicizzazione”.

Al di là delle scelte metodologiche sulla strutturazione terminologica delle gerarchie thesaurali e sulla rappresentativa del lessico del settore di riferimento, sono emerse da subito come necessarie riflessioni e scelte operative sulla politica d'uso dei termini del thesaurus per attribuire le parole chiave.

Se nella soggettazione di materiale bibliografico abbiamo l'operazione di sommarizzazione che ci porta alla realizzazione di un'unica stringa coestesa (che rappresenta il tema di base della pubblicazione) nel caso dei documenti d'archivio procedere a sommarizzazione non è quasi mai possibile oppure ci porterebbe a perdere aspetti molto rilevanti, facendo perdere il senso dell'operazione stessa di indicizzazione semantica.

Per utilizzare al meglio i descrittori del thesaurus, è necessaria una politica di indicizzazione condivisa che guidi tutti coloro che sono autorizzati ad inserire le parole chiave ai documenti che trattano, con un livello di dettaglio standardizzato nell'attribuzione dei termini descrittori.

Nella molteplicità di parole chiave attribuibili ad un'unica risorsa archivistica, serve, ad avviso di chi scrive, la guida da parte di quella che possiamo definire come “l'anima del progetto”: nel caso di PATH, poiché il progetto

stava dedicando molta attenzione alla parte materica – a tutti gli oggetti, a tutte le tecniche, a tutti i mestieri coinvolti per realizzarli – per far emergere tutto il mondo delle arti performative è servita un’attenzione soprattutto agli aspetti materici, squisitamente materiali. Il database e il thesaurus dovevano quindi dare garanzia di questo sguardo, incarnando questo nuovo approccio rispetto ai tradizionali studi sugli aspetti estetici e rendendo tangibile il nuovo approccio anche attraverso i termini thesaurali, orientando così il lavoro in modo molto concreto a quella che sarebbe stata la ricerca per contenuto degli utenti (Grillo e Berti 2024).

L’équipe di ricerca ha scandagliato gli archivi familiari trattati alla ricerca di notizie sulla realizzazione di eventi performativi, ma i medesimi oggetti documentali, indagati sotto lo sguardo con altra angolatura disciplinare, potrebbero suggerire anche parole chiave diverse: di fatto un documento può avere parole chiave diverse a seconda di qual è il tipo di informazione che il ricercatore/indicizzatore cerca: nel caso di un conto per l’allestimento di un evento, uno storico dei trasporti si concentrerà sui denari corrisposti per le carrozze, uno storico della musica si concentrerà sul prezzo delle prestazioni dei cantanti, etc.

In questo senso è stata preziosa la condivisione con tutta l’équipe di ricerca PATH dei principi base dei thesauri e dell’indicizzazione semantica, attraverso una formazione frontale di due ore e attraverso una guida di riferimento realizzata da chi scrive in formato video, con esemplificazioni del lavoro che potessero accompagnare l’operazione di attribuzione dei descrittori da parte dei ricercatori.

Nell’équipe PATH ogni ricercatore era autorizzato ad attribuire le parole chiave ai documenti inseriti nel database: la scelta di affidare ai ricercatori l’attribuzione dei descrittori è stata dettata dalla piena consapevolezza che gli specialisti di un certo dominio disciplinare hanno una maggiore capacità di interpretare i documenti. Il lavoro condiviso ha sottolineato come fosse fondamentale standardizzare il livello di dettaglio nell’attribuzione dei descrittori¹².

4. Conclusioni

La scelta di sviluppare PATH per l’indicizzazione di materiali di archivio antichi a partire da una base thesaurale preesistente e generalista è derivata da questioni di sostenibilità del lavoro, come detto, ma soprattutto dall’orienta-

¹² Il tema è stato trattato in una relazione, condivisa con José María Domínguez (Istituto del Teatro de Madrid ITEM – Universidad Complutense), dal titolo *Tesaurizando las artes performativas en Roma hacia 1700: una propuesta de colaboración interdisciplinar* per il III Seminario Italo-Spagnolo di Biblioteconomia e documentazione Los retos en la gestion y comunicacion del patrimonio bibliográfico y documental: digitalización e IA, Universitat de Barcelona, 31 ottobre 2024, i cui atti sono in corso di pubblicazione.

mento generale della ricerca e le politiche dell'Unione Europea sulla condivisione della conoscenza, sul riutilizzo dei dati. Ci è sembrato un buon inizio quello di partire da quanto già strutturato sulla base degli standard internazionali, sacrificando in alcuni casi delle specificità o a volte discostandoci nel caso di qualche gerarchia troppo modernamente intesa ai nostri fini, ma usando di fatto una lingua di indicizzazione comune che possa permettere una migliore esperienza dell'utente finale, sulla scia del modello proposto nell'interfaccia dei termini del Thesaurus Nuovo Soggettario con il campo Risorse di Archivi e Musei, ove uno o più link permettono di accedere alle risorse presenti in banche dati di archivi e/o cataloghi online di archivi e musei.

Sperabilmente il suggerimento di link alle varie risorse che trattano di un dato argomento potrebbe essere proposto in contesti che siano dedicati non ad indicizzatori – come lo è tendenzialmente l'interfaccia di consultazione del Thesaurus Nuovo Soggettario – ma ad utenti. La prospettiva che ci si auspica è che gli utenti possano partire dall'argomento della loro ricerca e reperire le fonti utili, a prescindere da quale sia la tipologia di istituto di conservazione, sulla base di analisi semantica espressa attraverso linguaggi comuni che i thesauri possono garantire.

Ringraziamenti

Un sentito ringraziamento alla PI del Progetto ERC PerformArt, Anne-Madeleine Goulet (Tours, Centre d'études supérieures de la Renaissance), per tutti gli insegnamenti metodologici e per il sostegno ricevuto, e alla coordinatrice delle attività scientifiche, Michela Berti, per la fattiva collaborazione interdisciplinare.

Riferimenti bibliografici

- BNCF (Biblioteca Nazionale Centrale di Firenze). 2021. *Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto*. 2. ed. interamente rivista e aggiornata. Associazione italiana biblioteche, Biblioteca nazionale centrale di Firenze.
- Bruni, Silvia, Francesca Capetta, Anna Lucarelli, Maria Grazia Pepe, Susanna Peruginelli, e Marco Rulent. 2016. "Verso l'integrazione tra archivi, biblioteche e musei: alcune riflessioni." *JLIS.it* 1: 225-44. <https://doi.org/10.4403/jlis.it-11482>.
- Becherucci, Andrea, Silvia Bruni, Benedetta Calonaci, et al. 2019. "Libri per gli internati militari italiani durante la Seconda guerra mondiale: un inedito di Ernesto Rossi." *Biblioteche oggi* 4: 24-48. <https://www.doi.org/10.3302/0392-8586-201904-024-1>.

- Capetta, Francesca. 2015. "Il nome delle cose. Resoconto della giornata MAB dell'8 ottobre 2015." Consultato il 4 novembre 2025. <http://anaitoscana.org/2015/10/28/il-nome-delle-cose-resoconto-della-giornata-mab-dell8-ottobre-2015/>.
- Capetta, Francesca. 2017. "Il progetto 'Pascoli nello specchio delle sue carte' e i nuovi sviluppi". *Progressus* 1: 39-48.
- Daquino, Marilena. 2021. "Linked Open Data native cataloguing and archival description." *JLIS.it* 12 (3): 91-104. <https://doi.org/10.4403/jlis.it-12703>.
- Gnoli, Claudio, Vittorio Marino, e Luca Rosati. 2006. *Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il Web*. Tecniche nuove.
- Goulet, Anne-Madeleine. 2020. "Un outil pour étudier les spectacles de l'Ancien Régime: la base de données PerformArt." *Revue d'historiographie du théâtre, Écrire l'histoire des spectacles avec des bases de données* 5.
- Grillo, Manuela. 2015. *Indicizzazione semantica di bandi, manifesti e fogli volanti*. Editoriale documenta.
- Grillo, Manuela. 2021. "La costruzione di un thesaurus per la performance." In *Spectacles et performances artistiques à Rome (1644-1740). Une analyse historique à partir des archives familiales*, a cura di Anne-Madeleine Goulet, Élodie Oriol e José Maria Dominguez. Publications de l'École française de Rome.
- Grillo, Manuela. 2023. "Indicizzazione semantica di opere antiche." Presentazione delle Linee guida sull'indicizzazione per soggetto e la classificazione nel servizio bibliotecario nazionale (SBN), Roma, marzo. Pubblicato il 27 aprile 2023, da ICCU (Istituto Centrale per il Catalogo Unico). Youtube, 00:19:00. <https://www.youtube.com/watch?v=S8RG-Vyw3xo>.
- Grillo, Manuela, and Michela Berti. 2024. "From scientific research to a difficult definition, and back again: 'Magnificenza' as a Paradigm in the Development of a Thesaurus". In *Noble Magnificence. Cultures of the Performing Arts in Rome, 1644-1740*, edited by Anne-Madeleine Goulet and Michela Berti. Brepols.
- Guarasci, Roberto, e Mauro Guerrini. 2022. *Cos'è l'indicizzazione*. Editrice Bibliografica.
- Guerrini, Mauro, e Tiziana Possemato. 2015. *Linked data per biblioteche, archivi e musei: perché l'informazione sia del web e non solo nel web*. Editrice Bibliografica.

- ICCU (Istituto Centrale per il Catalogo Unico). 2023. "Linee guida sull'indicizzazione per soggetto e la classificazione nel Servizio Bibliotecario Nazionale (SBN)". A cura del Gruppo di lavoro per la catalogazione semantica in SBN. Consultato il 4 novembre 2025. https://norme.iccu.sbn.it/index.php?title=Norme_comuni/Linee_guida_sull%27indicizzazione.
- Lucarelli, Anna. 2019. "Un'opera è un'opera: dialogo non troppo immaginario sull'indicizzazione dell'antico." In *Viaggi a bordo di una parola. Scritti sull'indicizzazione semantica in onore di Alberto Cheti*, a cura di Anna Lucarelli, Alberto Petrucciani e Elisabetta Viti. Associazione Italiana Biblioteche (AIB).
- Lucarelli, Anna. 2022a. "Thesauri in the Digital Ecosystem." *JLIS.it* 13 (1): 156-76. <https://doi.org/10.4403/jlis.it-12744>.
- Lucarelli, Anna. 2022b. "A new edition of the Guida al sistema italiano di indicizzazione per soggetto." *JLIS.it* 13 (3): 103-14. <https://doi.org/10.36253/jlis.it-494>.
- Marinelli, Ester. 2018. *Library linked data: verso l'interoperabilità tra le istituzioni culturali*. Mimesis.
- PerformArt. 2023. Aggiornato il 23 marzo 2023. <https://performart.huma-num.fr/>.
- Piccini, Silvia, Andrea Bellandi, et Matteo Abrate. 2019. "Diaterm: un modèle pour représenter l'évolution diachronique des terminologies dans le web sémantique". In *Actes de la conférence TOTh 2019, Le Bourget du Lac, 6-7 juin 2019*. Presses Universitaires Savoie Mont Blanc, 55-68.
- Valacchi, Federico. 2023. "Dai metadati ai dati, dai contesti ai contenuti: aumentare la descrizione archivistica." In *La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD*. Università degli Studi di Siena, 214-19.
- Valacchi, Federico. 2015. *Diventare archivisti. Competenze tecniche di un mestiere sul confine*. Editrice Bibliografica.
- Valacchi, Federico. 2022. "Le parti e il tutto. Integrare la conoscenza." *JLIS.it* 13 (3): 1-11. <https://doi.org/10.36253/jlis.it-477>.
- Tomasi, Francesca. 2022. *Organizzare la conoscenza: Digital Humanities e Web semantico. Un percorso tra archivi, biblioteche e musei*. Editrice bibliografica.
- Tomasi, Francesca, e Marilena Daquino. 2015. "Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare", *JLIS.it* 6 (3): 13-38. <https://doi.org/10.4403/jlis.it-11133>.

Rappresentare gli atti notarili medievali in Linked Open Data (LOD)

Modello descrittivo ed estrazione automatica di entità dal latino medievale per l'archivio digitale di Notariorum Itinera

Bianca La Manna*

Abstract: Medieval notarial deeds, as legal documents, are a valuable source for increasing knowledge about people and places during that time and for tracing relationships between them. This project has a double objective: to publish in Linked Open Data (LOD) the medieval notarial deeds held by the center Notariorum Itinera, which are already contained in a relational database (DB) and accessible from the website of the center; to enrich both the original DB and the Knowledge Graph (KG) with new data automatically extracted from medieval Latin's texts through the application of Named-Entity Recognition (NER) techniques.

Keywords: Linked Open Data, Semantic Web, Named-Entity Recognition, Medieval Latin, Digital Archive.

1. Introduzione

Gli atti notarili sono una fonte rilevante per raccogliere dati e sapere relativi ad un determinato periodo storico e, in particolar modo, per epoche come il medioevo, che richiedono ancora un'attenzione e una maggiore rappresentazione nella ricerca, soprattutto nel campo dell'umanistica digitale. Dagli atti notarili medievali, è possibile estrarre entità, quali luoghi, persone, oggetti e relazioni che, una volta pubblicate nel web semantico, danno la possibilità di creare connessioni tra entità identificate univocamente e rese interoperabili e accessibili senza intermediazione tra le istituzioni, arricchendo la conoscenza del dominio. Raccogliere questo tipo di dati diventa, pertanto, ancora più importante per ricostruire la storia e gli attori di una città come quella di Genova, che ha avuto una significativa influenza sul commercio marittimo del tempo.

Questo progetto si pone un duplice obiettivo: da una parte creare un grafo di conoscenza, a partire dai dati già contenuti nella base di dati dell'archivio digitale di *Notariorum Itinera* (Notarium Itinera n.d.a), per poterli rendere

* Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi (DIBRIS), Università di Genova, Genova, Italia. bianca.lamanna@edu.unige.it. ORCID: 0009-0000-6834-6583.

disponibili come Linked Open Data (LOD); dall'altro di estrarre automaticamente questo tipo di informazioni, relative alle entità citate, dal latino medievale e, più in generale, capire quali tecniche possono essere applicate per ottenere informazioni aggiuntive in maniera automatica da questi dati testuali non strutturati.

Il primo obiettivo è stato ottenuto ricreando lo schema descrittivo sul quale è stata strutturata la base di dati relazionale, per avere un unico modello dal quale poter derivare la rappresentazione in LOD. Riutilizzando ontologie già esistenti, in accordo con i principi Findable, Accessible, Interoperable, and Reusable (FAIR) (Wilkinson et al. 2016), è stato realizzato un modello concettuale, secondo il quale è stato popolato il grafo di conoscenza da pubblicare nel web semantico. Per quanto riguarda l'estrazione automatica, tecniche di riconoscimento automatico delle entità Name Entity Recognition (NER) sono state applicate ai testi digitalizzati e ripuliti. I risultati sono ancora ad uno stadio sperimentale che si propone, però, di dare una panoramica esaustiva sugli strumenti, le metodologie e le prospettive per l'applicazione di tecniche di elaborazione del linguaggio naturale (Natural Language Processing - NLP) per il latino medievale.

2. Rappresentare e pubblicare gli atti notarili in LOD

Il centro studi *Notariorum Itinera* possiede una biblioteca digitale di documentazione archivistica che permette l'accesso alle riproduzioni digitali di una parte dei fondi *Notai antichi* e *Notai ignoti* dell'Archivio di Stato di Genova (Notariorum Itinera n.d.b) e una biblioteca digitale bibliografica che raccoglie gli oggetti digitali gestiti dal sistema informativo del centro (Notariorum Itinera n.d.c). I dati sono raccolti in una base di dati relazionale, per la quale era già stata definita una struttura logica fatta di classi e relazioni, utilizzando una terminologia standard di settore. Per questo motivo non è stato necessario creare un modello descrittivo dell'oggetto archivistico, in quanto già definito, ma riadattare quello esistente ad una struttura a grafo che riconducesse le classi ad entità, adoperando ontologie già esistenti, in linea con i principi FAIR di riutilizzo (*reuse*). Per la creazione del modello descrittivo per gli atti notarili medievali, per prima cosa è stato formalizzato in un diagramma entità/relazione (E/R) il modello logico sottostante alla base di dati relazionale, ovvero una struttura descrittiva in grado di rappresentare i dati indipendentemente dal tipo di implementazione o struttura del database (Fig. 1).

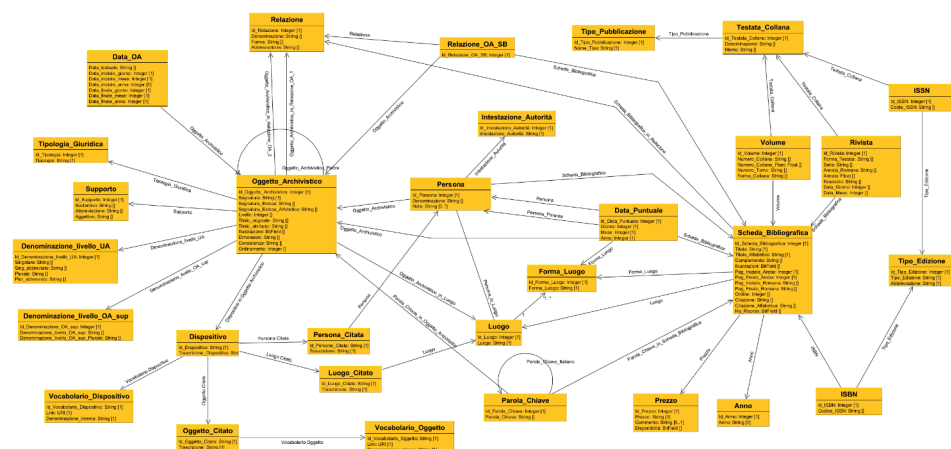


Figura 1. Elementi principali del diagramma E/R ricostruiti dalle classi e relazioni presenti nella base di dati relazionale.

Questo passaggio è stato necessario per ricostruire il modello teorico dietro l'implementazione e per porre le basi per la mappatura tra la base di dati relazionale e quella a grafo. Partendo da questa rappresentazione teorica, è stato possibile formalizzare il modello concettuale, riutilizzando ontologie esistenti in grado di descrivere i due concetti principali sui quali è stata costruita la base di dati relazionale: *oggetto archivistico* e *scheda bibliografica*.

Per la descrizione degli oggetti archivistici, le ontologie di settore principali selezionate e implementate sono:

- Ontology for Archival Description (OAD): realizzata dall'Istituto beni artistici culturali e naturali della Regione Emilia-Romagna, basata sullo standard archivistico ISAD (G) e promossa dall'International Council on Archives (ICA) (Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna e Regesta.exe 2018);
- Sistema Archivistico Nazionale (SAN): realizzato dal centro Metodologie e Applicazioni per Archivi Storici (MAAS) per l'Istituto Centrale per gli Archivi (ICAR) (ICAR e Centro MAAS 2015).

Dal momento che la base di dati non era esclusivamente focalizzata sui materiali archivistici, queste ontologie sono state selezionate in quanto sufficientemente esaustive per la descrizione di dominio, ma anche adatte all'interoperabilità con ontologie di dominio diverso. Tuttavia, bisogna considerare che per una rappresentazione prettamente archivistica, un'ontologia rinomata e solida quale Records in Context Ontology (RiC-O) (Felicati 2021; International Council on Archives Expert Group on Archival Description 2024) potrebbe essere preferibile.

Per la descrizione delle schede bibliografiche, le principali ontologie utilizzate sono:

La descrizione sintetica dell'oggetto archivistico, senza tenere conto delle sue relazioni più ampie, si basa su cinque entità:

- oad:ArchivalResource: indica un corpus documentale oggetto di descrizione (Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna 2018);
- oad:ArchivalResource: indica un corpus documentale oggetto di descrizione (ICAR - Centro MAAS 2015);
- san:dateComplarc: data dell'oggetto archivistico (ICAR - Centro MAAS 2015);
- san:Luogo: luogo di pubblicazione della risorsa (Russo et al. 2023);
- a-archive:ArchivalResourceLevel: rappresenta il livello archivistico di una risorsa, ad es. unità archivistica, fondo etc. (Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna 2018)
- oad:ExtentAndMedium: rappresenta la descrizione dell'estensione e del supporto ().

A queste entità si aggiungono altre relazioni con valori letterali, *rdfs:Literal* (Cyganiak et al. 2014), che danno informazioni aggiuntive quali: le date, la tipologia di complesso, il titolo e la denominazione, le parole chiave (Fig. 3).

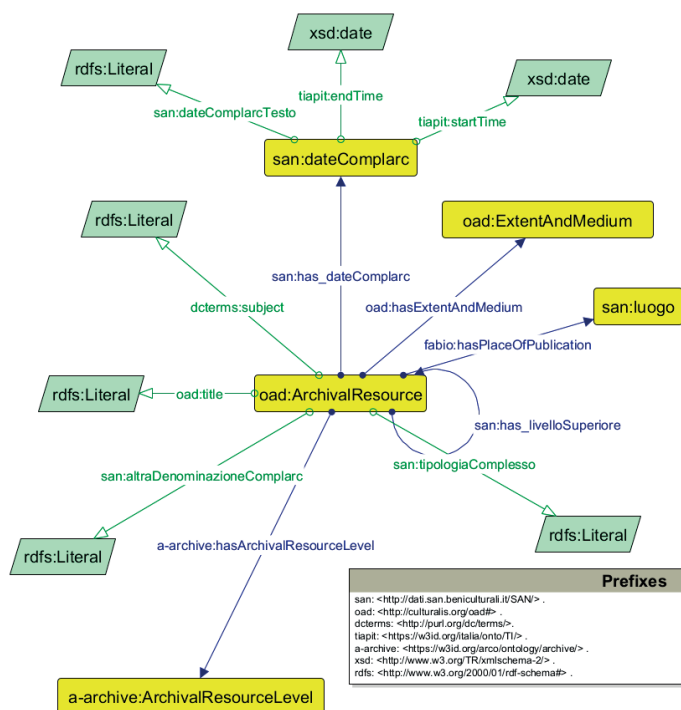


Figura 3. Descrizione dell'oggetto archivistico sulla base di ontologie esistenti, secondo la sintassi di Graffoo (in giallo le entità e i valori nominali).

In seguito, il modello è stato implementato in un grafo RDF, popolato con i dati provenienti dalla base di dati relazionale. Questo passaggio è stato effettuato tramite un codice per la mappatura, realizzato ad hoc attraverso il linguaggio di programmazione Python utilizzando la libreria *pyodbc* (Klehammer 2024) per la connessione alla base di dati tramite linguaggio SQL, e *rdflib* (RDFLib 2025) per la conversione dei dati in triple RDF, esportando il risultato in un file nel formato Turtle, attualmente pubblico e consultabile su Zenodo (La Manna 2025)¹. Di seguito è riportata una parte della descrizione di un'istanza del grafo di un'entità di tipo *oad:ArchivalResource* (Fig. 4, Fig. 5).

```
@prefix san: <http://dati.san.beniculturali.it/SAN/> .
@prefix oad: <http://culturalis.org/oad#> .
@prefix dcterms: <http://purl.org/dc/terms/>. tiapit: <https://w3id.org/italia/onto/II/> .
@prefix arco-lite: <https://w3id.org/arco/ontology/arco-lite/> .
@prefix a-archive: <https://w3id.org/arco/ontology/archive/> .
@prefix xsd: <http://www.w3.org/TR/xmlschema-2/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
```

Figura 4. Esempio di risorsa in RDF, secondo la sintassi Turtle: prefissi delle ontologie.

```
ni:OggettoArchivistico-6 a oad:ArchivalResource ;
  oad:hasExtentionAndMaterial "mm. 215x155 , Carta" ;
  oad:title "Relazione di Genova" ;
  san:altraDenominazioneComplar "[Ordinamento dello stato, vicende storiche del sec. XVI, notizie finanziarie relative agli anni 1596-1597]" ;
  san:has_dateComplarc ni:DataRedazione-OA-6 ;
  san:has_livelloSuperiore ni:OggettoArchivistico-2 ;
  san:tipologiaComplesso "volume" .
```

Figura 5. Esempio di risorsa RDF in sintassi Turtle: oggetto archivistico “relazione di Genova”.

La stessa istanza di oggetto archivistico descritta sopra in linguaggio RDF può essere visualizzata nella sua struttura a grafo, che rende con evidenza visuale l'entità e le sue relazioni (Fig. 6).

¹ Per completare la pubblicazione dei dati in LOD, è necessario depositare il grafo di conoscenza in una base di dati per triple e renderlo accessibile tramite un *endpoint*, interrogabile con linguaggio SPARQL, ovvero un punto di accesso connesso alla rete dal quale è possibile estrarre o richiedere in maniera diretta le risorse presenti nella base di dati. Si prospetta, come risultato finale, la creazione di una pagina web all'interno di Notariorum Itinera, ancora in via di sviluppo, per consentire l'interrogazione tramite linguaggio SPARQL.

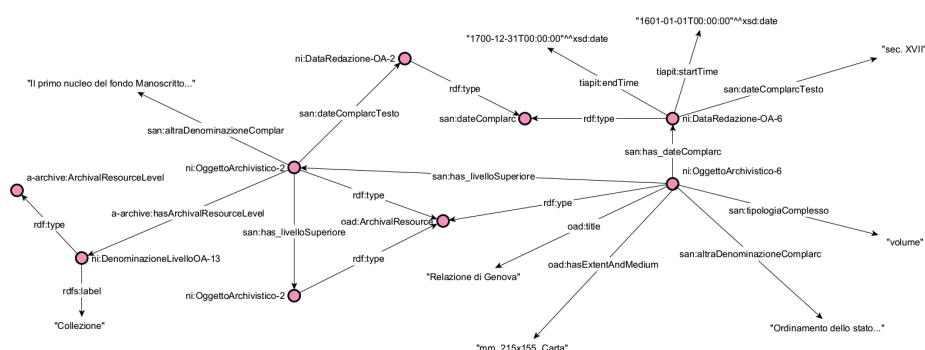


Figura 6. Rappresentazione visiva della stessa istanza di oggetto archivistico, “Relazione di Genova”, realizzata tramite Graffoo: i punti rappresentano le istanze di entità, gli altri nodi rappresentano i valori letterali.

Questo tipo di rappresentazione apre nuove prospettive per la diffusione dei dati, soprattutto in merito ad una più facile integrazione e disambiguazione delle entità rispetto ad altre basi di dati del settore.

3. Il latino medievale nell’ambiente digitale

Realizzato il grafo di conoscenza, il secondo obiettivo è di arricchirlo con nuove entità estratte automaticamente. Questo passaggio ha come primo requisito necessario la disponibilità di testi, in questo caso di atti notarili, trascritti in formato digitale di testo semplice, in una quantità tale da essere rappresentativa della lingua. Uno degli obiettivi a lungo termine a cui il centro di studi sta lavorando è quello di realizzare un modello linguistico per la trascrizione automatica dei manoscritti, tramite l'utilizzo della piattaforma Transkribus (READ-COOP SCE 2023), che permetterebbe di ottenere i dati necessari per procedere con gli strumenti di NLP. Nella prospettiva di poter lavorare in futuro su questi dati, questa sezione si prefigge lo scopo di fornire la descrizione di uno studio preliminare volto ad analizzare i limiti e le possibilità dell'applicazione dell’NLP al latino medievale².

² Sebbene esistano progetti di grande interesse per il settore, con finalità simili (García-González et al. 2020) la metodologia adottata in questa ricerca non è comparabile dal momento che il modello descrittivo era già esistente e conforme agli standard di settore (cap. 1) ed è stato, quindi, esclusivamente trasposto nei corrispettivi riscontrabili nelle ontologie già esistenti e conformi agli stessi standard. L'estrazione delle entità, invece, è stata eseguita con metodologie di riconoscimento automatico (NER) su testo non strutturato, senza finalità di pubblicazione del testo originale. Per questo motivo, l'unica analisi comparativa e di stato dell'arte si concentrerà in questa sezione sugli strumenti di NLP per il latino medievale e non su elementi di codifica del testo, quali Extensible Markup Language/Text Encoding Initiative (XML/TEI), che, invece, potranno essere di interesse della fase di trascrizione tramite Handwriting Text Recognition (HTR).

I modelli linguistici di grandi dimensioni (Large Language Model - LLM) sono ad oggi il principale strumento per l'estrazione automatica delle entità dai testi, in quanto, essendo modelli già allenati su grandi quantità di dati, richiedono un minore utilizzo di potenza di calcolo e, allo stesso tempo, possono essere aggiornati su dati inediti. Tuttavia, poiché allenare un modello richiede a sua volta grandi quantità di dati, questi strumenti hanno una maggiore resa nell'applicazione con lingue tuttora in uso, ovvero lingue che continuano a produrre materiale ogni giorno e per cui si dispongono maggiori risorse digitalizzate.

Questa differenza tra lingue definite premoderne e lingue vive risiede anche nelle modalità con cui sono state preservate che portano le prime a disporre di un corpus finito, a presentare una variazione maggiore, ovvero caratteristiche non standardizzate, e ad avere risorse disponibili limitate (Johnson et al. 2021, 20). Per questo strumenti di NLP, come gli LLM, che in linea di principio possono essere estesi anche a lingue storiche, nella pratica trovano difficoltà e limiti di applicazione proprio dovuti alla progettazione dei modelli stessi, pensata per corpora di lingue vive e parlate.

Un altro limite, tuttavia, consiste nella difficoltà di stabilire quale strumento sia preferibile adottare, a causa della carenza di strumenti di valutazione comparativi in questo settore che, quando presenti, sono concentrati sulle lingue moderne (Won et al. 2018, 3; Ehrmann et al. 2023).

I principali sistemi di riconoscimento automatico delle entità nominate (NER) si basano in linea di massima su tre tipi di algoritmi: basati su regole (*rule-based*); supervisionati (*supervised*); non supervisionati (*unsupervised*). Diversi studi sostengono che i metodi supervisionati ottengano la migliore misura F (*F-measure*)³ (Won et al. 2018); altri esperimenti hanno ottenuto risultati migliori applicando Conditional Random Fields (CRF), una tecnica di modellazione statistica (Boros et al. 2020, 80). Tuttavia, in entrambi i casi il limite nell'applicazione di questi strumenti a testi redatti in lingue storiche e, in questo caso, al latino medievale, risiede nella difficoltà di trovare grandi quantità di dati annotati o semplicemente digitalizzati in formato di testo libero (Chastang et al. 2021).

Secondo Boros et al. (2020, 80), la soluzione migliore per evitare di doversi servire di grandi quantità di corpora annotati sarebbe ricorrere a una combinazione di *gazetteers* e dizionari e strategie di apprendimento attivo (*active learning*) (Erdmann et al. 2019). Sebbene questo potrebbe essere effettivamente l'espedito vincente, richiede che gli studiosi intenzionati a digitalizzare il patrimonio testuale in latino medievale contribuiscano all'arricchimento o alla creazione di queste raccolte di dati nominali, accessibili e affidabili.

³ Una misura di accuratezza che si basa sulla precisione (*precision*) e il recupero (*recall*) nella valutazione dei risultati dell'estrazione di informazioni.

Posti i limiti quantitativi dei dati a nostra disposizione nell'approcciarsi a queste soluzioni, in questa sezione verranno trattati gli strumenti che si possono utilizzare per l'estrazione delle entità relative a persone, luoghi, date e oggetti a partire da un corpus ridotto e da una potenza computazionale limitata. Per quanto riguarda persone e luoghi, è stato tentato l'approccio del riconoscimento automatico delle entità nominate (NER), ovvero l'estrazione automatica dal testo delle entità in base ad una generalizzazione probabilistica dell'algoritmo su dati non osservati in precedenza, in base ad un modello precedentemente addestrato. Per le date e gli oggetti, abbiamo provato un approccio misto, basato su regole (*rule-based*) e tramite l'utilizzo di vocabolari per il riconoscimento degli oggetti (Boros et al. 2020, 80).

3.1. Il corpus

Il corpus selezionato consiste in 163 atti notarili redatti da Lamberto di Sambuceto, un notaio genovese, in un periodo che va dall'11 ottobre 1296 al 23 giugno 1299. I testi, già editi (Balard 1983), sono stati ripuliti in maniera semi-automatica, rimuovendo tutti i segni editoriali (es. note a piè di pagina, riferimenti al cartolario, etc.) e normalizzando la notazione numerica, attraverso un programma Python realizzato appositamente.

La normalizzazione dei numeri è stata effettuata allo scopo di rendere le date maggiormente identificabili ed estraibili successivamente, ed è stata applicata convertendo la forma estesa, quando presente, ad esempio "millesimo", in numerale romano, ad esempio "M", e questa forma è stata convertita nella notazione standard del sistema numerico arabo, in questo caso "1000". In questo passaggio è stata impiegata una piccola libreria Python *roman* (Pilgrim e Zope Foundation 2025), personalizzata nel riconoscimento di alcune combinazioni numeriche. In questo modo è stato possibile, una volta riconosciute ed estratte le varie parti che compongono la data (giorno, mese, anno), codificare il risultato in qualsiasi tipo di formato per le datazioni conforme allo standard ISO 8601-1:2019 Date and time – Representations for information interchange – Part 1: Basic rules.

Per avere dei dati verificati (*ground truth*) su cui poter valutare il rendimento dei vari strumenti di estrazione automatica delle entità, inoltre, sono state annotate manualmente le entità presenti nei vari atti notarili.

Il corpus, in numeri, consiste di:

- 44175 tokens;
- 2444 entità, di cui: 599 luoghi; 1575 persone; 96 anni; 174 date.

3.2. Strumenti esistenti per il latino

Una volta definito il corpus, è stata considerata la varietà di modelli linguistici già esistenti per l'estrazione automatica di entità nominali da testi latini. Le difficoltà, già esposte, nell'approccio a lingue premoderne non hanno impedito negli ultimi anni la creazione di strumenti finalizzati all'applicazione dell'NLP a testi storici. Le principali librerie e framework basati su modelli linguistici di grandi dimensioni per l'NLP applicato al latino sono:

- Classical Language Toolkit (CLTK): una libreria Python che prende spunto dai principali programmi di NLP (Johnson et al. 2021, 20; Johnson 2021), tra cui spaCy (Honnibal et al. 2023), per fornire un approccio modulare per l'elaborazione dei linguaggi premoderni che vada incontro alle loro necessità⁴. Il codice è disponibile su Github (Johnson 2025), sotto la licenza MIT;
- spaCy (LatinCy): una serie di operazioni spaCy per l'NLP applicato al latino (Burns 2023). Si tratta di uno strumento *general purpose*, ovvero non specializzato in nessuna fase specifica della lingua latina.

Altri strumenti che offrono un approccio diverso per l'analisi della lingua latina, che va oltre l'analisi grammaticale e l'estrazione delle entità sono:

- LiLa (Università Cattolica del Sacro Cuore n.d.): una base di conoscenza per le forme canoniche del latino (Mambrini e Passarotti 2023). Tra le varie funzionalità incorpora strumenti di NLP per il riconoscimento automatico dei lemmi sulla banca di dati da loro raccolti;
- quanteda (Benoit et al. 2021): un pacchetto per il linguaggio di programmazione R che lavora in combinazione con altri pacchetti come *stm* (Roberts et al. 2018) per la modellazione del contenuto e *text2vec* (Selivanov 2016) per l'incorporamento di parole (*word embedding*) (Benoit et al. 2018, 1).

3.3. NLP applicata al latino medievale

Per poter applicare gli strumenti selezionati sul nostro corpus è stato necessario definire il flusso di operazioni (*pipeline*) per il riconoscimento e l'estrazione automatica delle entità. Sono state definite tre fasi:

- preelaborazione dei dati: questa fase consiste nella preparazione dei dati prima di essere trattati da qualsiasi algoritmo e contiene le fasi di normalizzazione del testo; tokenizzazione; rimozione delle parole non significative (*stopwords*); assegnazione delle parti del discorso (*POS-tag*); lemmatizzazione. Incorporare queste informazioni aiuta successivamente il modello a interpretare in maniera più corretta le entità;

⁴ Lo standard ISO di riferimento per le lingue antiche, estinte e storiche è ISO 639.

- riconoscimento delle entità (NER): sperimentazione dei modelli precedentemente allenati applicati al riconoscimento delle entità nel nostro corpus;
- valutazione: verifica dell'accuratezza dell'estrazione delle entità sulla base dei dati verificati, ovvero delle entità annotate manualmente sul nostro corpus.

3.3.1. Preelaborazione dei dati

Per quanto riguarda la preelaborazione dei dati, modelli generici quali CLTK e spaCy risultano estremamente performanti, questo perché molte regole grammaticali e logiche del latino non sono cambiate significativamente nel tempo e, quando si sono verificate variazioni, come ad esempio differenze grafiche di alcune lettere, poiché si tratta di variazioni fisse e limitate, sono state integrate con successo nell'algoritmo. Per questo, la prima fase è stata portata a termine tramite librerie Python disponibili, come definito nella tabella (Tab. 1).

Funzione	Strumento
Normalizzazione del testo	Codice Python ad hoc; libreria Python: roman.
Tokenizzazione	CLTK/spaCy (LatinCy)
Rimozione delle parole non significative	CLTK
Assegnazione delle parti del discorso	CLTK
Lemmatizzazione	CLTK

Tabella 1. Preelaborazione dei dati: fasi e strumenti utilizzati.

Relativamente alle versioni adottate, per l'intero progetto è stato utilizzato Python 3.9, pip 24.0. Per quanto riguarda le librerie, la versione di CLTK è 1.3.0, compatibile con spaCy versione 3.7.4⁵.

3.3.2. Estrazione automatica delle entità

Il secondo procedimento della *pipeline* riguarda l'estrazione delle entità. In base a quanto affermato finora, le opzioni percorribili in questo caso sono le seguenti:

- riutilizzare modelli linguistici già esistenti (CLTK/spaCy);

⁵ Il modello linguistico allenato, `la_core_web_lg` (vedi https://huggingface.co/latincy/la_core_web_lg), è compatibile esclusivamente con la versione spaCy $\geq 3.7.5$. Si può scaricare manualmente il modello con il comando `pip install "la-core-web-lg @ https://huggingface.co/latincy/la_core_web_lg/resolve/main/la_core_web_lg-any-py3-none-any.whl"`. Tuttavia, trattandosi di una forzatura, possono verificarsi incompatibilità del modello con il codice spaCy che influiscono nell'esecuzione.

- perfezionare un modello esistente (*fine-tuning*);
- utilizzare un approccio basato sulla definizione di regole fisse (*rule-based*).

Nel prospetto seguente è possibile osservare i pro e i contro delle varie opzioni (Fig. 7).

Opzione 1: NER attraverso librerie esistenti (CLTK/SpaCy)

Opzione 2: personalizzare un modello già allenato

PRO: • Generalizzazione • Efficienza • Scalabilità	CONTRO: • Costo computazionale • Scarsità di dati
---	---

Opzione 3: approccio basato su regole

PRO: • Pochi dati richiesti • Precisione	CONTRO: • Problemi di scalabilità • Generalizzazione limitata
--	---

Figura 7. Metodi di estrazione delle entità ed eventuali pro e contro.

Trovare una libreria Python con un modello perfettamente funzionante, come avvenuto per la preelaborazione del testo, è certamente la soluzione ideale. L'estrazione delle entità di luogo e persona tramite il codice fornito dalle librerie CLTK e spaCy, tuttavia, ha portato dei risultati poco soddisfacenti. La sfida principale che emerge immediatamente nel tentativo di applicare i modelli linguistici esistenti agli atti notarili presi in considerazione risiede nel fatto che la maggior parte di questi modelli è stata addestrata su testi di latino classico. La ragione, precedentemente trattata, è la differenza di rappresentanza del latino classico rispetto al latino medievale nell'ambiente digitale. Pertanto, questi modelli, una volta applicati al latino medievale, decrescono nei parametri di resa. Questo viene ancora più accentuato nel caso in questione, dal momento che oggetto del progetto sono manoscritti di latino medievale tardo (ca. XII-XV sec.).

Provando ad applicare gli strumenti precedentemente citati su testi medievali è stato subito evidente che la funzione di riconoscimento delle entità era limitata da due circostanze, ampiamente descritte da Chastang et al. (2021):

- la maggior parte dei nomi comuni del medioevo non esistevano o erano utilizzati raramente nel periodo classico e, per questa ragione, non erano mai stati osservati precedentemente dal modello linguistico che, pertanto, non è in grado di identificarli. Nei casi in cui, attraverso l'utilizzo della struttura logica e grammaticale sottostante, il modello riesce ad isolare l'entità e riconoscerla come tale, spesso assegna un'etichetta erranea, considerando ad esempio "persona" un "luogo";
- le convenzioni di denominazione nel latino si sono evolute nel tempo (Ehrmann et al. 2023, 8), da una struttura *Nomen Prenomen Cognomen* ad una *Nomen + de + toponimo*. Tenendo conto del fatto che l'estrazione di entità composte da più parole sotto un'unica etichetta rimane una delle funzioni più complesse, se in casi di denominazioni classiche, ad esempio "Marco Aurelio", poteva essere affrontata con una ricostruzione a posteriori dell'entità, in questo caso viene complicata ulteriormente dall'errore di etichettatura sopra descritto. Per cui un toponimo, riconosciuto come entità, viene classificato come "luogo" e non come parte di una denominazione di persona.

Questo porta ad una estrazione di luoghi e persone completamente inaffidabile e rende poco percorribile una soluzione quale post-elaborare il risultato in base a regole fisse che ricostruiscono la struttura *Nomen + de + toponimo* combinandolo nuovamente in un'entità unica.

Prendendo in considerazione le altre due opzioni, ovvero perfezionare un modello esistente o realizzare un codice che definisca le regole di estrazione delle entità, i punti principali che fanno la differenza nella preferenza di una soluzione rispetto ad un'altra riguardano:

- la quantità di dati: sia per costruire che per perfezionare un modello probabilistico servono grandi quantità di dati annotati, mentre per un approccio a regole questo elemento è quasi irrilevante;
- scalabilità: l'adattabilità e l'efficienza dell'algoritmo all'aumentare delle risorse e, quindi, del carico di lavoro.

Nel nostro processo abbiamo provato a personalizzare il modello di LatinCy, dal momento che si tratta di un procedimento accessibile e facile da applicare con il codice di spaCy (spaCy n.d.), ma la quantità di entità annotate a nostra disposizione era talmente insufficiente da aver riportato risultati irrilevanti.

L'estrazione delle entità di luogo e di persona, pertanto, è ancora in fase di sperimentazione. I limiti evidenziati fino ad ora possono essere superati in futuro digitalizzando e rendendo accessibili questi dati inediti o tramite dizionari controllati o tramite immissione massiva di testi per il perfezionamento dei modelli linguistici esistenti.

Le entità temporali, ovvero le date, non sono state considerate nella fase di allenamento dei modelli linguistici oggetto di studio, questo perché sono stati

prevalentemente allenati su testi narrativi. La tipologia documentale dell'atto notarile, invece, vede in queste informazioni uno degli elementi costitutivi e imprescindibili del testo, dove spesso vengono segnate le date ufficiali di redazione del documento. Per l'estrazione di queste entità, dal momento che si tratta di una notazione formalmente fissa e facilmente identificabile, è stato pensato un codice specifico in Python per l'estrazione automatica che, nel complesso, ha dato risultati accettabili, riuscendo ad estrarre tutte le entità temporali presenti nei testi in oggetto.

Bisogna sottolineare nuovamente, però, che questa fase di sperimentazione è stata effettuata su 163 atti notarili redatti dallo stesso notaio, un numero esiguo di testi per poter effettuare una concreta ed efficace valutazione dei vari modelli linguistici e codici. Tuttavia, già dall'applicazione degli strumenti di NLP principali per la lingua latina è stato possibile intuire le criticità dell'utilizzo di modelli linguistici di grandi dimensioni ad uso generale e la necessità di dare una maggiore rappresentanza sia alla lingua latina medievale e tardo medievale, che a tipologie documentali diverse per poterli adattare al di fuori di un contesto letterario classico.

3.4. LiLa

Uno dei progetti più ampi e recenti per l'analisi linguistica del latino è *LiLa: Linking Latin* (Mambrini e Passarotti 2023). Questo progetto nasce con l'intenzione di creare una banca dati di lemmi latini, pubblicati in LOD. Tra le funzionalità messe a disposizione dalla piattaforma c'è il *Text Linker* che consente, attraverso strumenti integrati di NLP, di inserire un testo in latino che viene automaticamente analizzato sulla base dei lemmi presenti nella banca dati. Si ottiene come risultato finale per ogni parola del testo il link al lemma corrispondente, quando esistente, con tutte le informazioni ad esso collegate.

Sicuramente LiLa è un grande esempio di sperimentazione e di utilizzo efficace delle nuove tecnologie sulle lingue storiche, consentendo l'integrazione e l'interoperabilità delle risorse. Per quanto riguarda l'utilizzo di questo strumento nel nostro caso specifico, bisogna sottolineare ancora una volta come i testi medievali digitalizzati siano pochi e per lo più di natura narrativa o religiosa. La carenza, pertanto, della terminologia specifica del linguaggio notarile insieme alla mancanza nei testi utilizzati per la creazione della banca dati di riferimenti ai luoghi e ai nomi meno frequenti o in uso nel tempo e nel luogo in cui gli atti del nostro caso studio sono stati redatti, rende qualsiasi strumento inadatto al riconoscimento di quelle che sono le parti più significative per la classificazione del testo, fondamentali per uno studio storico e rappresentativo del tema.

3.5. Quanteda

Un approccio diverso al testo, sempre finalizzato all'estrazione automatica di dati per l'arricchimento della descrizione semantica degli oggetti archivistici, è stato quello sperimentato tramite la libreria *quanteda*, utilizzata nell'ambiente R studio (R Foundation 2025), per l'analisi quantitativa dei dati testuali.

L'obiettivo di questa sperimentazione è stato quello di provare a classificare in maniera automatica gli atti notarili in base alla loro tipologia, allo scopo di inserire questa ulteriore informazione semantica all'interno del modello descrittivo e, di conseguenza, nella base di dati semantica e a grafo. Definite 15 categorie⁶ e le parole chiave che ne caratterizzavano ciascuna, è stato creato un dizionario e, tramite la libreria *quanteda.textstats*, è stato possibile assegnare in maniera automatica ad ogni atto una o più categorie. Questa assegnazione viene effettuata sulla corrispondenza esatta della stringa della parola chiave contro il testo dell'atto. La sperimentazione è stata effettuata sempre sui 163 atti di Sambuceto. 124 atti su 156 sono stati classificati correttamente (7 non erano classificabili in quanto incompleti o frammentari).

Questo tipo di approccio, che può rientrare nella categoria di algoritmi basati su regole, presenta le limitazioni esplicate sopra, in particolare, è difficilmente generalizzabile ad atti notarili che sono stati redatti in un contesto diverso. Questo è stato dimostrato provando ad utilizzare lo stesso codice per classificare alcuni atti di Antonio de Inghibertis (Ruzzin 2020), di epoca successiva, ovvero risalenti al periodo dal 1330 al 1346, ottenendo risultati di gran lunga peggiori (solo 26 atti su 100 sono stati classificati). È stato interessante notare come la differenza nella terminologia non fosse data, in questo caso, dalla distanza temporale o spaziale, quanto dal contesto di redazione, dal momento che Inghibertis lavorava per la curia.

Un altro elemento interessante emerso è che la categoria più difficile da individuare con questa tecnica è stata quella della quietanza, atto che definisce un pagamento al debitore e che, per questo, può trattare relazioni di diversa natura ed ha una terminologia meno marcata.

4. Conclusioni

L'atto notarile e, in generale, gli oggetti archivistici, in quanto documenti formali, hanno una terminologia di settore specifica e ben definita che rende identificabili le varie componenti nella rappresentazione strutturata dell'oggetto. Partendo da una base di dati relazionale che aveva già portato a termine

⁶ Questo passaggio è stato effettuato grazie all'aiuto di Luca Filangieri, assegnista di ricerca presso Notariorum Itinera. Le 15 categorie sono: mutuo, vendita, quietanza, *acomendacio*, dote, nolo, procura, dichiarazione di debito, cambio, testamento, inventario, promessa, cessione di diritti, manomissione, locazione.

la definizione delle classi e delle relazioni, è stato più semplice poter definire il modello concettuale per la pubblicazione in LOD. La parte più interessante, in questo passaggio, è stata constatare quanto lavoro sia stato svolto negli ultimi anni per la rappresentazione degli oggetti archivistici nel web semantico, impegno che ha avuto come esito la realizzazione di ontologie di settore esaustive che utilizzano il linguaggio specifico secondo gli standard concordati e, allo stesso tempo, garantiscono l'interoperabilità con le ontologie culturali, aprendo la possibilità ad una integrazione dei dati permeante e interdisciplinare. Se dal punto di vista del dato strutturato il panorama è promettente, dal punto di vista testuale, invece, le difficoltà sono maggiori e le sfide ancora aperte.

Con l'applicazione dei modelli probabilistici nell'ambito dell'informazione testuale e la creazione di modelli linguistici di grandi dimensioni già allenati e resi accessibili alla comunità digitale, le frontiere del riconoscimento automatico delle entità e dell'analisi testuale si sono aperte a nuovi fronti di sperimentazione. Se da un lato questo ha reso possibile ideare progetti di gestione e classificazione automatica di grandi quantità di dati testuali non strutturati, dall'altra parte, proprio perché ideati per i *big data*, questi strumenti applicati a lingue sottorappresentate e con risorse limitate hanno minori possibilità concrete di utilizzo. Nel caso di lingue come il latino, le cui fonti sono esclusivamente testuali, la cui quantità varia in base alla produzione dell'epoca e alle modalità di conservazione e tradizione dei testi, il limite quantitativo è un elemento connaturale da affrontare. Il latino medievale e, soprattutto, tardo medievale, sembra essere ancora meno rappresentato nell'ambiente digitale, sia per un limite nelle risorse, ma anche per una minore digitalizzazione dei testi. Per questo modelli linguistici come quelli messi a disposizione da CLTK e spaCy non possono ottenere risultati soddisfacenti, in quanto non sono mai stati allenati su testi dell'epoca. Ancor più difficile il cambio di contesto, da quello letterario-narrativo a quello notarile e, all'interno di quest'ultimo, nei vari ambienti (pubblico, privato, curia etc.), in quanto questo spostamento comporta dei cambiamenti totali di registro e linguaggio sui quali il modello necessita di essere allenato e perfezionato per essere in grado di comprendere e analizzare il testo, nonché l'emergere di entità di persone e luoghi locali che non hanno precedente rappresentanza in alcun documento.

Progetti come LiLa segnano una strada da percorrere in futuro: basi di dati concentrate sul lemma e interconnesse, per permettere l'accessibilità e la rappresentazione della lingua nella sua interezza. Per procedere in questo senso, però, è necessaria la collaborazione delle istituzioni archivistiche che devono procedere con la digitalizzazione e diffusione dei testi, totale o parziale.

Nella fase successiva, il progetto prevede l'impiego di metodi di riconoscimento automatico dei caratteri manoscritti (Handwriting Text Recognition - HTR) condotti tramite *Transkribus* direttamente sulla riproduzione digitale

dell'immagine del manoscritto originale. Questo porterebbe alla digitalizzazione di una quantità di testi sufficiente per poter sperimentare maggiormente nell'applicazione dell'NLP al latino medievale.

Esistono, infine, strumenti come il pacchetto R *quanteda* o la libreria Python *roman* che con poche semplici istruzioni consentono di eseguire delle operazioni di NLP sui testi digitalizzati anche di latino medievale. Queste opzioni hanno lo svantaggio di essere meno generalizzabili, ma in base allo scopo della ricerca, possono essere delle soluzioni facili e immediate da applicare e, con l'aggiunta di porzioni di codice specifiche, possono portare a risultati soddisfacenti, come l'estrazione delle date di redazione nel caso degli atti notarili.

Ringraziamenti

Questo progetto è stato finanziato dal PRIN 2022 ON (Objects in Network), e portato avanti in collaborazione con il Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi (DIBRIS) dell'Università di Genova e la Società Ligure di Storia Patria.

Riferimenti bibliografici

- Balard, Michel. 1983. *Notai genovesi in oltremare: atti rogati a Cipro da Lamberto di Sambuceto (11 ottobre 1296 – 23 giugno 1299)*. Istituto di Medievistica.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, et al. 2018. “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software* 3 (30): 774. <https://doi.org/10.21105/joss.00774>.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Muller, e Akitaka Matsuo. 2021. “quanteda: Quantitative Analysis of Textual Data.” <https://quanteda.io/index.html>.
- Boros, Emanuela, Veronica Romero, Martin Maarand, et al. 2020. “A Comparison of Sequential and Combined Approaches for Named Entity Recognition in a Corpus of Handwritten Medieval Charters.” In *17th International Conference on Frontiers in Handwriting Recognition (ICFHR), 8-10 settembre, Dortmund, Germania*, edited by Juan Guerrero. Washington, DC: IEEE Computer Society Conference Publishing Services, 79–84. <https://doi.org/10.1109/ICFHR2020.2020.00025>.
- Burns, Patrick J. 2023. “LatinCy: Synthetic Trained Pipelines for Latin NLP. Consultato il 5 novembre 2025. <https://arxiv.org/abs/2305.04365v1>.

- Chastang, Pierre, Sergio Torres Aguilar, and Xavier Tannier. 2021. "A Named Entity Recognition Model for Medieval Latin Charters." *Digital Humanities Quarterly*, 15 (4).
- CIDOC CRM Special Interest Group. 2024. *Volume A: Definition of the CIDOC Conceptual Reference Model version 7.1.3*. https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.3.pdf.
- Cyganiak, Richard, David Hyland-Wood, and Markus Lanthaler. 2014. "RDF 1.1 Concepts and Abstract Syntax. W3C Proposed Recommendation." Consultato il 5 novembre 2025. <https://www.w3.org/TR/rdf11-concepts>.
- Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. "Named Entity Recognition and Classification on Historical Documents: A Survey." *ACM Computing Surveys* 56 (2): 1-47. <https://doi.org/10.1145/3604931>.
- Erdmann, Alexander, David Joseph Wrisley, Benjamin Allen, et al. 2019. "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2-7 giugno 2019 Minneapolis, Minnesota*, edited by Jill Burstein et al. Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2223-34.
- Feliciati, Pierluigi. 2021. "Archives in a Graph: the Records in Contexts Ontology within the framework of standards and practices of Archival Description." *JLIS.It*, 12 (1): 92-101. <https://doi.org/10.4403/jlis.it-12675>.
- Gangemi, Aldo, Margherita Porena, Maria Chiara Frangipane, Maria Letizia Mancinelli, and Valentina Presutti. 2023. *ArCo Ontology (ArCo network) version 2.0*. <http://wit.istc.cnr.it/arco/lode/extract?lang=en&url=https://raw.githubusercontent.com/ICCD-MiBACT/ArCo/master/ArCo-release/ontology/arco/arco.owl>.
- García-González, Herminio, Elena Albarrán Fernández, Jose Labra Gayo, and Miguel Calleja Puerta. 2020. "Converting Asturian Notaries Public deeds to Linked Data using TEI and ShExML." In *Workshop on Humanities in the Semantic Web (WHiSe) Proceedings 2 giugno 2020 Heraklion, Greece*, edited by Alessandro Adamou, Enrico Daga, and Albert Meroño-Peñuela. CEUR Workshop Proceedings. Aachen, Germany: CEUR-WS, 41-46. <https://ceur-ws.org/Vol-2695/>.
- ICA/EGAD (International Council on Archives Expert Group on Archival Description). 2024. *International Council on Archives Records in Contexts Ontology (ICAR RiC-O) version 1.0.2*. Parigi: ICA (International Council on Archives).

- ICAR - Centro MAAS.2015. Archived January 19, 2022 at <https://web.archive.org/web/20220119183102/http://dati.san.beniculturali.it/lode/aggiornato.htm#d4e1643>.
- ICAR - Centro MAAS.2015. Archived January 19, 2022 at <https://web.archive.org/web/20220119183102/http://dati.san.beniculturali.it/lode/aggiornato.htm#d4e1781>.
- ISO (International Organization for Standardization). 2019. *Date and time – Representations for information interchange – Part 1: Basic rules*. ISO 8601-1:2019. Parigi: ISO.
- Istituto Centrale per gli Archivi, e Centro Metologie e Applicazioni per Archivi Storici. 2015. “SAN Ontologia version 1.1”. <http://dati.san.beniculturali.it/lode/aggiornato.htm>
- Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna e Regesta.exe. 2018. “OAD Vocabulary Specification version 1.3.” Consultato il 5 novembre 2025. <http://culturalis.org/oad/>.
- Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna.2025. “OAD - Ontology for Archival Description” Archived October 17, 2025 at <https://web.archive.org/web/20250820013801/http://culturalis.org/oad/#n>.
- Johnson, Kyle P. 2021. “CLTK: website.” Consultato il 5 novembre 2025. <http://cltk.org/>.
- Johnson, Kyle P. 2025. “CLTK: github repository.” Consultato il 5 novembre 2025. <https://github.com/cltk/cltk>.
- Johnson, Kyle P., Patrick Burns, John Stewart, Todd Cook, Clément Besnier, and William Mattingl. 2021. “The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations 1-6 agosto 2021 Bangkok, Thailand (online)*, edited by Heng Ji, Jong C. Park, and Rui Xia. Association for Computational Linguistics, 20-29. <https://doi.org/10.18653/v1%2F2021.acl-demo.3>.
- Kleehammer, Michael. 2024. “pyodbc 5.2.0.” Modificato il 17 ottobre 2025. <https://pypi.org/project/pyodbc/>.
- La Manna, Bianca. 2025. “Grafo di conoscenza di Notariorum Itinera serializzato in formato Turtle (1.0).” Zenodo. <https://doi.org/10.5281/zenodo.17339380>.
- Library of Congress. 2016. *BIBFRAME 2 Ontology, version 2.6.0*. <https://id.loc.gov/ontologies/bibframe.html>.

- Mambrini, Francesco, and Marco Carlo Passarotti. 2023. "The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms." *Journal of Open Humanities Data* 9 (28): 1–5. <https://doi.org/10.5334/johd.145>.
- Montani, Ines, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. "explosion/spaCy: v3.7.2: Fixes for APIs and requirements (v3.7.2). Zenodo". <https://zenodo.org/records/10009823>.
- Notarium Itinera. n.d.a. "Notarium Itinera." Consultato il 5 novembre 2025. <https://notariorumitinera.eu/>.
- Notariorum Itinera. n.d.b. "Digital library archivistica." Consultato il 5 novembre 2025. https://notariorumitinera.eu/Digital_Library_Archivistica.aspx.
- Notariorum Itinera. n.d.c. "Digital Library bibliografica." Consultato il 5 novembre 2025. https://notariorumitinera.eu/Digital_Library_Bibliografica.aspx.
- Peroni, Silvio. 2013. "Graffoo." <https://essepuntato.it/graffoo/specification/>.
- Peroni, Silvio, and David Shotton. 2012. "FaBiO and CiTO: ontologies for describing bibliographic resources and citations." *Journal of Web Semantics* 17: 33–43. <https://doi.org/10.1016/j.websem.2012.08.001>.
- Peroni, Silvio, and David Shotton. 2018. "The SPAR Ontologies." In *The Semantic Web - ISWC 2018 8-12 ottobre 2018 Monterey, California*, edited by D. Vrandećić et al. Switzerland: Springer International Publishing AG, 119–36. https://doi.org/10.1007/978-3-030-00668-6_8.
- Pilgrim, Mark, and Zope Foundation. 2025. "roman 5.0." Consultato il 18 luglio 2025. <https://pypi.org/project/roman/>.
- R Foundation. 2025. "The R Project for Statistical Computing." Modificato il 13 giugno 2025. <https://www.r-project.org/>.
- RDFLib Team. 2025. "rdflib 7.1.4." Consultato il 5 novembre 2025. <https://rdflib.readthedocs.io/en/stable/>.
- READ-COOP SCE. 2023. "Transkribus website." Consultato il 5 novembre 2025. <https://www.transkribus.org/it>.
- Roberts, Molly, Brandon Stewart, and Dustin Tingley. 2018. "stm. An R Package for the Structural Topic Model." Modificato il 31 luglio 2018. <https://www.structuraltopicmodel.com/>.
- Russo, Alessandro, Giorgia Lodi, and Margherita Porena. 2023. Archived November, 9 2025 at <https://web.archive.org/web/20250327055217/http://wit.istc.cnr.it/arco/lode/extract?lang=en&url=https://raw.githubusercontent.com/ICCD-MiBACT/ArCo/master/ArCo-release/ontologie/archive/archive.owl#d4e834>.

- Ruzzin, Valentina, ed. 2020. *Antonio de Inghibertis de Castro (Genova, 1330-1346)*. Genova: Società ligure di storia patria.
- Selivanov, Dmitry. 2016. "text2vec." Consultato il 5 novembre 2025. <https://text2vec.org/>.
- Sequeda, Juan, Syed Ahmed Tirmizi, Oscar Corcho, and Daniel P. Miranker. 2011. "Survey of Directly Mapping SQL Databases to the Semantic Web." *The Knowledge Engineering Review* 26 (4): 445–86. <https://doi.org/10.1017/S0269888911000208>.
- spaCy. n.d. "Training Pipelines & Models." Consultato il 5 novembre 2025. <https://spacy.io/usage/training>.
- Università Cattolica del Sacro Cuore. n.d. "LiLa: Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin." Consultato il 5 novembre 2025. <https://lila-erc.eu/#page-top>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*, 3 (1). <https://doi.org/10.1038/sdata.2016.18>.
- Won, Miguel, Patricia Murrieta-Flores, and Bruno Martins. 2018. "Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora." *Frontiers in Digital Humanities* 5. <https://doi.org/10.3389/fdigh.2018.00002>.

L'univocità e la variabilità terminologica a supporto della rappresentatività delle risorse semantiche di organizzazione e rappresentazione della conoscenza specialistica

Claudia Lanza*

Abstract: This paper aims at presenting a reflection upon some of the main mechanisms with which the semantic representativeness can be guaranteed by organization and representation knowledge systems with reference to a specialized domain of study. The analysis will dive into the concept of term uniqueness and update within these semantic resources. The perspective is to offer a methodological approach that could be employed over other domains characterized by a specialized terminological asset extracted from a source documentation strictly compliant with the sector of interest.

Keywords: Knowledge Organization Systems, Corpus-based representativeness, Information retrieval, Thesauri, Semantic coverage.

1. Introduzione

L'analisi verterà dapprima sull'importanza rivestita dal livello di specificità terminologica all'interno dei sistemi di organizzazione e gestione della conoscenza (Knowledge Organization Systems - KOSs). Verranno approfondite le fasi proprie dello sviluppo degli strumenti semantici, i.e., (i) la costruzione del corpus di partenza specialistico, connotato dai concetti chiave dei settori oggetto di studio selezionati sulla base di un'accurata estrazione all'interno della documentazione di riferimento, (ii) la trasformazione dei concetti chiave in dati strutturati, termini candidati per far parte delle risorse di organizzazione della conoscenza, (iii) la validazione di determinati asset terminologici nella prospettiva di garantire una copertura semantica rispetto al dominio di riferimento. Successivamente si presenteranno alcune tecniche qualitative e statisti-

* Dipartimento di Culture, Educazione e Società, Università della Calabria, Rende (CS), Italia. claudia.lanza@unical.it. ORCID: 0000-0002-3018-1987.

Il lavoro presenta un approfondimento in italiano di parte delle tecniche illustrate nella monografia pubblicata con l'editore CRC Press: Lanza, Claudia. 2022. *Semantic control for the cybersecurity domain: investigation on the representativeness of a domain-specific terminology referring to lexical variation*. CRC Press.

che per la determinazione della soglia di rappresentatività di uno strumento di controllo semantico. Queste analisi permetteranno di valutare la qualità delle risorse semantiche considerate come sistemi per guidare verso la comprensione del dominio specialistico (Albertoni et. al. 2016). Per quanto riguarda la prima tipologia di valutazione per la misurazione della rappresentatività semantica verranno presi in considerazione indici quali il grado di specificità dei termini, l'aderenza al dominio di studio, e la competenza degli esperti del campo di studi oggetto d'analisi. Il secondo gruppo di tecniche riguarda, invece, il rilevamento della rappresentatività in rapporto alla copertura semantica rispetto al corpus di partenza attraverso tecniche di mappatura con i sistemi di disseminazione informativa del settore, i.e., i *gold standard*, misure di frequenza, di richiamo e precisione, Type/token ratio e hapax, e infine la variabilità semantica.

2. Strutturare la conoscenza specialistica

I sistemi organizzazione della conoscenza specialistica hanno come presupposto fondante quello di consentire una gestione ottimale e strutturata di un determinato dominio di studio per finalità di recupero delle informazioni, fungendo da strumenti di mediazione tra i bisogni informativi espressi da un utente e la base di conoscenza del settore di riferimento (Hjørland 2008). La maggior parte degli studi presenti in letteratura che hanno descritto e approfondito le caratteristiche dei KOSs pone l'attenzione sulla loro funzione di esplicitazione delle informazioni dei settori specialistici in forma strutturata, in riferimento all'organizzazione della conoscenza in linearità con il dominio a cui quest'ultima appartiene (per citarne alcuni Hjørland 2003;2008;2016; Smiraglia 2014; Giunchiglia et al. 2014; Pieterse e Kourie 2014; Zeng 2008; Mazzocchi 2018; Souza et al. 2012; Alavi e Leidner 2001; Soergel 2009). La specificità terminologica è un elemento che caratterizza comunemente le analisi dei lavori presenti sui sistemi di organizzazione della conoscenza in quanto tale caratteristica conferisce aderenza al dominio. Un ulteriore tema, centrale nei lavori dei suddetti autori, è relativo alla diversificazione degli strumenti di controllo semantico che possono essere sviluppati per strutturare la conoscenza di un dato dominio di studio. Sigel (2016) passa in rassegna, così come altri contributi in letteratura, (tra cui Amirhosseini 2021; Souza et al. 2012; Zeng 2008), le diverse tipologie di KOSs partendo dalle liste di autorità, fino ad arrivare ai thesauri e alle ontologie come forme di massima espandibilità conoscitiva semantica per guidare l'utente verso un corretto recupero dell'informazione specialistica. Il processo di costruzione di una risorsa semantica parte generalmente da esigenze specifiche di determinati gruppi di utenti, per cui è necessario fornire un mezzo attraverso cui questi ultimi possano identificare il quadro informativo aggiornato del dominio. Pavel e Nolet (2002) evidenzia-

va come il principale punto d'analisi nell'ambito della ricerca terminologica dovesse essere orientato allo sviluppo di un modello di raggruppamento di termini che possano riflettere i concetti di domini strutturati della conoscenza specialistica, intesa come veicolo di diffusione dell'informazione tecnica verso determinati utenti.

2.1. Compilazione del corpus di partenza specialistico

La costruzione delle risorse di organizzazione e rappresentazione della conoscenza prevede una fase preliminare di costruzione del corpus di partenza specialistico (Broughton 2006), connotato dai concetti chiave dei settori oggetto di studio selezionati sulla base di un'accurata estrazione all'interno della documentazione di riferimento.

Questo primo passaggio implica sia una fase di trasformazione dei concetti chiave in dati strutturati, termini candidati per far parte delle risorse di organizzazione della conoscenza, che di validazione da parte di esperti del campo di studi oggetto d'analisi di determinati asset terminologici nella prospettiva di garantire una copertura semantica rispetto al dominio di riferimento.

Più specificatamente, la compilazione del corpus di partenza prevede una fase iniziale di raggruppamento della documentazione del settore oggetto d'analisi, e da questo insieme l'avvio delle successive fasi approfondite da ampi studi presenti nella letteratura (tra cui Kageura e Marshman 2019; Vivaldi e Rodríguez 2007; Pazienza et al. 2005; Jacquemin et al. 2005; Peñas et al. 2001) che coinvolgono l'estrazione dei termini di dominio per essere inclusi nel sistema di organizzazione e gestione della conoscenza prescelto e la validazione dell'estrazione terminologica per garantire la copertura semantica rispetto al dominio di riferimento.

La compilazione di un corpus documentale di partenza è un passaggio fondamentale per consentire la realizzazione di strumenti di organizzazione e rappresentazione della conoscenza al fine di classificare e sistematizzare le informazioni tecniche interne.

2.2. La costruzione di un sistema terminologico

La costruzione di un sistema terminologico efficiente implica solitamente l'acquisizione di informazioni orientate al dominio a partire dai testi del corpus, in particolare quelli in grado di fornire densità conoscitiva semantica e granularità sul lessico che si intende rappresentare (Barrière 2006). Queste strutture sono conosciute in letteratura come Basi di Conoscenza Terminologica (Terminological Knowledge Bases - TKB) (Condamines 2018). Nel dettaglio, queste ultime supportano i processi di strutturazione delle conoscenze

specialistiche fondendo le competenze proprie della linguistica computazionale. Lo scopo di creare risorse terminologiche solide, attraverso una selezione di termini candidati rappresentativi, è legato al tentativo di fornire un'aderenza alle conoscenze specialistiche incluse nel corpus di fonti consultate (Bowker e Pearson 2002).

I documenti selezionati per far parte del corpus di partenza, destinato a fornire un dataset informativo in input per le attività di pre/post-elaborazione linguistica, sono scelti nella prospettiva di recupero delle informazioni riferite a domini specifici e popolamento terminologico nelle risorse semantiche. Il presupposto da cui muove la raccolta dei documenti di un corpus di partenza è fondato sul principio di autorevolezza delle fonti al fine di portare a termine il processo di costruzione di uno strumento semantico secondo un principio di affidabilità nelle informazioni elaborate e condivise. Infatti, la scelta dei documenti che rientrano all'interno di un corpus specialistico è caratterizzata da principi relativi (i) all'affidabilità degli editori, (ii) al rispecchiamento delle tematiche del dominio, (iii) all'osservazione di un preciso arco temporale, (iv) o allo status ufficiale dei documenti riconosciuto da esperti del settore.

Nello sviluppo di tali strumenti di organizzazione e gestione della conoscenza per determinati domini di studio specialistici, le unità terminologiche generalmente provengono dai risultati di un'estrazione automatizzata dai documenti di partenza. I termini inclusi nelle risorse semantiche devono mantenere un valore univoco, come affermato nello standard *NISO TR-06-2017, Issues in Vocabulary Management* in riferimento alla definizione dei thesauri: «Vocabolario controllato: un elenco di termini che sono stati enumerati esplicitamente. Questo elenco è controllato ed è disponibile presso un'autorità di registrazione del vocabolario controllato. Tutti i termini di un vocabolario controllato devono avere una definizione univoca e non ridondante».

In questo contributo verrà presentata l'analisi sulla rappresentatività di una risorsa tesaurale, poi reingenirizzata in ontologia (Damerchiloo e Beheshti 2022; Kless et al. 2012) sul dominio della Cybersecurity (Lanza 2022; Lanza e Daille 2019)¹. Il thesaurus contiene attualmente 243 termini di dominio strutturati secondo le relazioni di gerarchia, associazione ed equivalenza seguendo le procedure di costruzione dei rapporti semantici stabiliti nella norma ISO 25964 del 2011 e del 2013.

¹ Si vedano i lavori pubblicati per il progetto dell'Osservatorio di Cybersecurity (OCS) dell'Istituto di Informatica e Telematica (IIT) – Consiglio Nazionale delle Ricerche (CNR): (Lanza 2022; Lanza e Daille 2019).

3. Misure qualitative della rappresentatività delle risorse di organizzazione e gestione della conoscenza

3.1. Specificità e univocità terminologica

Condamines e Picton (2022) hanno approfondito dettagliatamente i presupposti di base della cosiddetta *terminologia testuale* che pone come prospettiva quella di accedere al corpus tramite i termini specialistici del dominio oggetto di studi. Per concretizzare questa forma di accesso gli autori propongono tre passaggi da seguire: (i) la redazione di un elenco dei termini candidati, studiando il significato in base alla distribuzione dei termini nel corpus (secondo un approccio di tipo bottom-up); (ii) l'analisi dei contesti in cui compaiono i termini, ricercando quelli in cui sono presenti elementi informativi rilevanti per lo studio (secondo un approccio di tipo top-down); e (iii) la convalida da parte di esperti del settore.

Il primo passaggio implica la categorizzazione dei contesti dei termini e la loro successiva interpretazione e prende le basi dalla sociolinguistica (Firth 1957) e dalla visione matematica del linguaggio (Harris 1968). L'idea principale alla base di questo approccio è che esista una correlazione tra similarità distributiva e similarità di significato. È quindi possibile comprendere il significato di una parola o di un gruppo di parole considerando tutti i suoi contesti. All'interno del secondo passaggio, invece, vengono definiti e specificati gli indicatori di relazione concettuale o elementi lessico-sintattici a cui è associata un'interpretazione preliminare, ovvero gli indicatori associati a contesti ricchi di conoscenza hanno meno probabilità di variare da un corpus specializzato all'altro. Per ogni analisi, gli elementi direttamente correlati all'obiettivo dello studio possono essere definiti a priori, ovvero la presenza di elementi che indicano uno spostamento di significato, una variazione, ecc.

Pertanto, la specificità del dominio dovrebbe essere considerata una delle caratteristiche principali da tenere in considerazione per calcolare la soglia di copertura degli strumenti semantici e la variazione della terminologia nel tempo.

Sulla stessa linea del contributo precedentemente menzionato, Pulizzotto et al. (2022) nel framework dell'analisi concettuale prendono in considerazione il concetto lessicale. Nel contributo degli autori si fa specifico riferimento al campo della *Conceptual analysis (CA)* per cui il significato di una parola veicola sempre un concetto e un concetto ha sempre la sua espressione lessicale, la forma lessicale standard di un concetto. Non sempre, però, seguendo le riflessioni degli autori, i concetti hanno una chiara corrispondenza nelle lingue naturali e possono essere espressi anche da un gruppo di parole. La visione offerta dagli autori presuppone la considerazione della terminologia basata sulla produzio-

ne del testo in stretto rapporto con il suo contesto, il corpus documentale di partenza. I testi vengono raccolti per formare un corpus perché appartengono a generi pertinenti allo scopo dello studio e l'interpretazione dei dati linguistici deve tenere in considerazione la natura dei testi e l'oggetto dello studio.

La maggior parte delle considerazioni esposte nell'ambito della terminologia testuale si concentra sull'analisi dei corpora di partenza e sul confronto tra sub-corpora, dove le differenze di distribuzione sono associate a differenze ravvisate nelle situazioni comunicative che precedono la costruzione del sub-corpus. Nello specifico, nei casi in cui vi è la necessità di identificare l'utente finale, un gruppo di utenti esperti o non esperti del settore per i quali si procede alla realizzazione di sistemi di rappresentazione semantica, possono emergere differenze nella distribuzione dei termini, ad esempio quando vi è un utilizzo di alcuni di essi nel linguaggio generale e può essere necessario avviare un processo di normalizzazione terminologica per rappresentare univocamente il concetto veicolato. Tutto ciò viene osservato nella prospettiva di garantire una rappresentatività adeguata alle risorse semantiche proprie di un determinato campo di studi specialistico che è connotato da documenti che presentano lessici settoriali.

Condamines (2018) sottolinea l'importanza per i TKB di essere allineati alla specificità del corpus, nel modo in cui dovrebbero aderire alla natura inequivocabile dei domini specializzati e ai contesti applicativi in cui sono destinati ad essere utilizzati (Dancette e L'Homme 2004).

L'accuratezza e l'appropriatezza dei termini rispetto alla conoscenza di un dominio specialistico sono temi ampiamente approfonditi negli studi di Cabré et al. (2001) quando esplora i molteplici strumenti di estrazione terminologica. L'autrice fa specifico riferimento all'applicazione pratica delle risorse semantiche, siano esse appartenenti ad ambiti generali di studio o specialistici. Ciò che viene estratto è infatti un elenco di termini singoli o composti, che devono presentare un elevato livello di unità, concetto introdotto da Kageura e Umino (1996) che indica il grado di forza delle collocazioni sintagmatiche.

3.2. Copertura semantica rispetto alla documentazione del dominio di studi

Nel rapporto che si instaura tra la specificità e univocità e i termini di un determinato settore di studio gioca un ruolo fondamentale l'analisi del dominio di studi specialistico. Per Santini (2018) nell'analisi della qualità dei corpora nell'ambito del web possono essere individuate come dimensioni della variazione linguistica variabili quali dominio, genere, registro, stile, mezzo. Seguendo le sue affermazioni il dominio è il "campo tematico", "area" in cui viene utilizzato il documento.

Kageura e Umino (1996) esplorano il concetto di univocità in termini di connessione lessicale stabile che si instaura rispetto ai concetti specifici del dominio. Il grado di specificità può essere calcolato confrontando il corpus specializzato con uno generico, ciò può evidenziare il peso dei termini in contesti diversi e aiutare la selezione di quelli più tecnici per il dominio ai fini di garantire una maggiore rappresentatività rispetto al campo della conoscenza oggetto d'interesse. Nella riflessione portata avanti da Faber e Rodriguez (2012) la specificità terminologica fa parte di domini settoriali in contesti in cui la conoscenza specializzata è permeata da una presentazione di concetti che hanno la necessità di acquisire una forma standardizzata terminologica per la loro designazione univoca, diventando così punti di accesso a strutture informative più scientifiche e tecniche. Per ottenere questa forma di sistematizzazione della conoscenza gli autori evidenziano l'importanza di affidarsi a risorse documentali aderenti al grado di settorialità del dominio, cioè che possano allineare il lessico del settore con le attività di normalizzazione terminologica e stabilire una connessione interlinguistica con le strutture della conoscenza. Un recente studio che può essere incluso nelle riflessioni sul rapporto tra univocità terminologica e aderenza al dominio di studi è quello di Barité e Rauch (2022) che presentano un approfondimento agli studi sulla terminologia riferendosi all'analisi del dominio, nella considerazione macrostrutturale della mappatura dei termini su un dominio di studi.

La specificità terminologica in rapporto al dominio di studi è in larga parte osservata da studi che analizzano la documentazione da cui i termini provengono, il corpus di partenza, quale « Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi.» (Barbera et al. 2007, 26). Nello specifico, il corpus documentale creato per la realizzazione di KOSs, e composto in massima parte da testi specialistici, deve seguire delle fasi di compilazione che prendano in considerazione l'allineamento della terminologia utilizzata nella comunità di riferimento circoscritta a un periodo di tempo delimitato a seconda delle esigenze degli utenti finali (Pearson 1998).

Sulla base dell'aderenza terminologica al dominio di studi e all'univocità che i termini riescono a mostrare in linearità con la documentazione di partenza, diversi studi in letteratura hanno posto l'attenzione ai principi di rappresentatività del corpus documentale rispetto all'ambito della conoscenza da strutturare. Sinclair (2005), Leech (1991), Biber (1993) condividono l'idea che il valore di un corpus possa essere attribuito alla sua proporzione numerica dei testi posizionando il processo di compilazione del sistema documentale di partenza in un determinato arco temporale che possa delimitare l'osservazione

della variabilità semantica all'interno della documentazione collazionata. Studi sulla rappresentatività dei corpora, tra cui Biber (1993), Condamines (2007), Hunston (2002) suggeriscono che la copertura dei corpora possa essere correlata al contesto applicativo al quale le risorse semantiche saranno applicate in modo efficace.

Per Indurkha (2010) il numero di campioni nelle diverse categorie di testo dovrebbe essere proporzionale alle loro frequenze e/o pesi nella popolazione target affinché il corpus risultante sia considerato rappresentativo, «The proportion and number of samples for each text category. The numbers of samples across text categories should be proportional to their frequencies and/or weights in the target population in order for the resulting corpus to be considered as representative» (Indurkha 2010, 153). Per Condamines e Picton (2022) la rappresentatività del corpus è correlata all'obiettivo dello studio, pertanto la tipologia di testi attraverso cui i termini vengono diffusi gioca un ruolo importante. Occorre inoltre prestare attenzione alla presenza di sottocorpora e ai livelli di difficoltà di accesso ai dati. Per l'autrice la rappresentatività del corpus è legata alla nozione di dominio e sembra più facile da raggiungere in corpora specializzati, ma la conoscenza non è organizzata a priori in domini in cui si presume che discorso e terminologia siano strutturati; la conoscenza, infatti, non è naturalmente discontinua (Rogers 2013).

Hunston (2002) suggerisce l'ipotesi di considerare il corpus come un riferimento statico o dinamico. Nel primo caso possiamo riferirci a un corpus come a un campione di documenti che possono rappresentare un dominio, nel secondo la compilazione del corpus si riferisce al monitoraggio del cambiamento semantico. Come chiarito nei lavori pubblicati dalla sezione dedicata allo studio dei corpora dell'Università di Lancaster (McEnery e Hardie 2012), i modelli statici possono presentare diverse opportunità per supportare un'analisi molto efficiente della variazione linguistica interna ai documenti utilizzati, trasformandoli in frame di campionamento, che a loro volta divengono strumenti essenziali per comprendere la granularità di un corpus specialistico: i corpora basati su una documentazione non specialistica tendono a calcolare la loro rappresentatività sulla base dell'incorporazione di diversi tipi di testi, mentre per quanto riguarda quelli specializzati, l'attenzione dovrebbe essere mantenuta sull'intervallo di "chiusura" o "saturazione" del corpus.

Barbera et al. (2007, 50) chiariscono che quando si analizza la rappresentatività delle risorse semantiche si pone l'attenzione alla nozione di "bilanciamento":

Una discussione particolarmente efficace, che mette in luce i vari problemi della nozione, è data da Tognini-Bonelli 2001, p. 55-57, che si pone direttamente il problema di come e cosa campionare ("sampling") per conseguire tale rappresentatività. Alla rappresentatività è strettamente legato anche il dibattito sulle dimensioni di un corpus (più il corpus è ampio, più lo si potrebbe rite-

nere rappresentativo di una lingua. E se rappresentatività è bilanciamento, si tratta anche, per dirla alla Engwall (1994), di una questione di “choice e non di chance”, in cui la fase di progettazione e finalizzazione non è dunque secondaria: l'ipotesi della scelta casuale implica infatti che la varietà dei testi, qualunque sia la sua estensione, non sarà rappresentativa dell'intera popolazione, né di uno specifico gruppo; la selezione programmata ed organizzata dei materiali, invece, deve aderire alla definizione della lingua che il corpus si propone di rappresentare ed i confini entro i quali le scelte avvengono devono essere ben tracciati. Rappresentatività, finalizzazione e dimension sono pertanto parametri interrelati: è infatti chiaro però che l'attenzione per la rappresentatività di un corpus significa attenzione per la funzione cui il corpus stesso è destinato; ed ai fini di ricerche specifiche, una raccolta di dimensioni limitate ma in sé equilibrata, ben etichettata ed accuratamente verificata, comprendente in sé la dimensione linguistica che interessa studiare, risulterà più efficace di una raccolta ampia ma non proporzionata al suo interno.

4. Misure quantitative della rappresentatività delle risorse di organizzazione e gestione della conoscenza

Nella seguente sezione verranno presentate le principali tecniche di misurazione adottate per l'attività di ricerca eseguita sul domino di studi della sicurezza informatica².

4.1. Frequenza terminologica

Facendo riferimento ai lavori di Rapp (2014) e Francom (2010) la rappresentatività garantita dalle risorse semantiche viene calcolata dalla frequenza delle parole in un corpus, dalla loro co-occorrenza e dalle parole comuni nei testi. «A corpus whose word frequencies are highly correlated to the familiarity norms is more likely to be a good surrogate for everyday language» (Rapp, 2014, 2118).

La determinazione dei valori di precisione dei concetti rappresentati dai termini in una risorsa semantica può essere calcolata attraverso diverse tecniche tra cui la misura della frequenza, che indica la rappresentatività di un termine rispetto a un concetto specializzato di un dominio.

Questa misurazione statistica assistita a livello computazionale si ottiene mettendo in relazione il rapporto tra due liste terminologiche (una specifica per il dominio e una per la documentazione generale), ottenute calcolando le frequenze relative per ciascuna di esse. I punteggi di diversi termini conte-

² Il corpus creato per la realizzazione del thesaurus sulla Cybersecurity è composto da 560 documenti che coprono un arco temporale dal 2002 al 2017 e include sia documentazione legislativa che divulgativa (riviste di settore specialistiche) in lingua italiana, così come report e documentazione tecnici.

nuti nella lista ottenuta dalla fase di estrazione terminologica³ possono essere messi a confronto con un elenco più generico di termini estratti da un corpus linguistico generale per evidenziare come alcuni di essi siano caratterizzati da un livello di specificità più elevato. Questa comparazione ha permesso di selezionare i termini ad altri come termini di ingresso nel thesaurus. I risultati hanno dimostrato che alcuni dei termini candidati rappresentativi, selezionati in base alla validazione dei gruppi di esperti, presentavano un punteggio più elevato nel calcolo della specificità, come ad esempio *spam* o *malware*, mentre altri venivano recuperati molto frequentemente in contesti generici e influenzavano da un punto di vista terminologico alcune scelte terminologiche da fare nelle fasi di validazione verso l'inclusione in modo manuale, come nel caso, ad esempio, di *virus*, *sicurezza*, *accesso*, *protezione*.

4.2. Richiamo e Precisione

Le metriche di Richiamo e Precisione rilevano la specificità dei concetti come valore discriminante per garantire la copertura di uno strumento semantico rispetto a determinati documenti di partenza. Nell'analisi portata avanti sul dominio della sicurezza informatica il risultato è stato ottenuto stabilendo un confronto con corpora esistenti sul settore di studio, ovvero i *gold standard* (Shadbolt et al. 1999) attraverso una procedura di corrispondenza nell'allineare termini rappresentativi scelti da quelli estratti con le unità semantiche ufficialmente accettate dai gruppi di esperti. Nel dettaglio, sono state messe a confronto le liste di termini estratti con i software impiegati per questa attività (T2K e Termsuite) con quelle estratte dai principali asset terminologici presenti sul dominio della Cybersecurity. In questo modo è stato possibile identificare il livello della copertura che le fonti selezionate per compilare il corpus di partenza e, di conseguenza, la risorsa semantica possono garantire rispetto alle informazioni condivise all'interno di comunità di esperti. È anche possibile stabilire così la presenza di discrepanze e procedere, seguendo la supervisione degli utenti esperti del dominio di studi, con una diversa sistematizzazione delle unità terminologiche all'interno delle risorse semantiche al fine di garantire una forma di rappresentatività sempre più alta.

4.3. Type/token Ratio e Hapaxes

Tra le altre misurazioni statistiche a supporto della definizione di un modello di rappresentatività semantica rientrano quelle incentrate sul calcolo della

³ L'estrazione terminologica da cui è stata avviata la fase di scelta dei termini candidati a far parte della risorsa semantica è stata eseguita con l'ausilio di software automatici Text2Knowledge (Dell'Orletta et al. 2014) e Termsuite (Daille 2016).

ricchezza lessicale e della diversità dei vocabolari controllati in determinati corpora, cioè il rapporto *type/token ratio* (Kettunen 2014), e quelle riferite all'individuazione della proporzione delle parole che ricorrono molto raramente e una sola volta in un corpus di partenza, e che possono evidenziare quanto possa essere valutato specifico un elenco di termini candidati da includere in una risorsa semantica, gli *hapax*. Il risultato ottenuto dal rapporto di tutte le parole presenti in un corpus (token, unità minime in cui è diviso il testo (Cfr. Piscopiello e Beraccini 2009) e quelle uniche (type) indica

in valori percentuali, il numero di parole diverse all'interno di un determinato testo rispetto al numero di parole complessive dello stesso. A titolo d'esempio, affermare che, a parità di tokens, $TTR1 > TTR2$ (dove $TTR1$ e $TTR2$ rappresentano rispettivamente la Type/Token Ratio di un ipotetico testo 1 e di un ipotetico testo 2) equivale a dire che il numero di parole diverse del testo 1 è superiore rispetto a quello del testo 2. Pertanto, applicando tale parametro ai concetti di densità e ricchezza terminologica, si potrebbe affermare che la TTR è direttamente proporzionale al carattere divulgativo di un testo: ciò vuol dire che quanto più elevato è il valore della TTR (e, quindi, quanto più grande è il numero di types rispetto ai tokens), tanto più il testo sarà ricco a livello lessicale/terminologico e, di conseguenza, tante più probabilità si avranno che lo stesso abbia un carattere divulgativo piuttosto che scientifico. In definitiva, tale ricchezza si traduce generalmente in un numero significativo di varianti e sinonimi, condizione questa – come si è visto – tipica dei testi divulgativi”, la formula che invece fornisce l'elenco degli hapax può essere considerata una misura efficiente per calcolare il grado di specificità e accuratezza terminologica in rapporto ai documenti di partenza (Piscopiello e Bertaccini 2009, 5).

Per quanto riguarda il corpus costruito per la Cybersecurity, quasi il 70% delle parole ricorre una volta, le più frequenti sono, ad esempio: *rischi, controllo, tecnologie, Internet, dispositivi*, e nelle prime centinaia è incluso anche il termine *Cybersecurity*. Per quanto riguarda le parole uniche recuperate con il calcolo 1 e 2 gram, le più frequenti sono - ad eccezione delle stopwords che in questa fase vengono fornite in output e poi saranno rimosse utilizzando tecniche di estrazione dei termini più sofisticate, rispettivamente: 1-gram: *sicurezza, servizi, protezione, controllo, software*; 2-grams: *Parlamento europeo, decreto legislativo, dati personali, accesso, comunicazioni elettroniche*.

4.4. Variabilità semantica

L'evoluzione linguistica che interessa i diversi campi del sapere specialistico va di pari passo con gli aggiornamenti dal punto di vista terminologico, che, caratterizzati da fattori di continuità, puntano a rappresentare in modo onnicomprensivo lo spettro informativo di un determinato dominio. Come affermato da L'Homme (2015) i termini sono strettamente dipendenti dai conte-

sti in cui si suppone essi siano impiegati. Le risorse semantiche contenenti i termini rappresentativi che riflettono il quadro concettuale di specifici ambiti di studio sono destinate ad essere aggiornate nella prospettiva di seguirne la variabilità semantica e, di conseguenza, agevolare il miglioramento della specificità e dell'univocità terminologica nella rappresentazione della conoscenza specialistica (Cabr  2009).

La variabilit  dei termini in determinati strumenti di controllo semantico (Daille 2005)   una caratteristica essenziale da tenere in conto per lo sviluppo di risorse volte a rappresentare in forma standardizzata uno specifico campo della conoscenza. L'aggiornamento rispetto ai cambiamenti presenti nei contenuti pubblicati da una serie di fonti prestabilite negli ambiti specialistici d'indagine pu  garantire una versione attuale, e di conseguenza pi  affidabile, di uno strumento semantico se collegato a flussi di nuove informazioni. La misurazione che pu  essere effettuata sulla lista dei termini inclusi all'interno di una risorsa semantica per garantirne la sua rappresentativit  rispetto al dominio d'interesse specialistico prevede una mappatura temporale con i sistemi scientificamente utilizzati per la disseminazione della conoscenza settoriale.

5. Conclusioni

La prospettiva di questo contributo   di fornire un approfondimento teorico sui principi di accuratezza e appropriatezza dei termini rispetto alla conoscenza di un dominio specialistico, tema ampiamente approfondito negli studi di Cabr  et al. (2001) quando esplora i molteplici strumenti di estrazione terminologica. Le risorse semantiche di organizzazione della conoscenza specialistica contengono al loro interno un elenco di termini singoli o composti in forma strutturata che devono presentare un elevato livello di *unit *, concetto introdotto da Kageura e Umino (1996). Per garantire una forma di copertura semantica di tali risorse rispetto al campo del sapere che queste ultime sistematizzano   stata presentata una panoramica di alcune tecniche adoperabili in differenti contesti applicativi. Le valutazioni qualitative relative alla specificit  e univit  terminologica si sommano alle misurazioni statistiche sulla copertura semantica rispetto all'informazione contenuta nei *gold standard* del dominio di studi in base a criteri di frequenza terminologica e ricchezza lessicale. La possibilit  di disporre delle competenze nel dominio di studio, caratteristica spesso resa possibile dalla presenza della figura del gruppo di esperti del dominio (Schultz et al. 1968), rappresenta un ulteriore supporto essenziale per le fasi di validazione verso l'inclusione dei termini candidati tali da garantirne una rappresentativit  aderente ai contesti d'uso. In ultima istanza, risulta fondamentale porre l'attenzione sui processi di aggiornamento dei KOSs nell'asset terminologico che offrono prendendo in considerazione il principio della

variabilità semantica per poter assicurare una forma di allineamento accurato rispetto alle variazioni avvenute nello spettro conoscitivo del settore di studio.

Riferimenti bibliografici

- Alavi, Maryam, and Dorothy E Leidner. 2001. "Knowledge management and knowledge management systems: Conceptual foundations and research issues." *MIS quarterly*, 25:107-36. <https://doi.org/10.2307/3250961>.
- Albertoni, Riccardo, Monica De Martino, and Alfonso Quarati. 2016. "Integrated Quality Assessment of Linked Thesauri for the Environment." In *Electronic Government and the Information Systems Perspective. EGOVIS 2016*, edited by A. Kö and E. Francesconi, *Lecture Notes in Computer Science* 9831:221-35. Springer, Cham. https://doi.org/10.1007/978-3-319-44159-7_16.
- Amirhosseini, Maziar. 2021. "A Dialectic Perspective on the Evolution of Thesauri and Ontologies." *Knowledge organization* 48(6): 403-29. <https://doi.org/10.5771/0943-7444-2021-6-403>.
- Barbera, Manuel, Elisa Corino, e Cristina Onesti. 2007. *Corpora e linguistica in rete*. Guerra Edizioni.
- Barité, Mario, and Mirtha Rauch. 2022. "Terminological studies as domain analysis: a critical exploration." *Brazilian Journal of Information Science: Research trends* 16. <https://doi.org/10.36311/1981-1640.2022.v16.e02140>.
- Barrière, Caroline. 2006. "Semi-automatic corpus construction from informative texts." In *Lexicography, Terminology and Translation, Text-Based Studies in honour of Ingrid Meyer*, edited by Lynne Bowker. University of Ottawa Press. <https://doi.org/10.2307/j.ctt1ckpgs3.9>.
- Biber, Douglas. 1993. "Representativeness in corpus design." *Literary and linguistic computing* 8 (4): 243-57. <https://doi.org/10.1093/llc/8.4.243>.
- Bowker, Lynne, and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.
- Broughton, Vanda. 2006. *Essential thesaurus construction*. Facet publishing.
- Cabré, María Teresa. 2009. "La Teoría Comunicativa de la Terminología, una aproximación lingüística a los términos." *Revue Francaise De Linguistique Appliquee* 14 (2009): 9-15.

- Cabrè, Maria Teresa, Rosa Estopà, and Jordi Vivaldi Palatresi. 2001. "Automatic term detection: A review of current systems." In *Natural Language Processing 2 - Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme. John Benjamins Publishing Company.
- Condamines, Anne and Aurélie Picton. 2022. "Textual Terminology: Origins, principles and new challenges." In *Theoretical Perspectives on Terminology*, John Benjamins Publishing Company.
- Condamines, Anne. 2007. "L'interprétation en sémantique de corpus : le cas de la construction de terminologies." *Revue Française de Linguistique Appliquée* XII (1):39-52. <https://doi.org/10.3917/rfla.121.0039>.
- Condamines, Anne. 2018. "Terminological knowledge bases From Texts to Terms, from Terms to Texts." In *The Routledge Handbook of Lexicography*, edited by Pedro A. Fuertes-Olivera. Routledge.
- Cram, Damien, and Béatrice Daille. 2016. "Terminology extraction with term variant detection." In *Proceedings of ACL-2016 System Demonstrations, August 2016 Berlin, Germany*, edited by Sameer Pradhan and Marianna Apidianaki. Association for Computational Linguistics. Kerrville (Texas), 13-18.
- Daille, Béatrice. 2005 "Variations and application-oriented terminology engineering." *Terminology International Journal of Theoretical and Applied Issues in Specialized Communication*, 11(1): 181-97. <https://doi.org/10.1075/term.11.1.08dai>.
- Damerchiloo, Mansoureh, and M. Hosseini Beheshti. 2022 "Converting thesaurus to ontology (a systematic review)." *Library and Information Science Research* 11 (2):105-27.
- Dancette, Jeanne Eugénie, and Marie-Claude L'Homme. 2004. "Building specialized dictionaries using lexical functions." *Linguistica Antverpiensia*.
- Dell'Orletta, Felice, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. "T2K²: a system for automatically extracting and organizing knowledge from texts." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, Reykjavik, Iceland, 2014, 2062-70.

- Faber, Pamela, and Clara Inés López Rodríguez. 2012. "2.1 terminology and specialized language." *A cognitive linguistics view of terminology and specialized language* 20.9.
- Firth, John Rupert. 1957. *Applications of General Linguistics*. <https://doi.org/10.1111/J.1467-968X.1957.TB00568.X>.
- Francom, Jerid, Amy LaCross, and Adam Ussishkin. 2010. "How Specialized are Specialized Corpora? Behavioral Evaluation of Corpus Representativeness for Maltese." In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, edited by Nicoletta Calzolari, Khalid Choukr, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias European Language Resources Association (ELRA).
- Giunchiglia, Fausto, Biswanath Dutta, and Vincenzo Maltese 2014. "From knowledge organization to knowledge representation." *Knowledge Organization* 41 (1):44-56. <https://doi.org/10.5771/0943-7444-2014-1-44>.
- Harris, Zellig. 1968. *Mathematical Structures of Language. Interscience Tracts in Pure and Applied Mathematics* 21. New Yor: Interscience Publishers.
- Hjørland, Birger. 2003. "Fundamentals of knowledge organization." *Knowledge organization* 30 (2):87-111. <https://doi.org/10.5771/0943-7444-2003-2-87>.
- Hjørland, Birger. 2008. "What is knowledge organization (KO)?" *Knowledge Organization*, 35(2-3): 86-101. <https://doi.org/10.5771/0943-7444-2008-2-3-86>.
- Hjørland, Birger. 2016. "Knowledge organization (KO)." *Knowledge Organization*, 43(6):475-84. <https://doi.org/10.5771/0943-7444-2016-6-475>.
- Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge University Press, <https://doi.org/10.1017/CBO9781139524773>.
- Kageura, Kyo and Bin Umino. 1996. "Methods of automatic term recognition: A review." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259-89. <https://doi.org/10.1075/term.3.2.03kag>.
- Kageura, Kyo, and Elizabeth Marshman. 2019. "Terminology extraction and management." In *The Routledge handbook of translation and technology*, edited by Minako O'Hagan. Routledge. 61-77.
- Kettunen, Kimmo. 2014. "Can Type-Token Ratio Be Used to Show Morphological Complexity of Languages?" *Journal of Quantitative Linguistics*, 21(3): 223-45. <https://doi.org/10.1080/09296174.2014.911506>.

- Kless, Daniel, Ludger Jansen, Jutta Lindenthal, and Jens Wiebensohn. 2012. "A method for re-engineering a thesaurus into an ontology". In *Formal Ontology in Information Systems. Proceedings of the Seventh International Conference (FOIS 2012)*, IOS Press.133-46.
- Jacquemin, Christian, and Didier Bourigault. 2005. "Term extraction and automatic indexing." In *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov. Oxford University Press. 599-615.
- Indurkha, Nitin, and Fred J. Damerau (Eds.). 2010. *Handbook of Natural Language Processing* (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781420085938>.
- ISO (International Organization for Standardization). 2011. *Information and documentation -Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval*. ISO 25964-1:2011.
- ISO (International Organization for Standardization). 2013. *Information and documentation – Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies*. ISO 25964- 2:2013.
- Lanza, Claudia. 2022. *Semantic control for the cybersecurity domain: investigation on the representativeness of a domain-specific terminology referring to lexical variation*. CRC Press.
- Lanza, Claudia, and Béatrice Daille. 2019. "Terminology systematization for Cybersecurity domain in Italian Language." In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019, Terminologie et Intelligence Artificielle (atelier TALN-RECITAL& IC) Toulouse, France*, edited by Emmanuel Morin, Sophie Rosset and Pierre Zweigenbaum. ATALA, 7-18.
- Leech, Geoffrey. 1991. "The state of the art in corpus linguistics." In *English Corpus Linguistics*, Routledge.
- L'Homme, Marie-Claude. 2015. "Terminologies and Taxonomies." In *The Oxford Handbook of the Word*, edited by John R Taylor, Oxford University Press.
- Mazzocchi, Fulvio 2018. "Knowledge organization system (KOS): an introductory critical account." *Knowledge Organization* 45(1): 54-78.
- McEnery, Tony, and Andrew Hardie. 2012. "Answers to exercises: Chapter One discussion questions." In *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- NISO (National Information Standards Organization). 2017. *TR-06-2017Issues in Vocabulary Management*.

- Pavel, Silvia, et Diane Nolet. 2002. *Précis de Terminologie*. Terminologie et normalisation, Bureau de la traduction.
- Pazienza, Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. "Terminology extraction: an analysis of linguistic and statistical approaches." In *Knowledge mining: Proceedings of the NEMIS 2004 final conference*. Berlin, Springer-Verlag Berlin and Heidelberg GmbH & Co. KG., 185:255-79.
- Pearson, Jennifer. 1998. *Terms in Context*. Amsterdam: John Benjamins Publishing Company.
- Peñas, Anselmo, Felisa Verdejo, and Julio Gonzalo. 2001. "Corpus-based terminology extraction applied to information access." *Proceedings of corpus linguistics*.
- Pieterse, Vreda, and Derrick G. Kourie. 2014. "Lists, taxonomies, lattices, thesauri and ontologies: paving a pathway through a terminological jungle." *Knowledge Organization* 41 (3): 217-29. <https://doi.org/10.5771/0943-7444-2014-3-217>.
- Piscopiello, Maria Grazia, e Franco Bertaccini. 2009. "Densità e ricchezza terminologica: generi testuali a confronto." *mediazioni* 7. <https://mediazioni.sitlec.unibo.it/index.php/no7-anno2009.html>.
- Pulizzotto, Davide, Jean-François Chartier, Francis Lareau, Jean-Guy Meunier, and Louis Chartrand. 2018. "Conceptual Analysis in a Computer-Assisted Framework: Mind in Peirce." *Umanistica Digitale* 2 (2). <https://doi.org/10.6092/issn.2532-8816/7305>.
- Reinhard Rapp. 2014. "Using Collections of Human Language Intuitions to Measure Corpus Representativeness." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2117–2128, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Rogers, Margaret. 2013. "What is a 'Domain' and is this a Useful Question?" *Asp, revue du GERAS* 64: 5-16. <https://doi.org/10.4000/asp.3810>.
- Santini, Marina, Wiktor Strandqvist, Mikael Nyström, Marjan Alirezai and Arne Jönsson. 2018. "Can We Quantify Domainhood? Exploring Measures to Assess Domain-Specificity in Web Corpora." In *Database and Expert Systems Applications. DEXA 2018. Communications in Computer and Information Science*, edited by M. Elloumi et al. 903. Springer, Cham. https://doi.org/10.1007/978-3-319-99133-7_17.

- Sinclair, John. 2005. "Corpus and Text - Basic Principles". In *Developing Linguistic Corpora: A Guide to Good Practice*, edited by Martin Wynne.
- Schultz, Claire K. *et al.* 1968. *Evaluation of Indexing by Group Consensus. Final Report*. ED025288.
- Shadbolt, Nigel, Kieron O'hara, and Louise Crow. 1999. "The experimental evaluation of knowledge acquisition techniques and methods: history, problems and new directions." *International Journal of Human-Computer Studies* 51(4):729-55, <https://doi.org/10.1006/ijhc.1999.0327>.
- Sigel, Alexander. 2006. "From traditional Knowledge Organization Systems (authority files, classifications, thesauri) towards ontologies on the web." in *Introducing Terminology-based Ontologies* edited by Winfried Schmitz-Esser & Alexander Sigel.
- Smiraglia, Richard P. 2014. *The elements of knowledge organization*. Springer Cham Heidelberg.
- Soergel, Dagobert. 2009. *Knowledge organization systems: overview*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/<https://www.dsoergel.com/SoergelKOSOverview.pdf>
- Souza, Renato R., Douglas Tudhope, and Mauricio B. Almeida. 2012. "Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems." *Knowledge organization* 39(3):179-92. <https://doi.org/10.5771/0943-7444-2012-3-179>.
- Vivaldi, Jorge, and Horacio Rodríguez. 2007. "Evaluation of terms and term extraction systems: A practical approach. Terminology". *International Journal of Theoretical and Applied Issues in Specialized Communication* 13 (2): 225-48. <https://doi.org/10.1075/term.13.2.06viv>.
- Zeng, Marcia. 2008. "Knowledge organization systems (KOS)." *Knowledge Organization* 35 (2-3):160-82. <https://doi.org/10.5771/0943-7444-2008-2-3-160>.

Possibilities for using cultural warrant to complement literary work:

An analysis of the thesaurus of Brazilian Folklore and Popular Culture

Larissa Moraes Martins, Carlos Eduardo da Silva Carvalho,
Ana Cristina de Albuquerque*, Mario Barité**

Abstract: This study analyzes the Brazilian Folklore and Popular Culture Thesaurus from the perspective of literary warrant, proposing possibilities for enrichment through cultural warrant. Using a theoretical framework focused on the organization of knowledge, it examines the structure of the thesaurus, its terminological categories and the bibliographic corpus on which it is based. Although the instrument shows methodological coherence and a consistent empirical basis, there are limitations in terms of the transparency of editorial processes and the lack of participatory strategies with the cultural communities involved. It is therefore proposed to adopt cultural assurance as a complementary approach, in order to broaden the semantic depth and representativeness of the vocabulary. Such integration can strengthen the relevance and accessibility of the thesaurus, especially with regard to the interface, inclusion mechanisms and usability for different audiences. The study highlights the potential of articulating literary and cultural collateral in the organization of knowledge in culturally complex domains.

Keywords: Literary Warrant, Cultural Warrant, Thesaurus, Knowledge Organization Systems, Popular Culture.

1. Introduction

This paper is the result of reflections based on Barité et al. (2010) and aims to analyze the Thesaurus of Brazilian Folklore and Popular Culture (Centro Nacional de Folclore e Cultura Popular 2006a), its structure and presentation from the perspective of cultural warrant.

According to Barité (2015), a thesaurus is a knowledge organization system (KOS) made up of terms that have both functional and semantic relationships with each other. The authors also state that the thesaurus is organized «under

* Centro de Educação, Comunicação e Artes; Universidade Estadual de Londrina (PPGCI/UUEL); Londrina; Brazil. larissa.moraes.martins@uel.br. ORCID: 0000-0002-4732-1747; eduardo.carvalho98@uel.br. ORCID: 0009-0005-7467-0474; albuani@uel.br. ORCID: 0000-0003-3506-0479.

** Facultad de Información y Comunicación; Universidad de la Republica (UDELAR); Montevideo; Uruguay. mario.barite@fic.edu.uy. ORCID: 0000-0002-2992-6582.

strong terminological control, so as to constitute an ideal tool for storing and retrieving information in specialized areas» (Barité 2015, 156).

The thesaurus of Brazilian folklore and popular culture began to be conceived in the 1980s by the National Centre for Folklore and Popular Culture (CNFCP) together with the Edison Carneiro Folklore Museum and the Amadeu Amaral Library, both sectors of the institution. However, due to various issues, it only became a reality in 2004, with UNESCO sponsorship. It is worth highlighting the importance of a thesaurus linked to the theme, since folklore and popular culture are fundamental bases of Brazilian cultural heritage.

Throughout the introduction or presentation of the system, the objectives of the thesaurus are not presented, however, they are mentioned in the CNFCP activity report when presenting the second version of the thesaurus as part of the «project Preserving Traditions: Multimedia Thesaurus of Brazilian Popular Culture, which aims to encourage research, documentation and dissemination of sound, visual and textual collections on the country's popular culture. The project is sponsored by Caixa Econômica Federal, through the Caixa Program for the Adoption of Cultural Entities» (Centro Nacional de Folclore e Cultura Popular 2006b, 50).

The thesaurus has two forms of structure, always presenting the term with an explanatory note as to the relationships. In one, the terms are presented in alphabetical order regardless of classification and in the other, it has a systematic structure in which the terms are organized hierarchically based on their relationships in 15 classes: Food; Artifact; Association; Productive activity; Rural activity; Craft construction; Popular expression; Individual; Popular language; Popular literature; Place; Raw material; Medicine; Belief system; and Time. The last seven classes were included in the second version of the thesaurus.

For each of the classes there is also an explanatory note, as well as their specific terms and/or associated terms. The first class, food, according to its note refers to a product derived from «plants and/or animals and/or minerals, whether fresh or processed. These substances, when ingested by an individual, should nourish and satisfy hunger» (Centro Nacional de Folclore e Cultura Popular 2006a). The second class, called artifact, notes that it is an «individual form of material culture» (Centro Nacional de Folclore e Cultura Popular 2006a). The third class presented by the thesaurus, called association, refers to the union of individuals due to a specific objective, whereby «they remain cohesive thanks to procedures, routines and sanctions that they accept and approve of in a conscious and rational manner» (Centro Nacional de Folclore e Cultura Popular 2006a).

As for productive activity, the fourth class in the thesaurus, it deals with the production of goods to meet society's needs, through economic, social or cultural activities. Despite being related to it, the craft construction class has

its own specificities, since it deals with the construction of goods using either a craft technique or materials available in the environment (Centro Nacional de Folclore e Cultura Popular 2006a).

Another class that has some specificities is rural activity, which deals with the

«process that demarcates, in the social rhythm, a special moment of a sacred and/or profane nature, in which the temporary replacement of certain symbolic and/or behavioral paradigms by other paradigms whose function is to reinforce the social structure and cultural standards. It is characterized by a specific space-time configuration, by the use of objects, systems of languages and behaviors, and by emblematic signs, marked by tradition, whose meaning is common to the group» (Centro Nacional de Folclore e Cultura Popular 2006a).

In relation to popular Expression, this refers to a locution or phrase specific to certain social or cultural groups.

As for the class of individual, this refers not only to the human species, but to specimens of species, «organic or inorganic, which constitute a distinct unit» (Centro Nacional de Folclore e Cultura Popular 2006a). The raw material class is about the «main substance that is used in the manufacture of something» (Centro Nacional de Folclore e Cultura Popular 2006a).

As for the class of medicine, it refers to a single substance or set of substances intended to treat both physical and spiritual ailments. In relation to the belief system, this is described as a set of practices, representations and values that encompass the «cosmological, philosophical, historical, sociological and psychological dimensions in their approach to the essential questions of human life» (Centro Nacional de Folclore e Cultura Popular 2006a). This set is organized in such a way as to divide the world into the sacred and the profane.

Two other classes that are very close are popular language and popular literature. The former is described as a process by which humans express their mental states, used by different groups within a linguistic community, «has a dynamic and non-normative character, which differentiates it from formal language» (Centro Nacional de Folclore e Cultura Popular 2006a) and popular literature, which encompasses more than one literary genre, such as tales, short stories, novels, and others, is defined as a «collection of narratives in prose and/or verse, of a traditional or non-traditional nature, transmitted orally» (Centro Nacional de Folclore e Cultura Popular 2006a).

As for the class of place, this is presented as the «physical and/or symbolic space, to which identity, relational and historical characteristics are attributed; for carrying out or practicing various activities» (Centro Nacional de Folclore e Cultura Popular 2006a). And finally, the class of time which «comprises milestones and cycles of social life, expressed by belief systems and social practices (religiosity, work, leisure, etc.)» (Centro Nacional de Folclore e Cultura Popular 2006a).

It was not possible to see the total number of terms, or the number by class, which could be interesting for different studies or for users of the system. Then, we discuss the warrants, with special attention to the literary warrant, identified as being used in this instrument.

2. On the use of the Literary Warrant

The Thesaurus in question does not explain in its presentation and introduction texts (either in the first or second version) which warrant was used in the composition of the instrument. Lourenço (2022), when presenting the UFRGS Thesaurus of Brazilian Folklore (an instrument based, among other sources, on the thesaurus explored in this paper), explains that the principle of literary warrant was used specifically, basing the work on the search for terms in articles, books, dictionaries and documents, in a deductive approach (creating categories before searching for the terms). It is understood that the Thesaurus of Folklore and Brazilian Popular Culture follows a similar path, since it establishes 283 bibliographic sources in its most recent version (Centro Nacional de Folclore e Cultura Popular 2006a). It is important to pay attention to the definitions and characteristics of a literary warrant in order to correctly approach the Thesaurus of Brazilian Folklore and Popular Culture as an example of an instrument based on a literary warrant.

Beghtol (1995) explains that literary assurance is concerned with describing the content of certain documents. Barité (2018) recalls that warrants are authorities that the subject involved in the activity of organizing knowledge will invoke to justify and verify decisions, whether they are explicit about making a warrant. Even if sometimes the organizers of an instrument or system do not make clear the choices made and the authorities that guide their work, the presence of some kind of warrant is what differentiates a Knowledge Organization System from an arbitrary proposal (Barité 2018, 530). Along with Beghtol's (1995) argument above, Barité (2018) recalls that Hulme's proposal in 1911 involved an empirical approach to the study of literature, proposing a distinction in classification standards from other forms of work. Hulme disagreed with the use of the authority of philosophers and thinkers or scientific consensus as simple justifications for the establishment of a classification system with its subjective elements, proposing the transfer of authority in the establishment of a system to knowledge itself in its recorded and socialized materiality (Barité 2018).

The basic foundation of literary assurance must therefore be the literature linked to the domain in question, both for validation and for the extraction of terminology, moving from the state-of-the-art to a structured conceptual reconfiguration (Barité et al. 2010). In this sense, the principle of literary warrant goes against «an aprioristic organization of knowledge or one based

on formal aspects of classification theory» (Barité et al. 2010, 125), with the documentation on which it is based being the source of validation itself, since the object of classification has its substance in this material (Barité et al. 2010). This perspective reinforces the empirical dimension of the activity of organizing knowledge based on literary warrants in relation to other warrants that can be anchored, for example, in ethical or philosophical principles.

The next topic reflects on the potential enrichment of the type of instrument analyzed here, based on the use of mechanisms linked to cultural warrants.

3. Analysis of the thesaurus and possibilities for enrichment based on the Cultural Warrant proposal

As explained above, in the presentation and introduction texts of the thesaurus analyzed there is no stated mention of a specific objective, and it can be inferred from the CNFCPs activity report that this thesaurus aims to contribute to the research, documentation and dissemination of collections on the country's popular culture.

The terms are organized alphabetically and hierarchically, based on a relationship with some fundamental categories. In the first version published in 2004, there were seven classes: Food, Artifact, Association, Productive activity, Ritual activity, Individual and Raw material. In the latest version available (Centro Nacional de Folclore e Cultura Popular 2006a), the categories Oral Literature, Medicine, Artisanal Construction, Belief System, Place and Time were included, as well as expanding the categories of Ritual Activity and Productive Activity with the inclusion of the subcategories Musical Activity, Narrative Activity, Religious Practice and Popular Pharmacopoeia.

It is interesting to note the inclusion of new categories and the expansion of some that have shown the need for expansion to include the organized terms. Barité et al. (2010) suggests that in addition to using literary assurance to formulate knowledge organization systems, it is possible to use it to validate or re-observe them, and it is common for those responsible for these systems, when needing to prepare new editions, to make decisions regarding terms that need to be excluded or incorporated. The process and result of this work are summarized in Figure 1.

Thesaurus of Cultural Heritage Objects in Brazilian Museums	
Structure	Attributes
General plan: • 16 categories; Ex.: Personal use objects; • 77 subcategories Ex.: Clothing. Systematic part: Development of the general plan; Hierarchical relations of genus/species of all preferred terms are evidenced at greater levels of depth. Alphabetical part: Structured alphabetical index, with the complete list of preferred and non-preferred terms, in which each term is accompanied by its logical and semantic relations with other terms.	Monolingual - Portuguese (BR); 4,558 terms; 3,412 preferred terms; 1,146 non-preferred terms.
	Methodology • Collection of terms in thesauri, dictionaries, glossaries and encyclopedias, specialized or not, published in Brazil and abroad; • Inclusion of terms directly requested by Brazilian museums; • Inclusions throughout the construction of categories during consultations in literature and museums with specific collections such as the Delgado de Carvalho Museum and the Fundação do Teatro Municipal do Rio de Janeiro.

Figure 1. Characterization. Source: Adapted from Ferrez (2016).

From the perspective of literary assurance, the thesaurus organizing team’s attention to reformulating categories and incorporating new ones is something positive. Therefore, to enable a better understanding of the discussion and presentation of the thesaurus in question, we provide below a representation of its categories and subcategories in Figure 2.

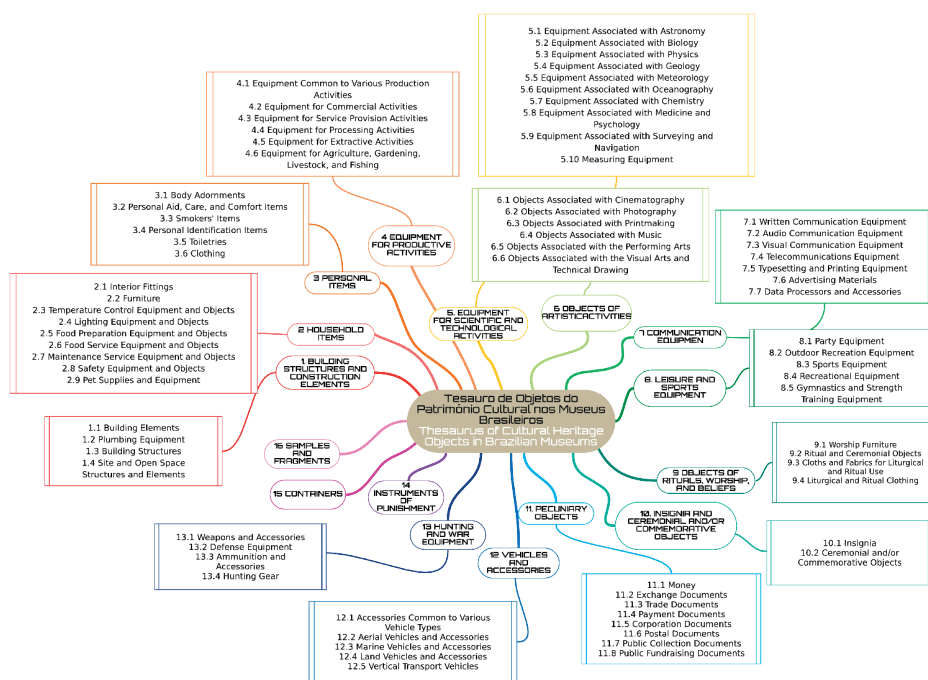


Figure 2. Categories and Subcategories. Source: Adapted from Ferrez (2016).

As explained by Beghtol (1995) and Barité (2010; 2018), Hulme understood that what is fundamental in the use of literary warrant is the constant reference to the empirical concreteness of the observed material, in this case, materials linked mainly to academic publications in folklore and popular culture studies. However, the lack of transparency in relation to the processes of alteration in the thesaurus, as well as the lack of explanation of the materials that solidified the changes, is a weak point in the material.

Likewise, the process of inserting or removing terms is not made explicit by the organizers of the device. While the first version had 2092 terms divided into seven categories, it is not known exactly how many new terms became part of the updated version, nor is it known how the subordination of certain terms changed with the inclusion of new categories and subcategories. In view of the 283 works and documents referred to in the thesaurus as sources consulted, it would be interesting to observe the determinants that suggested the changes. A brief comparison of the structure of the two versions is represented in Figure 3.

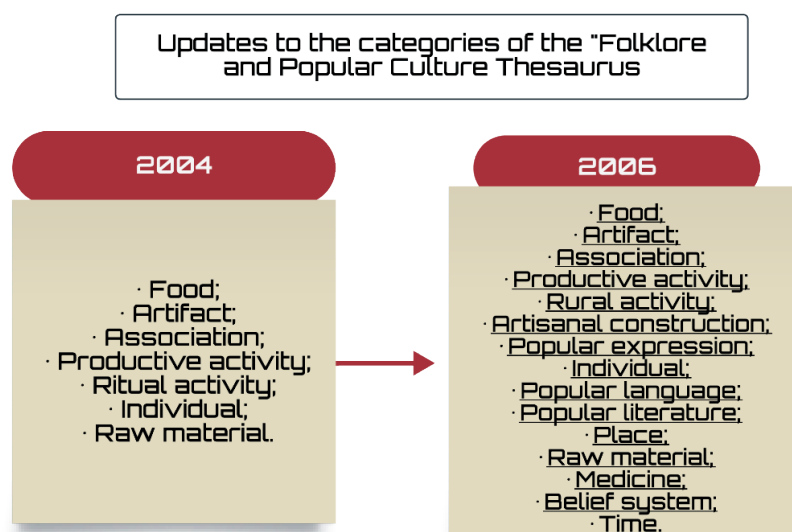


Figure 3. Comparison first and second version. Source: Adapted from Ferrez (2016).

The extent of the bibliography on which the thesaurus is based is remarkable, with almost three hundred academic materials and the presence of government documents, consisting of a considerable corpus of literary collateral, despite not declaring the use of this collateral at any time.

For educational purposes, we present below two terms as examples of how they appear in the thesaurus analyzed here, in order to better illustrate their relationships and description within the thesaurus. Along with these terms, we provide examples of how these same terms appear in the *Thesaurus des Objects Mobiliers* (Verdier and Magnien 2001) organized by H  l  ne Verdier.

The *Thesaurus des Objects Mobiliers* is based on the Palissy Base and the collection *Principes d'analyse scientifique*. Although it is not aimed at museums and their collections, it still has some similarities to the thesaurus analyzed here, so this relationship can allow for a better understanding of the organization in Ferrez (2016). It is important to emphasize that this is not a comparison, as this would be inappropriate.

In Table 1 we present the terms "accordion" and chandelier and how they appear in each of the thesauri, the differences and similarities in the relationships, organization of the terms and even in the order in which the information is presented.

Field	Thesaurus of Cultural Heritage Objects in Brazilian Museums	<i>Thesaurus des Objects Furniture</i>
1st	Term: Accordion	Term and notation: Accordéon 14.04.02.01.
2nd	NA: A wind instrument whose sound is produced by the actuation of bellows, which produce a current of air that passes through a set of reeds inside the instrument's body. Its structure is usually made of wood, with a box and a keyboard for the high notes, connected by a pleated bellows to another box with buttons that activate the low notes.	TG: free reed aerophone
3rd	UP: Accordion Accordion Concertina Bagpipes Harmonica Accordion	NE: Portable keyboard instrument consisting of a set of free metal reeds controlled by keys or buttons. The air is actuated by bellows.
4th	TG: Aerophone	NA: Created in Austria in 1829, based on the principles of the harmonica. The accordion can be diatonic or chromatic, and this information should be indicated in the PDEN field, if possible.
5th	SCA: 0604 Objects Associated with Music	Bibliographic reference
1st	Term: Chandelier	Term and notation: Chandelier 08.10.03.
2nd	NA: Light fixture made of metal, glass, wood, etc., not necessarily connected to the electrical system, attached to the ceiling and consisting, in principle, of a central column and many lights distributed in arms.	TG: lamp
3rd	TG: Lighting objects and equipment	VA: pendant (pendant light) synagogue chandelier church chandelier crown of lights chandelier pendants (crystals)
4th	TA: Ceiling light	NE: A light fixture suspended from the ceiling and composed of multiple light sources, which may be shaped like a ring (crown of lights) or a wheel, with arms (spokes). It can be powered by candles, oil, kerosene, gas, or electricity.
5th	SCA: 0204 Lighting Objects and Equipment	Bibliographic reference

Table 1. Organization of terms in thesauri. Source: Adapted from Verdier and Magnien (2001) and Ferrez (2016).

So, even when dealing with the same object using the same type of guarantee, differences are present even in the order in which the elements are presented. The terms were chosen randomly, but the organization follows the same pattern in both, including or removing elements according to the term's needs.

Interesting differences worth noting are the explanatory and application notes, which in the Brazilian thesaurus were constructed based on bibliographic references, but without specifying which. While the French thesaurus specifies in its introduction that both the explanatory and application notes come from the Palissy database and are accompanied by a bibliographic reference specific to that object. They are not grouped by categories at the end of the whole as in the Brazilian case. These are merely highlighted curiosities without establishing a judgment of better or worse between them.

Resuming our analysis after the example of the presentation of thesauri, Ferrez (2016) in addition to the thesaurus in pdf format, the recent version of the thesaurus is operated on a World Wide Web page, allowing access to the items via hyperlinks, making both the systematic and alphabetical presentation very fluid - the interface is clean and the direction to a term or category is immediate, depending on just one click. There are interesting mechanisms for linking images and bibliographic materials linked to the term that can be observed, to enable the user to access the thesaurus.

However, there were some limitations imposed by the technologies available at the time of the latest version.

Some examples of these limitations include a lack of features for international users, such as translation tools, as well as a lack of accessibility for people with disabilities (PWD). This includes missing tools like high-contrast viewing, the ability to enlarge and/or change the font, screen reader functionality, and other resources that could improve the thesaurus's accessibility and usability.

Another important point is related to the system's flexibility, its compatibility with different systems and browsers, since there is no thesaurus format for smartphones, only the desktop version, which can hinder access by certain audiences and comfortable use of the system (individuals with limited vision, the elderly, among others).

There is also a specific concern about the use of warrants in this context. Among the different perspectives on culture, it can be said that it is the totality of relationships established between a group, structured in hierarchical sets, in the distinctive characters manifested in the behavior of the individuals of this group in relation to others (Laplantine 2003). Also, according to Laplantine (2003), the experience of contact with other cultural forms, an expression of otherness, leads the observer of other cultural forms to see what their own references allow them to observe and understand. In terms close to our problems,

«culture may influence what data is represented and what is not in any given knowledge organization system» (Lee 2021, 6).

This question is relevant to any topic to be discussed, but in the context of Brazilian folklore, which is fed by a series of structures from different cultural matrices, such as the various indigenous, African and Iberian influences (Brandão 1982), it is a discussion that is constantly in evidence. The choice to use literary collateral is an interesting one, as it allows for the approximation of qualified productions that are crossed by academic reflection for the composition of the thesaurus, which is of great value. However, there are certain limitations that are inherent to the empirical proposal.

The empiricism can never be used in a pure form, because the way we select sources for our empirical studies and the way we see the world is always influenced by our subjectivity and cultural background. Therefore, our position today is to recognize the empirical, rational, historical and pragmatic/political elements that underpin our SOC. Because the historical and pragmatic/critical elements have been under-represented in the history of OC, the need is in particular for their integration. Hulme's vision is important (and in line with the growing trend towards bibliometric studies), but it should be recognized that there will always be a tension between different epistemological positions and that a pure empiricist position is an illusion (Barité 2018, 531).

It is also understood that it would be possible to approach the use of cultural collateral as complementary collateral in order to further enrich the thesaurus. Barité et al. (2010) reports on the proposal made by various authors to improve existing classification systems or controlled vocabularies by using other warrants (such as cultural) alongside literary ones.

By way of example, some of the works included in the sources used by the thesaurus are classic texts by renowned "interpreters of Brazil", such as Gilberto Freyre and Luís da Câmara Cascudo, basic authors for social and cultural theory in Brazil, with particular interpretations and their own image of our reality (Albuquerque Júnior 2015; Couto 2015; Novais and Arruda 1998). As Lee (2021) explains, cultural assurance is a conscious and deliberate movement with reference to culture carried out by classification and information professionals to build their systems. Considering the different practical possibilities of cultural assurance (Lee 2021), it might be possible to establish dialogues with communities associated with the terms in the thesaurus in order to deepen the work, with a view to enriching the relationships between terms, the presence of synonyms or possible hierarchical restructuring, for example. This is, however, a very complex job that would require even more literary depth (including the dimension of literary warrant) and the development of specific and functional methodologies for working groups to operate in conjunction with communities that share the terms present in folklore (which is also multicultural, involving different communities and perspectives that

do not necessarily coincide). One could think of mechanisms for bringing together communities that experience the elements of folklore and popular culture dealt with by the instrument, the creation of commissions including members of communities, or carrying out targeted anthropological work, for example.

Taking two concrete examples from the observed thesaurus, we can mention the terms “Tapiri” and “Tucupi”. The former is presented as «A place used as a temporary shelter or storage for raw materials; it consists of a small hut with a round wooden frame and a thatched roof. It does not require a wall»¹ (Centro Nacional de Folclore e Cultura Popular 2006c). Its generic term constitutes “Place”, and its associated term “Artisan work”. In this case, the use of cultural guarantee could provide an understanding with a broader meaning; note the understanding that “Tapiri” gains for the culture of Tupi-speaking peoples. Meaning in Tupi language: «“Tapiri” is an indigenous word in the Tupi language that is used by various traditional peoples in the North region of Brazil and means, among other things, a place where people gather to share ideas, food, rest, and experiences» (Associação Brasileira de Agroecologia 2023).

In this case, “Tapiri” leaps from a strict sense, associated solely with a place, to the cultural enrichment that connects Tapiri to a shared environment, where people live, converse, and share experiences. Similarly, we can mention the case of “Tucupi”, which is presented as «processed food, extracted from the juice of peeled cassava and squeezed into a tipiti»² (Centro Nacional de Folclore e Cultura Popular 2006d) a definition that could be enriched through cultural guarantee, adding elements of its preparation and the full range of possible dishes derived from its culinary base – whether in its definition or in the construction of associated terms and expansion of the thesaurus.

Both examples could be profoundly enriched if accompanied by mechanisms based on cultural guarantee, as in the aforementioned hypotheses of building a committee of community members who use the terms that appear in the thesaurus – where these participants would add meaning and associations to the thesaurus terms.

4. Conclusions

The Thesaurus of Brazilian Folklore and Popular Culture is a significant initiative in the field of knowledge organization, especially considering the complexity and diversity inherent in popular knowledge. The structured on the basis of literary warrant, the instrument is based on a robust bibliographic

¹ Original definition in Portuguese [N.d.A].

² Original definition in Portuguese [N.d.A].

corpus, made up mostly of academic works, technical documents and productions by recognized interpreters of Brazilian culture. This methodological choice, although not formally explained in the texts presenting the thesaurus, denotes adherence to an empirical perspective based on textual materiality. This foundation warrants, on the one hand, the legitimacy of the terminological selection and, on the other, shows the concern with fidelity to the documentation on the domain covered.

The organization of the thesaurus into systematic and alphabetical categories, as well as the expansion of classes between the first and second versions, point to the attention paid by the team responsible to the evolution of folklore studies and the need to accommodate new terms and cultural expressions. The inclusion of subcategories and the reformulation of some classes show that the instrument's management team has worked to update it in line with the proposed thematic scope. However, this evolution is not accompanied by transparent records of the criteria and decision-making processes used, which limits the traceability of changes and, consequently, the full reliability of the instrument for certain scientific and professional uses. The lack of data on the current number of terms and their distribution by category also compromises quantitative or comparative analyses between versions.

Furthermore, although the use of literary warrant is a positive differential in anchoring terminological selection in established documents, the opening of explicit relations with procedures linked to cultural warrants with the approximation of the social groups that produce and experience the cultural expressions represented by the thesaurus would be very enriching. In this sense, the proposal to complement this warrant with a cultural warrant is promising. As authors such as Lee (2021) and Barité et al. (2010) point out, the adoption of other warrants – especially those that consider the cultural and community dimension – can broaden representativeness, semantically enrich the vocabulary and strengthen its social function. Incorporating the voices of the communities that hold popular knowledge, through ethnographic listening, collaborative working groups or consultative committees, could contribute to the instrument's greater adherence to the cultural reality it seeks to represent.

Finally, the technological environment that supports the thesaurus could be improved to better meet the principles of accessibility and usability. The lack of mechanisms aimed at foreign users, people with disabilities or with different levels of digital literacy limits the system's reach and effectiveness. Improvements in this regard – such as the inclusion of translations, contrast tools and automated reading, as well as versions adapted for mobile devices – would be consistent with the thesaurus' purpose of promoting the documentation and dissemination of Brazilian popular culture in its multiple manifestations. Thus, while recognizing the merits of the proposal based on literary assurance, it reinforces the need for an expanded and integrated approach, which

also includes the foundations and practices of cultural assurance, favoring the creation of more inclusive, dynamic and representative systems for organizing knowledge.

References

- Albuquerque Júnior, Durval Muniz de. 2015. "Consultando o Cascudo: gêneros textuais, escrita de si e interpretação do Brasil no Dicionário do Folclore Brasileiro de Luís da Câmara Cascudo." *Revista Escrita Da História* 2(3): 14–30. <https://www.escritadahistoria.com/2023/12/consultando-o-cascudo-generos-textuais.html>.
- Associação Brasileira de Agroecologia. 2023. "TAPIRIS de Saberes: tire dúvidas sobre as apresentações de trabalhos no 12º CBA." Congresso Brasileiro de Agroecologia. Accessed November 17, 2025. <https://cba.aba-agroecologia.org.br/tapis-de-saberes/>.
- Barité, Mario, Juan Carlos Fernández-Molina, José Augusto Chaves Guimarães, and João Batista Ernesto De Moraes. 2010. "Garantia literária: elementos para uma revisão crítica após um século." *Transinformação* 22 (2): 123–38. <https://www.scielo.br/j/tinf/a/prtXbCcMkLD48hdnHR4tcfS/?format=pdf&lang=pt>.
- Barité, Mario. 2015. *Diccionario de organización del conocimiento: Clasificación, Indización, Terminología*. 6. ed. Comisión Sectorial de Investigación Científica (CSIC) de la Universidad de la República.
- Barité, Mario. 2018. "Literary Warrant." *Knowledge Organization* 45 (6): 517–36. <https://doi.org/10.5771/0943-7444-2018-6-517>.
- Beghtol, Claire. 1995. "Domain analysis, literary warrant, and consensus: The case of fiction studies." *Journal of the American Society for Information Science* 46 (1): 30–44. [https://doi.org/10.1002/\(SICI\)1097-4571\(199501\)46:1%3C30::AID-ASI4%3E3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-4571(199501)46:1%3C30::AID-ASI4%3E3.0.CO;2-F).
- Brandão, Carlos Rodrigues. 1982. *O que é folclore?*. Brasiliense.
- Centro Nacional de Folclore e Cultura Popular. 2006a. "CNFCP." Tesouro de Folclore e Cultura Popular Brasileira. Accessed November 17, 2025. <http://www.cnfcp.gov.br/tesouro/apresentacao.html>.
- Centro Nacional de Folclore e Cultura Popular. 2006b. "Relatório de Atividades." Accessed November 17, 2025. http://www.cnfcp.gov.br/pdf/relatorio_atividades_2006.pdf.
- Centro Nacional de Folclore e Cultura Popular. 2006c. "Tapiri." Tesouro de Folclore e Cultura Popular Brasileira. Accessed November 17, 2025. <https://antigo.cnfcp.gov.br/tesouro/00000974.htm>.

- Centro Nacional de Folclore e Cultura Popular. 2006d. "Tucupi." *Tesouro de Folclore e Cultura Popular Brasileira*. Accessed November 17, 2025. <https://antigo.cnfcp.gov.br/tesouro/00001472.htm>.
- Couto, Elvis Paulo. 2015. "Imagens do Escravismo Colonial: Gilberto Freyre em contraste com outros intérpretes do Brasil." *Revista Sem Aspas* 4 (1): 19-25. <https://doi.org/10.29373/sas.v4i0.8029>.
- Ferrez, Helena Dodd. 2016. *Tesouro de Objetos do Patrimônio Cultural nos Museus Brasileiros*. Rio de Janeiro: Prefeitura do Rio. <https://drive.google.com/file/d/1RAjy3740ejJAtyaN0FlB8heHZp7dq97a/view>.
- Laplantine, François. 2003. *Aprender Antropologia*. Editora Brasiliense.
- Lee, Wan-Chen. 2021. "Operationalizing Cultural Warrant in Knowledge Organization." PhD diss., University of Washington. <https://digital.lib.washington.edu:443/researchworks/handle/1773/47600>.
- Lourenço, Gustavo Emanuel. 2022. "Folclore Brasileiro." *Tesouro Brasileiro*. Accessed November 17, 2025. https://www.ufrgs.br/tesauros/index.php/thesa/terms/458/?lan_g=pt_br.
- Novais, Fernando Antonio, and Maria Arminda do Nascimento Arruda. 1998. "Apresentação: revisitando intérpretes do Brasil." *Revista USP* (38): 6-9. <https://www.revistas.usp.br/revusp/article/view/28389>.
- Verdier, Hélène, and Aline Magnien. 2001. *Thesaurus des objets mobiliers*. Éditions du patrimoine.

Identification of types of documents and knowledge organization: transparency and access in records management

Bianca Therezinha Carvalho Panisset*, Lucia Maria Velloso de Oliveira**,
Leticia Souza da Costa Sampaio***, Camilla Campoi de Sobral****

Abstract: This study focuses on archival knowledge organization practices applied to the management of electronic records at Fundação Casa de Rui Barbosa (FCRB), a Brazilian public institution dedicated to preservation, research, and education. This research presents results from the project “Electronic document management in FCRB”, developed under the Program to Incentive the Production of Technical and Scientific Knowledge in the Area of Culture. The methodology involved document analysis, literature review, and the survey of active types of records in the Electronic Information System (SEI). As an illustrative case, six types of archival records related to scientific research development activities were selected. The results include the identification of the producing activity and the proposal of concepts for 237 types of records. The study concludes that the adopted approach contributes to record standardization, transparency, and public information access.

Keywords: Knowledge organization, Records management, Archival records, Typology of records, Fundação Casa de Rui Barbosa.

1. Introduction

In Brazil, the use of electronic systems for conducting administrative processes within the federal public administration began in 2015. Since then, a government software called the Electronic Information System (SEI), which was originally developed for use in the Judicial System, and afterwards its use was expanded to the Federal Executive Branch. The implementation of this system has led to the digitization of administrative procedures in public institutions and has significantly increased the production of digital records,

* Historical and Institutional Archives Service at Fundação Casa de Rui Barbosa. bianca.panisset@rb.gov.br. ORCID: 0000-0002-5568-5225.

** Memory and Information Center at Fundação Casa de Rui Barbosa. luciemarie@gmail.com. ORCID: 0000-0003-0546-2602.

*** Fundação Casa de Rui Barbosa. souzaleticia@id.uff.br. ORCID: 0000-0001-6385-5012.

**** Universidade Federal Fluminense (UFF). camillacampoi@id.uff.br. ORCID: 0000-0002-6725-2929.

prompting reflections on how to organize, describe, and retrieve these records in an efficient, transparent, and accessible way. In this context, the articulation between Knowledge Organization (KO) and archival records management practices becomes strategic for structuring information flows and strengthening public transparency. This study aims KO analyze how the identification and standardization of types of archival records, based on theoretical foundations of KO, can contribute to the improvement of born-digital records management, focusing on expanding access to information and promoting user autonomy.

To this end, the article presents results from the project “Electronic Document Management in FCRB”, developed under the Program to Incentive the Production of Technical and Scientific Knowledge in the Area of Culture of Fundação Casa de Rui Barbosa (FCRB), a Brazilian public institution of preservation, research and education. The project proposes a typological analysis of the records produced in SEI, aiming to align archival and classification concepts with the reality of Brazilian public administration. As an illustrative case, the study details the types of archival records related to scientific research development activities at the institution, demonstrating how conceptual structuring and standardization of these records can enhance process traceability, institutional efficiency, and commitment to information access.

2. Knowledge Organization and Archival Records Management Practices

Recent studies align KO with archival practices, such as appraisal, classification, and description of archival records, aiming to promote transparency and access (Barros and Sousa 2019; Oliveira et al. 2020). In this study, we analyze how the typological analysis of records, associated with knowledge organization, strengthens transparency and access to archival records. The research is based on the understanding of archival records as persistent representations of the activities that generate them (Yeo 2008).

The study by Silva and Tognoli (2022) highlights the emergence of a Brazilian discourse community dedicated to the articulation between Archival Science and KO, proposing a significant shift in the recognition of the archival domain as a legitimate and productive part of the KO field. The authors demonstrate that, although KO initially developed under the strong influence of Librarianship, focusing on content-oriented systems, Archival Science – with its emphasis on the context of production and the principles of provenance and original order – offers its own theoretical and methodological foundations that enrich this interdisciplinary field.

Their analysis of International Society of Knowledge Organization (ISKO) events points to the consolidation of Brazilian research that has contributed to this theoretical and practical convergence, especially regarding archival classification and description as knowledge organization processes. This perspective broadens the possibilities for integrating archival systems with representation models, providing a foundation for more contextualized, interoperable, and sensitive practices aligned with the specificities of archival records (Silva and Tognoli 2022).

Tognoli et al. (2019) define the conceptual structures of archival knowledge based on three elements: the concept of fonds, which refers to the grouping of records by shared provenance; the documentary form, which encompasses the intrinsic and extrinsic elements of the records; and the context of production, which considers the activities and functions that originated the records. Understanding these conceptual structures contributes to the study of records typology, since this analysis is carried out through the examination of the producer and the production context (Oliveira 2009).

According to Hjørland (2022), a KO should be understood as an activity deeply rooted in epistemological, historical, and social contexts, being inseparable from the theoretical perspectives that underpin it. The author argues that every organization system carries with it assumptions about what is relevant, true, or valid within a given domain (Hjørland 2022).

This line of thought reinforces the idea that KO processes are not neutral, but rather constructions that reflect scientific paradigms and social interests. In the context of the FCRB project, this approach helps justify typological analysis as a practice that must consider not only the administrative functionality of the records but also the assumptions guiding their representation and retrieval.

Gnoli (2008) proposes a conceptual expansion of KO by suggesting that classification systems should incorporate ontological, epistemological, cultural, and temporal dimensions. This approach goes beyond the traditional disciplinary framework and favors the representation of phenomena in their complexity, allowing greater flexibility in information retrieval. By considering that the content of records results from multiple factors and perspectives, it is also possible to reflect on the contributions of this perspective to the management of digital records, since understanding these dimensions supports a more dynamic, plural, and contemporary methodology for the identification and classification aligned with the demands of public administration.

KO, when structuring systems for information representation and retrieval, often adopts philosophical assumptions about the nature of meaning and language. Astaneh (2023) draws attention to the predominant representationalist character in many theoretical models, in which meaning is conceived as a direct correspondence between mental concepts and elements of reality.

This approach, although still present in traditional archival practices, has been criticized for ignoring the dynamic role of context, communicative intentions, and social practices in the production and interpretation of records (Astaneh 2023).

In the field of records management, this approach implies considering classification and description as constructions influenced by multiple semantic, technical, and institutional layers. Thus, the integration of KO with archival practices demands continuous reflection on the adopted representation models, to encompass not only the logical structure of the records but also the pragmatic conditions of their production, use, and retrieval.

The integration between archival practices and knowledge organization models has benefited from the development of standards that promote semantic interoperability and the reuse of classificatory structures. As an example, the Simple Knowledge Organization System (SKOS), as discussed by Smith (2022), presents itself as a flexible and accessible solution for representing controlled vocabularies, allowing archival concepts to be formalized as entities endowed with unique identifiers and relational properties.

By adopting a logic based on hierarchical and simple associative relationships, SKOS enables the description of record types with multiple linguistic and contextual labels, which broadens its potential use in digital records management environments (Smith 2022). This approach can be relevant in the context of institutions dealing with born-digital collections and seeking to implement standardized information representation systems, promoting access, preservation, and transparency through structures that respect both the complexity of archival arrangements and the demands of contemporary technological interoperability.

The understanding of information as a construction that depends on the mediation between concepts, language, and social context is essential for the development of effective knowledge organization systems applied to records management. Fugmann (2022) argues that information cannot be reduced to a set of mechanically transmitted data, as it acquires meaning only when considered relevant by a situated receiver.

This perspective directly relates to the reality of archives, where the representation of records requires identifying their formal properties, their records-producing activity, and the relationship of this activity to the functions performed by their creator, while also considering the multiple interpretations and uses these records may assume over time. Indeed, by recognizing that different social groups associate diverse meanings to the same informational object, archival practice is challenged to adopt more flexible and contextually sensitive approaches in the naming and conceptualization of record types.

The discussions outlined throughout this section allow us to reflect that the intersection between KO and records management within archival practices

demands an approach that is both conceptually critical and sensitive to the multiple dimensions of information. These theoretical contributions collectively guide the proposal presented in this study, which explores an applied experience of identifying and standardizing record types within the FCRB. By bridging theory and practice, the project seeks to contribute to the promotion of transparency, access, and intelligibility of the institution's administrative and scientific processes.

3. Methodology

This study is characterized as exploratory research, based on qualitative methods of data analysis and processing. The methodological strategy adopted was bibliographic research, with a literature review of articles and books, aiming to theoretically support the discussions presented.

Within the scope of the project developed at FCRB, the record types active in the SEI were identified, aiming to enhance control over their creation, processing, and use. This approach aligns with Geoffrey Yeo's (2008) understanding, which conceives records as persistent representations of the producing activities.

Initially, a preliminary study about institutional functions and goals was conducted on the producing agency, as well as an analysis of technical documents as the classification scheme, the retention schedule, the Vocabulary Control rules and an analysis of the record types available in the SEI were done. After surveying the record types produced and received by the system, bibliographic research was carried out using reference papers, legislation, scientific articles, government manuals, and institutional materials, with the purpose of supporting the identification of the respective producing activities and the proposal of concepts for each record type. For this, specialized dictionaries and glossaries related to the analyzed records were used, complemented by searches in institutional portals and glossaries prepared by public entities with recognized authority, such as public archives, courts, and institutions linked to public administration.

4. Results and discussion

This text presents the results of the project "Electronic Document Management in FCRB", part of the Program to Incentive the Production of Technical and Scientific Knowledge in the Area of Culture of FCRB, a Brazilian public institution of Preservation, Research and Education (PIPC). As said before, the institution is dedicated to preservation, research, and education, affiliated with the Ministry of Culture. The project seeks to improve the archival ma-

nagement of the institution's born-digital records through the study of record types produced in the SEI, a government software adopted in Brazilian public administration for the creation, processing and maintenance of administrative procedures.

Since November 2023, a preliminary study was conducted on the 784 active records in the SEI, considering FCRB's classification scheme and retention schedule. To identify the records' producing activities and develop concept proposals, bibliographic and documental research was carried out, using specialized glossaries, dictionaries, institutional materials, and Brazilian legislation as primary sources.

By July 2025, the project's results include the development of 237 concept proposals associated with the producing activities of the records, in addition to 87 suggestions for reduction (exclusion or merging) of redundant record types, types incompatible with the producing activity, or those whose legislation justifies their deactivation. The project's development aims to broaden access to information, as well as to promote greater knowledge about the institution and its activities.

By defining the activity, the type, and establishing its concept, it is possible to standardize its production and use, improving communication with users and their autonomy, thereby increasing administrative transparency and access. It is important to emphasize that incorrect selection of record types hinders information retrieval and results in lack of access. The standardization of record types by public administration has been a recent demand aimed at enhancing the transparency of public actions.

As an example, the document types identified during the research within the scope of FCRB research programs are presented here. The types in question represent the distinct phases of implementation, monitoring, and development of research grants.

The research activities at FCRB occur through programs such as the Program to Incentive the Production of Technical and Scientific Knowledge in the Area of Culture which aims to educate, train and qualify human resources in various areas of research. Its activities encompass institutional, scientific and technological development in different areas of knowledge such as outreach activities, preservation and technical processing of materials from the institution's museum, archives, libraries and architectural collections. Each project and scholarship hiring generates an administrative process.

Among the project's results, the development of concept proposals for record types linked to the scientific research development activities at FCRB stands out. These record types reflect the different phases of the research grant awarding process and representationally express the archival records that materialize the institution's administrative and academic actions. Based on the identification of the producing activity and the construction

of specific concepts, the aim was to highlight the institutional functions associated with each type, in order to promote their standardization, recover their production logic, and ensure greater transparency and traceability of the processes. The table below presents six representative record types of this workflow, along with their respective producing activities and concept proposals, illustrating the exemplary case of the research.

The six record types presented in Table 1 are persistent representations of activities carried out within the phases of entry, management and monitoring, and completion of research scholarships. These are: *FCRB Scholarship Application Form*, *Letter of Interest to Participate in Research*, *Certificate of Activity Performance*, *Partial Research Report*, *Research Scholarship Performance Review* and *Final Research Report*.

Record Type	Producing Activity	Concept Proposal
Letter of Interest to Participate in Research	Statement of justification by a researcher or student expressing interest in participating in a research project with or without a scholarship.	Archival record that presents the justification of interest from a researcher or student to join a research project with or without a scholarship. It should include the candidate's name, project identification, explanation of reasons, and academic and/or professional experience in the area.
Certificate of fulfilment of research obligations	Certification of fulfilment of research obligations and activities carried out by a scholarship holder at the Foundation.	Archival record that certifies the fulfilment of research activities by a scholarship holder, required for scholarship payment. It must include the project name, type of scholarship, and the signature of the advisor or legal substitute.
FCRB Scholarship Application Form	Registration of a candidate's application in the call for scholarship selection for the Foundation's research programs.	Archival record that registers a candidate's application for the FCRB scholarship selection. It must contain personal and academic data, chosen project, and description of supporting documents required by the call. It is completed via the FCRB GOV.BR portal.

Record Type	Producing Activity	Concept Proposal
Research Scholarship Performance Review	Advisor's assessment of the partial or final report and the scholarship holder's performance, submitted to the Advisory Committee.	Archival record in which the advisor evaluates the scholarship holder's report and performance during the scholarship period. It should include scholarship holder's data, project information, period, advisor's assessment, signature, and date.
Final Research Report	Presentation of the scientific research journey, activities, and results achieved at the end of the scholarship.	Archival record prepared by the scholarship holder presenting the research journey, including project title, period, names of the scholarship holder and advisor, structured with introduction, development, and conclusion, methodology, results, productions, events, and final considerations.
Partial Research Report	Presentation of scientific research progress, activities, and partial results every six months.	Archival record prepared by the scholarship holder, mandatory for project continuation.

Table 1. Record types related to scientific research development activities within the research programs at FCRB (Authors' own).

The entry process into the research programs at FCRB begins with the submission of the *FCRB Scholarship Application Form*, which registers candidates' applications to research projects. This form is completed with personal and academic data and is generated through the FCRB GOV.BR portal.

The form (Fig. 1) is a record type generated from the activity of applying to scholarship selection calls. It records the candidate's personal, academic, and supporting data, serving as the initial document in the administrative process for scholarship granting. Attached to this record are various documents requested in the project's call, including the *Letter of Interest to Participate in Research*, where the candidate presents justification and reasons for participating in the research project, detailing their experiences.

07/08/2023, 15:41

Lecom BPM

[FCRB] Inscrição na Seleção de Bolsas - PIPC por Cidadão		Status Em Andamento	Código 000.792
---	--	-------------------------------	--------------------------

Analisar Solicitação

Ciclo: 01

Início da Atividade
04/08/2023

FCRB - Inscrever-se no Programa de Bolsas PIPC

N.º do Protocolo da Solicitação
000261.0000787/2023

Dados do Candidato

Nome do Solicitante

CPF do Solicitante

Documento de Identificação

Órgão Expeditor
DETRAN RJ

E-mail do Solicitante

E-mail (Opcional)

Telefone

Data de Nascimento

Título Acadêmico (Mais Alto Nível)
Mestre em Ciência da Informação

É agente público da ativa?

CEP

Não

20060-080

Logradouro

Número

Complemento

Bairro
Centro

UF
RJ

Município
Rio de Janeiro

Dados Acadêmicos

Link para o Currículo Lattes
<https://buscatextual.cnpq.br/buscatextual/visualizacv.do>

É Bolsista de Agência Governamental?

Não

☒ Existe Vínculo Acadêmico?

Instituição

https://fcrb.servicos.gov.br/bpm/carrega_etapa?action=carrega&sCodProcesso=792&sCodEtapas=7&sCodCiclo=1&sCodForm=20&sCodVersao=1

1/3

Figure 1. Model of the record type *FCRB Scholarship Application Form* (SEI/FCRB 2025).

As described in Table 1, the letter of interest (Fig. 2) is a record type produced as part of the application process. This letter formalizes the candidate’s expression of interest, presenting motivations, affinities with the project, and academic background, and is essential for evaluating the merit of the application.

Eu, [REDACTED], estudante do segundo período do curso de Letras, na universidade PUC do Rio de Janeiro, tenho grande interesse em participar do projeto "A gramatização no Brasil - obras de referência: 1822-1930".

O projeto dialoga com áreas de conhecimento em que pretendo me especializar, principalmente a Linguística e segmentos similares. Acredito que o estudo lexicográfico comparado de um período relativamente extenso possa ser proveitoso para a minha graduação.

Sobre a segunda opção, o projeto "História e Memória da AMLB: 50 anos", oferece a oportunidade de ter contato com o rico acervo da Fundação Casa Rui Barbosa. Representa um trabalho de extrema importância, já que, segundo Jacques Le Goff, historiador francês, uma das bases da História é a constituição da memória. Se, para a língua, cinquenta anos podem não representar mudanças radicais, para a historiografia são outros os parâmetros.

Reafirmo a minha vontade em participar de todas as etapas dos estudos mencionados, e declaro que minha primeira opção seria o projeto "A gramatização no Brasil", visto que gostaria de aprofundar meus conhecimentos da língua.

Figure 2. Model of the record type *Letter of Interest to Participate in Research* (SEI/FCRB 2025).

After the evaluation and implementation process, the management and monitoring of the scholarships is carried out through the preparation of reports by the scholarship researchers, the advisor's evaluation of the project, and the review by the Advisory Committee. According to FCRB Ordinance No. 03/2024, the PIPC Advisory Committee at FCRB is composed of researchers, technologists, and analysts from the FCRB staff. This committee coordinates the internal selection process for projects requiring scholarship holders, assesses demands regarding academic suitability and merit, and organizes the candidate selection process. Additionally, the committee monitors the academic performance of scholarship holders, approves partial and final reports, organizes the presentation of results, and resolves issues not foreseen in contracts.

Activities arising from this stage include the *Partial Research Report*. The document shown in Figure 3 is prepared by the scholarship holder after the first six months of activities, reporting the progress of the research. It constitutes a mandatory step for continuing in the program and expresses institutional control over the project's development.

Based on the analysis of the partial or final report, the advisor prepares the *Research Scholarship Performance Review* (Fig. 4), assessing the scholarship holder's performance. This document formalizes that evaluation and is submitted to the Advisory Committee as part of the scholarship program's governance.

FUNDAÇÃO Casa de Rui Barbosa

ATESTADO DE DESEMPENHO DE ATIVIDADES

PROCESSO Nº 00000.000000/2023-00

Atesto que XXXXXXXXXXXXXXXX (nome do bolsista), bolsista tipo XXXXX do programa XXXXXXXXXXXXXXXX (nome do programa), da Fundação Casa de Rui Barbosa - FCRB, tem desenvolvido regularmente suas atribuições no projeto XXXXXXXXXXXXXXXX (nome do projeto), sob minha orientação.

Solicito o pagamento da respectiva bolsa de pesquisa relativa ao mês de agosto de 2014, e posterior devolução do processo ao XXXXXXXX (setor responsável) para acompanhamento.

Figure 5. Model of the record type *Certificate of Activity Performance* (SEI/FCRB 2025).

At the closing stage of the scholarship period, the scholarship holder must submit the *Final Research Report* (Fig. 6), where they present the entire research journey, activities carried out, and results achieved. With a technical-scientific structure, the document formalizes the results obtained and becomes part of the institutional collection.

Gestão de documentos eletrônicos na Fundação Casa de Rui Barbosa

SUMÁRIO

1. INTRODUÇÃO 3

2. REFERENCIAIS TEÓRICOS E METODOLÓGICOS 4

3. ATIVIDADES DESENVOLVIDAS DURANTE O PERÍODO 6

4. RESULTADOS OBTIDOS 11

5. CONSIDERAÇÕES FINAIS 13

REFERÊNCIAS 15

ANEXO I 18

ANEXO II 42

por

Relatório final apresentado ao Programa de Iniciação Científica da FCRB, relativo às atividades desenvolvidas no período de novembro de 2023 a outubro de 2024, sob a orientação de

Bolsista: FCRB

Processo nº: 01550.000342/2023-01

Rio de Janeiro,
outubro 2024

Figure 6. Model of the record type *Final Research Report* (SEI/FCRB 2025).

The standardization and organization of record types at FCRB promote transparency in the stages of scholarship selection, implementation, monitoring, and completion. This records flow ensures that all phases of the process are clear and traceable, from the initial application to the submission of the final report, facilitating management, transparency, and access to information. Thus, the integration of conceptual knowledge organization frameworks with

archival practices enhances administrative efficiency and strengthens the institutional mission of FCRB.

5. Conclusions

This study highlighted the importance of integrating the foundations of KO with archival practices to enhance the management of born-digital records in public institutions. Through the analysis of record types produced in the SEI at FCRB, it was possible to propose concepts based on the study of the producing activities, thereby contributing to the standardization, transparency, and traceability of institutional records. The experience demonstrated that the employed typological identification methodology expands access to information and strengthens the role of archives in promoting institutional memory and ensuring rights. In addition to presenting a proposal applicable to the reality of public administration, the results discussed here reinforce the relevance of projects that articulate theory and practice, contributing to the consolidation of a critical, contextualized, and interdisciplinary approach to archival knowledge organization.

While this case study is rooted in the specific context of a Brazilian public institution, the core principles and methodological approach are highly transferable to other international contexts. Public administration and cultural heritage institutions globally face similar challenges in managing the deluge of born-digital records. The strategy of analyzing producing activities to standardize record types is a universally applicable framework that can be adapted to different environments to enhance administrative efficiency and promote public trust.

In addition to presenting a proposal applicable to the reality of public administration, the results discussed here reinforce the relevance of projects that articulate theory and practice, contributing to the consolidation of a contextualized and interdisciplinary approach to archival knowledge organization.

References

- Astaneh, Mahboobeh Farashbashi. 2023. "Representationalism in Knowledge Organization and Information Retrieval: Cheer or Criticism?" *Knowledge Organization* 50 (7): 475–88. <https://doi.org/10.5771/0943-7444-2023-7-475>.
- Barros, Thiago Henrique Bragato, e Renato Tarciso Barbosa de Sousa. 2019. "Organização do Concepimento e Arquivologia: abordagens metodológicas." *Informação & Informação* 24 (2): 76–92. <https://doi.org/10.5433/1981-8920.2019v24n2p76>.

- Fugmann, Robert. 2022. "What Is Information? An Information Veteran Looks Back." *Knowledge Organization* 49 (1): 3–5. <https://doi.org/10.5771/0943-7444-2022-1-3>.
- Gnoli, Claudio. 2008. "Ten Long-Term Research Questions in Knowledge Organization." *Knowledge Organization* 35 (2–3): 137–49. <https://doi.org/10.5771/0943-7444-2008-2-3-137>.
- Hjørland, Birger. 2022. "Science, Part I: Basic Conceptions of Science and the Scientific Method." *Knowledge Organization* 48 (7–8): 473–98. <https://philpapers.org/rec/HJRSPI>.
- Oliveira, Lucia Maria Velloso de. 2009. *Projeto Análise tipológica dos documentos em arquivos pessoais: uma representação do código social*. Rio de Janeiro: Serviço de Arquivo Histórico e Institucional/Centro de memória e Informação/Fundação Casa de Rui Barbosa.
- Oliveira, Lucia Maria Velloso de, Bianca Therezinha C. Panisset, and José Antonio da Silva. 2020. "Mediation in Archives Organization, Classification and Transparency." In *Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark*, edited by International Society for Knowledge Organization (ISKO), 1st ed., 338–47. *Advances in Knowledge Organization* 17. Baden-Baden: Ergon – ein Verlag in der Nomos Verlagsgesellschaft. <https://doi.org/10.5771/9783956507762>.
- Silva, Amanda Marissa Soares da, e Natália Tognoli. 2022. "A organização do Conhecimento Arquivístico: A Emergência de uma Comunidade Discursiva Brasileira." *Acervo* 35 (2): 1–17. <https://revista.arquivonacional.gov.br/index.php/revistaacervo/article/view/1794>.
- Smith, Arthur. 2022. "Simple Knowledge Organization System (SKOS)." *Knowledge Organization* 49 (5): 371–84. <https://doi.org/10.5771/0943-7444-2022-5-371>.
- Tognoli, Natália Bolfarini, Ana Célia Rodrigues, and José Augusto Chaves Guimarães. 2019. "Archival Knowledge: Conceptual Frameworks for Recent Terminology in the KO Domain." *Knowledge Organization*, 46 (7), 522–29. <https://doi.org/10.5771/0943-7444-2019-7-522>.
- Yeo, Geoffrey. 2008. "Concepts of Record (1): Evidence, Information, and Persistent Representations." *American Archivist* 70 (2): 315–43. <https://doi.org/10.17723/AARC.70.2.U327764V1036756Q>.

Applying Integrative Levels Classification to Korean KOS Registry

Bridging Phenomena and Knowledge Systems

SoYoung Yoon*, Claudio Gnoli**, Ziyong Park*

Abstract: This study investigates the applicability of the phenomenon-based classification system ILC2 to the classification of the Korean KOS Registry. It also examines the potential of generative AI to supplement human expert validation in the classification process. To this end, approximately 450 KOS entries were classified through a three-stage cross-validation by human experts, along with classification code suggestions generated by ChatGPT. The results indicate that the phenomenon-centered ILC effectively reflects KOS topics organized around applied domains, and it can further support the identification of needs for new categories in emerging applied fields, as well as Korea- and Asia-specific subjects. Additionally, generative AI contributed to both suggesting classification notations and facilitating expert cross-checking; however, limitations were observed, such as hallucinations where the AI suggested notations not present in the ILC2 schedules. This study provides both practical and theoretical implications for the management of KOSs and the development of classification systems.

Keywords: Korean Knowledge Organization System (K-KOS), Knowledge Organization System (KOS) Registry, Integrative Levels Classification (ILC), Phenomenon-based Classification, Generative AI-assisted Classification.

1. Introduction

Knowledge organization systems (KOSs) encompass a broad range of structures designed to represent and manage knowledge resources. Depending on historical and disciplinary contexts, similar systems have been referred to using different terms (Voß 2016; Gnoli 2020). The Networked Knowledge Organization Systems (NKOS) community has contributed to clarifying such terminology by developing a dedicated KOS vocabulary, including the KOS Types Vocabulary (Zeng and DCMI NKOS Task Group 2015).

* College of Humanities & Arts, Hansung University, Seoul, Republic of Korea.
corba99@gmail.com. ORCID: 0000-0001-7677-4158; zgpark@hansung.ac.kr. ORCID: 0000-0002-8985-9991.

** Science and Technology Library, University of Pavia, Pavia, Italia. claudio.gnoli@unipv.it. ORCID: 0000-0002-4721-7448.

A wide range of institutions and individuals – including those in the public sector – utilize KOSs to structure and manage knowledge resources. In today's digital information environment, nearly all platforms that provide information rely on some form of KOS, varying in type and level of complexity. Korea's KOS landscape, like those in other countries, encompasses a variety of systems developed and applied across different domains. However, there remains a lack of overarching classification frameworks that can comprehensively encompass and interlink these diverse KOSs.

To address this gap, the present study explores the applicability and suitability of Integrative Level Classification version 2 (ILC2) for organizing Korean KOSs. As a phenomenon-based classification scheme, ILC2 supports multidimensional indexing and is designed to accommodate both natural and cultural phenomena. To test its applicability, we examine 458 KOS entries included in the K-KOS Registry and evaluate their classification under ILC2. The K-KOS Registry – used as the dataset for this study – is a curated collection of KOS entries primarily developed by Korean public institutions across diverse domains.

In addition, the study explores the potential role of generative AI – specifically ChatGPT-4o – as a support tool in the classification process. AI assistance is evaluated in terms of its ability to suggest candidate classmarks, provide interpretive rationales, and mitigate inconsistencies in expert-based classification. By combining a phenomenon-based framework with generative AI support, this study proposes a scalable and flexible approach to managing complex national-level KOS datasets.

This paper is structured as follows. Section 2 reviews the theoretical background, including KOS types, phenomenon-based classification, and recent developments in AI-assisted indexing. Section 3 outlines the data collection and classification process. Section 4 presents and analyzes the classification results. Section 5 discusses the implications of our findings and concludes with directions for future work.

2. Literature Review

2.1. Classifying Knowledge Organization Systems in KOS Registries

During the 20th century, classification schemes have been developed especially for use in library collections and bibliographies. However, already at the beginning of the century Paul Otlet also planned to classify articles, reports, films, exhibits, documentation internal to organizations etc. by the Decimal Classification (Folino 2024).

KOSs themselves are in need of being indexed for easy browsing, search and retrieval in KOS registries. As a KOS usually covers a broad domain or a set of domains, as opposed to the very specific subject matter of a book or an article, what is needed to index KOSs seems to be a general system of the whole of knowledge; assigned classes are not expected to be very specific, and associations of more than one class for the same KOS are likely (e.g. food, agriculture and health, or engineering and physics).

The Basic Register of Thesauri, Ontologies and Classifications (BARTOC) has been indexed using the Dewey Decimal Classification (DDC), the Integrative Levels Classification (ILC) and the EuroVoc thesaurus, which all have general coverage (Ledl and Gnoli 2017; Gnoli et al. 2019). Building on this approach, Park (2024) applied DDC, ILC, and the Korean Decimal Classification (KDC) to classify KOSs in the K-KOS Open Archives. Of these, ILC is the most innovative scheme.

Some studies have focused on classifying KOS types themselves. Souza, Tudhope, and Almeida (2012) analyzed various forms of KOSs by identifying multiple dimensions and organizing them into a taxonomy. Similarly, Voß (2016) extracted KOS types from Wikidata and proposed a typology-based classification of KOSs.

2.2. The Integrative Levels Classification

ILC is a general classification scheme that has been developing since 2004 by an international team coordinated by Claudio Gnoli (ISKO Italia 2024). While building on the tradition of general bibliographic classifications, it is original in that its main classes are phenomena, that is the objects of knowledge such as celestial objects, plants or languages, instead of disciplines, such as astronomy, botany or linguistics, as is usual in bibliographic classifications (Gnoli et al. 2024).

Each ILC class includes several ranks of subclasses, expressed by small caps: for example, class *q* “language” includes subclasses *qm* “morphs”, *qo* “words”, *qp* “phrases” etc., which in turn are subdivided into further ranks. Classes can be specified by facets, as in classical faceted classifications. Facets are indicated by digits and belong to a set of ten fundamental categories: 9 quality, 8 quantity, 7 part, 6 property, 5 transformation, 4 opposition, 3 agent, 2 place, 1 time, 0 perspective. Thus e.g. in class *q*, facet *q7sf* “phonemes” is introduced by digit 7 because phonemes are a part of language. This facet can be combined with subclasses in the example above to give *qm7sf* “morphs, phonemes”, *qo7sf* “words, phonemes” and so on.

While the scheme is developing continuously, edition 1 (ILC1) was frozen and published in 2011, edition 2 (ILC2) was in 2019 (Park, Gnoli, and Morelli 2020; Park et al. 2020), and edition 3 is expected to be in the coming

years. ILC2 is also available in SKOS format (<https://www.iskoi.org/ilc/skos.php>) (Binding et al. 2020).

The basic structure is similar in all editions. However, in each edition some classes are moved or developed, and the syntax of facets is partly different. In particular:

- bound facets – that is facets that are only applied within a specific class, such as phonemes (as there are no phonemes of celestial bodies or phonemes of polities: only phonemes of languages) – are introduced by *-a-* in ILC2, and by a single digit in ILC3;
- parallel facets – that is facets that take their values from another class – are introduced by *-91-*, *-92-* etc. in ILC2, and by multiple digits in ILC3;
- free facets – that is facets that may connect any two classes – are introduced by single digits in ILC2, and by *-0-*, *-10-*, *-20-* etc. in ILC3.

ILC3 can be considered the most advanced edition from both the technical and the content points of view. However, as it is not freezed yet, before its publication indexing projects and linked data applications need to use ILC2 for the purposes of information consistency.

2.3. AI-assisted Classification

Classification is a complex intellectual activity that involves both subject analysis and symbolic transformation of the classified entity. Recent research has explored the application of AI to support this process. In particular, AI has been used to semi-automatically assist human experts in classification and subject indexing. Examples include machine learning–based approaches to assigning subject headings (Ahmed et al. 2023), as well as pilot studies using generative AI to suggest classification codes and subject terms (Noruzi 2024) once a silent repository of knowledge, is undergoing a digital metamorphosis. At the heart of this transformation is the integration of artificial intelligence (AI). Both studies utilized standard knowledge organization systems (KOS), such as the Library of Congress Subject Headings (LCSH), for AI training, and relied on MARC-formatted bibliographic records annotated by human experts for model validation. Martin (2024) also supports this view, noting that generative AI can ease the mental demands of cataloging by assisting with subject analysis and code assignment. Rather than pursuing fully automated systems, these studies aimed to reduce the time required for indexing and to complement human subjectivity with AI-generated recommendations. Given their use of external structured vocabularies like LCSH and DDC, these approaches may be seen as a form of retrieval-augmented generation (RAG).

In this context, Qin (2020) offers a more theoretical perspective, emphasizing that although knowledge organization (KO) and knowledge representation

(KR) have traditionally functioned as distinct paradigms, the advancement of AI is fostering their convergence, positioning KOSs as both organizational tools and foundational assets for reasoning and knowledge acquisition in AI environments.

3. Methodology

To provide an overview of the research process, Figure 1 summarizes the methodology in three main stages: dataset selection, data classification, and data visualization and sharing. First, 458 K-KOS entries were selected from *K-KOS Open Archives* as the dataset. Second, the data were classified through a three-step process led by two domain experts, who independently assigned and reviewed ILC codes to ensure consistency. Throughout this process, generative AI (ChatGPT) was used as a supplementary tool—providing suggestions for potentially relevant categories and highlighting comparable classification patterns—thus supporting expert decision-making without replacing it. Finally, the classified KOS resources were indexed and made publicly accessible via K-KOS Open Archives, supporting visualization, search, and reuse.

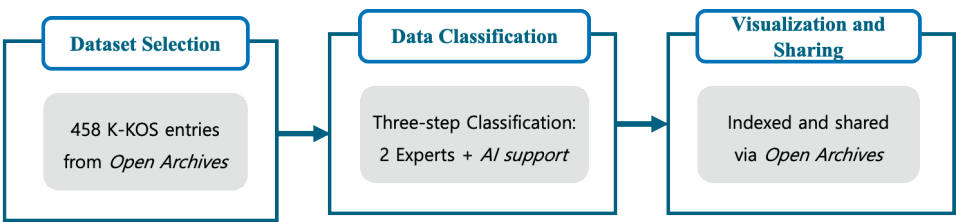


Figure 1. Overview of the methodology: dataset selection, classification, and sharing via Open Archives.

3.1. Dataset Selection

In this study, a total of 458 KOSs recorded in *K-KOS Open Archives* – the registry of the Korean KOSs (K-KOS) – were selected for analysis. K-KOS Open Archives is a publicly accessible platform that collects, curates, and shares KOSs developed by government agencies, public institutions, cultural heritage organizations, research institutes, and individual scholars. The registry encompasses a wide range of KOS types, including classification schemes and terminological resources. Each entry is accompanied by structured meta-data describing the KOS’s title, description, managing organization, and relevant reference sources.

Given the applied and practice-oriented nature of many KOSs included in the registry, K-KOS provides a useful empirical context for examining the principles and applicability of ILC2. The dataset is analyzed with respect to KOS type (Table 1), managing organization (Table 2), and online service platform (Table 3).

K-KOS KOS Type	Number of KOS	ratio	NKOS KOS Types
glossary (용어집)	205	44.76%	glossary, terminology
classification scheme (분류체계)	168	36.68%	classification scheme
encyclopedia (백과사전)	34	7.43%	-
list (목록)	28	6.11%	list
dictionary (언어사전)	6	1.31%	dictionary
thesaurus (시소러스)	6	1.31%	thesaurus
name authority list (인명사전)	5	1.09%	name authority list
cataloging rules (목록규칙)	4	0.87%	-
gazetteer (지명사전)	2	0.44%	gazetteer
Total	458	100.0%	

Table 1. Distribution of KOS Types in K-KOS and Their Correspondence to the NKOS Vocabulary.

Organization	Number of KOSs
Statistics Korea	42
National Folk Museum of Korea	31
Rural Development Administration	18
National Research Institute of Cultural Heritage	12
Korea Customs Service	11
Ministry of Environment	11
Ministry of the Interior and Safety	10
Health Insurance Review & Assessment Service	8
Ministry of National Defense	8
Korea Health Industry Development Institute	8

Table 2. Distribution of Classified K-KOS by Managing Institutions (Top 10 Organizations).

KOS Service Platform	Number of KOSs
Korean Law Information Center	56
Korean Statistical Classification	42
Encyclopedia of Korean Folk Culture	25
National Heritage Knowledge Connection	15
National Crop Pest Management System	13

Table 3. Distribution of Classified K-KOS by Service Platforms (Top 5 Plat).

KOS types were primarily categorized according to the NKOS KOS Types Vocabulary (Zeng and DCMI NKOS Task Group 2015), with certain modifications introduced to reflect the specific characteristics of K-KOS. In addition to classification schemes and glossaries, other types of resource – such as encyclopedias – also represent a notable share of the overall distribution (see Table 1). Table 2 presents the top ten managing institutions among the 198 organizations surveyed, all of which maintain at least eight KOSs. These include central government agencies such as Statistics Korea, which manages 42 KOSs primarily focused on national standard classifications, and cultural heritage institutions like the National Folk Museum of Korea, which manages 31 KOSs including glossaries and terminologies related to Korean folk culture. In terms of digital service provision, Table 3 summarizes the top five online platforms hosting K-KOSs. The Korean Law Information Center, which provides access to legal information and terminology, ranks first, followed by the Korean Statistical Classification platform, which serves as a major gateway to national statistical standards and datasets. However, unlike the Korean Statistical Classification platform operated directly by Statistics Korea, the National Library of Korea does not serve as a dedicated platform that both manages and distributes specific KOSs. Although it provides access to various glossaries and dictionaries, it does not function as the primary service or administrative body for those KOSs and is therefore excluded from the rankings.

The results indicate that classification schemes and terminological resources constitute a significant proportion of the registry. Moreover, the majority of KOSs were developed by public-sector entities, particularly government ministries and government-funded research institutes, and are disseminated via dedicated online platforms. This trend underscores the strong presence of public-sector KOSs and their alignment with national administrative functions. In light of the dominance of public-sector contributions, the overall distribution of K-KOS resources is further examined using the Business Reference Model (BRM), which provides a high-level categorization of government functions. Table 4 illustrates the distribution of KOSs across government

functions, with particularly high concentrations in cultural, industrial, and health-related domains.

BRM	Number of KOS	ratio
Culture, Sports, and Tourism	129	28.17%
Industry, Trade, and SMEs	51	11.14%
Health	48	10.48%
Agriculture and Forestry	38	8.30%
General Public Administration	37	8.08%
Environment	29	6.33%
Science and Technology	21	4.59%
Social Welfare	21	4.59%
National Defense	17	3.71%
Public Order and Safety	16	3.49%
Transportation and Logistics	14	3.06%
Education	12	2.62%
Fiscal, Tax, and Financial Affairs	12	2.62%
Communication	5	1.09%
Marine and Fisheries	5	1.09%
Regional Development	3	0.66%
Total	458	100.0%

Table 4. Functional Distribution of K-KOS According to Business Reference Model.

3.2. Data Classification

This study adopted the Integrative Levels Classification, 2nd edition (ILC2), as the classification framework. A free faceted approach was employed, allowing each KOS to be assigned multiple classification codes as appropriate. This method aligns with the general indexing needs of KOS registries, where classification tends to be broad rather than specific. As KOSs typically cover wide or cross-cutting domains – unlike books or articles that deal with narrowly defined topics – they are better indexed using a general system that accommodates multiple perspectives. BARTOC, for instance, applies ILC in a similarly free-faceted way to support browsing and retrieval without over-specification.

Given the scale and diversity of the dataset, which included 458 KOS entries, the application of full facet formulas was intentionally limited. While ILC2 theoretically supports the construction of syntagmatic “freely-faceted” combinations, fully expressing all relevant aspects of a concept could result in excessively long and potentially opaque classmarks, in absence of an automatic translator of faceted notation as experimented in other applications (Gnoli et al. 2023a, 2023b). For practical reasons, the classification process prioritized conciseness and interpretability, especially to maintain consistency across entries and support comparative analysis. Nonetheless, certain compound notations that are already enumerated in the official ILC schedules – particularly

those commonly used in web-based public versions – were selectively incorporated. These pre-formed combinations allowed for limited facet expression without necessitating complex syntactic construction.

Importantly, while this free classification approach preserves the phenomenon-based structure that is central to ILC, it does not fully leverage the expressive potential of its facet syntax. In this study, the choice of a simplified approach reflects both the practical demands of managing a heterogeneous dataset and the conceptual fit of general, high-level classifications for KOSs. At the same time, it acknowledges the trade-off involved: while the phenomenon-based logic of ILC is retained, the richness of facet-based articulation is necessarily constrained.

To ensure the consistency and reliability of the classification results, a three-step expert-based cross-review procedure was implemented. The process involved three stages – (1) initial coding by Expert 1, (2) reclassification by Expert 2 with support from generative AI (ChatGPT-4o), and (3) final review by Expert 1 – as illustrated in Figure 2. Each step was designed to complement the others, aiming to integrate multiple perspectives into the classification process.

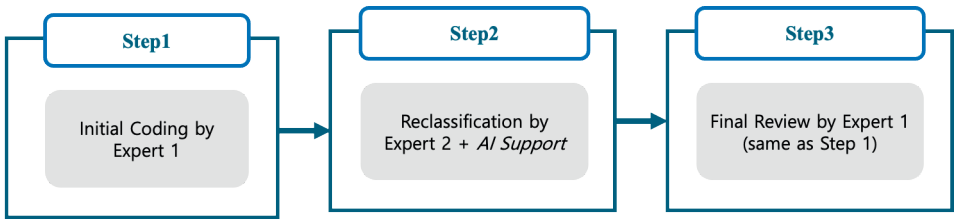


Figure 2. Three-step classification process involving expert review and generative AI support.

Step 1. Initial Coding by Expert 1

In the first step, Expert 1 assigned initial ILC2 classification codes to all K-KOS entries. The process emphasized a broad yet conservative interpretive stance, minimizing the use of multiple codes in cases involving ambiguous or polysemous subjects. Special attention was given to the structural consistency of notations, with careful consideration of the thematic domain to which each KOS entry belongs.

Step 2. Reclassification by Expert 2 with Generative AI Support

In the second step, Expert 2 reviewed the initial classifications and performed reclassification as needed. This stage allowed for more specific interpretations and the integration of multiple subject dimensions, resulting in the addition of secondary ILC2 codes where appropriate. Generative AI was em-

ployed to assist in this process by suggesting classification options and providing justifications, which facilitated alternative perspectives not initially considered. Importantly, Expert 2 evaluated all AI-generated suggestions and made the final decisions, thereby mitigating the potential risk of AI hallucinations.

To support AI-based classification, a dataset of ILC2 classmarks and captions was first compiled from the official ILC2 web service (ISKO Italia 2026). When a main class is selected, a hierarchy of subclasses appears on the landing page; the compiled dataset included most entries from rank 1 to 3, with some extending to ranks 4 or 5. A total of 2,503 ILC2 classes were collected. As illustrated in Figure 3, the ILC2 class list (left) and a translated KOS entry (right) were used to generate classification suggestions via GenAI. While the English class names provided by ILC2 served as the primary reference, Korean captions were also included to aid interpretation in the K-KOS classification context. The dataset was uploaded in CSV format to the ChatGPT-4o interface, enabling the model to suggest relevant classmarks along with justification based on each KOS record's metadata. Expert 2 consulted these AI-generated suggestions especially in cases where classification was ambiguous or diverged from Expert 1's judgment.

ILC2 Classmarks and Labels (English/Korean)

ILC_no	caption_eng	caption_kor	DDC_no	main cls
a	forms; mathematical objects [formal sc	형식; 수학적 객체 [형식 과학; 논리; 수학]	DDC 510 160	a
a92	in neighbourhood	근처에 있는		a
a93	implied by; as a consequence of premi	전제로부터 암시됨		a
a94	with error	오류와 함께		a
a95	through transformation; operation; fun	변환, 작용, 함수, 사상 등을 통해		a
a96	having property	속성을 가진		a
a97	with part; element; component	부분, 요소, 구성요소의 함께		a
a98	equal to result	결과와 동일한		a
aat	elements	요소		a
aau	subsets	부분집합		a
aaw	cosets	잉여집합		a
ab	individuals; instances	개체; 사례		a
ac	classes; sets; collections [set theory]	클래스; 집합; 모음 [집합 이론]		a

KOS Entry Title and Description (auto-translated for demonstration)

Ancient wooden tablets: List of wooden tablets by site

About KOS

Ancient Wooden Tablets: List of Wooden Tablets by Historical Site is a list of wooden tablets organized by historical site that were registered on the National Gaya Cultural Research Institute's Ancient Korean Wooden Tablets' website (Currently National Heritage Knowledge > Research 338 > Ancient Wooden Tablets). A total of 312 wooden tablets excavated from 22 historical sites are registered. The list covers not only those excavated from Seongsuamseong Fortress in Haman but also the entire country, and provides basic information on the excavation site, shape, size, etc. of the wooden tablets, as well as image and drawing data.

Figure 3. ILC2 classes and a KOS example used for AI-assisted classification.

Step 3. Final Review by Expert 1

In the final step, Expert 1 re-evaluated the revised classifications produced in Step 2. This review focused not only on validating the appropriateness of additional codes but also on ensuring the overall coherence and semantic clarity of the assigned notations. Rather than making extensive modifications,

this phase emphasized interpretive harmony and alignment across the dataset. Cases requiring further discussion or refinement were flagged and documented for consideration during the final analysis.

4. Results and Findings

4.1. Results of classification

4.1.1. Overview of Classification Results

In this study, a total of 458 K-KOS entries were classified using the Integrative Levels Classification, 2nd edition (ILC2). The classification process was divided into three sequential steps, and the results at each stage were summarized according to the top-level ILC2 classes, as shown in Table 5 and Table 6. While the overall patterns across Step 1, Step 2, and Step 3 remained broadly similar, closer examination reveals certain variations. These similarities and differences can be analyzed by comparing cases in which the assigned class remained consistent throughout the process versus those in which reclassification occurred between steps.

Class_no	Class_name	Step 1	Step 2	Step 3
<i>a</i>	forms	1	1	1
<i>d</i>	energy	9	9	11
<i>f</i>	molecules	4	4	4
<i>h</i>	celestial bodies	3	3	3
<i>i</i>	rocks	1	1	1
<i>j</i>	land	21	25	29
<i>m</i>	organisms	12	34	41
<i>n</i>	populations	17	18	18
<i>o</i>	instincts	1	1	1
<i>p</i>	consciousness	-	1	1
<i>q</i>	language	17	17	19
<i>r</i>	rituals	16	15	15
<i>s</i>	communities	30	59	70
<i>t</i>	polities	60	50	44
<i>u</i>	enterprises	48	46	38
<i>v</i>	technologies	89	61	54
<i>w</i>	artifacts	31	38	38
<i>x</i>	artworks	65	41	41
<i>y</i>	knowledge	33	34	29
Total		458	458	458

Table 5. K-KOS classification results by top-level ILC2 classes at each step (excluding multi-class assignments).

Class_no	Class_name	Step 1	Step 2	Step 3
<i>a</i>	forms	1	3	3
<i>d</i>	energy	9	11	11
<i>f</i>	molecules	4	4	4
<i>h</i>	celestial bodies	3	3	3
<i>i</i>	rocks	1	1	1
<i>j</i>	land	21	28	29
<i>m</i>	organisms	17	48	48
<i>n</i>	populations	21	26	27
<i>o</i>	instincts	1	1	1
<i>p</i>	consciousness	-	8	8
<i>q</i>	language	17	18	21
<i>r</i>	rituals	16	16	16
<i>s</i>	communities	33	125	123
<i>t</i>	polities	61	62	62
<i>u</i>	enterprises	49	81	92
<i>v</i>	technologies	96	84	83
<i>w</i>	artifacts	31	53	54
<i>x</i>	artworks	65	45	45
<i>y</i>	knowledge	35	56	54
Total		481	673	685

Table 6. K-KOS classification results by top-level ILC2 classes at each step (including multi-class assignments).

4.1.2. Cases with Consistent Classification Across Steps

During the classification process from Step 1 to Step 3, a total of 187 entries retained the same classification code between Step 1 and Step 2, and 438 entries showed consistency between Step 2 and Step 3. Notably, 179 KOS entries maintained an identical classification code across all three steps. Consistency was particularly high for KOS entries dealing with specific applied domains or well-defined subjects. This group included KOSs related to astronomy, mathematics, physics, satellites, minerals, climate, agriculture, national defense, law, fine arts, architecture, and performing arts. Representative examples of such cases are presented in Table 7.

KOS name	Assigned ILC code
Minerals Classification (광물분류)	<i>ib</i> “minerals” [mineralogy]
Climate Technology Classification (국가 기후기술 분류체계)	<i>jal</i> “climates” [climatology]
Legal Terms (법률용어)	<i>tay</i> “regulations; institutions; legal order; law; norms; rules; acts” [jurisprudence; legal theory]
Korea Performing arts box office Information System Genre Classification System (공연예술통합전산망 장르분류 체계)	<i>xmWr</i> “performing arts; shows”

Table 7. Examples of classification changes and ILC entries (Step 1–2).

In addition, an analysis of the distribution of KOS entries based on the top-level ILC categories reveals a clear tendency toward classifications in applied and socially oriented domains. Specifically, a greater number of KOS entries were assigned to categories such as *s* “communities”, *t* “polities”, *u* “enterprises”, and *v* “technologies”, rather than to more abstract or theoretical categories like *a* “forms”, *d* “energy”, and *f* “molecules”, which appear earlier in the ILC schedules. Among these, *s* “communities” and *v* “technologies” consistently emerged as the most frequently assigned categories across all classification steps. These findings suggest that KOSs are more closely aligned with the applied and societal domains in ILC.

4.1.3. Cases of Classification Changes Across Steps

In cases where classification codes were changed throughout the three steps, it was observed that Step 2 introduced a more expansive and integrative perspective compared to the conservative and inclusive approach of Step 1. This shift led to a more active incorporation of multiple thematic aspects. As shown in Step 3 of Table 6, which includes all assigned notations, an average of approximately 1.5 ILC codes was assigned per KOS. Some classes experienced particularly significant increases when multiple codes were allowed. For instance, the class *s* “communities” included 70 entries under single-code assignments, but this number rose to 123 when multiple codes were considered. Similarly, *u* “enterprises” increased from 38 to 92 entries. This trend suggests that certain sets of ILC classes are better suited to capturing the complexity of hybrid or interdisciplinary KOS themes.

By actively embracing multiple notations and leveraging generative AI suggestions in Step 2, a greater number of classification codes could be applied than in Step 1. As shown in Table 2, Step 1 resulted in 458 code assignments, while Step 2 yielded 673, and Step 3 reached 685 in total. A breakdown of the 271 changed notations between Step 1 and Step 2 is summarized in Table 8.

Type of Change	Frequency (Step 1→Step 2)	Frequency (Step 2→Step 3)
Change in the number of codes (Single → Multiple)	95	18
Specification within the same class	99	-
Change to a different upper-level class	86	3
Total	271	20*

Table 8. Types and Frequencies of Classification Code Changes Between Step 1 and Step 2.

* One case involved both a change in class and a change in the number of codes. The total count excludes this overlap.

As shown in Table 8, a total of 271 classification codes were changed from Step 1 to Step 2. Among these, 95 cases involved changes from a single classification code to multiple codes, 99 cases involved the addition of more specific codes within the same main class, and 86 cases involved changes to entirely different classes based on a revised perspective. Representative examples of each type are presented below.

- 1) change in the number of codes (Single → Multiple)
 - *Business Reference Model* (정부기능분류체계)
 [step 1] *tal* “civil service; public sector; community services; public administration”
 [step 2] *tal + sp* “public institutions; public bodies; services of general interest”
- 2) specification within the same class
 - *Terms of vehicle emission gas* (배출가스 용어사전)
 [step 1] *ny* “ecosystems; biomes; biocores” [environmental sciences]
 [step 2] *nyal* “pollution”
- 3) change to a different upper-level class
 - *Anatomical Terminology* (해부학 용어)
 [step 1] *yite* “medical science”
 [step 2, 3] *mao* “organs; systems” [anatomy]

Next, the changes from Step 2 to Step 3, as shown in Table 8, indicate that an additional 23 classification codes were modified. Specifically, 19 entries were assigned multiple codes, 3 entries were further specified within the same class, and 2 entries were reassigned to different classes. These results suggest that the active expansion of classification codes in Step 2 was further extended in Step 3. Representative examples of each change type are summarized below.

- 1) change in the number of codes (Single → Multiple)
 - *Classification of Automobile Parts* (자동차 분류맵)
 [step 1] *wver* “cars”
 [step 2] *wver + uaqst* “tariffs” Some hallucinations required
- 2) change to a different upper-level class

- *Glossary of National Land* (국토용어사전)
 [step 1] *ux* “transport”
 [step 2] *j* “land; landforms; regions; geographical features”

4.2. Web-Based Access to Classified KOS Resources

The classified entries were published via the K-KOS Open Archives to promote transparent reuse, offering searchable indexes and multilingual access. Originally developed in Korean, the platform now includes an English-language interface to enhance accessibility and facilitate international interoperability. Figure 4 illustrates how K-KOS resources are organized using ILC2 within the platform.

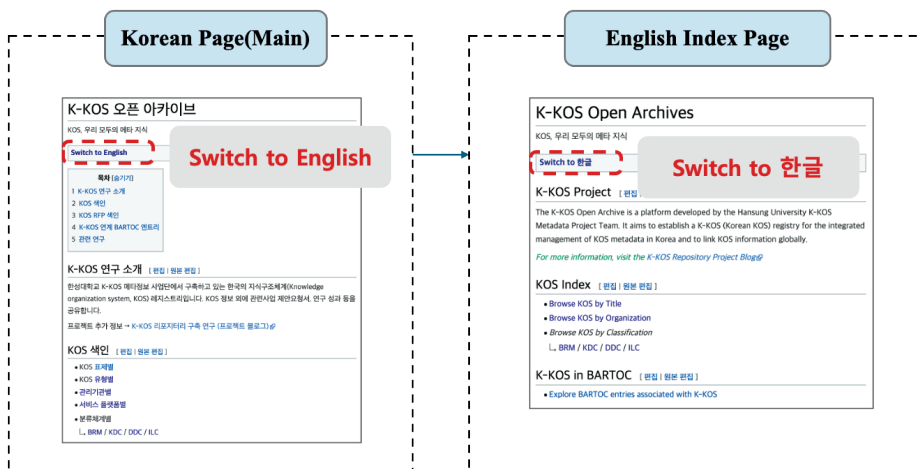


Figure 4. Language toggle interface between the Korean main page and the English index page of K-KOS Open Archives (K-KOS Metadata Project Team 2026).

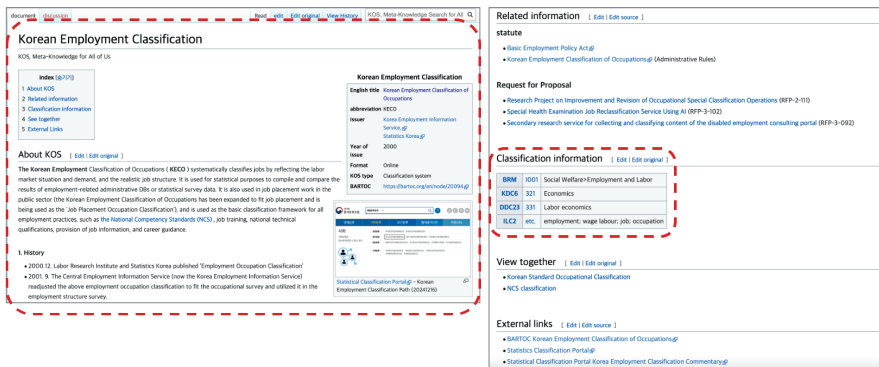


Figure 5. Screenshot of *Korean Employment Classification of Occupations* on K-KOS Open Archives (Automatically translated by Google Chrome) (K-KOS Metadata Project Team 2025).

Figure 5 provides an example of a detailed KOS entry—Korean Employment Classification of Occupations—in which classification information is displayed alongside KOS metadata. The classification index supports browsing by BRM, KDC, DDC, and ILC schemes. The page also includes external links to corresponding KOS entries in BARTOC, thereby enhancing interoperability and visibility within the broader KOS ecosystem.

4.3. Discussion

This section discusses the implications of the classification process and results from three interrelated perspectives. First, it examines how well the principles of ILC align with the inherent characteristics of Knowledge Organization Systems (KOSs) (4.3.1). Second, it evaluates the applicability of ILC to Korean KOSs and emerging technological domains, while identifying areas for potential refinement (4.3.2). Third, it reflects on the use of generative AI as a support tool in the classification workflow, highlighting both its contributions and limitations (4.3.3).

4.3.1. Phenomenon-Based and Multi-Dimensional Classification in ILC

This section discusses two key aspects of applying ILC to KOS classification: (1) the conceptual suitability of ILC's phenomenon-based structure, and (2) the implications of assigning multiple class codes in a free classification manner.

First, the ILC framework enabled the classification of KOS entries based on the phenomena or entities they describe, rather than the disciplinary fields to which they belong. This contrasts with traditional academic classification systems – such as Dewey Decimal Classification (DDC) – which typically impose a subject hierarchy aligned with academic domains. For instance, a KOS on “disaster management” could be directly assigned to *spr* “emergency services; rescue services” in ILC, without requiring a disciplinary framing under meteorology, civil engineering, or public safety. The results shown in Table 9 – based on single-class assignments only – highlight this conceptual alignment. A significant number of KOS entries were classified under *s* “communities” (12.9%), *t* “politics” (10.9%), and *v* “technologies” (13.3%). These classes reflect real-world phenomena and functions of KOSs in applied contexts such as governance, technological infrastructure, and social organization. In contrast, more abstract or scientific classes like *a* “forms”, *d* “energy”, or *f* “molecules” were rarely used.

Second, the use of multiple ILC class assignments reveals a distinct strategy for representing the multidimensionality of KOS content. In disciplinary-based

systems such as the DDC, assigning multiple numbers often implies coordination among disciplinary perspectives. In contrast, within ILC, multiple assignments indicate coexisting aspects of the same KOS entry, consistent with a free classification approach. As shown in Table 10, when multiple codes are included, the share of entries classified under *s* “communities” increases to 18.6%, while *v* “technologies” accounts for 12.5%, and *u* “enterprises” for 12%. These increases reflect the prevalence of hybrid or context-rich KOSs that span social, technological, and economic domains. However, the current study employed a simplified, free approach that allowed juxtaposition of classes without constructing fully faceted classmarks. While this method enhanced interpretability and practical feasibility, it limited the expressive potential of ILC’s facet syntax. Thus, although ILC supports multidimensional representation in principle, its full capacity was only partially realized under the current approach.

Class No.	Class Name	Step 3 (Count)	Step 3 (%)
<i>a</i>	forms	1	0.2
<i>d</i>	energy	11	2
<i>f</i>	molecules	4	0.9
<i>h</i>	celestial bodies	3	0.7
<i>i</i>	rocks	1	0.2
<i>j</i>	land	29	5.5
<i>m</i>	organisms	41	7.4
<i>n</i>	populations	18	3.9
<i>o</i>	instincts	1	0.2
<i>p</i>	consciousness	1	0.2
<i>q</i>	language	19	3.7
<i>r</i>	rituals	15	3.3
<i>s</i>	communities	70	12.9
<i>t</i>	polities	44	10.9
<i>u</i>	enterprises	38	10
<i>v</i>	technologies	54	13.3
<i>w</i>	artifacts	38	8.3
<i>x</i>	artworks	41	9
<i>y</i>	knowledge	29	7.4
Total		458	100

Table 9. Step 3 Classification Results of K-KOS by ILC2 Main Classes (Excluding Multiple Assignments).

Class No.	Class Name	Step 3 (Count)	Step 3 (%)
<i>a</i>	forms	3	0.4
<i>d</i>	energy	11	1.6
<i>f</i>	molecules	4	0.6
<i>h</i>	celestial bodies	3	0.4
<i>i</i>	rocks	1	0.1
<i>j</i>	land	29	4.2
<i>m</i>	organisms	48	7.1
<i>n</i>	populations	27	3.9
<i>o</i>	instincts	1	0.1
<i>p</i>	consciousness	8	1.2
<i>q</i>	language	21	2.7
<i>r</i>	rituals	16	2.4
<i>s</i>	communities	123	18.6
<i>t</i>	polities	62	9.2
<i>u</i>	enterprises	92	12
<i>v</i>	technologies	83	12.5
<i>w</i>	artifacts	54	7.9
<i>x</i>	artworks	45	6.7
<i>y</i>	knowledge	54	8.3
Total		685	100

Table 10. Step 3 Classification Results of K-KOS by ILC2 Main Classes (Including Multiple Assignments).

The multidimensionality of KOSs, as revealed through ILC, can be illustrated by cases in which multiple classification codes were assigned. In these cases, two to four ILC classes were used, reflecting the choice not to synthesize facet elements into a single notation. Rather than combining classes into compound codes using facet indicators, distinct ILC classes were enumerated as needed. The following examples present KOS entries with multiple assigned classes.

Two-Class Assignments

Korean folk religion, rites (한국의 민속신앙)

- *ral* “rites” + *sykk* “Korean civilization”;
- Assigned to 11 entries, including *Folk religion of villages in Korea* (한국의 마을신앙), *Encyclopedia of Korean Folk Beliefs* (한국민속신앙사전).

Military supplies (군수품), *force support* (전력 지원)

- *tat* “military” + *uag* “goods distribution”;
- assigned to 6 entries, including *Classification of military supplies* (군수품의 분류), *Types of Force Support Systems* (전력지원체계 분류).

Social insurance (사회보험)

- *spas* “social insurance” + *spl* “welfare; social services”;
- assigned to 5 entries, including *National Pension Terms* (국민연금 용어사전), *Long-term Care Insurance Terms* (장기요양보험 용어사전).

Three-Class Assignments*Electronic components tariff* (전자부품 관세)

- *vtr* “electrical manufacturing” + *uaqst* “tariffs” + *uak* “trade; commerce”;
- assigned to 3 entries, including *Semiconductor Guide* (반도체 가이드), *Mobile Phone Parts Guide* (휴대전화부품 가이드).

Artists, Groups (예술인 단체)

- *x* “artworks” + *px98x* “artists” + *uo* “corporations; companies”;
- assigned to 3 entries, including *Seoul Art Guide Name Authority List* (서울아트가이드 인명사전), *Professional Art Organization List* (전문예술법인·단체 리스트).

Four-Class Assignment*Aviation* (항공)

- *vx926y* “aviation” + *vx* “cargo; freight transportation” + *wvr* “aircrafts” + *wux* “airports”;
- assigned to 4 entries, including *Terms of Aviation* (항공용어사전), *Korea Institute of Aviation Safety Technology Glossary* (KIAST 용어사전).

4.3.2. Addressing Gaps in Cultural and Topical Coverage

During the classification process, several KOS entries lacked clear counterparts within existing ILC2 classes or could only be matched at a very general level. This was particularly evident for topics reflecting recent technological developments or contemporary social trends. Although these topics are inherently phenomenon-based, they could not be precisely aligned with existing ILC categories due to limitations in specificity and coverage. For instance, KOSs related to renewable energy, secondary batteries, and green taxonomy were difficult to place due to insufficiently granular entries. Similarly, digital innovation domains such as blockchain, big data, and the Internet of Things had to be mapped to the most proximate existing classes, thereby limiting classification precision. Representative KOS examples falling into this category were classified under the following ILC2 classes.

- *Classification of IoT* (사물인터넷 분류체계) - *wnw* “Internet”;
- *Korea New and Renewable Energy Center Terms* (신·재생에너지센터 용어사전) - *dam* “energy form”;
- *E-learning Industry Classification* (이러닝산업분류) - *spe* “education services; schools”;

- *e-sports classification*(이 스포츠 등급분류) - *xw* “sport; games” *xy9ytm* “videogames”.

In addition, the cultural specificity of many Korean KOSs revealed underdeveloped domains in ILC. Certain entries contain culturally bound concepts—such as regional proper names (e.g., Jeju Island), indigenous religious movements (e.g., Donghak), or traditional knowledge systems—that are not easily rendered into a single English expression. In such cases, classification fidelity may be improved by incorporating Romanized Korean labels or introducing culturally specific entries.

- *Korean Sign Language Dictionaries*(한국수어사전) - *qvh* “Korean language”;
- *Folk religion of villages in Korea:Jeju-do*(한국의 마을신앙:제주특별자치도) - *ral* “rites; practices; worship; uses; habits; ceremonies; liturgies; divine service” *sykk* “Korean civilization”;
- *Encyclopedia of Korean Folk Arts:Nongak* (한국민속예술사전:농악) - *xmavd* “traditional; root; ethnic” *sykk* “Korean civilization”.

These observations suggest that future versions of ILC, including the upcoming ILC3, would benefit from more flexible inclusion of regionally specific and topically emergent categories, enabling improved applicability to diverse national contexts. The following examples illustrate KOS entries that reflect such needs.

4.3.3. Assessing the Role of Generative AI in Faceted Classification

In this study, generative AI – specifically ChatGPT-4o – was used as a support tool during the second phase of the classification process. The AI was tasked with suggesting one or more ILC2 class codes for each KOS entry, based on contextual descriptions, and provided brief justifications for its choices. This allowed the human classifier to explore alternative perspectives, particularly for entries involving semantic ambiguity or interpretive disagreement between experts. In cases involving complex or multidimensional topics, AI-assisted suggestions often included multiple codes, thereby contributing to broader conceptual coverage.

Example1: Classification of *List of Wooden Tablets by Historical Site* (고대 목간: 유적별 목간목록)

Recommended ILC2 Classmarks by Generative A

Based on the description of the KOS, the generative AI (ChatGPT-4o) proposed the following classmarks:

- *yt* – writing → The primary function of wooden tablets is written communication; therefore, this class is appropriate;

- *mpat* – wooden materials → Since the items are wooden artifacts, this class represents the material aspect accurately;
- *sykk* – Korean civilization → As the KOS focuses on archaeological finds within Korea, this class reflects the cultural and geographic scope.

Corresponding ILC2 Classmarks and Captions

The suggested codes correspond to the following official ILC2 classmarks:

- *yt* “documents; instantiations; knowledge items”;
- *mpat* “trees; arboreous; arboreal plants”;
- *sykk* “Korean civilization”.

Final Classmarks Assigned by Expert Review

After evaluating the material, function, and cultural context of the KOS, the following two classmarks were selected:

- *yt* “documents; instantiations; knowledge items”;
- *sykk* “Korean civilization”.

The classmark *mpat* was excluded, as the material aspect was deemed less central than the functional and cultural dimensions in this case. However, the limitations of generative AI were also apparent. In some instances, the AI produced hallucinated codes that did not exist in ILC2 or made structurally inconsistent recommendations. As a result, every AI-generated suggestion required expert verification, and final decisions rested entirely with human experts. While the tool proved useful in generating ideas and highlighting cross-category linkages, its outputs could not be used uncritically. Future use of generative AI in classification will require improved training on structured vocabularies and controlled integration into expert workflows.

In another case, for the KOS entry *Disaster and Safety Industry Classification*, the combined use of *spr* “emergency services” and *tas* “social safety” was proposed to better capture its complex thematic scope. However, the additional suggestion of *uaq* – interpreted by the AI as representing “industrial sectors” – was in fact inaccurate, as *uaq* in ILC2 denotes “taxes; levies.” This reflects a case of hallucination in generative AI, underscoring the need for expert validation of AI-generated classification results.

Example2: Disaster and Safety Industry Classification (재난안전산업 분류)

Recommended ILC2 Classmarks by Generative AI

- *spr* – emergency services → Relevant to disaster response, recovery, and preparedness functions described in the KOS;
- *tas* – social safety → Captures the broader societal goal of disaster and safety-related classifications;

- *uaq* – industrial sectors → Since the KOS itself defines sectors within an industry, this high-level class contextualizes the classification's scope.

Corresponding ILC2 Classmarks and Captions

- *spr* “emergency services; rescue services”;
- *tas* “police; constabulary; civil guard; protective services; cops”;
- *uaq* “taxes; levies”.

Final Classmarks Assigned by Expert Review

After evaluating the material, function, and cultural context of the KOS, the following two classmarks were selected:

- *spr* “emergency services; rescue services”;
- *tas* “police; constabulary; civil guard; protective services; cops”.

After identifying *uaq* “taxes; levies” as a hallucinated recommendation, the case was re-submitted to the ChatGPT prompt with a clarification. The AI subsequently returned a revised set of ILC2 classmarks, none of which involved non-existent or misinterpreted codes. This illustrates the importance of not only assessing the relevance of AI-generated suggestions, but also verifying that the proposed classmarks correspond to actual entries in the classification system.

Re-recommended ILC2 Classmarks by Generative AI

- *spr* – emergency services → Relevant to disaster response, recovery, and preparedness functions described in the KOS;
- *sprf* – fire and rescue services → Specifically captures firefighting and rescue operations, which are key components of the industry;
- *uaqi* – social-security contributions → Reflects the institutional and policy-related dimension of disaster safety support, though its relevance may be secondary.

When appropriate classmarks were not available, it was also possible to prompt the AI to suggest new classmarks that could potentially be added to ILC2. For example, in the case of the *Classification of Blockchain* KOS, the AI not only recommended two existing ILC2 codes, but also proposed additional classmarks along with justifications for their inclusion in future updates of the classification system.

- KOS title - *Classification of Blockchain* (블록체인 분류체계);
- KOS description - Classification of Blockchain is one of the new software domain classification schemes used in the *Software Industry Survey* published by the Software Policy & Research Institute. The classification system is described in the December 2018 report titled *2018 Software Industry Survey*, issued by the same institute. This new set of software classifications includes schemes for cloud services, big data,

Internet of Things (IoT), artificial intelligence, VR/AR/MR, software-convergent services, and blockchain (*translated from Korean*).

The classmarks suggested by the generative AI are as follows. As shown below, both of the recommended classmarks currently listed in the ILC2 schedule were found to be valid and supported by reasonable justification. The newly proposed classmarks, on the other hand, were not part of the existing ILC2 scheme and appear to be conceptual extensions of *wnv* “computer networks.”

Suitable ILC2 Classmarks

- *wnv* – computer networks [telematics] → Blockchain is fundamentally a decentralized network technology, which can be categorized under computer networks;
- *wnuas* – software → As blockchain is classified as part of the software industry, this classmark reflects its technological foundation.

Suggested New Classmarks Not in ILC2

- *wnvbx* – distributed ledger technology → A proposed subclass under computer networks to represent the distinctive architecture of blockchain systems;
- *wnvas* – crypto-assets or smart contracts → Reflects emerging applications of blockchain beyond infrastructure, including digital asset exchange and contract automation.

Notably, the *ILC developing version* (ISKO Italia 2026), which is currently under development as the successor to ILC2, includes a specific entry for “blockchain.” In this version, the previous *wnv* class for “computer networks” has been updated to *suv*, under which *suvvl* refers to “blockchains,” categorized as a subclass of “distributed ledgers; shared ledgers.” This comparison suggests that the generative AI’s proposed extensions – although not present in ILC2 – are aligned with the direction of ongoing classification development and may offer meaningful guidance.

Blockchain in ILC developing version

<i>suv</i>	computer networks [telematics]
<i>suvv</i>	distributed ledgers; shared ledgers
<i>suvvl</i>	blockchains

5. Conclusions

This study explored the applicability of the Integrative Levels Classification, second edition (ILC2), to the classification of Korean Knowledge Organization Systems (K-KOS), as well as the role of generative AI in supporting

the classification process. By applying ILC2 to 458 KOS entries and integrating AI-assisted suggestions, the research yielded several key insights.

First, the phenomenon-based structure of ILC2 proved conceptually well-suited to the classification of KOSs, which often reflect applied, interdisciplinary, and real-world domains. The flexibility of ILC2 allowed for multidimensional representations through multiple class assignments, especially in domains such as technology, governance, and communities. Second, the classification process revealed topical and cultural gaps in the current ILC2 schedules. Specific challenges were encountered when classifying KOSs related to emerging technologies and culturally specific content. These findings suggest the need for more granular and localized categories in future updates of ILC. Third, the use of generative AI, particularly ChatGPT-4o, demonstrated both potential and limitations. While AI provided valuable classification suggestions and alternative perspectives, some AI-generated suggestions exhibited hallucinations, requiring expert validation before final adoption. AI was most effective when used as a complementary tool rather than as a substitute for expert judgment.

Overall, the study contributes a practical model for applying ILC2 to a national KOS registry and demonstrates how AI can enhance classification workflows. It also identifies future directions for improving classification systems through iterative updates, culturally responsive expansions, and enhanced human–AI collaboration.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A03085177).

References

- Ahmed, Mustak, Mondrita Mukhopadhyay, and Parthasarathi Mukhopadhyay. 2023. "Automated Knowledge Organization AI ML Based Subject Indexing System for Libraries." *DESIDOC Journal of Library & Information Technology* 43 (March):45-54. <https://doi.org/10.14429/djlit.43.01.18619>.

- Binding, Ceri, Claudio Gnoli, Gabriele Merli, Marcin Trzmielewski, and Douglas Tudhope. 2020. "Integrative Levels Classification as a networked KOS: a SKOS representation of ILC2." In *Knowledge organization at the interface: Proceedings 16th International ISKO Conference*, eds. Marianne Lykke, Tanja Svarre, Mette Skov and Daniel Martínez-Ávila, Baden-Baden: Ergon, 49-58.
- Folino, Antonietta. 2024. *Alle origini della Classificazione decimale universale: la corrispondenza tra Paul Otlet e Melvil Dewey*. Milano: Editrice Bibliografica.
- Gnoli, Claudio. 2020. *Introduction to Knowledge Organization*. Facet Publishing.
- Gnoli, Claudio, Douglas Tudhope, Patrícia de Almeida, Romana Barbui, Ceri Binding, and Véronique Blot. 2023. "Simpler search in a complex world: browsing ethnographic videos by freely faceted classmarks." In *Proceedings 7th ISKO UK Conference*, Glasgow, 24-25 July 2023. CEUR Workshops, <https://ceur-ws.org/Vol-3661/06-SHORT-CGnoli-ISKOUK2023.pdf>.
- Gnoli, Claudio, Ziyoun Park, and Andreas Ledl. 2019. "Dimensional analysis of subjects: indexing KOSs in BARTOC by phenomena, perspectives, documents and collections." *1st ISKO Low Countries Conference*, Bruxelles 21-22 June 2019.
- ISKO Italy. 2019. "ILC as SKOS." Last updated October, 22. <https://www.iskoi.org/ilc/skos.php>
- ISKO Italia. 2024. "Integrative Levels Classification." Last updated September 24. <https://www.iskoi.org/ilc/>.
- ISKO Italia. 2026. "ILC developing version." Last updated February, 24. <https://www.iskoi.org/ilc/2/ilc.php>.
- K-KOS Metadata Project Team. (2025). "Korean Employment Classification of Occupations". June 20, 2025. https://openarchives.net/wiki/index.php?title=Korean_Employment_Classification_of_Occupations.
- K-KOS Metadata Project Team. (2026). "K-KOS Open Archives". Last updated February, 25. https://openarchives.net/kos/K-KOS_Open_Archives.
- Ledl, Andreas, and Claudio Gnoli. 2017. "Indexing KOSs in BARTOC by a disciplinary and a phenomenon-based classification: preliminary considerations." In *Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar*, 14-15 September 2017, London, eds. Aida Slavic, Claudio Gnoli. Würzburg: Ergon, 109-17.
- Martins, Sugabsen. 2024. "Artificial Intelligence-Assisted Classification of Library Resources: The Case of Claude AI." *Library Philosophy and Practice* (e-Journal). <https://digitalcommons.unl.edu/libphilprac/8159>.

- Mazzocchi, Fulvio. 2018. "Knowledge organization system (KOS): an introductory critical account." *Knowledge Organization* 45 (1): 54-78. Also available in ISKO Encyclopedia of Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli, <https://www.isko.org/cyclo/kos>.
- Noruzi, Alireza. 2024. "The Use of Artificial Intelligence in Knowledge Organization and Subject Indexing." *Informology*. Informology Center, <http://eprints.rclis.org/45949/>.
- Park, Ziyong. 2023. "Aggregating Distributed KOSs: Enriching with Meta Information and Linking to the Multilingual KOS Registry." *The Electronic Library* 41 (6): 755-69. <https://doi.org/10.1108/EL-06-2023-0168>.
- Park, Ziyong. 2024. "A Comparative Study of Classification Systems for Organizing a KOS Registry." *Journal of the Korean Society for Library and Information Science* 58 (2): 269-88. <https://doi.org/10.4275/KS-LIS.2024.58.2.269>.
- Park, Ziyong, Claudio Gnoli, and Daniele P. Morelli. 2020. "The second edition of the Integrative Levels Classification." *Journal of Data and Information Science* 5 (1): 39-50.
- Qin, Jian. 2020. "Knowledge Organization and Representation under the AI Lens." *Journal of Data and Information Science* 5 (1): 3-17. <https://doi.org/10.2478/jdis-2020-0002>.
- Souza, Renato Rocha, Douglas Tudhope, and Maurício Barcellos Almeida. 2012. "Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems." *Knowledge Organization* 39 (3): 179-92. <https://doi.org/10.5771/0943-7444-2012-3-179>.
- Voß, Jakob. 2016. "Classification of Knowledge Organization Systems with Wikidata." In *CEUR Workshop Proceedings* 1676: 15-22. Proceedings of the 15th European NKOS Workshop, Hannover, Germany. <https://ceur-ws.org/Vol-1676/#paper2>.
- Zeng, Marcia Lei and DCMI NKOS Task Group. 2015. "KOS Types Vocabulary (Networked Knowledge Organization Systems Dublin Core Application Profile)." *DCMI NKOS Task Group*, <https://nkos.dublincore.org/nkos-type.html>.

L'evoluzione della struttura documentale nelle cartelle cliniche psichiatriche dell'archivio storico del manicomio di Girifalco (1881-1978)

Emanuela Nicole Donato*, Grazia Serratore**,
Maria Taverniti*, Maria Teresa Chiaravalloti*

Abstract: In the 19th and 20th centuries significant methodological and scientific changes transformed the psychiatric field. In this context, medical records represent a primary documentary source, as their structure reflects the changes in scientific, cultural, and social approaches over the decades. This study focused on the analysis of medical records from the historical archive of the former psychiatric hospital of Girifalco (CZ, Italy), which includes approximately 15,800 medical records. Through five-year interval sampling 2,290 medical records were analyzed. Results show that structural and content-related changes in medical records were gradual transformations while, in some cases, different versions of the same type of document coexisted for years. The research highlights how these documents transcend the mere function of recording clinical data, serving instead as historical testimonies of the various ways mental illnesses were understood and classified, as well as of the changes that occurred in the medical practice of this discipline.

Keywords: Medical records, Psychiatric hospital, Knowledge organization, Document structure, Girifalco.

1. Introduzione

Il periodo compreso tra il XIX e il XX secolo ha registrato significativi cambiamenti metodologici e scientifici nella disciplina psichiatrica (Babini 2011), influenzando profondamente la comprensione e il trattamento dei disturbi

* Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Rende (CS), Italia. emanuela.donato@iit.cnr.it. ORCID: 0009-0003-1350-1429; grazia.serratore@iit.cnr.it. ORCID: 0009-0009-9481-2213; maria.chiaravalloti@iit.cnr.it. ORCID: 0000-0003-4695-2026; maria.taverniti@cnr.it. ORCID: 0000-0001-7000-5817.

** Dipartimento di Culture, Educazione e Società, Università della Calabria, Rende (CS). srrgrz-97r52m208c@studenti.unical.it. ORCID: 0009-0009-9481-2213

Il testo è stato scritto in totale collaborazione e accordo tra le autrici, tuttavia vanno ascritti a Emanuela Nicole Donato il paragrafo 3 e 5; a Grazia Serratore il paragrafo 2 e 2.1; a Maria Teresa Chiaravalloti i paragrafi 1 e 4. Le autrici Emanuela Nicole Donato e Grazia Serratore hanno lavorato congiuntamente per i paragrafi 4.1, 4.2 e 4.3, mentre Maria Taverniti ha coordinato la stesura e le revisioni del contributo.

mentali. In questo contesto di continua e profonda evoluzione, le cartelle cliniche storiche dei pazienti ricoverati negli ospedali psichiatrici costituiscono una fonte documentale primaria della pratica medico-psichiatrica, ma anche una fonte documentale secondaria dell'evoluzione degli approcci scientifici, culturali e sociali nei confronti della disciplina (Tamburini e Benelli 1885; Caffaratto 2010). Le informazioni in esse presenti sono direttamente influenzate dagli eventi storici e scientifici contingenti alla loro stesura (Greco 2018). Gli archivi storici manicomiali hanno, pertanto, suscitato nel tempo l'interesse di studiosi di molteplici discipline e numerosi sono stati i progetti volti al loro recupero e alla loro valorizzazione (Trovato 2018; Baggio 2020; Ricucci 2025), ma anche gli sforzi atti a creare studi comparativi (Valeriano 2017). Su tutti merita menzione il progetto *Carte da legare* della Direzione generale archivi del Ministero della Cultura, il quale si articola su tre livelli di intervento: il censimento degli archivi, il riordinamento e l'inventariazione della documentazione. Grazie all'utilizzo di un comune schema di metadati, sul sito di progetto (Direzione Generale Archivi n.d.) è possibile condurre analisi statistiche basate su diversi parametri connessi al ricovero dei pazienti.

Qualsiasi sia la prospettiva di indagine non si può prescindere dalla considerazione che le cartelle cliniche sono evidenza dell'organizzazione e rappresentazione delle conoscenze mediche nel corso dei decenni, nonché specchio di correnti ideologiche e del contesto storico-sociale (Fiorino 2002; Carrino 2018). La scelta delle denominazioni delle etichette di sezione e dei campi nelle cartelle cliniche, quindi, non risponde soltanto all'esigenza di registrare le informazioni, ma incide direttamente su di essa, determinando ciò che viene annotato, ciò che viene preso in considerazione per la cura del paziente e, infine, ciò che viene conservato nel tempo.

Al di là del loro contenuto clinico, è pertanto interessante valutare la strutturazione delle informazioni all'interno delle cartelle cliniche, in termini di organizzazione e denominazione delle sezioni e sottosezioni, e come esse si modifichino nel corso del tempo, in quanto «if we want to be able to search across versions of concept schemes, we have to track these changes» (Tennis 2006, 13). Mediante il richiamo alla metafora ontogenetica lo sviluppo dei vocabolari viene accomunato allo sviluppo di un organismo vivente, che cambia costantemente durante il corso della sua vita. Ma mentre per un organismo vivente è intrinsecamente possibile tracciare a ritroso i vari stadi evolutivi attraversati, lo stesso non è sempre fattibile per un vocabolario e questo ostacola il recupero dell'informazione nel corso delle varie versioni di uno stesso sistema di organizzazione della conoscenza, il cosiddetto *persistent retrieval* (Tennis 2006). Avere, invece, un meccanismo che permetta di tenere traccia dell'evoluzione di un concetto all'interno di un vocabolario consentirebbe non solo di non perdere informazione, ma anche di rendere evidente come cambiano parole e concetti, anche in relazione al contesto di utilizzo.

Il presente studio intende, pertanto, indagare l'evoluzione dell'organizzazione delle informazioni all'interno di documenti sanitari contenuti nelle cartelle cliniche dell'archivio storico dell'ex ospedale psichiatrico di Girifalco, piccolo paese collinare in provincia di Catanzaro, che rimase in attività dal 1881 al 1978 (Marcello 1995) e il cui archivio storico è, ad oggi, composto di circa 15.800 cartelle cliniche. Disporre di documentazione prodotta in un arco temporale così ampio offre la possibilità di effettuare uno studio diacronico delle denominazioni delle sezioni e sottosezioni delle cartelle cliniche, mettendo in relazione i cambiamenti strutturali e contenutistici con le teorie emerse nel periodo di massimo fervore ideologico della psichiatria.

Questo studio è stato condotto nell'ambito del progetto PRIN 2022 PNRR DILIGO, acronimo di *D*igital *p*reservation, *L*inguistic *a*nalysis and *v*alorization *of* *h*istorical *a*rchive *of* *t*he *f*ormer *p*sychediatric *h*ospital *of* *G*irifalco, che ha l'obiettivo di recuperare e valorizzare l'accesso alle cartelle cliniche del manicomio di Girifalco attraverso processi di digitalizzazione, trascrizione, annotazione e analisi linguistica per indagare aspetti culturali, politici, sociali e sociolinguistici di fine XIX secolo e inizio XX secolo. La sottoscrizione di un accordo di collaborazione con l'ASP di Catanzaro ed il nulla osta della Soprintendenza Archivistica e Bibliografica della Calabria hanno permesso di avere accesso ai locali dell'archivio storico del manicomio di Girifalco per consultare le cartelle cliniche ivi conservate.

I paragrafi successivi sono articolati come segue: il secondo paragrafo illustra le fasi salienti della storia del manicomio di Girifalco, descrivendo il complesso archivistico costituito dalle cartelle cliniche e dettagliando le diverse tipologie documentali che le compongono. Nel terzo paragrafo viene descritta la metodologia di selezione e analisi del campione di cartelle cliniche, specificando le motivazioni della scelta di considerare solo i documenti clinici contenuti nelle cartelle cliniche. Infine, i paragrafi conclusivi riportano i risultati emersi dall'analisi effettuata, discutendone le implicazioni metodologiche e prospettando una possibile soluzione di rappresentazione evolutiva dei concetti clinici costitutivi della struttura dei documenti sanitari manicomiali e della conoscenza in essi veicolata.

2. Il manicomio di Girifalco ed il suo patrimonio archivistico

Nell'Italia meridionale post-unitaria si registravano scarsità di strutture dedicate all'accoglienza e alla cura dei pazienti affetti da patologie mentali ed inadeguatezza delle poche disponibili, specie in termini di capienza. I folli della regione venivano sistematicamente ricoverati presso il Real Manicomio di Aversa, in provincia di Caserta, fino a quando, nel luglio 1877, la struttura fu costretta a precludere ogni nuovo ricovero per mancanza di spazio. L'aumento delle richieste di ricovero in Calabria rese ancor più urgente la creazione di un

manicomio regionale. La scelta di Girifalco come sede del nuovo manicomio fu motivata sia da ragioni pratiche che da considerazioni teoriche. Dal punto di vista logistico, Girifalco occupava una posizione centrale nella Regione, agevolando il trasporto dei degenti; inoltre, la prossimità alla Corte di Appello di Catanzaro lasciava spazio alla prospettiva futura di istituire una sezione criminale del manicomio. Sul piano economico la presenza di un ex-convento dei frati Riformati, facilmente convertibile in un istituto di accoglienza, avrebbe permesso di contenere i costi iniziali di realizzazione della struttura. La scelta fu, poi, influenzata da alcune concezioni terapeutiche dell'epoca, secondo cui un «sorridente panorama campestre» (Marcello 1995, 69) era in grado di svolgere un'azione climatologica benefica sul sistema nervoso e regolarizzare il flusso delle idee. Così, il 18 settembre 1878 venne deliberata l'istituzione del Manicomio Provinciale di Girifalco, che aprì ufficialmente nel luglio del 1881 e rimase in attività fino al 1978, quando entrò in vigore la Legge Basaglia (Legge n. 180 del 13 maggio 1978) che sancì la definitiva chiusura dei manicomi in Italia. Questa istituzione fu un centro all'avanguardia nella cura della malattia mentale, grazie alla vivacità ideologica dei diversi luminari che vi prestarono servizio, fra gli altri si ricordano il direttore Silvio Venturi e l'alienista Marco Levi Bianchini. Quest'ultimo, in particolare, introdusse l'innovativa tecnica terapeutica dell'*open door*, che dava la possibilità ai malati di mente meno gravi di uscire dal manicomio durante il giorno, di integrarsi con la comunità cittadina locale e di lavorare nelle officine e nei laboratori che davano sostentamento al manicomio stesso, ribadendo così l'importante ruolo riabilitativo del lavoro.

Il manicomio di Girifalco fu operativo per quasi un secolo, registrando 22.415 ingressi e, attualmente, il suo archivio storico si compone di circa 15.800 cartelle cliniche. La disparità tra il numero totale di ingressi ed il numero delle cartelle cliniche che compongono l'archivio è legata ad ingressi multipli del medesimo paziente, allo smarrimento o al deterioramento di alcune di queste. Si presume che, prima dell'attuale ordinamento dell'archivio, le cartelle cliniche non fossero ordinate secondo il numero progressivo di matricola di ingresso del paziente, ma secondo un ordine progressivo distinto per "Uomini" e "Donne", seguendo lo stesso criterio adottato per la suddivisione dei reparti nel manicomio. Nonostante la parte più consistente dell'archivio sia rappresentata dalle cartelle cliniche, sono presenti anche quattro tipologie di registri, che rappresentano dei preziosi strumenti di corredo organizzati per comune e provincia di residenza, matricola d'ingresso e genere dei ricoverati. Inoltre, è possibile trovare i protocolli della corrispondenza, i fascicoli del personale sia in servizio che in quiescenza, la documentazione contabile e i carteggi riconducibili a vari uffici e settori dell'istituto, tra cui quello infermie-

ristico, la direzione sanitaria, la segreteria e i servizi tecnici¹. L'archivio storico, collocato nell'ex ospedale psichiatrico e conservato in buono stato, costituisce una risorsa documentaria di grande rilievo per la ricostruzione storica delle attività operative e della prassi documentale della struttura, in quanto raccoglie materiali prodotti nel corso dello svolgimento delle funzioni amministrative e cliniche dell'ente. Nel corso degli anni le cartelle cliniche del manicomio di Girifalco hanno suscitato un costante interesse da parte di studiosi e ricercatori provenienti da diversi settori della conoscenza. Fra i lavori di ricerca scientifica svolti, il progetto ALPHA (eAsy intelligent service Platform for Healthy Ageing), svolto in collaborazione tra l'IIT-CNR, l'ICAR (Istituto di Calcolo e Reti ad Alte Prestazioni)-CNR e il Centro Regionale di Neurogenetica (CRN) di Lamezia Terme (CZ) tra il 2014 ed il 2018, aveva l'obiettivo di individuare profili di rischio per l'insorgenza di patologie neurodegenerative attraverso l'analisi dei segni e dei sintomi riportati nelle cartelle cliniche di pazienti con lieve decadimento cognitivo. Contestualmente, è stata condotta un'analisi diacronica su pazienti affetti da demenza ricoverati a Girifalco che, nell'ambito di uno studio genealogico svolto dal CRN, ha permesso di tracciare la trasmissione ereditaria di alcune malattie neurodegenerative, tra cui la malattia di Alzheimer, contribuendo a retrodatare la prima attestazione clinica documentata di tale patologia (Borrello et al. 2016; Bruno et al. 2022). Per lo svolgimento di tali attività, sono state digitalizzate le cartelle cliniche relative al periodo 1881-1931, di cui le prime 550 (risalenti agli anni 1881-1894) sono state anche trascritte. In continuità rispetto al progetto ALPHA, si pongono le attività del PRIN DILIGO, nell'ambito del quale è stata svolta l'analisi descritta in questo contributo. DILIGO ambisce alla digitalizzazione di circa 3.000 cartelle cliniche relative al periodo 1931-1944 e all'esplorazione del metalinguaggio e del lessico relativo alla malattia e ai trattamenti durante la Prima guerra mondiale (1914-1918) ed il regime fascista (1922-1943). Le cartelle cliniche del manicomio di Girifalco sono delle testimonianze storiche dei cambiamenti intercorsi nella documentazione in relazione alle modalità di registrazione delle informazioni nel tempo e all'influenza delle teorie dell'epoca. Pertanto, lo studio della struttura delle tipologie documentali contenute nelle cartelle cliniche può offrire una prospettiva inedita per indagare come siano state concettualmente organizzate le informazioni, quali legami si siano instaurati tra i termini usati per identificare una sezione documentale e i concetti a questi connessi, e in che modo il linguaggio utilizzato abbia contribuito a veicolare alcune teorie.

¹ L'archivio storico risulta, ad oggi, parzialmente inventariato in Carte da Legare, grazie al lavoro di riconciliazione e riversamento dei metadati delle cartelle cliniche condotto dalla sede di Cosenza dell'Istituto di Informatica e Telematica del Consiglio Nazionale delle Ricerche (IIT-CNR).

2.1. Le cartelle cliniche e le tipologie documentali

Nel sistema manicomiale italiano, ogni evento legato al processo assistenziale dei pazienti doveva essere documentato in modo accurato. Questa prassi divenne norma con l'entrata in vigore della Legge n. 36 del 14 febbraio 1904 *Disposizioni sui manicomi e sugli alienati*. Tuttavia, anche senza riferimenti normativi precedenti, è plausibile ritenere che, già prima del 1904, le cartelle cliniche degli istituti manicomiali fossero compilate con criteri analoghi. A supporto di questa ipotesi si pone l'opera *L'antropologia nelle carceri*, pubblicata nel 1885 dagli psichiatri Tamburini e Benelli, nella quale veniva proposto un modulo per le notizie anamnestiche dei detenuti, somigliante per contenuti e struttura ai moduli clinici utilizzati nei manicomi. Nello stesso anno venne pubblicato anche il *Manuale di semeiotica delle malattie mentali* dello psichiatra Enrico Morselli, che si proponeva come guida per lo svolgimento dell'esame clinico del malato di mente. Infatti, il manuale articolava l'osservazione del paziente a partire dall'anamnesi, passando per l'indagine antropologica e fisiologica, fino ad arrivare all'analisi psicologica dei disturbi mentali. Di conseguenza, sia nelle cartelle cliniche proposte per i detenuti, sia in quelle redatte nei manicomi erano previste apposite sezioni dedicate alla rilevazione di misurazioni antropometriche degenerative, in accordo con le teorie lombrosiane dell'epoca (Lombroso 1876).

Al contempo, tuttavia, si delineava anche l'esigenza di adottare tipologie documentali simili all'interno delle strutture manicomiali in modo da uniformare le diagnosi e le informazioni raccolte durante il percorso diagnostico, favorendo così lo studio comparato di casi simili nella stessa struttura o afferenti a strutture diverse. Le cartelle cliniche svolgevano, quindi, in questo contesto, una funzione di supporto alla registrazione sistematica dei casi e alla rilevazione statistica dell'incidenza di specifici fenomeni patologici. Pertanto, anche se non è possibile attestare un utilizzo delle cartelle cliniche per fini didattici in senso stretto, la documentazione sanitaria in esse contenuta fu lungamente oggetto di consultazione da parte degli psichiatri interessati alla raccolta e all'analisi di casi di studio.

La cartella clinica rappresenta il contenitore di tutta la documentazione prodotta e acquisita dal momento dell'ammissione di un paziente e per tutta la durata della sua degenza nell'istituto manicomiale. Dunque, ogni tipologia documentale presente all'interno delle cartelle cliniche era generata nello svolgimento di una specifica funzione o attività, secondo una configurazione, data dalla specifica disposizione e natura delle informazioni veicolate (Bellotto 2004). Successivamente il Regio Decreto n. 615 del 16 agosto 1909, all'art. 63, lettera b, ribadì l'obbligo di redigere

un fascicolo personale per ciascun ricoverato, nel quale debbono essere conservati i documenti relativi all'ammissione, i provvedimenti, le comunicazioni

e la corrispondenza dell'autorità giudiziaria, di quella amministrativa e della famiglia, la diagnosi e il riassunto mensile delle condizioni dell'alienato, e gli atti relativi al licenziamento di esso per guarigione od in esperimento o per altra causa.

Queste tipologie di documenti sono contenute anche nelle cartelle cliniche del manicomio di Girifalco e la consistenza dei fascicoli personali dei pazienti varia in funzione della durata della degenza o della natura del ricovero. Tuttavia, ai fini del presente studio, vengono esclusivamente presi in considerazione ed analizzati i documenti clinici contenuti nelle cartelle, ovvero la modula informativa, la tabella nosologica e il diario clinico, di cui si riportano immagini esemplificative in Figura 1.

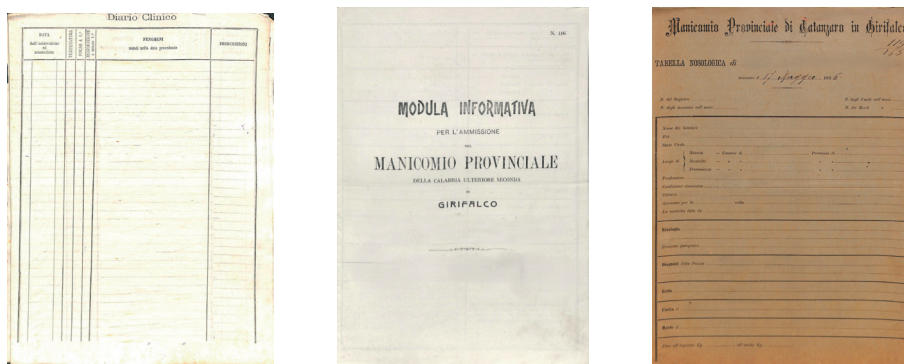


Figura 1. In ordine da sinistra, le prime pagine di modula informativa, tabella nosologica e diario clinico.

L'iter che conduceva al ricovero di un paziente aveva inizio con la presentazione del certificato medico e dell'atto di notorietà, che servivano per predisporre l'ammissione provvisoria. Le condizioni dell'alienato venivano attestate nella modula informativa, che era un questionario prestampato destinato a raccogliere, oltre alle generalità del paziente e alle sue condizioni economico-sociali, anche informazioni anamnestiche e relative all'ereditarietà della pazzia. Sebbene non sia possibile stabilire con esattezza quando le module informative siano state introdotte in modo sistematico nel panorama manicomiale italiano, per il manicomio di Girifalco si registra la presenza della modula informativa come documento accompagnatore al momento del ricovero dal 1884 in poi (Chiaravallotti e Taverniti 2021). Il ricovero del paziente doveva essere, poi, confermato o respinto entro trenta giorni dal momento dell'ammissione sulla base della relazione di valutazione del direttore del manicomio, come stabilito dall'art. 2 della Legge n. 36 del 1904.

Quando un malato di mente entrava in osservazione nel manicomio veniva eseguita una valutazione preliminare che dava luogo alla compilazione della tabella nosologica, o storia clinica, ovvero un documento interno del quale erano riportati i dati anagrafici del folle, arricchiti di ulteriori informazioni quali istruzione, professione, religione, numero di figli, condizione economica, eventuali ricoveri precedenti nella medesima struttura o in altre; a queste informazioni seguivano, poi, l'eziologia della malattia, la diagnosi, l'esito del ricovero. Le osservazioni di carattere sanitario comprendevano invece: un compendio storico della malattia, un breve resoconto dello stile di vita del paziente, indicazioni sul momento di insorgenza della malattia mentale ed episodi di recidività, l'indole del delirio, la possibile causa della follia. Una sezione particolarmente rilevante della tabella nosologica era, poi, quella dell'esame oggettivo, che comprendeva la registrazione delle misure antropometriche, l'esame di vari apparati, la valutazione delle facoltà intellettive, affettive, istintive e morali del paziente. Infine, il diario clinico fungeva da resoconto delle visite per tutte la durata della degenza del paziente all'interno della struttura manicomiale, consentendo di controllare e annotare periodicamente il comportamento dell'alienato, le cure e le terapie adottate, nonché le visite dei familiari.

Come per la modula informativa, anche per la tabella nosologica non è possibile definire con esattezza una data di introduzione; tuttavia la sua presenza tra le tipologie documentali di diversi archivi manicomiali consente di affermare che questa fosse già in uso nella seconda metà del XIX secolo². Le informazioni raccolte in questi documenti erano pressoché le stesse, cambiava la modalità di annotazione e l'analiticità delle informazioni registrate. Nel manicomio di Girifalco, invece, la tabella nosologica fu introdotta dal direttore Dario Maragliano nel 1880 e rientra, quindi, tra le tipologie documentali presenti sin dal primo ricovero registrato nella struttura nel 1881 (Chiaravalloti e Taverniti 2021). Anche il diario clinico, in quanto parte della tabella nosologica, è stato compilato già per i primi pazienti ammessi e, soprattutto all'interno di cartelle cliniche più antiche, si configura come una narrazione dettagliata della degenza dei pazienti all'interno della struttura ed è destinato alla registrazione giornaliera o periodica del comportamento del paziente (Chiaravalloti e Taverniti 2021).

Questa complessa stratificazione documentale indica come le cartelle cliniche e i documenti sanitari in esse contenuti abbiano rappresentato dei veri e propri dispositivi di organizzazione e formalizzazione della conoscenza, soprattutto nel dominio della psichiatria. I documenti sanitari conservati nelle cartelle del manicomio di Girifalco veicolano il sapere medico del XIX e del XX secolo, poiché traducono l'esperienza del ricovero e della cura in categorie

² Negli archivi storici dell'Ospedale Neuropsichiatrico di Bergamo e dell'ex ospedale psichiatrico San Servolo di Venezia, già nel XIX secolo si adottavano tipologie documentali come cedole e tabelle nosologiche, questionari anamnestici e diari clinici.

osservabili, registrabili e comparabili. In questo senso, dunque, la documentazione sanitaria considerata non solo riflette le pratiche cliniche in uso all'epoca, ma contribuisce a far emergere le modalità con cui la follia veniva descritta, interpretata e gestita.

3. Metodologia

I documenti storici, per loro natura, non sono mai statici e neutrali: la loro struttura rispecchia il modo in cui l'informazione è stata organizzata e utilizzata in un dato momento storico, rappresentando anche una testimonianza delle funzioni da essi assolte. Poiché queste funzioni cambiano nel tempo in risposta a nuovi bisogni istituzionali e culturali, anche la forma del documento si trasforma per rispondere a tali esigenze. In questo senso, la forma conserva e trasmette nel tempo la funzione stessa del documento, garantendone l'integrità del significato veicolato (Duranti 1997). Il carattere diacronico di queste mutazioni strutturali rende complessa l'analisi automatica del loro layout in periodi diversi, poiché tracciare tali cambiamenti non è semplice, soprattutto nel caso di strutture grafiche poco prevedibili e che non seguono dei modelli standardizzati comuni ad altri documenti storici. Infatti, i modelli automatici di analisi e riconoscimento dei layout funzionano in modo accurato con strutture regolari e prevedibili, come ad esempio quelle riscontrate nell'impaginazione dei giornali, e quando si prendono in considerazione periodi di tempo ben definiti (Gutehrle e Atanassova 2022). Gli strumenti di analisi automatica del layout risultano quindi poco adatti a un patrimonio documentale ampio ed eterogeneo, soggetto nel tempo a cambiamenti strutturali di varia natura, come le cartelle cliniche del manicomio di Girifalco. Per questa ragione, nel presente studio si è scelto di adottare un approccio manuale e qualitativo, in grado di valorizzare le peculiarità dei documenti clinici storici presi in esame. A differenza degli strumenti automatici, che mostrano una limitata flessibilità nell'adattarsi alle variazioni strutturali dei documenti nel tempo e non sono in grado di riconoscere possibili trasformazioni semantiche nelle denominazioni delle sezioni, l'approccio manuale offre la possibilità di prestare attenzione consapevole e contestualizzata ai cambiamenti. In questo contesto, l'impiego di un approccio manuale consente di cogliere non solo i cambiamenti formali, ma anche sfumature sottili di significato che possono emergere con il mutare delle pratiche mediche e del contesto d'uso. In tal modo, è possibile ridurre fraintendimenti o semplificazioni che gli strumenti automatici, ancorati a regole fisse e modelli standardizzati, tendono ad introdurre, migliorando conseguentemente l'analisi dei contenuti di questo patrimonio documentale così complesso e stratificato. L'analisi della struttura delle cartelle cliniche assume, quindi, un'importanza centrale per comprendere come la pratica medica abbia plasmato la documentazione clinica e psichiatrica tra il XIX ed il XX secolo,

rivelando così i mutamenti nelle modalità di rappresentazione e categorizzazione delle informazioni in questo dominio della conoscenza.

La metodologia adottata per la realizzazione del presente studio si è basata su un approccio manuale, mirato a derivare, ricostruire e comprendere la loro struttura logica. L'obiettivo primario è stato quello di identificare gli elementi strutturali e le categorie informative ricorrenti all'interno dei documenti clinici, come titoli, intestazioni, paragrafi e tabelle. Queste componenti, che concorrono a definire la forma del documento, non sono solo meri elementi grafici, poiché essi riflettono e supportano le funzioni che il documento era chiamato a svolgere. Attraverso l'analisi manuale del layout, è stato possibile ricostruire il ruolo svolto dalla struttura logica delle cartelle cliniche nello svolgimento delle loro funzioni e nell'organizzazione delle informazioni che esse contengono. L'attività di digitalizzazione delle cartelle cliniche, prevista nell'ambito del PRIN DILIGO, ha permesso di avere maggiore contezza delle loro variazioni strutturali. Ad esempio, si è rilevato che nel fascicolo dei pazienti con degenze multiple vi è la coesistenza di più versioni della stessa tipologia documentale, rispondenti ai modelli adottati nei diversi anni di ammissione, mentre nei fascicoli dei pazienti con degenze prolungate permane lo stesso modello di cartella clinica anche negli anni in cui ne sarebbe stato adottato uno diverso per le nuove ammissioni. Ciò avviene sostanzialmente perché nelle degenze prolungate vi era la tendenza ad aggiornare progressivamente la sezione di diario clinico già presente e si avviava all'esaurimento dello spazio a disposizione, aggiungendo ulteriori fogli mediante spille e graffette. Queste constatazioni hanno permesso di prestare maggiore attenzione nell'analizzare i documenti in relazione ai periodi di degenza o al numero e alla data dei ricoveri, rendendo tuttavia difficile stabilire con precisione il momento in cui è avvenuto il passaggio da un formato all'altro. Inoltre, poiché si è riscontrato che le cartelle cliniche mantengono una struttura relativamente stabile nel breve-medio periodo, un'analisi generale della serie cartelle cliniche si sarebbe rivelata di limitata utilità conoscitiva e ridondante, comportando un significativo dispendio di tempo per la loro revisione a fronte di una sostanziale omogeneità documentale. Tali considerazioni hanno evidenziato la necessità di introdurre un criterio di selezione delle cartelle cliniche da sottoporre ad analisi su base temporale, al fine di individuare con maggiore efficacia casi rappresentativi dei cambiamenti occorsi e quindi significativi ai fini dello studio. Si è deciso quindi di eseguire un campionamento delle cartelle cliniche su base quinquennale, a partire dal 1881, anno del primo ricovero. Al campione sono state inoltre aggiunte le cartelle cliniche dei ricoverati nell'anno 1978, anno dell'entrata in vigore della Legge Basaglia, in quanto ritenuto di particolare rilevanza storica e sociale per la psichiatria italiana. Questo arco cronologico è stato riconosciuto adeguato e sufficiente per tracciare eventuali mutamenti nel layout delle cartelle cliniche, riuscendo a rilevare i cambiamenti più significa-

tivi nella loro struttura. Il campione di analisi, così selezionato, consta quindi di 2.290 cartelle cliniche, con una media di 120 cartelle analizzate per anno, selezionate in maniera randomica a inizio, metà e fine dell'anno analizzato e prendendo in considerazione sia le cartelle cliniche di uomini che di donne, senza una predeterminazione sul rapporto numerico fra queste due categorie.

Per le finalità del presente lavoro, l'analisi si è focalizzata esclusivamente sulle tipologie documentali strettamente connesse alla pratica clinica: le module informative, le tabelle nosologiche e i diari clinici. Queste sono state considerate di particolare interesse e rilevanza, poiché capaci di far emergere i processi di gestione della conoscenza clinica nel dominio psichiatrico nel corso del tempo. Infatti, sebbene le altre tipologie documentali offrano informazioni preziose sull'iter amministrativo del ricovero, non consentono un'analisi sistematica dei processi di produzione, formalizzazione e trasformazione delle conoscenze mediche, principalmente per la limitata occorrenza di termini medici, usati per lo più a fini amministrativi e quindi non significativi per le finalità di questo studio.

Le cartelle cliniche oggetto del campionamento sono state esaminate e analizzate senza che sia stato effettuato alcun trattamento dei dati personali, in conformità con quanto previsto dalla normativa vigente in materia di protezione dei dati personali.

L'approccio manuale e qualitativo scelto contribuisce alla valorizzazione della conoscenza delle cartelle cliniche del manicomio di Girifalco, migliorando la comprensione dei processi di categorizzazione delle informazioni ivi contenute, favorendone la predisposizione di strumenti di accesso per garantire una maggiore fruibilità dei contenuti.

4. Risultati e discussione

L'analisi del layout dei documenti sanitari ha messo in luce le trasformazioni della struttura e dei contenuti delle cartelle cliniche del manicomio di Girifalco avvenute nel corso del tempo. Queste trasformazioni, al pari delle teorie scientifiche che vi sottendono, non sono state rapide o improvvise, ma si sono sviluppate e affermate in lenta progressione, determinando così l'utilizzo di diverse versioni della stessa tipologia documentale nello stesso lasso temporale.

L'analisi comparativa tra le diverse versioni dei documenti sanitari considerati ha permesso di identificare significative variazioni sia a livello di organizzazione che di rappresentazione lessicale delle informazioni cliniche, riscontrando la presenza di diversi formati e varianti della stessa tipologia di documento nello stesso periodo. Tuttavia, tali mutamenti non sono ascrivibili a meri aggiornamenti formali: in linea con quanto osservato negli studi critici sull'organizzazione della conoscenza, le modifiche nelle strutture e nelle etichette documentali riflettono dinamiche culturali, istituzionali e professionali,

nonché orientamenti teorici propri di ciascun periodo storico (Ibekwe 2024), sottolineando come la «Knowledge Organization deals with the inherent tension between the non-neutrality of representation of knowledge and the commitment to respond promptly and accurately to the demands and information needs of different knowledge domains and user communities» (Oliveira Milani et al. 2024, 610). Analizzare come cambia una tipologia documentale nel tempo significa, quindi, individuare strutture e modalità con cui i concetti sono stati categorizzati, nonché le connessioni tra le informazioni registrate ed il contesto storico. Pertanto, il processo di analisi e ricostruzione delle trasformazioni documentali costituisce un'operazione di organizzazione della conoscenza e, come tale, richiede un'attenta riflessione sui criteri di interpretazione delle fonti, poiché il rischio è quello di sovrapporre inconsapevolmente modelli interpretativi contemporanei a strutture documentali storiche, compromettendo la comprensione del loro significato originario e del contesto in cui sono state prodotte.

Lo studio condotto mira, quindi, ad elaborare esclusivamente una analisi qualitativa delle tipologie documentali selezionate con l'obiettivo di evidenziare la loro evoluzione strutturale nel tempo. Non sono stati estrapolati dati quantitativi relativi al campione esaminato – come, ad esempio, il numero totale di documenti riconducibili a uno stesso layout – poiché, avendo adottato un criterio di campionamento di tipo temporale, essi non risulterebbero rappresentativi dell'incidenza cronologica di una determinata versione documentale e non contribuirebbero a ricostruire le dinamiche evolutive delle diverse tipologie di documenti. Una analisi quantitativa per valutare l'incidenza delle varie versioni documentali potrebbe risultare esaustiva e rappresentativa solo mediante una revisione sistematica dell'intero corpus delle cartelle cliniche.

Dall'analisi delle cartelle cliniche con più ricoveri emerge, inoltre, una prassi di continuità nella documentazione sanitaria in cui ogni nuovo ingresso prevede la redazione di una nuova modula informativa, mentre tabella nosologica e diario clinico vengono solo aggiornati. Ciò avviene perché la modula informativa è redatta dal medico condotto per giustificare ogni nuovo ricovero, al contrario tabella nosologica e diario clinico, invece, sono documenti interni al manicomio e, quindi, il solo aggiornamento risulta propedeutico alla cura secondo una visione organica ed unitaria.

I risultati emersi dall'analisi delle tre tipologie documentali prese in considerazione – modula informativa, tabella nosologica e diario clinico – sono di seguito presentati, con l'obiettivo di mettere in luce i cambiamenti, più o meno marcati, intervenuti nel corso del tempo a livello strutturale, terminologico e di formalizzazione delle informazioni.

4.1. Modula informativa

Fino agli anni Venti del Novecento, la modula informativa era un bifoglio, in cui la prima pagina fungeva da frontespizio, riportando l'intestazione dell'istituto manicomiale e talvolta il nominativo del paziente, mentre le tre pagine successive erano destinate alla raccolta dei dati anagrafici, anamnestici e clinici del paziente. Queste hanno una struttura tabellare essenziale, articolata su due colonne: la colonna di sinistra conteneva le etichette prestampate dei campi, mentre quella di destra era riservata alla compilazione manuale del medico condotto del paese d'origine del paziente. Dalle immagini riportate in Figura 2 si evince che, per ragioni grafiche era stato assegnato uno spazio specifico per tutti i campi da compilare, il quale, però, non risulta sempre sufficiente o necessario per le informazioni inserite. Considerando che, in quegli stessi anni, la struttura delle cartelle cliniche da adottare nelle carceri fu teorizzata e ideata da psichiatri, è plausibile che anche queste tipologie documentali siano state strutturate nel dettaglio, senza tuttavia considerare le diverse esigenze espressive. Ad esempio, campi come *condizione economica* erano quasi sempre popolati con poche parole, mentre altri come *ereditarietà della pazzia* erano valorizzati con testi lunghi ed articolati. In alcuni casi, le descrizioni sono così dettagliate da assumere la forma di veri e propri racconti, che finiscono per riempire quasi interamente la colonna, lasciando poco o nessuno spazio ad altre informazioni, che vengono poi aggiunte successivamente o annotate ai margini. Al contrario, il campo *cura praticata, tolleranza ed esito avuto*, disponeva di uno spazio spesso superiore al necessario, presumibilmente perché, al momento dell'ammissione, raramente i pazienti avevano già seguito terapie specifiche e continuative, anche nel caso di ricoveri pregressi. Infatti, nel periodo compreso tra la fine del XIX e gli inizi del XX secolo, la follia costituiva un pesante stigma sociale e si registrava l'assenza di strumenti terapeutici idonei a garantire un'effettiva riabilitazione dei malati di mente, pertanto, l'internamento nei manicomi era finalizzato principalmente all'isolamento sociale e alla reclusione (Goffman 1968).

Nelle prime versioni della modula informativa, alcune etichette erano costituite da sequenze di quesiti, probabilmente concepite per orientare la raccolta delle informazioni di interesse. Si tratta spesso di formulazioni articolate e accompagnate da esempi, come nel caso del campo relativo all'ereditarietà familiare (Fig. 2). Le etichette diagnostiche rispondevano alla necessità di rendere la raccolta dei dati più sistematica ed efficiente, guidando la compilazione dei diversi medici condotti dei paesi di provenienza dei pazienti. Indistintamente dall'area territoriale, quindi, era fondamentale uniformare il più possibile la tipologia delle informazioni raccolte, affinché gli psichiatri della struttura potessero procedere agevolmente con la valutazione del paziente al momento dell'ammissione.

Se il paziente ha avuto precedenti disturbi di natura psichica	
Se il paziente ha avuto precedenti disturbi di natura fisica	
Se il paziente ha avuto precedenti disturbi di natura sociale	
Se il paziente ha avuto precedenti disturbi di natura familiare	
Se il paziente ha avuto precedenti disturbi di natura sessuale	
Se il paziente ha avuto precedenti disturbi di natura alimentare	
Se il paziente ha avuto precedenti disturbi di natura di altro tipo	
Se il paziente ha avuto precedenti disturbi di natura di altro tipo	
Se il paziente ha avuto precedenti disturbi di natura di altro tipo	
Se il paziente ha avuto precedenti disturbi di natura di altro tipo	

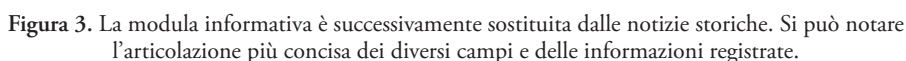
Se il paziente ha avuto precedenti disturbi di natura psichica	
Se il paziente ha avuto precedenti disturbi di natura fisica	
Se il paziente ha avuto precedenti disturbi di natura sociale	
Se il paziente ha avuto precedenti disturbi di natura familiare	

Se il paziente ha avuto precedenti disturbi di natura psichica	
Se il paziente ha avuto precedenti disturbi di natura fisica	
Se il paziente ha avuto precedenti disturbi di natura sociale	
Se il paziente ha avuto precedenti disturbi di natura familiare	
Se il paziente ha avuto precedenti disturbi di natura sessuale	

Figura 2. Le tre pagine interne della modula informativa, la cui compilazione era demandata al medico condotto del paese di provenienza del paziente.

Il linguaggio utilizzato, pur essendo medico-descrittivo, riflette pregiudizi culturali dell'epoca, come si evince dall'equiparare *beoni* o *delinquenti* a individui affetti da patologie nervose.

Sul finire degli anni Venti, la modula informativa assume la nuova denominazione di notizie storiche e, riducendosi a due sole pagine, denota un progressivo processo di sintesi e tecnicizzazione delle etichette dei campi (Fiorino 2010). Il confronto tra le diverse versioni della stessa tipologia documentale mette in evidenza sia elementi di continuità sia alcune trasformazioni. L'elemento di continuità più evidente è la natura essenziale e schematica del layout, che mantiene un impianto tabellare mostrando coerenza con la funzione originaria del documento di privilegiare l'efficacia comunicativa. Il cambiamento più evidente è la sintesi nelle etichette dei campi, che ha reso la compilazione più scarna, pur mantenendo invariata la sequenza e la tipologia delle informazioni richieste. Questo è evidente riprendendo proprio l'esempio del campo relativo all'ereditarietà della malattia mentale, poiché le stesse informazioni vengono raccolte sotto la nuova e più sintetica voce *Se altri in famiglia siano stati alienati, od abbiano sofferto di malattie nervose e quali*, come mostrato nella Figura 3.



Queste trasformazioni non sono dunque affatto neutre: la riformulazione delle etichette indica una mutata modalità con cui l'informazione viene raccolta, poiché si orienta il tipo di risposta attesa e, di conseguenza, si contribuisce alla diversa rappresentazione della conoscenza medica. L'abbandono delle descrizioni discorsive e l'adozione di etichette stereotipate comportano la perdita di una certa ricchezza narrativa, ma, al contempo, rispecchiano l'insorgenza di modelli interpretativi e operativi della psichiatria, che cerca di definire un approccio più sistematico e affermarsi come branca riconosciuta delle scienze mediche.

In sintesi, il passaggio da una modula informativa più descrittiva al modulo delle notizie storiche, più sintetico e tecnico, riflette l'evoluzione della disciplina stessa, che nel tempo affina i propri strumenti e chiarisce i criteri di rilevanza diagnostica.

4.2. Tabella nosologica

La tabella nosologica è un modulo composto da più sezioni, organizzate in sottosezioni, ognuna con intestazioni e campi prestampati, organizzati graficamente in spazi ben definiti da compilare. Questo strumento riflette i principi propri della nosologia, disciplina che classifica le malattie in base alla causa, ai meccanismi di insorgenza, ai sintomi e agli apparati coinvolti.

Dalle analisi a campione è emerso che le tre versioni del documento più rappresentative dei mutamenti intercorsi in questa tipologia documentale

sono risalenti agli anni 1881, 1931 e 1978. Si è, pertanto, deciso di focalizzare l'analisi comparativa sulle cartelle cliniche dei ricoveri avvenuti in questi anni. Esse evidenziano come la tabella nosologica sia il documento soggetto alle trasformazioni più rilevanti, sia in termini di estensione (raggiungendo sino alle dieci pagine), sia per modifiche nelle intestazioni e nella nomenclatura dei campi. I tre anni oggetto di analisi consentono di focalizzarsi su momenti chiave nella storia del manicomio di Girifalco: il 1881 corrisponde all'anno di apertura della struttura permettendo di riflettere su come le informazioni fossero raccolte e categorizzate all'inizio delle sue attività; il 1931 è stato scelto poiché intermedio nel periodo di attività del manicomio e, infine, il 1978 è l'anno di chiusura formale del manicomio e offre, quindi, la possibilità di valutare come la riforma psichiatrica abbiano mutato le modalità di considerare e trattare le malattie mentali.

La prima versione analizzata (1881) consta di quattro pagine distinte: nella prima venivano annotate le generalità del paziente, l'*eziologia*, la *diagnosi della pazzia* e l'*esito* del ricovero. Segue la sezione relativa all'*anamnesi della famiglia e del malato*, in cui, dopo un ampio spazio per annotazioni libere, vi è la sezione *stato presente*, dedicata all'esame clinico, contenente, a sua volta, la sezione *generalità ed antropometria*. Quest'ultima sezione è articolata in due parti: la prima relativa ad annotazioni circa *statura, costituzione, nutrizione, anomalie di sviluppo e di forma*; e la seconda dedicata alle misure craniometriche dettagliate in vari campi come *indice cefalico, altezza della fronte* e così via. Annotazioni di questo tipo consentono di riflettere su quanto, all'epoca, la psichiatria adottasse un approccio sperimentale, influenzato dalle teorie contingenti legate ad una concezione organicistica e biologica della malattia mentale. La terza pagina era dedicata alle *funzioni psichiche*, sezione articolata in otto campi accompagnati da diciture esplicative che ne orientavano la compilazione (Fig. 4).

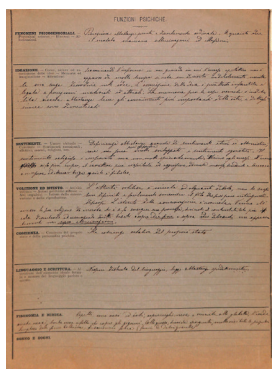


Figura 4. Terza pagina di una tabella nosologica del 1881. Si notano in dettaglio i campi contenuti nella sezione *funzioni psichiche*.

La quarta pagina era articolata in due parti, la prima dedicata all'annotazione delle *funzioni di relazione (sensibilità e motilità)* in cui si riscontra l'utilizzo di aggettivi modificatori caratterizzanti per la costruzione di campi che hanno una relazione gerarchica con la sezione di cui sono parte. E la seconda parte relativa alle *funzioni vegetative* contenente informazioni come *circolazione, respirazione, digestione*. Segue il diario clinico che, riconosciuto come documento a sé stante, verrà analizzato in seguito (si veda Paragrafo 4.3. Diario clinico).

In alcune cartelle cliniche, la versione appena descritta della tabella nosologica coesiste con un'altra probabilmente introdotta per ragioni politico-amministrative tra il 1882 ed il 1885. I due modelli si differenziano per l'intestazione: la versione sopra descritta, più antica, presenta l'intestazione *Manicomio Provinciale della Calabria Ulteriore Seconda in Girifalco*, mentre quella successiva è intestata *Manicomio Provinciale di Catanzaro in Girifalco*. Il modello rimane concettualmente uguale, anzi, il contenuto dei campi risulta essere spesso fedelmente copiato a mano.

La versione del 1931, intestata *Ospedale Psichiatrico Provinciale di Catanzaro in Girifalco*, mostra un cambiamento terminologico che riflette il riconoscimento della psichiatria come disciplina medica autonoma. Inoltre, rispetto alle versioni precedenti si nota un cambiamento di denominazione da *tabella nosologica* a *storia clinica*, indicando una maggiore attenzione all'anamnesi e all'evoluzione del quadro clinico. La storia clinica comprende, in questa versione, ben otto pagine che sottolineano l'importanza attribuita nel contesto diagnostico alla raccolta di dati relativi allo stile di vita e alle caratteristiche psico-fisiche del paziente. Nella prima pagina permangono le generalità del paziente con campi sintetici a compilazione standardizzata, si aggiunge però una sezione a testo libero intitolata *Anamnesi familiare ed individuale – inizio e corso della malattia – come si presenta l'infermo al momento dell'ammissione* e uno spazio dedicato all'apposizione delle fotografie del malato. Nella seconda pagina la sezione denominata *Stato attuale* riporta gli stessi campi di misurazione della sezione *Stato presente* del 1881 a cui seguono le due sezioni *cranio e faccia*. Segue una sezione inedita rispetto alla versione precedente *Note antropologiche degenerative*, che amplia e approfondisce l'esame delle caratteristiche somatiche ed organiche di matrice lombrosiana, testimoniando quanto il loro influsso fosse ancora pienamente radicato nella pratica psichiatrica. Vi rientrano campi che indagano le *degenerazioni* nelle varie parti del corpo. Segue la sezione dedicata alla *Vita vegetativa* (denominata *Funzioni vegetative* nel 1881), che si compone di vari campi quali *torace, addome, sistema linfatico e sanguigno* oltre al campo *siero diagnosi* riservato all'annotazione di test di diverse patologie, tra cui la reazione di *Wassermann* per sifilide e altri test eseguiti, in particolare, per tubercolosi e tifo. Segue poi la sezione *Vita di relazione* in cui sono annotate le diverse tipologie di sensibilità testate. In essa innovativa è la presenza di un disegno dello schema corporeo, atto a supportare graficamente

l'indicazione delle varie sensibilità corporee e un diagramma del campo visivo o grafico perimetrico accompagnato al campo sensibilità visiva per meglio annotare le informazioni dell'esame del campo visivo (Fig. 5).



Figura 5. Quinta pagina di una tabella nosologica del 1931. Si notano in dettaglio le illustrazioni mediche a corredo delle sezioni.

La pagina successiva contiene le sezioni *riflessi*, con tre sottosezioni e la sezione *motilità* che, rispetto alla versione precedente comprendente soli tre campi, si specializza ulteriormente espandendosi fino a dodici. L'ultima pagina comprende la sezione *funzioni psichiche* che diventa una sezione ad annotazione libera ad eccezione di tre campi: *diagnosi clinica*, *prognosi* e *cura*, i quali occupano uno spazio molto ridotto in basso nella pagina.

Nella tabella nosologica del 1978 le dimensioni tornano ad essere ridotte e consta infatti di quattro pagine. La prima pagina rimane dedicata alla *storia clinica* del paziente ed i campi rimangono invariati per significato, pur assumendo denominazioni diverse. Si segnala l'introduzione di due nuovi campi di *ammissione* e *dimissioni*, graficamente ampliati rispetto al passato, quando tali dati venivano annotati ai margini, conferendo così al documento maggiore ordine e chiarezza. La seconda pagina si apre con la sezione di anamnesi che cambia rispetto al passato: da spazio ad annotazione libera, viene esploso in tre campi standardizzati: *anamnesi familiare*, *anamnesi personale* e *modo di insorgenza dell'attuale malattia*. Segue poi nella stessa pagina la sezione dell'*esame clinico* contenente tre sottosezioni *esame antropologico*, in cui venivano annotate le misurazioni; *esame tegumentario*, dedicato ad annotazioni dell'apparato cutaneo e utile per rilevare la presenza di malattie infettive; e, infine, nella pagina successiva la parte relativa all'*esame delle funzioni vegetative*. La terza pagina contiene l'innovativa sezione *esame neurologico*, sinora mai prevista, e che rappresenta un passo significativo verso una valutazione più sistematica del paziente psichiatrico. Storicamente, la psichiatria aveva spesso trattato il disturbo mentale separandolo dalla funzione neurologica, mentre a partire dal secondo dopoguerra si afferma l'importanza di distinguere patologie neurologiche vere

e proprie da sindromi psichiatriche funzionali. Anche la sezione *esami di laboratorio*, che occupa l'ultima pagina, riveste un'importanza significativa (Fig. 6). Infatti, per la prima volta, gli esami ematochimici e sierologici trovano uno spazio strutturato all'interno del documento clinico, abbandonando la collocazione marginale che avevano nelle versioni precedenti, nelle quali erano prodotti a corredo della tabella nosologica, redatti su fogli volanti successivamente spillati o raccolti insieme. Questa scelta riflette una progressiva integrazione tra osservazione psichiatrica e strumenti diagnostici propri della medicina generale, coerente con l'evoluzione scientifica registrata nel dominio sanitario, a seguito dell'introduzione degli psicofarmaci negli anni Cinquanta.

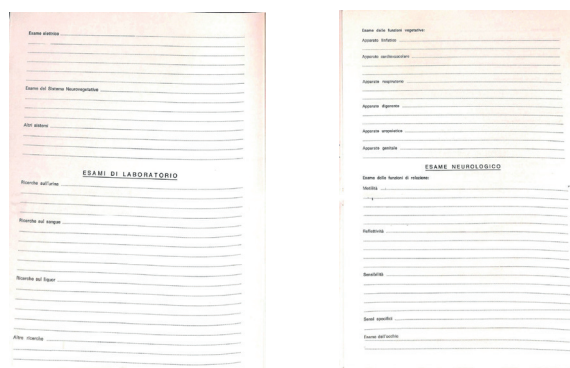


Figura 6. Sezioni esame neurologico ed esami di laboratorio in una tabella nosologica del 1978.

L'analisi dell'evoluzione delle tabelle nosologiche rivela un processo di maturazione ontogenetica che riflette lo sviluppo delle capacità organizzative e concettuali della disciplina stessa. Il passaggio dalle prime tabelle nosologiche alle versioni più mature corrisponde ad una progressiva strutturazione e progressione del pensiero medico-psichiatrico. Nelle versioni iniziali, molte informazioni erano raccolte in modo disorganizzato, spesso sovrapponendosi. Con il passare del tempo, queste vengono progressivamente incanalate in sezioni e campi precisi. Il processo più evidente riguarda, però, la trasformazione del linguaggio utilizzato nella documentazione clinica; infatti, le descrizioni inizialmente lunghe e dettagliate, si trasformano progressivamente in formulazioni concise. Emergono frasi predefinite e campi strutturati che permettono di veicolare la stessa quantità di informazione con maggiore efficienza linguistica (Fiorino 2010). Questo processo di progressiva contrazione presenta caratteristiche di economia lessicale legate ai tempi dettati dalla pratica e al consistente numero di campi da valorizzare: molti campi vengono lasciati sistematicamente in bianco o compilati con notazioni estremamente sintetiche, e spesso ambigue, come *normale* o *idem*. Questa riorganizzazione rende la struttura del documento più leggibile e razionale, facilitando il confronto tra casi clinici

simili. Parallelamente, si assiste allo sviluppo di una vera e propria ontogenesi delle capacità descrittive e categoriali del documento. Le informazioni non sono più semplicemente elencate, ma messe in relazione tra loro, favorendo la formazione di reti semantiche interne al linguaggio medico. Per esempio, all'interno del campo *ideazione* si distinguono sotto-aspetti come *corso delle idee*, *natura*, *associazione*, che non rappresentano categorie isolate, ma termini funzionali all'organizzazione di un quadro diagnostico coerente. La comparsa di una terminologia tecnica più specifica e condivisa consente di costruire connessioni tra concetti (sintomo-funzione-diagnosi), riducendo ambiguità e discontinuità.

4.3. Diario clinico

Il diario clinico si presenta come un documento a compilazione progressiva, aggiornato cronologicamente con osservazioni riguardo lo stato fisico e psichico del paziente. La sua struttura si è progressivamente evoluta assumendo layout differenti, pur restando essenziale e narrativa. In particolare, sono state rilevate cinque versioni di questa tipologia documentale e dal loro confronto, si riscontra una sostanziale continuità strutturale, caratterizzata da un layout che permane pressoché invariato nei suoi elementi fondamentali: una griglia semplice per la descrizione della degenza del paziente e, in alcune versioni, campi preimpostati per l'annotazione della data di registrazione dell'osservazione e delle cure praticate.

L'impaginazione è priva di illustrazioni e pensata per la raccolta di testo manoscritto. La prima versione analizzata, datata 1881 (Fig. 7) presenta intestazione *diario clinico* e denota un impianto articolato in colonne verticali con etichette ben definite. Il primo campo è dedicato alla *data*, a sua volta articolato in tre colonne verticali per giorno, mese ed anno di annotazione. Si trovano poi tre campi, *temperatura*, *polso* e *respirazione*, probabilmente così strutturati per consentire una rapida comparazione dei parametri vitali del paziente nel corso della degenza, ma questi campi non sono stati mai compilati. Il campo denominato *fenomeni notati nella data precedente*, al contrario, si configura come uno spazio aperto alla narrazione clinica, destinato all'annotazione libera delle osservazioni comportamentali e psichiche del paziente. L'utilizzo del termine *precedente* può trarre in inganno: questo si riferisce esclusivamente alla posizione della data all'interno della pagina rispetto al resoconto delle osservazioni effettuate e non vi è alcun riferimento ai fenomeni notati in precedenza. Le annotazioni e le descrizioni sullo stato di salute dei pazienti, in questa fase, si presentano come lunghe, articolate e ben dettagliate, estendendosi per diverse pagine. Il campo *prescrizioni*, infine, è destinato alla registrazione sintetica dei trattamenti somministrati nella data considerata, ma è compilato raramente.

Nella seconda versione del documento, riportata in Figura 7 e relativa all'anno 1886, si osserva una progressiva semplificazione del layout. Le etichette della prima versione vengono ridotte e permangono solo i riferimenti temporali *anno*, *mese* e *giorno*. La motivazione di questo cambiamento strutturale è probabilmente da ricondursi al mancato utilizzo e alla scarsa valorizzazione dei campi precedentemente previsti. Le informazioni più rilevanti sono ugualmente annotate sotto forma di testo manoscritto ed in linguaggio naturale. Inoltre, in questa nuova configurazione, per via della sua disposizione all'interno della pagina, l'etichetta *diario clinico* sembra avere la funzione di campo e non quella di intestazione della tipologia documentale. Per l'anno 1886 si nota la coesistenza di una seconda versione del diario clinico, come evidenziato in Figura 7. Questa presenta intestazione *Decorso e cura* e ha un layout ancora più essenziale, costituito da righe orizzontali, che guidano la scrittura, e da una riga verticale, dedicata alla separazione tra la data di compilazione e l'annotazione.

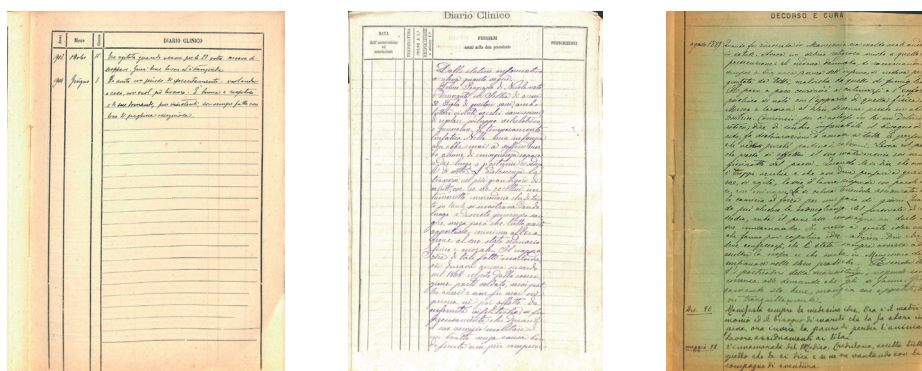


Figura 7. Esempi di diario clinico, uno del 1881 a sinistra e due del 1886 a destra.

Le versioni del documento relative agli anni 1931 e 1956 (Fig. 8) segnano un abbandono definitivo della struttura tabellare in favore di una struttura più libera per la gestione di annotazioni sintetiche nella pagina. Si assiste, infatti, a una progressiva trasformazione: dalle prime versioni, caratterizzate da lunghi resoconti narrativi e dettagliati, si passa nelle versioni più recenti a registrazioni sintetiche, spesso strutturate in elenchi di date, corredate da annotazioni brevi e sempre più orientate alla rilevazione di parametri vitali e alla registrazione delle somministrazioni farmacologiche.

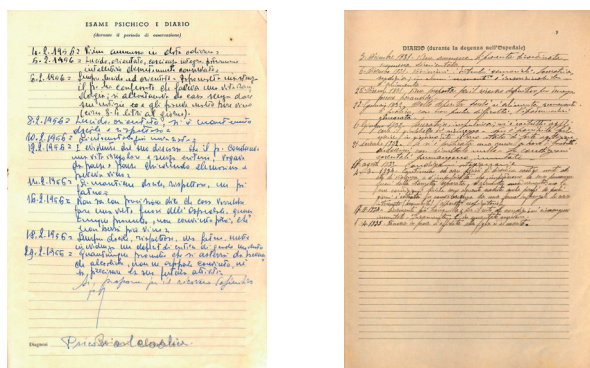


Figura 8: Esempi di diario clinico relativi rispettivamente agli anni 1931 e 1956.

Dall'esame delle diverse versioni del documento si evidenzia, in particolare, come le intestazioni siano cambiate nel tempo, riflettendo la diversa rilevanza e finalità attribuita al diario nel tempo. In ordine temporale, la progressione nella denominazione delle intestazioni va dal più generico *Diario clinico* (1881), alla coesistenza delle versioni *Decorso e cura* (1886) e *Decorso e cura della psicopatia* (1886), per arrivare poi a *Diario (durante la degenza in ospedale)* (1931) e, infine, a *Esame psichico e diario* (1956). Con quest'ultima evoluzione terminologica, il documento assume una duplice funzione, diagnostica e narrativa, testimoniando un cambiamento funzionale da spazio di osservazione soggettiva ad un documento che integra narrazione e dati clinici. Questa trasformazione è da intendersi non solo come un'evoluzione linguistica ma anche strutturale: nel tempo, si sviluppano e si affinano le capacità di rappresentare e utilizzare le parole in modo sempre più astratto, efficiente e coerente (Tennis 2006; Tognoli et al. 2019). In questo senso, l'evoluzione del diario clinico non comporta una perdita di contenuto, ma testimonia piuttosto una trasformazione verso una rappresentazione più efficace e operativa della conoscenza clinica.

5. Conclusioni

Il presente studio ha inteso indagare l'evoluzione della struttura dei documenti sanitari contenuti nelle cartelle cliniche del manicomio di Girifalco per rintracciare nelle evoluzioni formali della documentazione clinica anche l'evolversi dell'organizzazione delle conoscenze mediche, e psichiatriche in particolare, tra il XIX ed il XX secolo. In tal modo, lo studio contribuisce a sviluppare un innovativo percorso interpretativo nella rappresentazione della conoscenza di documenti clinici storici, contribuendo alla loro caratterizzazione.

I documenti sanitari analizzati – modula informativa, tabella nosologica e diario clinico – vanno quindi oltre la mera funzione di registrazione del dato clinico nella prassi manicomiale, configurandosi come testimonianze storiche delle diverse modalità di comprensione e classificazione delle malattie mentali e dei cambiamenti occorsi nella pratica medica della disciplina psichiatrica. Dall'analisi effettuata è emerso che la trasformazione della struttura dei documenti sanitari è stata lenta e graduale e, lungi dall'essere uniforme, è stata piuttosto frammentaria e progressiva, con la coesistenza, per lungo tempo, di più versioni della stessa tipologia documentale, soprattutto all'interno di cartelle cliniche con degenze prolungate, contribuendo così a rendere più sfumato il confine temporale del passaggio da un modulo ad un altro. Attraverso l'analisi del layout documentale, si è osservato, inoltre, che le modifiche spesso hanno interessato solo singole sezioni dei documenti analizzati, le loro stesse denominazioni o quelle di specifici campi, sottolineando un passaggio significativo da descrizioni dettagliate e narrative, a espressioni sintetiche, più concise e spesso riportate per punti. Pertanto, mentre il numero e la specializzazione dei campi aumenta, il loro contenuto diventa più sintetico.

Tra le tipologie documentali considerate, la tabella nosologica è il documento che ha mostrato maggiore evoluzione per struttura e contenuto, passando da sezioni narrative e descrittive a sezioni articolate e semanticamente più organizzate.

La conoscenza acquisita sulla struttura documentale dei documenti clinici analizzati offre spunti di ricerca futura. Infatti, per ogni layout identificato si potrebbero creare dei file XML-TEI per formalizzare le relazioni tra concetti, campi e contenuti, a supporto di più approfondite analisi computazionali. Le trascrizioni del contenuto delle cartelle cliniche potrebbero, infatti, essere innestate in uno schema XML-TEI appositamente predisposto, creando l'opportunità di condurre analisi comparative e studi più approfonditi, e al contempo riducendo il tempo necessario per la raccolta e la valutazione dei dati. I vari tag XML-TEI potrebbero, poi, generare le classi di un'ontologia della struttura documentale che consentirebbe di creare delle query per la navigazione dell'archivio. L'applicazione di standard internazionali, come XML-TEI per il tagging e la strutturazione dei layout, e Records in Contexts-Ontology (RiC-O) per la definizione di un'ontologia, consentirebbero di rendere tali strumenti semanticamente interoperabili.

Ringraziamenti

Questo lavoro si colloca nell'ambito delle attività del PRIN 2022P NRR P2022R5LJ7 "Digital preservation, Linguistic analysis and valorization of historical archive of the former psychiatric hospital of Girifalco (DILIGO)",

con il sostegno dell'Unione Europea – Next Generation EU, Missione 4 Componente 1, CUP B53D23030290001.

Riferimenti bibliografici

- Babini, Valeria Paola. 2011. *Liberi tutti: Manicomi e psichiatri in Italia, una storia del Novecento*. Bologna: il Mulino.
- Baggio, Serenella. 2020. "Memorie di guerra dagli archivi manicomiali del Trentino." In *Lingua e patologia. I sistemi instabili*, 203-34. Roma: Aracne Editrice – Gioacchino Onorati Editore.
- Bellotto, Heloísa Liberalli. 2004. *Arquivos permanentes: Tratamento documental*. 2nd ed. Rio de Janeiro: FGV.
- Borrello, Laura, Chiara Cupidi, Valentina Laganà, et al. 2016. "Angela R.: A Familial Alzheimer's Disease Case in the Days of Auguste D." *Journal of Neurology* 263 (12): 2494-98. <https://doi.org/10.1007/s00415-016-8294-x>.
- Bruno, Francesco, Valentina Laganà, Raffaele Di Lorenzo, Amalia C. Bruni, and Raffaele Maletta. 2022. "Calabria as a Genetic Isolate: A Model for the Study of Neurodegenerative Diseases". *Biomedicines* 10 (9): 9. <https://doi.org/10.3390/biomedicines10092288>.
- Caffaratto, Daniela. 2010. *Archivio dell'Ospedale neuropsichiatrico di Racconigi*. Torino: Hapax.
- Carrino, Candida. 2018. *Luride, agitate, criminali. Un secolo di internamento femminile (1850 1950)*. Roma: Carocci Editore.
- Chiaravalloti, Maria Teresa, e Maria Taverniti. 2021. "Sanus egredieris." *Mélanges de l'École française de Rome - Italie et Méditerranée modernes et contemporaines* 133 (1): 159-72. <https://doi.org/10.4000/mefrim.10674>.
- Direzione Generale Archivi. n.d. "Carte da legare. Archivi della psichiatria in Italia." Consultato il 13 luglio 2025. <https://cartedalegare.cultura.gov.it/home>.
- Duranti, Luciana. 1997. *I documenti archivistici: la gestione dell'archivio da parte dell'ente produttore*. Roma: Ufficio Centrale per i Beni Archivistici.
- Fiorino, Vinzia. 2002. *Matti, indemoniate e vagabondi. Dinamiche di internamento manicomiale tra Otto e Novecento*. Venezia: Marsilio.
- Fiorino, Vinzia. 2010. "Le cartella clinica: un'utile fonte storiografica?" In *Identità e rappresentazioni di genere in Italia tra Otto e Novecento*, a cura di Francesca Alberico et al. Genova: DISMEC.
- Goffman, Erving. 1968. *Asylums. Le istituzioni totali: i meccanismi dell'esclusione e della violenza*. Torino: Einaudi.

- Greco, Oscar. 2018. *I demoni del Mezzogiorno: Follia, pregiudizio e marginalità nel manicomio di Girifalco (1881-1921)*. Soveria Mannelli: Rubbettino.
- Gutehrle, Nicolas, and Iana Atanassova. 2022. "Processing the structure of documents: Logical Layout Analysis of historical newspapers in French." *Journal of Data Mining & Digital Humanities* NLP4DH. <https://doi.org/10.46298/jdmdh.9093>.
- Ibekwe, Fidelia. 2024. "Another Look at a Knowledge Organization Pioneer: Traces of Racism in Paul Otlet's Writings." *Knowledge Organization* 51 (1): 3-18. <https://doi.org/10.5771/0943-7444-2024-1>.
- Legge n. 36 del 14 febbraio 1904. *Disposizioni sui manicomi e sugli alienati*. Gazzetta Ufficiale del Regno d'Italia n. 43, 22 febbraio 1904.
- Legge n. 180 del 13 maggio 1978. "Norme per la riforma psichiatrica." Gazzetta Ufficiale, Serie Generale, n. 131, 2 giugno 1978.
- Lombroso, Cesare. 1876. *L'uomo delinquente in rapporto all'antropologia, alla giurisprudenza ed alla psichiatria*. Torino: Fratelli Bocca.
- Marcello, Domenico. 1995. *Un secolo di manicomio: Storia del Manicomio di Girifalco*. Catanzaro: Vincenzo Ursini Editore.
- Oliveira Milani, S., Widad Mustafa El Hadi, and Natália Tognoli. 2024. "Introduction to the Special Issue: Critical and Social Knowledge Organization." *Knowledge Organization* 51 (8).
- Regio Decreto 16 agosto 1909, n. 615. *Regolamento sui manicomi e sugli alienati*. Gazzetta Ufficiale del Regno d'Italia n. 210, 4 settembre 1909.
- Ricucci, Marina. 2025. "Presentazione dell'archivio digitale delle cartelle cliniche dell'ex ospedale psichiatrico di Maggiano." Intervento al seminario *Scrittura e malattia mentale*, Università di Pisa, 22 gennaio.
- Tamburini, Augusto, e Giulio Benelli. 1885. L'antropologia nelle carceri. *Rivista di discipline carcerarie in relazione con l'antropologia, col diritto penale, con la statistica* 15 (4): 136-47.
- Tennis, Joseph T. 2006. "Versioning Concept Schemes for Persistent Retrieval." *Bulletin of the American Society for Information Science and Technology* 32 (5): 13-16. <https://doi.org/10.1002/bult.2006.1720320506>.
- Tognoli, Natália Bolfarini, Ana Célia Rodrigues, and José Augusto Chaves Guimarães. 2019. "Archival Knowledge: Conceptual Frameworks for Recent Terminology in the KO Domain." *Knowledge Organization* 46 (7): 522-29. <https://doi.org/10.5771/0943-7444-2019-7-522>.
- Trovato, Silvia. 2018. "Guida all'Archivio dell'Ospedale Psichiatrico di Volterra." *Rassegna Volterrana* 95: 313-413.
- Valeriano, Annacarla. 2017. *Malacarne: Donne e manicomio nell'Italia fascista*. Roma: Donzelli.

Manuscripts in the Digital Age

Lorenza Talarico*

Abstract: This article explores the use of Linked Open Data (LOD) and International Image Interoperability Framework (IIIF) standard to link images and text, thus making one of the Library's fundamental manuscripts, *De educandis ingeniis* by Cardinal Federico Borromeo, accessible in an enhanced digital format. The project entailed the digitization of the manuscript, automatic transcription via Handwritten Text Recognition (HTR) using Transkribus and semantic encoding in Text Encoding Initiative (TEI) XML and Analyzed Layout and Text Object (ALTO) formats. A prototype developed on the Simple Annotation Server (SAS) integrates annotations, enabling an interactive engagement with the manuscript. The research outlines a replicable workflow aimed at enhancing accessibility, interoperability, usability and digital preservation. By addressing the lack of unified methodologies in the field, this work contributes to the broader discourse on digital philology and the sustainable integration of cultural heritage into the Semantic Web.

Keywords: Transkribus, HTR, Linked Open Data, IIIF, Digital library.

1. Introduction

In today's landscape, the integration of digital technologies into historical libraries is proving to be not just inevitable but essential, transforming these institutions from static knowledge repositories into dynamic instruments of cultural dissemination (Cusimano 2015). This integration is now demanded by both contemporary trends and the evolving needs of library users and scholars. The shift of digital libraries towards a more multimedia-oriented dimension is an incontrovertible reality. Since the early initiatives in the 1990s, there has been a continuous expansion in the range of materials incorporated into digital collections. This trend substantiates early theorists' predictions about digital convergence, demonstrating that these forecasts were accurate projections validated over time, not mere speculative exercises (Hanneken 2023). This change has been notably supported by advancements in techno-

* Università Cattolica del Sacro Cuore di Milano, lorenzatarico@gmail.com.

logical standards such as Linked Open Data (LOD) and the International Image Interoperability Framework (IIIF). These technologies have facilitated the creation of interoperable digital editions, making them accessible to scholars and researchers worldwide. Among the few institutions in Italy that have distinguished themselves in this process is the Veneranda Biblioteca Ambrosiana in Milan. As the first in the country to adopt the IIIF standard, the Ambrosiana, utilizing the image viewer Mirador and the image server Cantaloupe, has focused on preserving and promoting an invaluable collection of manuscripts and ancient works through its Digital Library, launched in 2019. Specifically, it aims to enhance the primary funds, namely Inferior, Superior, S.P. and Trotti, through digitization and open online access. This initiative ensures the preservation of these original manuscripts while concurrently promoting their use for study and research. Indeed, over the coming years, previously digitized manuscripts using various technical methods will gradually become available in digital format. At the same time, high-resolution digital copies of additional manuscripts, particularly those not accessible due to their storage conditions, are being produced.

Hence, this case study explores the application of LOD and IIIF standard to link images and text, making one of the Library's manuscripts, *De educandis ingeniis* by Federico Borromeo, accessible in a newly enhanced digital format. This brief treatise written in the seventeenth century reflects the founder's dedication to promoting culture and knowledge for the common good and represents an adequate starting point for testing the integration of digital images with textual transcriptions.

This study involves cutting-edge technology that is still evolving, but the hope is to provide the community and the library with a prototypical model workflow and data structuring system that may be replicated for other manuscripts, according to the principle of extensibility. This research topic aligns with the observations Federico Borromeo articulated in the tenth chapter, *On the lack of wealth*, of *De educandis ingeniis*:

Ma dall'altra parte i miei piccoli doni potrebbero parer alquanto più grandi se venissero paragonati non con le scienze, che sono tanto meritevoli, ed eziandio con coloro che gli ricevono, ma coi presenti tempi. I quali potranno almeno in ciò essermi favorevoli, perché sì come molto stimar si dee nelle sterilità e nelle grandi arsurre dei campi qualunque, benché piccolo, sostentamento e refrigerio, così in questo nostro secolo parmi di vedere che con gl'effetti e con le opere, e non con le sole parole e proferte, pochissimi siano stati coloro che soddisfatto abbiano alla fame e alla sete dei meschini letterati (Borromeo 2008, 143-144).

These words succinctly capture the reasons behind Borromeo's desire to write this brief treatise and highlight the patronage aspect of his work, positioning him as a "cultural organizer". His ambition was to «make Milan into a

second Rome» and he always rejected mediocrity, striving for excellence. Thus, with *De educandis ingeniis*, he outlined a cultural model for scholars: what the cardinal managed to accomplish in Milan over the course of a couple of decades has often been described with superlative adjectives, almost miraculous. The astonishment stemmed, and still does, from the resolution with which the project was implemented and the significant shift observed in the social and cultural fabric of early 17th-century Milan, which was a context that, despite recent and important historiographical revisions, can still be described as stagnant. Many centuries later, researchers in the fields of Digital Humanities and Linked Open Data have been trying to «far gran cose negli studi», namely to achieve great things in academic pursuits, much like Borromeo's primary ambition but, despite inevitable developments and flourishing activities, there remains a significant lack of homogeneity in the currently available Digital Scholarly Editions (DSEs) (Monella and Rosselli Del Turco 2020).

Hence, this project's approach, structured into three phases, aims to improve searchability, facilitate academic utility and establish a replicable workflow for other manuscripts. In essence, it seeks to answer a central research question: can modern technologies bridge the gap between ancient manuscripts and contemporary digital tools to enhance usability and preservation? The initial phase of the research contextualized the project by addressing the theoretical foundations of LOD and the Semantic Web. The second phase focused on technical implementation, encompassing the use of Transkribus to develop a Handwritten Text Recognition (HTR) model, the employment of TEI XML and ALTO formats and the adoption of IIIF standards. This process culminated in the creation of a final digital edition prototype on the Simple Annotation Server (SAS), enabling the side-by-side visualization of images and their corresponding transcriptions. In fact, the research involved uploading the text transcription annotations via SAS into the IIIF manifest, sourced directly from the Digital Library of the Ambrosiana, adhering to the principles of Linked Open Data (LOD). The LOD paradigm offers substantial advantages for language data, such as structural and conceptual interoperability, federation, the expressivity of semantic web models and the network effect. Given the context of this study and its placement within the realm of the Semantic Web, the use of the LOD paradigm was imperative for addressing the research problem. Therefore, the aim was to develop a model that can be replicated and adapted for other manuscripts, thereby enhancing cultural heritage through innovation and accessibility by bridging traditional practices with advanced technological tools and addressing the absence of a unified data-structuring model in the community.

2. Data and technologies employed

The core data for this case study comprises the IIIF images of *De educandis ingeniis*, accessible via the Digital Library of the Veneranda Biblioteca Ambrosiana website, alongside its corresponding IIIF manifest and the transcription of the manuscript. The Ambrosiana's Digital Library actively aims to enhance its foundational collections, namely Inferior, Superior, S.P. and Trotti, through systematic digitization and open online access. This strategic initiative serves a dual purpose: it safeguards these invaluable manuscripts for posterity while simultaneously broadening their utility for scholarly pursuits and research. As the first library in Italy to adopt the IIIF standard, the Ambrosiana leverages tools like the Mirador viewer and the image server Cantaloupe to ensure wide accessibility. Notably, the IIIF ecosystem is seamlessly integrated with the library's OPAC (On-line Public Access Catalog), establishing a direct link between a manuscript's catalog description and its corresponding digital resource (Cusimano 2019). IIIF, a set of open standards supported by a global consortium of cultural institutions, offers significant benefits, including high-quality image delivery with advanced zooming and manipulation options, compatibility with diverse tools for viewing and comparing images and the flexibility to «publish once and reuse often» across different platforms (Salarelli 2017).

The transcription of the manuscript was executed using Transkribus, a platform developed through two EU-funded research projects and managed by READ-COOP SCE since July 2019. In essence, Transkribus, built on Java technology, supports both manual and automatic transcription, facilitates the training of custom AI models, provides advanced search tools and allows document export in multiple formats. Nevertheless, Transkribus is a “closed environment”, presenting certain limitations, particularly compatibility issues when attempting to use externally the work produced within its ecosystem. These challenges, compounded by bureaucratic and authorization obstacles and platform copyright restrictions, collectively impede full adherence to LOD principles. Consequently, the goal of this specific study with *De educandis ingeniis* was to export the generated work and achieve comparable visualization results outside the Transkribus environment, thereby ensuring authentic interoperability and alignment with the LOD paradigm. For *De Educandis Ingeniis*, a HTR model was created on Transkribus by transcribing the entire manuscript to create a Gold Standard. The transcription was then exported in multiple formats, including TEI XML, a widely adopted format in Digital Humanities for encoding and managing digital data. Despite its flexibility and scholarly value, the integration of TEI with IIIF remains challenging and lacks standardization (Monella and Rosselli Del Turco 2020). As a result, after evaluating various approaches, ALTO files exported from Transkribus were then selected for further use.

ALTO, an open XML Schema developed under the EU-funded project Metadata Engine (METAe), offers a standardized format for describing the layout and content of physical text resources, such as book pages or newspaper articles. The distinct advantage of the ALTO format in this case study lies in its capacity to act as a pivotal connecting link, ensuring a streamlined workflow without data loss or redundancy during the annotation phase. It facilitates the efficient transformation, adaptation and linking of data, all within a LOD perspective. In fact, the ALTO files were converted into annotations using the Alto2Annotations transformer and subsequently uploaded to SAS, the final technology employed in this project to create a visualizable prototype via Mirador. SAS, a Java-based annotation server developed by Glen Robson, the Technical Coordinator of the IIIF Consortium, provides an accessible and open-source solution, enabling users to work locally on their computers while sharing the results via links as needed. It supports both IIIF and the Mirador viewer, allowing annotations stored as linked data in an Apache Jena triple store. This approach fully aligns with Linked Open Data principles, enhancing interoperability (Ide and Pustejovsky 2010) and semantic connections across the web, thereby underscoring the potential for linked data to serve as a foundation for future and similar projects. However, SAS can experience downtime, a vulnerability stemming from its reliance not only on its own stability but also on the stability of the website hosting the referenced IIIF manifest. Since SAS adheres to the LOD paradigm, its components are highly interoperable and interconnected, meaning that the failure of one element can potentially impact the entire system. Regardless, SAS provides a configurable backend capable of leveraging various technologies, including Jena and Sesame for Resource Description Framework (RDF) support, as well as SOLR. It also supports custom annotation bodies formatted in RDFa, allowing for greater flexibility in annotation structure and utilization. Therefore, future work could explore these server features to fully capitalize on Linked Open Data for enhanced data analysis, enrichment and information retrieval.

3. Workflow

The methodology and workflow of this project commenced with the study and application of Transkribus within the Veneranda Biblioteca Ambrosiana. However, this process has progressively evolved through dedicated research, continuous experimentation and the successful navigation of numerous challenges. This evolution is inherent to the project, given its reliance on cutting-edge technology that remains actively under development.

The initial phase involved the meticulous selection and review of the manuscript *De educandis ingeniis* and its historical context. These choices were guided by feasibility considerations, including the available timeframe and

the manageable volume of data. Subsequently, the manuscript images were accurately uploaded to the Digital Library in IIIF format, joining alongside the existing catalog on the platform. *De educandis ingeniis* model was then developed for the HTR of this particular handwriting style. All 92 pages of the manuscript underwent consistent and accurate line by line transcription, thereby creating a Gold Standard. Beyond the AI model itself, this effort yielded the ability to visualize the manuscript pages, uploaded as JPEG files derived from the “master files”, alongside their corresponding transcriptions directly within the platform, with the work also downloadable in various formats. Nevertheless, it became evident that for data preservation and long-term maintenance, and crucially, due to the inherent incompatibility of Transkribus’s workflow with the Ambrosiana Digital Library’s existing IIIF Mirador architecture, further direct use of Transkribus was deemed unnecessary. Despite this, the foundational work completed on the platform remains highly relevant, as it provided the essential HTR model for the manuscript, serving as the starting point for the entire case study.

Following this, an exploration of the various download formats led to an attempt to integrate the IIIF images from the Digital Library, specifically the IIIF manifest, with the TEI XML file downloaded from Transkribus containing the transcription. This involved extensive research into prior community efforts, notably the work of Paolo Monella and Fabio Cusimano (Monella and Cusimano 2019), as well as existing theoretical approaches. Theoretically, the TEI format inherently fulfills all interoperability and extensibility requirements (Burnard 2014), as the IIIF manifest of the manuscript could potentially reference it as an external resource through the referenced annotations technique (Monella and Rosselli Del Turco 2020). Within the TEI XML file, the *<facsimile>* element is pivotal, linking the transcription to the corresponding images of the manuscript pages. This connection is essential for aligning the digital representation of the manuscript with its physical counterpart, enabling users to view the original image alongside its transcription on compatible platforms, such as Edition Visualization Technology (EVT) (Monella and Rosselli Del Turco 2020). However, in the context of this case study, EVT was not adopted, despite being the current state-of-the-art within the community. This decision stemmed from EVT’s primary tailoring for critical editions of philological texts, designing it as a specialized system for a specific use case, thus rendering it less suitable for the broader needs of the Ambrosiana’s Digital Library. Consequently, while TEI and IIIF integration is theoretically feasible and actively being pursued by experts, a seamless or viable solution for this specific project has not yet been identified.

To address the initial research problem and achieve the project’s objectives, an alternative solution was imperative for leveraging Linked Open Data to display manuscript images alongside their transcriptions. This research led to

the discovery of SAS, which aligns with the LOD paradigm by storing annotations as linked data in an Apache Jena triple store. Its compatibility with both IIIF and the Mirador viewer, significantly enhances interoperability and semantic connections across the web, embodying LOD principles and maximizing data integration potential. Furthermore, this server streamlines the annotation of IIIF manifests and enables their automatic modification, while also supporting the upload of JSON files as annotations. This capability ensures the preservation of previous work and facilitates a smooth workflow without unnecessary delays. Specifically for *De educandis ingeniis*, 92 XML files, each representing a manuscript page, were exported from Transkribus in ALTO format, selected for its proven compatibility with SAS. ALTO, widely utilized in OCR contexts, integrates seamlessly with the Alto2annotations transformer, which facilitated the conversion of these files into IIIF-compliant JSON annotations. This process ensured the smooth transfer of work from Transkribus to SAS, effectively linking the transcriptions to each online manuscript page and maintaining an efficient workflow throughout the annotation process.

The primary objective was to establish a single prototype for this project utilizing the modified IIIF manifest generated through SAS. This involved structuring the data via the latter and revising the original IIIF manifest obtained through a link from the Digital Library. The modification employs the other content technique, which embeds links to Annotation Lists that point to external resources, namely, JSON files created by SAS containing the annotations written for each manuscript page. Before achieving this, all 92 ALTO files required conversion using the Alto2Annotations transformer. However, the initial visualization in Mirador revealed that the rectangles did not precisely correspond to the lines of text, leading to disorganized transcriptions. To rectify this, line level annotations were utilized (instead of word level ones) and the coordinates were removed from each file, allowing the annotation for each page to refer to the entire corresponding page. SAS necessitated JSON files compatible with Open Annotations IIIF v.2, requiring the application of the annotationListNoArt.xsl XSLT from the transformer. Subsequently, the output JSON files were manually edited to include the URI of the corresponding canvas from the published IIIF manifest in the Digital Library, ensuring proper communication and interoperability. Once the files were prepared, all annotations were successfully uploaded to SAS. Consequently, the server generated the modified IIIF manifest, enabling visualization in Mirador or Universal Viewer and facilitating work via a simple link.

As previously mentioned, SAS utilized Annotation Lists to incorporate annotations, specifically transcriptions, as this technique aligns with IIIF version 2. Annotation Lists serve as a method for grouping annotations, typically at the canvas or page level. They may encompass resources like page transcriptions or OCR data, formatted as annotations. These lists are linked to the

specific canvas they pertain to, functioning as independent resources meant to be accessed separately when required. This separation from the main manifest enables clients application to display images rapidly, thereby enhancing user experience by adding additional content and commentary as the user interacts with individual canvases. Moreover, Annotation Lists are generally resolvable, meaning their *@id* can be entered in a web browser to directly retrieve the list. As a result, they can be generated dynamically, while the manifest itself remains static and easily cacheable. Within the URI pattern, the *{name}* parameter must be unique to differentiate one Annotation List from others. While this parameter often matches the canvas name, it is not mandatory, as a single canvas might contain multiple lists categorized by resource type. Therefore, adherence to specific provided URIs, rather than pre-generating them, is crucial. Furthermore, each Annotation List must possess a unique HTTP(S) URI specified in the *@id* field, which should return a JSON representation upon access. Annotations are included in a *resources list* and the format should correspond to the media type returned upon dereferencing the resource. The *content* type should be selected from a list in the Open Annotation specification or other widely recognized ontologies. For resources directly rendered, such as images, transcriptions or musical performances, the motivation should be *sc:painting*. Content resources can also contain supplementary fields, including labels, descriptions, metadata, licenses and attribution.

In the context of this case study, the JSON file of the IIIF manifest published in the Digital Library contains the key elements which are characteristic of a manifest. Within the *sequences* property, the canvases are listed. Upon examining the modified and annotated IIIF manifest generated through SAS, it appears highly similar to the original one. However, a key distinction emerges when analyzing the *sequences*: a new property named *otherContent* is present, following the *thumbnail* property. This property has been utilized by SAS to link additional content, specifically an Annotation List, to the manifest. This represents the core addition to the new manifest, with the *@id* linking to another JSON file, specifically an Annotation List for each page of the manuscript. It is important to note that this structure is consistently applied across all 92 images, thus uniformly affecting every canvas within the manifest. Analyzing the JSON of the Annotation List reveals the page transcription text in the text/html format and the motivation of the annotation, specifically *sc:painting*, thereby confirming the theoretical framework discussed previously. Through this structured, interoperable, precise and hierarchical data organization, the desired and final prototype becomes accessible on Mirador.

4. Conclusions

Within Mirador, the manuscript can be virtually navigated page by page, with each image featuring an annotation linked to the text transcription initially completed in Transkribus. This approach successfully achieves the primary objective of this research: to integrate the transcription alongside the manuscript and link it to the IIIF manifest of the manuscript's images. This not only enhances interoperability but also provides the Ambrosiana Library with a methodology for potentially further advancing its digital library system. The overarching goal is, therefore, to render its invaluable collection even more accessible, interoperable and reusable for researchers, students and enthusiasts worldwide (Cusimano 2021).

Consequently, this work proposes a solution for integrating images and text within the Ambrosiana Digital Library. It leverages LOD and IIIF technologies to make the manuscript *De educandis ingeniis* available in a new, interactive digital format. The developed workflow was designed to overcome various bureaucratic, technical and copyright constraints associated with Transkribus, as well as addressing the current lack of necessary Mirador plugins for IIIF image annotation in the Ambrosiana Digital Library. By progressing from transcription in Transkribus to annotation in SAS through multiple stages of format conversion, a digital edition prototype was created, offering both enhanced accessibility and interactive research opportunities to scholars and enthusiasts. While the results currently remain within the research domain and the prototype is not yet available on the Ambrosiana website, they highlight how LOD and IIIF could significantly facilitate the navigation, search and consultation of digitized manuscripts while preserving the quality and integrity of the original materials. This model also demonstrates promise for replication with other manuscripts and potentially across other cultural institutions, thereby fostering a broader and more systematic dissemination of cultural heritage (Spina 2023). Nevertheless, certain limitations were encountered, primarily because these technologies represent cutting-edge advancements that continue their rapid evolution. Restrictions within Transkribus's closed ecosystem and the complexities of managing complex data partially limited the workflow's effectiveness in a standardized digitization context. As a consequence, the community shall prioritize research focused on the link between text and images. Future developments could include collaborations with other platforms and institutions to enhance interoperability and accessibility, alongside establishing a unified and shareable data structuring model. However, future works should also encompass additional technical improvements, such as supporting alternative encoding strategies to enable links to external resources, while ensuring that the entire edition, including XML, internal and external data and software, remains sustainable, durable and aligned

with the FAIR principles (Wilkinson, et al. 2016). Moreover, by continuing to utilize the SAS platform, it would be beneficial to conduct an in-depth study on the coordinates required for accurate annotations. It would be valuable to establish and standardize a method to convert image coordinates from ALTO format via Transkribus to IIIF format. That is to say, while this essential conversion is technically feasible, its practical implementation remains challenging. Successful completion of this would fulfill the parameters required by the transformer on SAS, enabling word level annotation. This, in turn, would lead to a more sophisticated, precise and efficient prototype that closely mirrors the output of Transkribus. Furthermore, it is clear that the future aim would be to apply this workflow to an increasing number of manuscripts, initially from the Ambrosiana Library and later from other institutions, continually refining the process to make it more streamlined and efficient. Integrating manuscript images with a searchable text edition represents a critical advancement for the academic world, as there currently appears to be no established standard for achieving this goal (Spina 2023). In this context, the adoption of LOD is not merely a conceptual ideal, but a crucial functional imperative: the project embraces structural and conceptual interoperability, the expressiveness of RDF-based models and the potential for data federation. In future developments, data enrichment could be enhanced, through cross-referencing with external authority datasets, thereby markedly improving information retrieval capabilities across the cultural heritage domain.

In conclusion, this project currently represents a small yet valuable contribution to the community and cultural resource management, fostering an ongoing dialogue between innovation and tradition. The developed prototype can facilitate the digital preservation of manuscripts while enhancing their accessibility and usability, echoing the ideal of *ad publicum commodum et utilitatem*, namely the common good, which inspired Federico Borromeo in founding the Veneranda Biblioteca Ambrosiana. Ultimately, history repeats itself: while contexts and technologies may evolve and change, the enduring spirit of public service and cultural sharing remains a constant in society.

References

- Borromeo, Federico. 2008. *De Educandis Ingeniis, Reproduction of the Original Manuscript and First Edition of the Text*. Busto Arsizio (VA): Nomos Edizioni.
- Burnard, Lou. 2014. *What Is the Text Encoding Initiative?* Marseille: OpenEdition Press.
- Cusimano, Fabio. 2015. "Bibliotecario Digitale, Umanista Informatico." *Bibelot. Notizie dalle biblioteche toscane* 21: 14-17.

- Cusimano, Fabio. 2019a. "Biblioteche di Conservazione & Data Curation: Dal Custos Catalogi al Digital Librarian. Il Caso della Veneranda Biblioteca Ambrosiana." *JLIS.it* 10 (1): 125-39.
- Cusimano, Fabio. 2019b. "La Nuova Biblioteca Digitale della Veneranda Biblioteca Ambrosiana." *L'Almanacco Bibliografico. Bollettino trimestrale di informazione sulla storia del libro e delle biblioteche in Italia, La questione*. Milano: Edizioni CUSL - Università Cattolica del Sacro Cuore, 1-2.
- Cusimano, Fabio. 2021. "The IIF-Based Digital Library of the Veneranda Biblioteca Ambrosiana." *Umanistica Digitale* 10: 423-32.
- Hanneken, Todd R. 2023. "What to Think About When Thinking About Digitization of Manuscripts." *Digital Philology* 12: 256-84. <https://doi.org/10.1353/dph.2023.a911846>.
- Ide, Nancy, and James Pustejovsky. 2010. "What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology." In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong, China.
- Monella, Paolo, and Fabio Cusimano. 2019. "Linking Text and Image: TEI XML and IIF." <http://www.paolomonella.it/reires2019/index.html>.
- Monella, Paolo, and Roberto Rosselli Del Turco. 2020. "Extending the DSE: LOD Support and TEI/IIF Integration in EVT." In *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La Svolta Inevitabile: Sfide e Prospettive per l'Informatica Umanistica*, edited by C. Marras, M.C. Passarotti, G. Franzini, and E. Litta. Milano: Università Cattolica del Sacro Cuore, AIUCD.
- Salarelli, Alberto. 2017. "International Image Interoperability Framework (IIF): Una Panoramica." *JLIS.it* 8: 50-66.
- Spina, Salvatore. 2023. "Handwritten Text Recognition as a Digital Perspective of Archival Science." *AIDAinformazioni* (1-2): 115-32.
- Wilkinson, Mark, Michel Dumontier, Ijsbrand Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. <https://doi.org/10.1038/sdata.2016.18>.

Vocabularies in the digital context:

Terminological plurality in the Knowledge Organization

Maria Lígia Triques*, Ananda Fernanda de Jesus **, Mario Barité***,
José Eduardo Santarem Segundo****, Ana Cristina de Albuquerque*****

Abstract: The study investigated the approach to the term “vocabulary” in the digital context by the Knowledge Organization (KO) community. An exploratory and bibliographic analysis was conducted based on a research protocol. The analysis drew on studies that discussed the vocabularies used for assigning computational semantics within the Knowledge Organization KO journal (2004-2024). The results indicate that most studies concentrate on describing particular objects or domains, and the relationship with the digital context is frequently superficial or implied. The hypothesis of terminological plurality was confirmed (54 terms/expressions), with only 5 being repeated, and a notable absence of explicit definitions. The term “controlled vocabularies” was the most frequent, without differentiating between traditional and digital/computational. The absence of precise definitions and the generic use of the term represents a significant barrier to interdisciplinary research advancement. The context of use, structure and functionality are essential to clarify which concept of vocabulary is in focus.

Keywords: Controlled Vocabularies, Structured semantic vocabularies, Machine-Processable Vocabularies, Digital Environments, Knowledge Organization.

1. Introduction

Vocabularies have traditionally been the subject of research in library and information science (LIS), particularly in relation to Knowledge Organization (KO). In this area, vocabularies are a controlled instrument known as “controlled vocabularies”. These instruments are artificial relational languages that

* Universidade Estadual de Londrina – UEL, Londrina, Brasil. mligia.triques@uel.br. ORCID: 0000-0002-0450-7138; albuani@uel.br. ORCID: 0000-0003-3506-0479..

** Universidade Estadual Paulista – Unesp, Câmpus de Marília, Brasil. af.jesus@unesp.br. ORCID: 0000-0001-7873-6040.

*** Universidad de la República – Udelar, Montevideo, Uruguay. mario.barite@fic.edu.uy. ORCID: 0000-0002-2992-6582.

**** Universidade Estadual Paulista – Unesp, Cmpus de Marília e Universidade de São Paulo – USP, Ribeirão Preto, Brasil. santarem@usp.br. ORCID: 0000-0003-3360-7872.

represent information and knowledge in a specific domain using a controlled and finite set of terms called descriptors (Kobashi 2008).

Studies on vocabularies have accompanied technological advances, especially those related to web development. These studies focus on tools in Linked Open Data (LOD) technologies, including value and metadata vocabularies (Zeng 2019). Vocabularies are also frequently referenced in discussions of Knowledge Organization Systems (KOS) in various contexts (Barité et al. 2024).

In this scenario, digital environments have become protagonists in the development of new ways of using and understanding vocabularies. These new approaches coexist with traditional notions that underpin this instrument, especially in its interdisciplinary nature with other fields, such as Linguistics, Lexicography, and Computing. These fields often act at the interface of Natural Language Processing (NLP), Artificial Intelligence (AI), and databases.

Technological possibilities introduce many variables into the creation and consumption of information. The dialogue about the use of vocabularies in digital contexts is influenced by various nomenclatures. Reaching a consensus in literature that allows for a precise definition is sometimes difficult. Various notions coexist and overlap with other terms, such as “set of metadata”, “ontologies” and “data/domain models”, to name a few. This makes it difficult to clearly identify and differentiate how these terms have been understood.

For the KO that works at the interface with traditional and emerging disciplines, this terminological imprecision makes it difficult to conduct interdisciplinary research that builds on assumptions related to vocabularies in analog and digital environments.

In this regard, we start from the premise that the KO literature offers few discussions that explicitly describe and differentiate the various perspectives on the use of “vocabulary” across interface areas. Although “vocabularies” is a basic term within KO, we argue that there is a plurality of terminologies shaped by the multiple uses introduced in the digital context. A context defined by the internet, digital technologies, and networked information systems.

To this end, a preliminary study was developed with the aim of identifying the definitions associated with the term “vocabularies” when applied to digital environments, more specifically in its relation to the attribution of computational semantics to data. The objective is to examine how the term “vocabularies,” when applied to digital contexts, has been addressed by the Knowledge Organization (KO) scientific community. Furthermore, the study seeks to confirm the hypothesis that there is a terminological plurality regarding the use of the term vocabulary in digital contexts.

Therefore, the purpose of this study is not to present an exhaustive analysis but rather to identify the terms that have been employed in relation to this theme. To achieve this, we conducted an initial exploratory study using

the journal *Knowledge Organization* (KO) as a reference, given its significant scientific output in the field and its status as a privileged venue for observing conceptual, terminological, and methodological trends. This choice makes it possible to map, although partially, how specialized literature has employed and consolidated different terms around the theme, providing a basis for future, more in-depth investigations.

In this sense, the proposal is justified so far as demonstrating the existing terminological plurality and systematizing the terms in use constitutes a fundamental step toward supporting and guiding subsequent research that investigates deeper into the discussions on such emerging usages, in coexistence with traditional ones, and identifying their correlations and distinctions.

2. Vocabularies: organizational structures

In a broad sense, the term “vocabulary” refers to a set of words belonging to a particular language or area of knowledge. In the specific context of areas such as Information Science that dialogue with KO, vocabularies take on a structured form, involving the selection and standardization of terms to accurately describe concepts, allowing for the standardized representation of concepts present in documents, objects or data (Zeng and Chan 2004).

KO has focused on the challenges and opportunities arising from the emerging contexts associated with information environments. These environments, especially digital ones, are characterized by a diversity of media, formats and data flows, requiring methods and instruments capable of guaranteeing the representation, retrieval and meaningful circulation of information. Historically associated with KO tools such as subject heading lists, classifications and thesauruses, vocabularies are organizational structures that accompany technological and socio-cultural transformations, reflecting the way they are conceived, implemented and understood.

The advent of the Semantic Web and the growing demand for interoperability between heterogeneous systems has led to the development of ontologies, which are formalized conceptual schemes and machine-processable models. These ontologies enable a richer and more connected representation of knowledge (Zeng 2008).

A brief review of the vocabularies evolution as organizational structures shows a trajectory marked by transformations that accompany the ways in which information is represented, structured and retrieved over time. These tools can be thought of from their consolidation as linguistic control mechanisms to their configuration as interoperable semantic structures in the digital sphere.

As Barité (2011) explains, although there are different understandings of what constitutes these instruments, such as the North American NISO stan-

dard and the ALA Glossary, there are broader conceptions that include all languages used to represent and organize knowledge as controlled vocabularies under a conceptual umbrella.

Vocabularies are therefore often understood within the broader scope of KOS, an expression that covers a diverse set of tools aimed at the structured representation of knowledge (Barité 2011). KOS are instruments that enable the organized interpretation of knowledge domains, contributing to terminological standardization, conceptual navigation and interoperability between information systems (Zeng and Chan 2004).

Zeng and Chan (2004, 1) highlight three fundamental categories among the main types of KOS based on characteristics such as structure and complexity, relationships between terms and historical function as defined in 2000 by NKOS (Zeng and Chan 2004). These categories are:

1. Term lists, which contain lists of words or phrases, often with definitions. Examples include authority files, glossaries, gazetteers, and dictionaries.
2. Classifications and categorization schemes, which emphasize the creation of subject sets. The most notable examples are library classification schemes, taxonomies, and categorization schemes.
3. Relational vocabularies, which emphasize the relationships and connections between terms and the concepts they represent, including lists of subject headings, thesauri, semantic networks, and ontologies.

Thus, the adoption of machine-processable vocabularies can be understood as an extension and reconfiguration of traditional knowledge representation practices, now mediated by specific languages and standards, such as the Resource Description Framework (RDF), the Web Ontology Language (OWL) and the Simple Knowledge Organization System (SKOS). These standards enable the formalization and interoperability of vocabularies in distributed digital environments, allowing not only the standardized description of concepts, but also the creation of semantic links between data from different sources. This contributes to the construction of more integrated and accessible knowledge networks and enables automatic inferences of relationships and links with other networked datasets (Zeng 2008; Jia 2020).

In this context, vocabularies provide the formal semantics of the concepts and relationships used to describe data and represent an area of interest, which makes them the foundation for Linked Data and Linked Open Data (Jia 2020). As Bizer, Heath and Berners-Lee (2009) discuss, while Linked Data is an approach based on best practices for making structured data interconnected, interoperable and accessible on the Web, using Semantic Web technologies, Linked Open Data (LOD) is the subset of Linked Data that adds the dimension of openness, i.e. the data is made freely available for use, reuse and redistribution.

In the Semantic Web, vocabularies then have their functions expanded to the various possible types of applications, which include classifying terms that can be used in a specific context, characterizing relationships and defining restrictions imposed on the use of terminologies (World Wide Web Consortium 2013).

As Jia (2020) explains, there are therefore various types of vocabulary that are generally recognized based on their structural and functional complexity and focus on fulfilling primary functions, which more complex structures generally perform most or all. Such functions can include the elimination of ambiguities, indexing and cataloging, as well as query expansion and the integration of information using identification, description and control of terms and conceptual relationships (Jia 2020).

The study by Jia (2020) proposes a comprehensive analysis of the categorization of semantic vocabularies, highlighting that they can be grouped according to different perspectives, depending on their function, structure and mode of availability. The first perspective considers RDF graph standards. In this approach, vocabularies are divided into sets of metadata elements, which define classes and attributes needed to describe resources such as people, organizations or places; and value vocabularies, such as thesauruses and subject classification systems, which provide controlled schemes for standardizing the values assigned to elements, ensuring uniformity and interoperability in data representation.

The second perspective refers to the degree of formalization. In this case, vocabularies are described using formal Semantic Web languages, such as RDF, which structures relationships between entities, OWL, which allows the definition of restrictions and complex logical inferences, SKOS, used to represent basic structures of controlled vocabularies, such as thesauruses, and RIF (Rule Interchange Format), which enables the exchange of rules between heterogeneous applications (Jia 2020). The third perspective proposed by Jia (2020) concerns the vocabularies openness. From this point of view, they can be classified as: vocabularies of limited openness, which make only part of their content available through licensing or restrictions on use; and fully open vocabularies, which offer unrestricted access through interfaces such as public APIs, favoring interoperability and the reuse of data by various applications.

In this sense, vocabularies range from linguistic control tools applied to indexing in analog environments to becoming elements of semantic infrastructure on the Web. The evolution of these systems is concomitant with the technological, conceptual, and methodological transformations of KO, reflecting the demands for interoperability, openness, granularity, and contextualized representation of information.

3. Methodological Procedures

An exploratory and bibliographical analysis is conducted in this study in order to identify how the term “vocabulary”, when applied to digital contexts, has been approached by the scientific community of knowledge organization.

To this end, this study provides, from an initial and exploratory survey of the proposed topic, a mapping and alignment of terminological plurality in the KO that permeates the concept of structured vocabularies in a machine-processable format. The objective of this study is not to provide an exhaustive analysis, but rather to identify the terms that have been employed to refer to this subject.

To capture this plurality, a research protocol (Tab. 1) was applied, based on the research question: *How has Knowledge Organization named and defined the term vocabularies when referring to structured vocabularies in machine-processable format, developed and applied with the aim of promoting semantic representation in digital contexts?*

Research question (main)	How has the Knowledge Organization community named and defined the term vocabularies, when referring to structured vocabularies in machine-processable format, developed and applied with the aim of promoting semantic representation in digital contexts?
Objectives	Identify the terms and definitions used by the Knowledge Organization community when referring to structured vocabularies in machine-processable format, developed and applied with the aim of promoting semantic representation in digital contexts
Search strategy	(“semantic” OR “data”) AND “vocabulary”; “vocabulary”
Databases consulted	Search in title and abstracts, applying the Source Title: KNOWLEDGE ORGANIZATION and Fields of Research (ANZSRC 2020) filters: 4610 Library and Information Studies; Source Title: KNOWLEDGE ORGANIZATION
Period covered	2004-2024
Idioms	English

Inclusion Criteria	(I) Discusses the use of applied vocabularies in establishing computational semantics in digital contexts
Exclusion Criteria	(E) Not in the languages established for the search; (E) Not in the period covered by the research (E) Only mentions the topic, without going into depth or presenting relevant definitions; (E) Does not address the topic of interest; (E) It was not possible to obtain access to the complete document; (E) Addresses vocabularies aimed at the analogical context (E) Addresses the application of traditional
Extraction form	1) Term used; 2) Definition; 3) Focus of the document.
Date of collection	21/11/2024

Table 1. Research Protocol Adopted.

The choice of both the database used for the Exploratory Analysis and the filters applied is related to the focus of the study of specific terminology to be carried out later, since its focus will be the analysis of the term in the context of knowledge organization. It was therefore assumed that the ISKO, in its international chapter, represents a relevant and sufficiently representative cross-section to make up the exploratory analysis, and that the Knowledge Organization KO journal, in addition to presenting the work presented at the international ISKO:

[...] is a forum for all those interested in the organization of knowledge on a universal or a domain-specific scale, using concept-analytical or concept-synthetic approaches, as well as quantitative and qualitative methodologies. KNOWLEDGE ORGANIZATION also addresses the intellectual and automatic compilation and use of classification systems and thesauri in all fields of knowledge, with special attention being given to the problems of terminology (Knowledge Organization 2025).

The time frame, which covers the last two decades, is justified by the emergence of the topic alongside the development of the World Wide Web. However, in the last two decades, approaches within and outside the digital sphere have coexisted and complemented each other, often leaving an implicit distinction between when the concept of structured vocabularies in computational semantics is or is not used in the study, suggesting a lack of conceptual rigor. Consequently, it was deemed necessary to employ exclusion criteria in instances where the study does not pertain to this specific context.

4. Results and Discussions

Based on the strategy presented, a total of 55 documents were retrieved, which underwent a selection process that identified 1 duplicate document, 36 rejected documents and 18 accepted documents to make up this preliminary research *corpus* (Tab. 2) shows the list of documents that make up the research *corpus* and their respective thematic focuses.

Id	Authors	Title	Focus
02	Baca e Gill (2015)	Encoding Multilingual Knowledge Systems in the Digital Age: the Getty Vocabularies	Present the history and evolution of the electronic thesauruses produced and maintained by the Getty Research Institute (GRI).
10	Chen (2019)	Semantic Enrichment of Linked Archival Materials	Enhanced a semantic data model for Chen Cheng-po's archival materials of paintings and documents.
18	Dobreski, Qin e Resnick (2020)	Depicting Historical Persons and Identities: A Faceted Approach	Developed a seven-facet structure of personal identity, based on a collection of 19th and 20th century images, aiming to contextualize terms and historical perspectives, improving user understanding.
27	Graf (2020)	Domain Analysis Applied to Online Graffiti Art Image Galleries to Reveal Knowledge Organization Structures Used Within an Outsider Art Community	Addresses a methodology based on domain analysis for the elucidation of knowledge organization systems used in image websites documenting graffiti, graffiti art and street art.
36	Lemos e Souza (2020)	Knowledge Organization Systems for the Representation of Multimedia Resources on the Web: A Comparative Analysis	Presents and compares existing metadata standards, vocabularies and ontologies for describing multimedia resources.
40	Marcondes (2020)	Towards a Vocabulary to Implement Culturally Relevant Relationships Between Digital Collections in Heritage Institutions	Discusses the development of Linked Open Data vocabulary to characterize the culturally relevant relationships of the collections of cultural heritage institutions.

47	Smith (2020)	Physics Subject Headings (PhySH)	Discusses the PhySH (Physics Subject Headings) vocabulary, a controlled faceted hierarchical vocabulary for physics, with some basic terms from related areas
12	Coladangelo (2021)	Organizing Controversy: Toward Cultural Hospitality in Controlled Vocabularies Through Semantic Annotation	Addresses the development of a controlled vocabulary to represent dance concepts for country dance choreography, focusing on discussing the implications of cultural and ethical issues related to the representation of gender and race.
30	Smith (2021)	RILM Index (Répertoire International de Littérature Musicale)	Discusses the creation of a thesaurus based on the Répertoire International de Littérature Musicale (RILM) Index, a partially controlled vocabulary designed to index academic writings on music and related subjects.
50	Velios e St.John (2021)	Linked Conservation Data: the Adoption and Use of Vocabularies in the Field of Heritage Conservation for Publishing Conservation Records as Linked Data	Describes an approach to improve the landscape of vocabularies in the field of conservation, providing guidelines for codification and alignment of vocabularies for future implementation of Linked Data.
14	Lemos et al. (2022)	A Proposal in Creating a Semantic Repository for Digital 3D Replicas: The Case of Modernist Sculptures in Public Spaces of Rio De Janeiro	It presents the results of the analysis of vocabularies for documents of physical artifacts and their digital counterparts on the Semantic Web.
26	Grabus et al. (2022)	Temporal Concept Drift and Alignment: An Empirical Approach to Comparing Knowledge Organization Systems Over Time	It presents a methodological contribution on how to approach changes in knowledge organization systems over time, improving the contextualization of historical humanities resources.

39	Marcondes (2022)	The Role of Vocabularies in the Age of Data: The Question of Research Data	Discusses how vocabularies can contribute to assigning computational semantics to digital research data in the context of Big Data.
48	Smith (2022)	Simple Knowledge Organization System (SKOS)	It discusses the Simple Knowledge Organization System (SKOS), a Linked Data vocabulary for representing knowledge organization systems.
05	Barité, Parentelli e Rauch (2024)	Knowledge Organization Systems in Reference Works Specialized in Knowledge Organization: A Study of Punctual Terminology	Identifies records of the term 'knowledge organization system' and its synonymous or similar expressions in dictionaries, glossaries, vocabularies and encyclopaedias specializing in knowledge organization.
08	Binding e Tudhope (2024)	KOS-based Enrichment of Archaeological Fieldwork Reports	It discusses automatic indexing techniques (for final intellectual judgment) based on knowledge organization systems.
22	Georgaki (2024)	A Maritime Heritage Thesaurus Based on a Greek Project Documentation Case	It proposes a thesaurus in the field of Maritime Heritage with the aim of improving methods of archiving oral sources, written accounts, images, 3D archaeology and other multimedia objects.
23	Gnoli et al. (2024)	Library Catalog's Search Interface: Making the Most of Subject Metadata	It addresses the underuse of elements of knowledge organization systems in online library catalogs, hindering effective search and discovery by subject.

Table 2. List of accepted documents and their respective approaches.

Although the last 20 years have been established as a time frame, discussions about vocabularies in the context discussed by the research are relatively recent in the Knowledge Organization scientific community and have only been explored in the last 10 years, with a significant increase in interest in the subject since 2020.

Most studies focus on presenting or discussing the construction of vocabularies for describing specific objects such as paintings, images, 3D works, graffiti and street art. They also focus on describing specific domains such as naval, cultural heritage and music for academic contexts. Although these studies deal with bringing vocabularies closer to the digital context through the proposed applications, this is not the main focus of the texts, and the issue is only mentioned in a superficial or subtextual way.

Among the accepted articles, Marcondes (2022) stands out for the purposes of this study, as he discusses the importance of vocabularies for assigning computational semantics to digital data. The article points out that when adapted to the context of Linked Data, vocabularies make it possible to express the semantics of data, both for humans and machines, facilitating processing by computer agents.

After identifying and analyzing the documents that make up the research *corpus*, we tried to identify the terms adopted by the authors to refer to vocabularies in the context analyzed. In some cases, authors use expressions to refer to the concept; these expressions were also collected.

The terms and expressions identified, along with the number of studies in which these terms appear, are presented (Tab. 3). A single text may utilize multiple terms or expressions.

Terms and Expressions	Occurrence
controlled vocabulary/controlled vocabularies	10
metadata vocabulary/vocabularies	4
value vocabulary	2
semantic vocabularies	2
KOS [knowledge organization system] vocabularies	2
vocabulary/vocabularies	2
semantically structured vocabularies	1
vocabularies as linked data	1
terminological enrichment of controlled vocabulary tools	1
data vocabularies	1
vocabularies published as LOD [linked open data]	1
data vocabularies enrichment	1
standardized vocabularies	1
formalized vocabulary	1
vocabularies (or ontologies)	1
KOS [knowledge organization system] terms	1
vocabularies of different types and dimensions	1
controlled vocabularies through semantic annotation	1
controlled vocabularies to enrich the data	1

KOS[knowledge organization system] entry vocabularies	1
specialized vocabularies	1
languages and vocabularies based on standards recommended by the World Wide Web Consortium (W3C)	1
structured vocabulary	1
LOD [linked open data] vocabularies	1
controlled vocabularies for semantic enrichment	1
LOD KOS [linked open data knowledge organization system]	1
vocabularies are available as structured data in a format which can be machine-processed	1
machine readable vocabularies	1
vocabularies expressed as semantic web technologies	1
vocabulary (concepts, terms)	1
vocabularies of reference	1
vocabulary in annotating digital heritage objects published as LOD [linked open data]	1
controlled/standardized vocabularies	1
RDF vocabularies	1
controlled vocabularies on the web	1
web of data context vocabularies	1
controlled languages	1
ontology vocabularies	1
Total	54

Table 3. List of terms/expressions.

As can be seen in Table 3, there is a wide range of terminology in the OC used to refer to the concept of vocabularies in the context studied. A total of 54 terms/expressions were identified, of which only 6 are repeated in more than one text. The other terms/expressions occur in only one of the texts analyzed.

The term with the highest occurrence is “controlled vocabularies”, which does not allow a distinction to be made between traditional controlled vocabularies and those created to refer to structured vocabularies in a machine-processable format, developed and applied with the aim of promoting semantic representation in digital contexts.

Among the single occurrence terms, there are a number of variations that mention the context of Linked Data, the semantic web and the web of data, terms such as: “semantic vocabulary” (2); “LOD vocabularies” (1); “vocabularies published as LOD” (1); “vocabularies expressed as semantic web technologies” (1); “Vocabularies as Linked Data” (1); “semantically structured vocabularies” (1); “languages and vocabularies based on standards recommended by

the World Wide Web Consortium” (1); “Web of Data context vocabularies” (1); “RDF Vocabularies” (1); “LOD KOS” (1). Figure 1 shows a diagram of the relationships between the terms/expressions and their IDs.

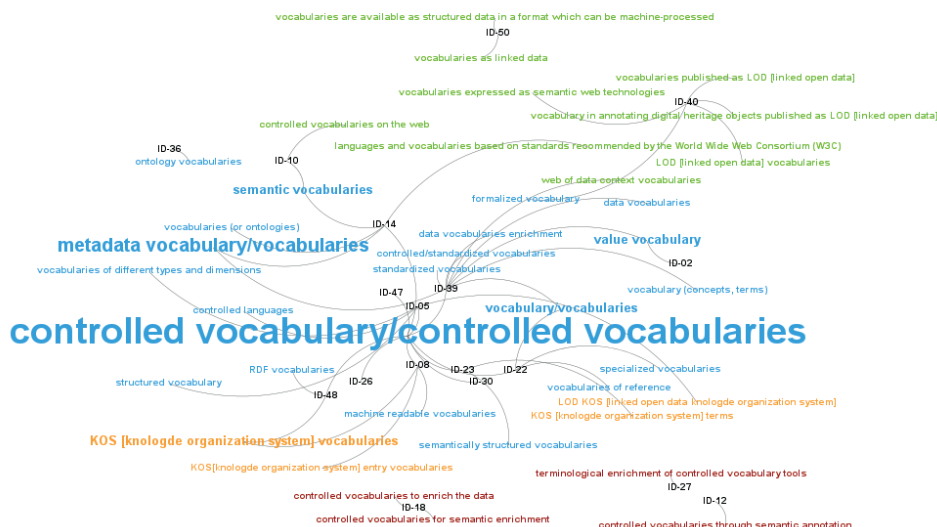


Figure 1. Diagram of the occurrence of terms.

The diagram assists in identifying terminological clusters, which can be categorized into four main groups. The first (blue) includes terms that express the relationship of vocabularies to specific types used to describe, organize, and interconnect data and metadata, such as “metadata vocabulary/vocabularies,” “semantic vocabularies,” “RDF vocabularies,” and “ontology vocabularies.”

The second group (green) highlights the link between vocabularies and the context of the Semantic Web, interoperability, and linked open data, including expressions such as “vocabularies as linked data,” “vocabularies expressed as semantic web technologies,” “vocabularies published as LOD (linked open data),” and “LOD vocabularies”. The third (orange), associate vocabularies with formal Knowledge Organization Systems (KOS), encompassing “KOS [knowledge organization system] vocabularies,” “KOS entry vocabularies,” and “LOD KOS”. The fourth group (red) underscores practical applications for semantic annotation, data enrichment, and terminological standardization, expressed through terms such as “controlled vocabularies through semantic annotation,” “controlled vocabularies to enrich the data,” “controlled vocabularies for semantic enrichment,” and “terminological enrichment of controlled vocabulary tools.”

The most frequently used term is “controlled vocabulary”, both in singular and plural forms, denoting any controlled list of terms to describe documents. Howev-

er, this definition only becomes explicit through semantic associations with terms such as “semantic,” “LOD,” “expressed as semantic web technologies,” and “web of data context vocabularies”. The analysis reveals a limited presence of explicit definitions for the concept of vocabulary, with only five instances clearly identified.

The first (ID-02) defines vocabularies, in the context of the Getty Vocabularies, as a central data structure mapped to a common schema and interconnected both technically and semantically. This definition also describes the term “value vocabulary” within the Semantic Web domain as patterns of semantically structured, machine-readable data values, such as authority files, thesauri, subject headings, and controlled lists, used to populate metadata elements. The second definition (ID-39) characterizes “vocabularies” as sets of descriptive metadata representing specific entities or resources in the Web context. In this sense, value vocabularies assign standardized data values to descriptive items associated with entity instances described by metadata vocabularies.

Another definition (ID-05) refers to the expression “LOD KOS,” citing Soergel (2009), as a broad term encompassing all value vocabularies and ontology levels within the Semantic Web framework. It further argues that KOS are also used by computer programs to reason over data and must represent formal knowledge about concepts. The fourth definition (ID-08) addresses vocabularies in the context of semantic enrichment techniques and tools based on KOS, applied to support information discovery, as exemplified by the Europeana experience. Finally, the fifth definition (ID-23) understands the term “controlled vocabulary,” in the context of KOS, as any controlled list of terms used in metadata to describe documents.

The low incidence of definitions makes it difficult to understand whether the studies really refer to the context of computational semantics. In most cases, it is through clarifications regarding structure and functionality that the intended meaning can be grasped. However, the distinction between terms such as “ontology vocabularies,” “data vocabularies,” “metadata vocabularies,” and “value vocabularies” is not made explicit in most cases.

5. Conclusions

The aim of this study was to identify how the term “vocabulary”, when applied to digital contexts, specifically in reference to structured vocabularies in machine-processable formats, has been approached by the scientific community of knowledge organization.

Based on the *corpus* constructed, it was possible to observe that most of the studies focus on the development and discussion of vocabularies for the description of specific information objects, as well as the description of different domains such as naval, cultural heritage and music for academic contexts.

No studies were identified that specifically aimed to discuss “vocabularies” in machine-processable formats, nor that examined the impact of this context on the structure and functioning of these tools; the topic was mentioned only superficially or implicitly. The study confirms the hypothesis that in the literature on KO there are few discussions that recognize and differentiate the various understandings of the uses of the term vocabulary.

The identification and analysis of the terms used by the authors to refer to vocabularies created for the digital context confirmed the hypothesis that there is a wide terminological plurality in relation to the term studied, with 39 variations observed in just 18 articles analyzed. In addition to the plurality, there was a lack of definitions for the term, making the scenario even more complex, especially in interdisciplinary studies with approaches related to Information Science and Computer Science.

The very generic use of the term carries conceptual imprecision, in that it is often used as a synonym for the most varied types of KOS, which in essence have differences. In addition, new KOS approaches are outlined in computational semantics, which are referred to simply as “vocabularies”. These encompass efforts undertaken within the Semantic Web framework, without a precise differentiation.

In this way, the context of use in terms of structure and functionality is what guarantees the explicitness of which vocabulary concept is in focus. However, the lack of an explicit conceptual definition demonstrates the emergence of this relational field and the need for a study of specific terminology in which vocabularies play a leading role.

In this sense, we highlight the need for clarity and in-depth study of the terminology and conceptualization of the term “vocabulary”. This domain is characterized by plurality and interdisciplinarity, and the lack of conceptual clarity can act as a substantial impediment to progress in teaching and research related to the subject.

As future studies, we hope to expand the *corpus* of studies, adding other databases related to Knowledge Organization. Based on this expansion, the next stage is to carry out a study of specific terminology, the aim of which is to promote greater clarity and in-depth study of the various understandings related to the application and structuring of vocabulary in the digital context. This includes adding discussions on overviews and definitions of other terms, such as ontologies and topic maps, in order to encompass this terminological universe.

Acknowledgments

We acknowledge the support of national funding agencies: *Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)* and *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)*.

References

- Barité, Mario. 2011. "Sistemas de organización del conocimiento: Una tipología actualizada." *Informação & Informação* 16 (3): 122–39. <https://ojs.uel.br/revistas/ucl/index.php/informacao/article/download/9952/pdf/0>.
- Barité, Mario, Varenka Parentelli, and Mirtha Rauch. 2024. "Knowledge Organization Systems in Reference Works Specialized in Knowledge Organization: A Study of Punctual Terminology." *Knowledge Organization* 51 (5): 340–61. <https://doi.org/10.5771/0943-7444-2024-5-355>.
- Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2009. "Linked Data – The Story So Far." *International Journal on Semantic Web and Information Systems* 5 (3): 1–22. <https://doi.org/10.4018/jswis.2009081901>.
- Jia, Junzhi. 2020. "From Data to Knowledge: The Relationships Between Vocabularies, Linked Data and Knowledge Graphs." *Journal of Documentation* 77 (1): 93–105. <https://doi.org/10.1108/JD-03-2020-0036>.
- Knowledge Organization*. 2025. IMR Press. Accessed July 11, 2025. <https://www.imrpress.com/journal/KO>.
- Kobashi, Nair Yumiko. 2008. *Vocabulário controlado: Estrutura e utilização*. Brasília: Escola Nacional de Administração Pública. <https://repositorio.enap.gov.br/bitstream/1/1289/41/Vocabul%C3%A1rio%20controlado%20-%20estrutura%20e%20utiliza%C3%A7%C3%A3o.pdf>.
- Marcondes, Carlos H. 2023. "The Role of Vocabularies in the Age of Data: The Question of Research Data." *Knowledge Organization* 49 (7): 467–82. <https://repositorio.ufmg.br/handle/1843/52263>.
- Soergel, Dagobert. 2009. *Knowledge Organization Systems*: Overview, 1–44.
- World Wide Web Consortium (W3C). 2013. *Linked Data Glossary*. Accessed July 11, 2025. <https://www.w3.org/TR/ld-glossary/>.
- Zeng, Marcia Lei. 2008. "Knowledge Organization Systems (KOS)." *Knowledge Organization* 35 (2–3): 160–82. <http://dx.doi.org/10.5771/0943-7444-2008-2-3-160>.
- Zeng, Marcia Lei. 2019. "Interoperability." *Knowledge Organization* 46 (2): 122–46. <https://www.nomos-elibrary.de/10.5771/0943-7444-2019-2-122.pdf>.
- Zeng, Marcia Lei, and Lois Mai Chan. 2004. "Trends and Issues in Establishing Interoperability among Knowledge Organization Systems." *Journal of the American Society for Information Science and Technology* 55 (5): 377–95. <https://doi.org/10.1002/asi.10387>.

Context-Driven Use of Automatic Tools in Medical Ontology Design: A Case Study in Allergology

Marcin Trzmielewski*

Abstract: The aim of this paper is to show a context-driven use of automatic tools for medical ontology design. Following an analysis of the limitations of the content-driven and quantitative approaches currently used by knowledge organization specialists, we propose a contextualized and qualitative approach. This approach enables the definition of user-oriented resources and criteria for conducting semi-automatic terminology extraction and annotation tasks, ultimately supporting the development of a useful allergology ontology. At the same time, it reveals a paradox: although automatic tools are intended to reduce time-consuming processes, qualitative methods require considerable time and effort to collect and analyze contextual data.

Keywords: Medical Ontologies, Terminology Extraction, Document Annotation, Document Indexing, Contextualized Approach.

1. Introduction

Due to the evolution of documentary environments and the information practices of actors in health, the use of automatic tools to design medical knowledge organization systems (MKOSs) appears to be a necessity. In numerous cases, automatic processes are *content-driven*, which can be reductive when developing systems intended to anticipate the needs of future users and effectively support their activities. Therefore, our aim is to show a *context-driven* use of extraction and annotation tools for processing allergology documentation, as part of the ontology design process for the Allergy Unit of Montpellier University Hospital. Our previous work has shown that the conceptualization and terminology of existing MKOSs in this domain – The Allergy Ontology, The Allergy Information Ontology for Enlightening People, The AllergyIntolerance, and The Allergy Detector II – do not adequately

* Laboratoire d'Etudes et de Recherches Appliquées en Sciences Sociales (LERASS), Université de Montpellier Paul-Valéry, France. marcin.trzmielewski@univ-montp3.fr. ORCID: 0000-0001-7436-2846.

represent the reality and knowledge of allergology, nor are they sufficiently adapted to the daily practices of specialists (Trzmielewski 2023; 2024). Thus, the need to design a new MKOS has become evident.

The present article focuses on the use of automatic tools in the design of MKOSs, discussing current methods and their limitations, as well as the theoretical and methodological approaches applied in the design of the ALLERGI-DOC Ontology. It also presents the results of our study. Finally, we discuss the contributions and limitations of the proposed approach and offer perspectives for the representation and organization of allergology knowledge.

2. The use of automatic tools in the design of medical knowledge organization systems

Following the integration of information and communication technologies into various professional and everyday-life social contexts, both human information interactions and health-related documentation have grown increasingly complex (Trzmielewski 2023). Consequently, information specialists must consider how to design MKOSs capable of representing and organizing vast and heterogeneous bodies of knowledge, while simultaneously supporting diverse information practices across multiple contexts and user groups (Trzmielewski 2024). Although intellectual techniques such as interviews, surveys, thematic analysis, user-oriented approaches, and concept theory-based analyses are widely employed for this purpose (Almeida and Farinelli 2017; Iyer and Raghavan 2018; Oh et al. 2015; Raghavan and Sajana 2010), the abundance and heterogeneity of the available data have led to the consideration of automated techniques – especially in ontology design – aimed at minimizing tedious, repetitive, time-consuming, and costly tasks (Hassan 2017; Richard 2017; Toro et al. 2024). For these reasons, over the past three decades, natural language processing, text mining, and machine learning have been employed to extract, compare, map, and structure medical terminology – i.e., candidate and validated terms and definitions – from large textual corpora, as well to match terms with corresponding words and definitions (Baneyx and Charlet 2006; Choi 2016; Lopes et al. 2009; Muresan and Klavans 2013; Soregel et al. 2004; Tapi Nzali et al. 2018). More recently, large language models and retrieval-augmented generation techniques have begun to be integrated into medical ontology editing tasks, including relationship prediction, term definition generation, and the formulation of logical and conceptual definitions (Liu et al. 2024; Toro et al. 2024).

However, the use of automatic tools is frequently associated with ontological (empirical) (Tennis 2012) and literary warrant (Barité 2018) approaches to knowledge organization and representation. These approaches typically in-

volve the design of MKOSs under laboratory conditions and methods, often based on limited samples that are not drawn from the actual document collections managed by identified organizations and intended to be organized and sought by intended users (Trzmielewski 2024). It also tends to focus on testing only the internal structure of the developed systems. These methods tend to simplify reality and overlook the human and socio-organizational aspects related to the use of documentation resources and info-communication tools in concrete situations of use. Designers provide a set of a priori, content-related quantitative assessment criteria, such as the relevance of extracted terms – measured by frequency, inverse document frequency (IDF), and C-value – or the performance of MKOSs in accomplishing document processing tasks, evaluated through recall and precision scores (Baneyx and Charlet 2006; Lopes et al. 2009; Tang et al. 2017; Tapi Nzali et al. 2018). Even though specialists still involve domain experts – most often physicians – and information professionals to validate certain content-related aspects of terminology and conceptualization (through consensus and quality assessments), potential and end users are consistently not included in this process, largely due to the assumption that they lack the necessary informational competencies (Baneyx and Charlet 2006; Tapi Nzali et al. 2018; Toro et al. 2024). As a result, little is known about their needs, and their criteria of usefulness are not taken into consideration.

Current practices appear to be reductive, and the use of automatic tools should be aligned with epistemological approaches to knowledge organization (Tennis 2012), as well as with qualitative methodology (Cisek 2013). Accordingly, a contextual and qualitative approach was adopted for the design of the ALLERGIDOC Ontology discussed in this paper. Instead of relying on a priori, content-driven quantitative indicators, we aimed to inductively identify criteria that were “external to the corpus” to propose representations that are both meaningful and useful to the actors involved (Clavier 2010, 111; Tennis 2012). However, such criteria are not available in advance. It was therefore necessary to actively explore the contextual environment and interact with the actors to find relevant external elements, to perform what Tennis (2012, §64) refers to as the “right action” – in this case, applying automatic techniques appropriately within the ontology design process. We therefore hypothesized that a contextual approach can assist knowledge organization specialists in identifying relevant criteria for the effective use of automatic tools in the development of MKOSs.

3. Theoretical and methodological framework

3.1. Production of contextual resources

We developed a utility-based contextualized approach (Trzmielewski 2024) by analyzing and actively participating in the interactions between allergy professionals and their informational and socio-organizational environments. This approach relied on three main components: (1) the analysis of the context of use of allergy knowledge, conducted through the study of the informational practices of allergy professionals who seek, produce, and mobilize knowledge in their domain; (2) the automatic analysis of a document corpus regularly used by these professionals in their daily activities; and (3) our involvement as a designer in interactions with actors, identified and considered both as experts and final users. We did not apprehend *context* as a reality out there, waiting to be analyzed and fully described, but rather as a theoretical and empirical construct shaped by the researcher's specific purpose in interaction with scientific object (Mucchielli and Noy 2005; Talja et al. 1999). Our aim was to produce contextual resources and criteria that would contribute to the development of a useful ontology to effectively support the work and information activities of allergy professionals. To this end, we sought to identify and construct criteria of *usefulness* by capturing the value of use that professionals assign to the ontology tool and related actions, through their interpretation and use (Mayère 1997; Simonnot 2002), as well as through our engagement with these actors, involving their interaction with the constructed ontology within the context of their practice. This approach aligns with human-centered design principles, which advocate for involving end users as the most qualified actors to assess and appropriate the developed systems, and for grounding the design process in the *context of use* – considered as a «combination of users, goals and tasks, resources, and environment» – rather than in technical or technological possibilities (ISO 2019, 3.1).

We began by studying the information practices of professionals working in the unit, as previously described in Trzmielewski (2024) and illustrated in Figure 1. We conducted 16 participant observations and 20 interviews with supervising and apprentice doctors, researchers, clinical study coordinators, nurses, childcare workers, and secretaries. This led to the construction of an informational corpus composed of observation reports and interview transcripts. We used this corpus to extract candidate terms and create allergy user-oriented facets, which were subsequently discussed and tested during validation meetings with professionals to assess their usefulness for document indexing and searching of allergy professionals' information targets.

The informational corpus also helped us to identify the documents used daily by professionals, which enabled the development of a documentary cor-

pus containing 4,363 documents (608,910 words). All documents were less than three years old, reflecting the professionals' preference for recent documentation, and were entirely written in French – the working language of the unit. This documentary corpus was divided into three subcorpora – scientific¹, generalist², and clinical³ – and analyzed using automatic tools. Our goal was to extract allergy-related terminology as well as to evaluate the ALLERGIDOC Ontology as a tool supporting document processing tasks.

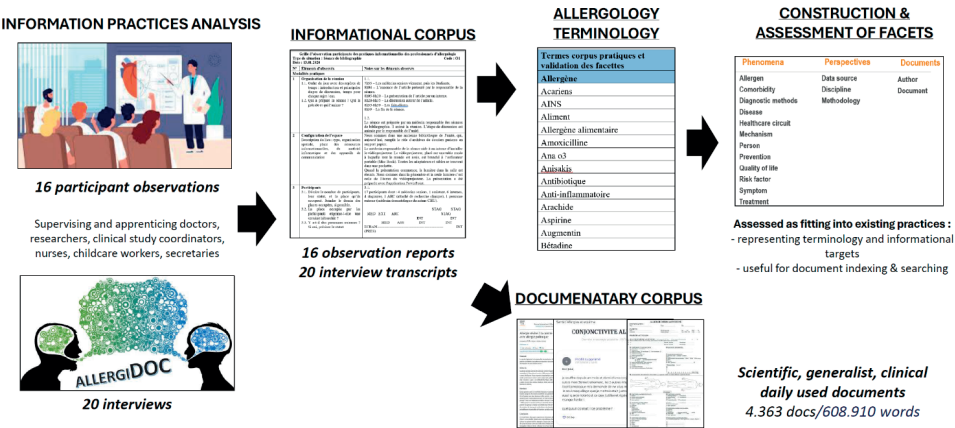


Figure 1. Process of producing contextual resources.

3.2. Ethical and legal obligations

The clinical and generalist documents from our corpus contained sensible data on people health condition, as well as personal data such as name, surname, age, sex, etc. We had to carry out few steps to use these data according to the French *Information Technology and Liberties Law* (Présidence de la République Française 1978) and the European *General Data Protection Regulation* (European Parliament, and Council of the European Union 2016, arts. 14-20). The collection of generalist data was approved by the university hospital's ethics committee and registered on *ClinicalTrials.gov* and the *French*

¹ The scientific subcorpus was composed of 864 titles and 761 scientific articles abstracts from the *Revue Française d'Allergologie* journal. As shown by the analysis of information practices, titles and abstracts are useful to professionals for representing and understanding the subject of documents (Trzmielewski 2023).

² The generalist subcorpus contained 403 titles of threads and 2,175 messages from *Doctissimo* generalist online health forum.

³ The clinical subcorpus included 70 clinical activity documents, including models of standardized protocols, medical reports, standardized questionnaires, check and patients lists, 89 anonymized medical reports, and 1 text document containing terms used in a document management system to categorize documents.

Health Data Hub. Considering this legislation, all personal data that appeared in clinical documents had to be erased. Although we erased personal data, we collected sensitive data, particularly related to individuals' health. Even though sensitive data are considered as a special category of personal data, the public interest inherent in our research justified their use (Cnil 2023). Moreover, the data were not collected directly from the individuals concerned, but rather secondarily. In this context, according to the French and European Law, we were required to inform *Doctissimo* (Doctissimo 2025) users, as well as Allergy Unit patients about the purpose of the data processing, the types of personal data collected, and the conditions of data storage. We were also obligated to guarantee their rights to rectification, erasure, etc., and to explain how to exercise them.

Doctissimo, the source of the collected generalist documents, is a public forum that does not require prior registration to access discussion threads. The content is clearly made public, and we are therefore authorized to reuse it for our research (Cnil 2023). Following the advice of a hospital methodologist and in accordance with the various legislative frameworks previously mentioned, we collectively informed forum users about our study. We posted a message on the first discussion thread in the "Allergies and Eczéma" section, titled *Modération sur Doctissimo*, which is the most visible thread in this forum category. On the forum, user profiles had already been anonymized, and all personal data had been removed by administrators in compliance with data protection regulations. Account names were pseudonymized, i.e., replaced by automatically generated alphanumeric strings, such as "sev93ydy". We did not collect any data on names, first names, pseudonyms, age, phone numbers, and email addresses that may have appeared in the message content.

Regarding the clinical corpus, to prevent the identification of individuals in medical records, a supervising doctor who assisted us with data collection removed the names of physicians and patients, hospital identification numbers, date of birth and consultation, as well as sex and geographical location-related data. He also used a tool available at the hospital to send individual notifications to patients whose data were included in the collection. No objections were received in response. At our end, we anonymized the documents sourced from document management systems used in the unit by removing the names of physicians, phone numbers, and email addresses.

The data, prepared accordingly, were securely stored on a university server at the University of Montpellier Paul-Valéry.

3.3. Use of semi-automatic tools

To collect data relevant to ontology design, we used the terminology extractor TermStat Web 3.0 (Drouin n.d.) and the text annotation tool Anno-

tator (Whetzel et al. 2011) to process documents from the documentary corpus. TermoStat is an automatic term acquisition tool developed and hosted at the University of Montreal. Its mechanism uses a method based on contrasting specialized and non-specialized textual corpora, combined with text tagging, character string extraction rules, and weighting and selection scores – such as frequency, specificity, log-likelihood, and weirdness – to identify terms (Drouin 2010). The tool takes a text as input and outputs a ranked list of candidate terms extracted from the text. Terms can be simple or complex, with each term assigned a score based on its frequency both in the analyzed corpus and in a reference corpus (Figure 2). The process of allergy terminology extraction (carried out in 2022) required converting the content into raw text and processing the three subcorpora separately. From each subcorpus, we extracted the 497 most frequently occurring candidate terms, corresponding to the number of terms manually extracted from the informational corpus. We compared the automatic extraction results for each subcorpus with the facets and terms identified through manual analysis of the informational corpus, to identify new useful terms and eliminate irrelevant or redundant ones. Allergy facets were particularly helpful in this task, as they provided a contextual semantic frame within which candidate terms could be used. The extraction results were validated by an allergist from the unit during a 1 hour and 30 minutes interview. Further validation was conducted by 10 other allergy professionals during the assessment of the 1.0 prototype of the ontology.

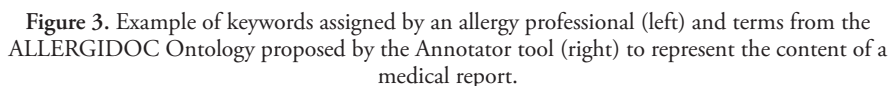
Liste des termes	Nuage	Statistiques	Structuration	Bigrammes
Candidat de regroupement				Fréquence
patient				1851
cas				1071
allergie				1031
étude				864
résultat				827
asthme				784
réaction				766
test				745
enfant				734
méthode				585
conclusion				569
traitement				546
p				510
introduction				490
ige				410
allergène				404
anaphylaxie				396
diagnostic				365
sensibilisation				349
prise				342
âge				314

Figure 2. A partial list of the terminological extraction of the scientific corpus with TernoStat.

Then, we assessed the utility of 3.0 version of the ALLERGIDOC Ontology to support automatic document processing – a need identified through the analysis of information practices (Trzmielewski 2024) – in situation of use. Seven professionals, i.e., one medical secretary, four attending doctors and two residents participated in meetings held in the Allergy Unit in 2025. Most of them reported having little or no prior experience with document annotation or indexing, as they are only accustomed to organizing work-related digital documents into folders. Some of them knew indexing principles from publicly available tools such as the PubMed search browser, consulted for scientific research.

Participants analyzed six selected daily-used digital documents, chosen among 15 pre-selected by the design team from the documentary corpus. The first participant – a medical secretary – chose two titles and abstract of scientific articles, two medical reports, and two messages from the generalist health

To generate annotations, participants copied and pasted textual content of documents in the online box, selected the ALLERGIDOC Ontology as dictionary⁶, and pressed the submit button. Annotator carried out text pre-processing task (such as eliminating stop words), matched words in the text to terms in the ontology by performing an exact string comparison (a “direct” match) between the text and ontology preferred and non-preferred terms (NCBO 2025). To perform the term recognition by using string matching on the dictionary, Annotator uses the Mgrep tool (Dai et al. 2008).



⁶ The ALLERGIDOC Ontology is hosted and available on BioPortal (Whetzel et al. 2011).

their informational, bodily and verbal behavior, and took notes. The content of the paper forms, as well as the observation and interview notes, were analyzed. We calculated an average number of keywords proposed by participants and the annotation tool, and we assessed their matching score and related patterns. Furthermore, we conducted a thematic analysis of their discourses on the proposed keywords and ontology terms, as well as on the perceived utility of the ALLERGIDOC Ontology to support document processing.

4. Results

Our study reveals that the data on the context of use of knowledge in the allergy unit, obtained through qualitative methods, leads to the identification of a set of useful human, informational, and socio-organizational elements. These elements guided the automatic extraction and annotation processes carried out by the design team and final users.

4.1. The semi-automatically extracted terminology

The specialized vocabulary used by professionals, which was extracted from the content of the observation reports and interview transcripts, as well as the user-oriented facets and validation meetings with professionals, guided the process of semi-automatic extraction of useful candidate terms, the discrimination of irrelevant ones, and the assessment of the performance of the extraction software. Almost all selected terms were accepted by the allergist during validation meetings. The involved professional excluded only 12 candidate terms that did not represent his practice. Ten other employees, participating in the assessment of the first version of the ALLERGIDOC Ontology, highlighted the relevance and utility of the extracted and further organized terminology. They emphasized that the ontology reflects “well”, “largely” and “completely” their professional reality and that the represented terms are used by them “almost every day”. The selected terms made sense to them, and they suggested that such terms could be used to index and search for allergy documentation, as well as to help to introduce new professionals to domain terminology.

Furthermore, we could observe above-average performance of TermoStat in relation to our extraction aims. From the total of 1,491 of the most frequently occurring candidate terms extracted to construct the first version of the ontology, on average, 63% were actually useful, while 37% were out of context (Table 1). This performance score demonstrates the relevance of the TermoStat tool for extracting medical terminology. At the same time, it highlights the need for contextual resources – i.e., daily-used documentation to

be represented and organized, contextual semantic frames, experts, and final users – to create a favorable environment and intellectual resources to guide the automated extraction process. However, the extraction tool also had its limitations, especially in processing complex terms, composed of two or three lexical units. We had to complete several single terms to adapt them to the allergy domain context. For example: “Arrêt” → “Arrêt cardio-respiratoire”, “Eosinophile” → “Œsophagite à éosinophiles”, “Anticorps” → “Anticorps monoclonaux”. To do so, we consulted the occurrence of terms in their phrase context within TermoStat.

	Scientific corpus	Clinical corpus	Generalist corpus	Total of extracted terms	Performance of the tool
Useful terms	192 (39%)	138 (27%)	240 (48%)	570 (38%)	63% of useful terms
Useful but redundant	119	171	77	367	
Not useful	186 (37%)	188 (38%)	180 (36%)	554 (37%)	37% terms out of context
Total	497	497	497	1491	

Table 1. Results of the semi-automatic term extraction supported by TermoStat, after assessment using contextual resources.

4.2. The utility of the ontology to support automatic document processing

The assessment meetings on the utility of the ALLERGIDOC Ontology version 3.0 to perform automatic document annotation revealed low performance. Among the average of 11 keywords proposed by professionals per document (Table 2), we noted only 2 exact matches per document with the ontology terms automatically proposed by Annotator (Table 3). This low efficiency is mostly due to three factors: 1) missing terms or terms that were not specific enough in the ontology; 2) the absence of contextual knowledge in the tools – knowledge that matters to allergy professionals and goes beyond string matching-based document processing; 3) the limitations of the annotation tool itself.

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Total	Mean
Participant 1	5	5	3	3	4	6	26	4,33
Participant 2	3	5	3	6	8	8	35	5,83
Participant 3	14	15	13	10	19	15	86	14,33

Participant 4	4	24	25	10	21	11	95	15,83
Participant 5	11	8	7	8	14	9	57	9,5
Participant 6	6	14	15	15	21	18	79	13,16
Participant 7	12	14	11	8	13	15	73	12,16
TOTAL	55	85	77	60	100	82	451	X
Mean (participants)	7,85	12,14	11	8,57	14,28	11,71	64,42	10,79
Mean (Annotator)	19	37	16	13	1	12	98	16,33

Table 2. The number of descriptors assigned by participants (green) and by annotation tool (orange) to represent the content of documents.

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Total	
Participant 1	2	0	0	0	0	1	3	
Participant 2	1	1	1	2	0	2	7	
Participant 3	5	2	4	5	0	4	20	
Participant 4	2	13	5	4	0	1	25	
Participant 5	0	5	0	1	0	1	7	
Participant 6	0	4	2	4	0	2	12	
Participant 7	4	3	1	6	0	3	17	
TOTAL	14	28	13	22	0	14	91	
Mean	2	4	1,86	3,14	0	2	X	2,16

Table 3. Scores showing the number of exact matches between keywords assigned by professionals and ontology terms proposed by the Annotator.

At the semantic level, the ontology terms lacking specificity included document section titles (e.g., “Introduction”, “Résultats”, “Conclusion”), dynamic phenomena (e.g., “Maladie”, “Traitement”, “Sensibilisation”, “Hospitalisation”), individuals (e.g., “Médecin”, “Patient”, “Allergologue”), and quantities (e.g., “Dose”, “Heure”), among others.

At the lexical level, employees suggested several keywords that were missing in the ontology: commercial drug names (e.g., “Zyrtec”, “Telfast”, “Wystam”, “Atarax”), Latin binomial names (e.g., *Solenopsis invicta*), names of events leading to hypersensitivity reactions (e.g., “Envenimation”, “Morsure”), as well as metonymic expressions such as “Tryptase” (a substance) used to designate “Dosage de la tryptase” (a blood test).

At the morphological level, professionals frequently used acronyms, such as “TC” for “Test cutané”, “AO” for “Angioœdème”, “AHI” for “Antihistaminique”, “ETP” for “Education thérapeutique”; abbreviations (e.g., “Allergo”, “Dermato”); hyphenated terms (e.g., “Angio-œdème”); unaccented terms

(e.g., “Eczema”); and plural forms designating allergen sources, diseases, and symptoms – many of which are currently absent from the ontology.

Also, many keywords proposed by professionals were present in the developed MKOS but literally absent from the corpus. This was frequently the case of hypernyms – such as the keyword “Pneumallergène” which was used to group various standard aeroallergens tested in patients: “Chat”, “Cyprès”, “Olivier”, “Platane”, “Ambroisie”, “Pariétaire”, etc., as reported in medical documents. The same observation applied to other allergens, diseases, and drugs classes. Some terms proposed by the employees were also semantically related to words appearing in the text, for example, the keyword “Pollen” was associated with terms referring to plants such as “Cyprès”, “Olivier”, “Platane”. Moreover, the annotation tool did not match words in the text when a plural form was missing in the ontology (e.g., “Hyménoptère” instead of “Hyménoptères”), or when ontology terms were morphologically close but either too broad (“Anaphylaxie” instead of “Choc anaphylactique”) or too narrow (“Eczéma atopique” instead of “Eczéma”).

Some issues were difficult to identify because, although the ontology terms corresponded to the keywords, they were not extracted by the tool. We assumed this could result from the absence of “rdfs:label” on some class names, missing underscores between words in certain complex terms within the ontology (e.g., “Syndrome oral”), typographical issues (e.g., “Angioedème” instead of “Angioœdème”); the presence of an apostrophe (as in “Trousse d’urgence”), or even the tool’s inability to extract complex lexical units at all. In addition, one document received only a single matching term, possibly due to a formatting problem. Therefore, we can question whether we were truly assessing the ontology itself, the annotator tool, or the combination of both as a single tool. Indeed, without the use of Annotator, the number of exact matches between keywords assigned by professionals and terms present in the ontology increased. There were, on average, more than 3 exact matches per document between the keywords assigned by professionals and the terms present in the ontology (Table 4).

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Total	
Participant 1	2	3	1	0	0	2	8	
Participant 2	2	4	3	4	2	2	17	
Participant 3	5	6	7	4	2	5	29	
Participant 4	0	14	9	5	7	2	37	
Participant 5	3	5	2	2	5	3	20	
Participant 6	1	7	5	6	5	2	26	
Participant 7	3	4	2	5	2	2	18	
TOTAL	16	43	29	26	22	18	154	
Mean	2,28	6,14	4,14	3,71	3,14	2,57	X	3,66

Table 4. Number of exact matches between keywords assigned by professionals and terms present in the ontology.

Nevertheless, professionals had no issues manipulating the annotation tool supported by the 3.0 version of the ontology, and most of them found the tool “hyper advanced”, “top”, “smart”, and stated that it “would make life easier”. The tool provided them with “interesting” descriptor suggestions they had not previously considered, such as those related to alternative medicine or disease stages. Professionals also shared with the design team some suggestions regarding potential applications of such a tool. All of them found the tool useful for annotating and indexing medical reports in electronic health records system, documents on appointment platforms, as well as social media messages in databases for their research—types of informational resources that currently lack adequate knowledge organization tools in both the hospital and the research laboratory where the employees work. They also shared with us their informational targets, such as common and rare skin tests (for allergens such as mites, drugs by pharmaceutical class and commercial name, exotic foods, etc.), blood tests, allergies, symptoms as expressed by specialists and patients (e.g., “Démangeaisons” vs. “Grattage”), affected body parts, alternative therapies, emotions/suffering, and photos – particularly emphasizing the need for detailed instructions on indexing and searching for allergens, which remains one of their most important informational priorities. In fact, they most frequently search for medical reports containing positive test results to support their clinical and research practices – for example to identify which types of patients reacted to a particular allergen – and make decisions on how to treat another patient with a similar medical history, or to compile statistics on the characteristics of patients allergic to the same agent. The medical secretary highlighted the need for interoperability between the Allergy Unit and the Pneumology Unit information systems, as these units are located close to each other and may share the same patients.

The annotation of allergens contributed to the gap between the number of annotations generated by Annotator – 16 terms per document – and the nearly 11 keywords proposed by the allergy professionals (Table 2). To improve this aspect, participants suggested including ancestor classes to group classes of allergens discussed in processed documents and prioritizing the indexing of only positive allergen test results. The potential broader classes already exist in the ontology as subclasses of the separated class “Allergen”, but they are not defined as direct broader classes of specific plants or foods, which are currently classified under “Living beings” and “Agricultural or manufactured products”, respectively. In addition, when (most) patients in the Allergy Unit undergo testing with standard baseline series – for example for aeroallergens or food allergens – the names of all allergens, whether negative or positive as results, are documented in medical reports. This distinction was not recognized by the annotation tool.

Moreover, professionals suggested introducing coordination rules to represent and search for complex subjects, particularly those involving static, dynamic, and abstract phenomena – such as agents, objects, events, processes, temporal aspects – and various attributes, especially within medical reports:

- allergy agent-disease, e.g., “SOA-fruits”, “Anaphylaxie au blé”,
- allergen-allergen cases of cross-reactivity, e.g., “Acariens-Crevette”,
- allergic test-allergic agent, e.g., “Test-Pneumallergène”
- symptom-severity (e.g., “Diarrhée-Sévère”), event (or disease)-severity (“Réaction allergique-Sévère”, as well as body part-clinical characteristics (“Main-Rouge”).

Accordingly, professionals also recommended introducing temporality-based indexing (e.g., allergy test-allergen-date) to reconstruct events from patients’ allergy histories.

Finally, during the tool assessment, professionals give us many suggestions regarding new terms (preferred terms, synonyms) and relationships to be added to the ontology, as well as terms and relationships to be removed.

5. Conclusions

This study shows that the effective use of automatic tools in medical ontology design may rely on resources and criteria that are external to the documentary corpus itself. In our case, such elements emerged from the analysis of the *context of knowledge use* by social actors – composed of human, informational, and socio-organizational environments – as well as through active interactions between the designer and the actors involved in validating and potentially using the system. By adopting an ecological and situated approach, we were able to observe and engage with allergy professionals in their natural work environments, analyze their information practices, and select documents representative of their daily tasks – all while respecting ethical and legal obligations related to the use of personal and sensitive data. This approach enabled us to assess and identify new elements requiring adjustment concerning the content, intentionality, and operability of the ALLERGIDOC Ontology. It also allowed us to identify new challenges, such as representing and searching for information related to allergens, temporality, and complex informational subjects within allergology documentation, as well as ensuring interoperability between hospital internal information systems. Such an approach goes beyond data-driven (Brewster et al. 2004) and satisfaction-based user evaluations (Sure et al. 2009), fostering instead *intelligent action* grounded in *search for intervention*, according to constructivist, complex systems modeling (Le Moigne 1999). Rather than predefining problems to solve, the contextual ontology design process itself became a space for formulating new, relevant problems

to address – to ultimately support allergology professionals’ practices more *effectively*.

Thus, a contextual approach can assist knowledge organization specialists in identifying relevant elements for using automatic tools in the development of MKOSs. In this framework, the context of use becomes a dynamic “frame of reference” (Vakkari 1997, 8), constructed throughout the research process, within which designers can interpret situations and take appropriate actions (Broudoux 2012), while leveraging automatic tools to build and evaluate useful MKOSs for intended environments. Our approach is illustrated in Figure 4 through the Utility in Context Model (UCM), a framework that can be applied to design complex and useful medical ontologies, as well as other types of classificatory and terminological systems.

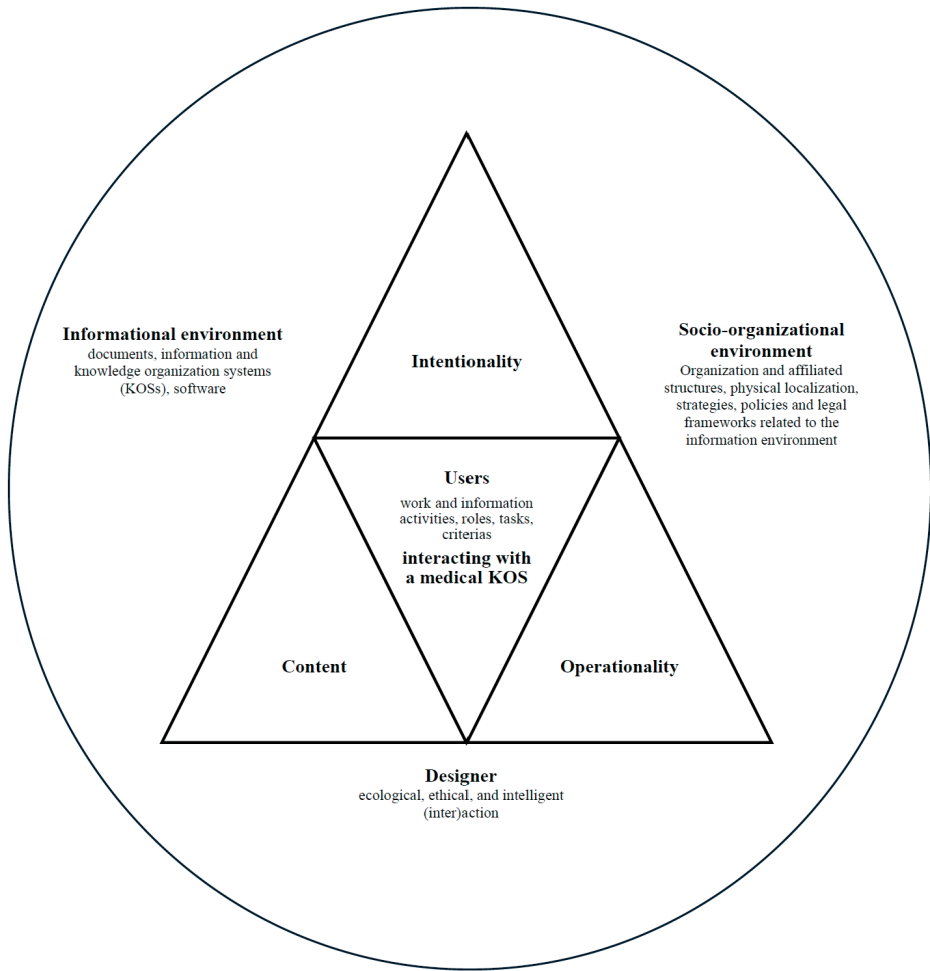


Figure 4. Utility in Context Model (UCM).

Nevertheless, our approach also reveals an inherent paradox. While automatic tools are generally employed to save time by automating tasks, their use within a qualitative, context-driven framework demands significant time investments for data collection, analysis, and validation. Furthermore, although we aimed to simulate real-world conditions, the experimental setting differed from the actual document organization environment of the Allergy Unit. Currently, professionals organize work-related digital documents in digital folders, with no established indexing system. Additional limitations stem from our reliance on external tools, such as the Annotator (which experienced some technical issues during our sessions) and TermoStat, which were not locally hosted – even though we used secure, university-provided services and excluded personal and sensitive data from the study. Moreover, we primarily evaluated the ontology representational capabilities, with limited assessment of its search functionalities in operational settings. A key limitation lies in the experimental nature of the study: it remains still a simulation rather than a full evaluation of the ontology integrated into a fully operational system. Assessment through the annotation tool was also constrained by the tool's limitations, such as its inability to extract complex terms from the ontology. The same issue arose during terminology extraction, suggesting that issues with complex term extraction may be common with certain automatic processing tools.

This research opens several perspectives for future work. Refining the ontology based on user feedback will be essential. Developing approaches to represent allergens, temporality, and complex subjects also remains a significant challenge. Ultimately, this study reinforces the need for more work emphasizing that medical ontologies for information processing and searching should not only be technically functional and logically consistent, but also meaningful and useful to their intended users.

Acknowledgments

I would like to thank the allergy professionals from the Allergy Unit of the Montpellier University Hospital – Selsabil Ayeb, Yanis Bouderbala, Jean-Luc Bourrain, Waffa Benderdour, Davide Caimmi, Pascal Demoly, and Snedjana Tintari – for their involvement and feedback regarding the present study. Special thanks also to the Information-Documentation students (referred to as “the design team”) at the University of Montpellier Paul-Valéry – Nada Bouhsina, Claire Casasso, Sabrina Duc, and Anne Guerre – for their valuable contribution to data collection and analysis.

References

- Almeida, Mauricio Barcellos, and Fernanda Farinelli. 2017. "Ontologies for the representation of electronic medical records: The obstetric and neonatal ontology." *Journal of the American Society for Information Science and Technology* 68 (11): 2529-42. <https://doi.org/10.1002/asi.23900>.
- Baneyx, Audrey, and Jean Charlet. 2006. "Évaluation, évolution et maintenance d'une ontologie en médecine : état des lieux et expérimentation." *Revue I3-Information, Interaction, Intelligence* special issue on *Ressources termino-ontologiques*: 147-73. Cépaduès Éditions.
- Barité, Mario. 2018. "Literary warrant." *Knowledge Organization* 45 (6): 517-36. <https://doi.org/10.5771/0943-7444-2018-6-517>.
- Brewster, Christopher, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. 2004. "Data-Driven Ontology Evaluation." In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, edited by Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2004/pdf/737.pdf>.
- Bringay, Sandra, Catherine Barry, and Jean Charlet. 2005. "Les annotations pour gérer les connaissances du dossier patient." In *Actes de la conférence IC 2005 – 16es journées francophones d'Ingénierie des Connaissances – 1-3 juin 2005 Nice*, edited by Marie-Christine Jaulent. <https://inria.hal.science/hal-01025416/document>.
- Broudoux, Evelyne. 2012. "Vers l'objet documentaire (re)contextualisé," paper presented at *9ème congrès des enseignants documentalistes de l'Éducation Nationale : Objets documentaires numériques : nouvel enseignement ?*, Paris, March 22. https://archivesic.ccsd.cnrs.fr/sic_00715868/document.
- Choi, Yunseon. 2016. "Supporting better treatments for meeting health consumers' needs: extracting semantics in social data for representing a consumer health ontology." *Information Research* 21 (4). <https://informationr.net/ir/21-4/paper731.html>.
- Cisek, Sabina. 2013. "Metodologia jakościowa we współczesnej informatologii: wybrane aspekty." *Przegląd Biblioteczny* 81 (3): 299-310. <https://doi.org/10.36702/pb.248>.

- Clavier, Viviane. 2010. "Indexer des parcours thématiques pour valoriser les collections de presse numérisée." In *Document numérique, entre permanence et mutations. Actes du 13^{ème} Colloque International sur le Document Électronique (CIDE.13), 16-17 décembre 2010, INHA, Paris*, edited by Madjid Ihadjadene, Manuel Zacklad and Khaldoun Zreik. Europia Productions.
- Cnil. 2023. "Données sensibles." Last modified August 2023. <https://www.cnil.fr/fr/definition/donnee-sensible>.
- Dai, Manhong, Nigam H. Shah, W. Xuan, et al. 2008. "An Efficient Solution for Mapping Free Text to Ontology Terms." In *American Medical Informatics Association Symposium on Translational BioInformatics, AMLA-TBI'08 12-15 March 2008 San Francisco, United States*.
- Doctissimo. 2025. "Forum santé – Allergies et eczéma." Last modified July 3. https://forum.doctissimo.fr/sante/allergies-eczema/liste_sujet-1.htm.
- Drouin, Patrick. 2010. "TermoStat – Guide de l'utilisateur." Last modified December 2010. https://termostat.ling.umontreal.ca/doc_termostat/doc_termostat_en.html.
- Drouin, Patrick. n.d. "TermoStat Web 3.0." Last access October 2025. <https://termostat.ling.umontreal.ca/>.
- European Parliament, and Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Official Journal of the European Union. Approved 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>.
- Hassan, Mohsen. 2017. "Knowledge Discovery Considering Domain Literature and Ontologies: Application to rare Diseases." PhD diss., Université de Lorraine.
- International Organization for Standardization. *Ergonomics and human-system interaction – Part 210: Human-centred design for interactive systems*. ISO 9241-2010:2019. ISO. Published 2019; confirmed 2025. <https://www.iso.org/standard/77520.html>.
- Iyer, Hemalata, and K. S. Raghavan. 2018. "Medical ontology: Siddha System of Medicine." In *Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal*, edited by Fernanda Ribeiro and Maria Elisa Cerveira. Ergon. <https://doi.org/10.5771/9783956504211-347>.
- Lardera, Marco, and Birger Hjørland. 2021. "Keyword". *Knowledge Organization* 48 (6): 430-56. <https://doi.org/10.5771/0943-7444-2021-6-430>.

- Le Moigne, Jean-Louis. 1999. *La modélisation des systèmes complexes*. Dunod.
- Liu, Hao, Shuxin Zhou, Zhehuan Chen, Yehoshua Perl, and Jiayin Wang. 2024. "Using Generative Large Language Models for Hierarchical Relationship Prediction in Medical Ontologies." In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, Orlando, FL, USA, June 03-06. Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/ICHI61247.2024.00040>.
- Lopes, Lucelene, Renata Vieira, Maria José Finatto, Daniel Martins, Adriano Zanette, and Luiz Carlos Ribeiro Jr. 2009. "Automatic extraction of composite terms for construction of ontologies: an experiment in health care area." *Electronic Journal of Communication, Information & Innovation in Health* 3 (1): 72-84. <https://doi.org/10.3395/RECIIS.V3I1.244EN>.
- Mayère, Anne. 1997. "Produits et services d'information : proposition de dépassement de la théorie standard de l'information." In *Société informationnelle : enjeux sociaux et approches économiques*, edited by Anne Mayère. L'Harmattan.
- Mucchielli, Alex, and Clair Noy. 2005. *Etudes des communications – Approches constructivistes*. Armand Collin.
- Muresan, Smaranda, and Judith L. Klavans. 2013. "Inducing terminologies from text: A case study for the consumer health domain." *Journal of the American Society for Information Science and Technology* 64 (4): 727-44. <https://doi.org/10.1002/asi.22787>.
- NCBO. 2025. "BioPortal Help – Annotator." Last modified January 7 2005. https://www.bioontology.org/wiki/BioPortal_Help#Annotator.
- Oh, Kyong Eun, Soohyung Joo, and Eun-Ja Jeong. 2015. "Online Consumer Health Information Organization: Users' Perspectives on Faceted Navigation." *Knowledge Organization* 42 (3): 176-86. <https://doi.org/10.5771/0943-7444-2015-3-176>.
- Présidence de la République Française. 1978. *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. In *Journal officiel de la République française*. Published 1978; confirmed July 2, 2025. <https://www.legifrance.gouv.fr/loda/id/JORFTEXT0000000886460>.
- Raghavan, K. S., and C. Sajana. 2010. "NeurOn. Modeling ontology for neurosurgery." In *Paradigms and conceptual systems in knowledge organization: Proceedings of the Eleventh International ISKO Conference 23-26 February 2010 Rome, Italy*, edited by Claudio Gnoli and Fulvio Mazzocchi. Ergon.
- Richard, Marion. 2017. "Apport de la modélisation ontologique pour le partage des connaissances en psychiatrie." PhD diss., Université Pierre et Marie Curie (Paris 6).

- Simonnot, Brigitte. 2002. "De la pertinence à l'utilité en recherche d'information : le cas du Web." In *Recherche récentes en sciences de l'information. Convergences et dynamiques – Actes du colloque MICS-LERASS 21-22 mars 2002 Toulouse*, edited by Viviane Couzinet and Gérard Régimbeau. ADBS Editions.
- Soergel, Dagobert, Tony Tse, and Laura Slaughter. 2004. "Helping Healthcare Consumers Understand: An 'Interpretative Layer' for Finding and Making Sense of Medical Information." In *MEDINFO Proceedings of the 11th World Congress on Medical Informatics*, edited by Marius Fieschi, Enrico Coiera and Yu-Chan Jack Li. IOS Press. <https://ebooks.iospress.nl/publication/21125>.
- Sure, York, Steffen Staab, and Rudi Studer. 2009. "Ontology Engineering Methodology." In *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer. Springer. https://doi.org/10.1007/978-3-540-92673-3_6.
- Talja, Sanna, Heidi Keso, and Tarja Pietiläinen. 1999. "The production of 'context' in information seeking research: a methodological view." *Information Processing and Management* 35 (6): 751-63. [https://doi.org/10.1016/S0306-4573\(99\)00024-2](https://doi.org/10.1016/S0306-4573(99)00024-2).
- Tang, Xiao-Bo, Wei Wei, Guang-Chao Liu, and Juan Zhu. 2017. "An Inference Model of Medical Insurance Fraud Detection: Based on Ontology and SWRL." *Knowledge Organization* 44 (2): 84-96. <https://doi.org/10.5771/0943-7444-2017-2-84>.
- Tapi Nzali, Mike Donald, Jérôme Azé, Sandra Bringay, Christian Lavergne, Caroline Mollevi, and Thomas Optiz. 2018. "Reconciliation of patient/doctor vocabulary in a structured resource." *Health Informatics Journal* 25 (4): 1219-31. <https://doi.org/10.1177/1460458217751014>.
- Tennis, Joseph T. 2012. "Le poids du langage et de l'action dans l'organisation des connaissances : Position épistémologique, action méthodologique et perspective théorique." *Etudes de communication* 39: 15-40. <https://doi.org/10.4000/edc.4010>.
- Toro, Sabrina, Anna V. Anagnostopoulos, Sue Bello, et al. 2024. "Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI)." *Journal of Biomedical Semantics* 15. <https://doi.org/10.1186/s13326-024-00320-3>.
- Trzmielewski, Marcin. 2023. "Vers l'élaboration d'un système d'organisation des connaissances en allergologie : l'analyse des documents et des pratiques informationnelles des acteurs." PhD diss., Université Paul-Valéry Montpellier 3.

- Trzmielewski, Marcin. 2024. "Towards an Allergy Ontology: A Utility-Based Contextual Approach." *Knowledge Organization* 51 (3): 151-68. <https://doi.org/10.5771/0943-7444-2024-3-151>.
- Vakkari, Pertti. 1997. "Information seeking in context: A challenging metatheory." In *ISIC '96: Proceedings of an international conference on Information seeking in context*, edited by Pertti Vakkari, Reijo Savolainen and Brenda Dervin. Taylor Graham. <https://dl.acm.org/doi/10.5555/267190.267221>.
- Whetzel, Patricia L., Natalya F. Noy, Nigam H. Shah, et al. 2011. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." *Nucleic Acids Res* 39 (Web Server issue): W541-5. <https://doi.org/10.1093/nar/gkr469>.

Exploring the relationships between the Cross-Cutting Women's Agenda - Multi-Year Plan 2024-2027 and the Knowledge Organization of legislative proposals under processing on women in Brazil

Carla Maria Martellote Viola*, Carla Beatriz Marques Felipe**

Abstract: Law N° 14.802 (2024) established *Brazil's Multi-Year Plan (2024–2027)*, highlighting five cross-cutting agendas, including the *Cross-Cutting Women's Agenda*, which guides this study. The research aims to represent and organize legislative proposals concerning women under analysis in the Chamber of Deputies. It proposes a classification aligned with the agenda's goals and examines how these proposals correlate with their objectives. The study combines bibliographic and documentary research, using the keyword *mulher* to retrieve proposals from the *Sileg-Tramitação system*, adopting both qualitative and quantitative approaches. A total of 713 proposals were identified; 691 were classified into five dimensions: living free from violence (287), economic autonomy (135), equality in decision-making and power (58), health and environment (136), and institutional capacity (75). The research highlights the importance of organizing legislative data to support inclusive public policies and promote *Sustainable Legislative Infodefeminism* – a framework for transparency, data access, and gender equality.

Keywords: Knowledge Organization, Cross-Cutting Women's Agenda - Multi-Year Plan 2024-2027, Legislative proposals, Women in Brazil, Sustainable Legislative Infodefeminism.

1. Introduction

Law N° 14.802, enacted on January 10, 2024, established the *Brazilian Multi-Year Plan* for the period 2024–2027 (Presidência da República 2024a). This normative act of the Executive Branch defined four cross-cutting priority agendas: (1) Children and Adolescents; (2) Women; (3) Racial equality; (4) Indigenous peoples; and (5) The environment. The priority agenda guiding this research is the Cross-Cutting Women's Agenda, which encompasses various objectives and targets to be achieved through collaborative efforts among fed-

* Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Coordenação de Tecnologias para Informação (COTEC), Brasília-DF, Brasil. viola.carla@gmail.com. ORCID: 0000-0002-0895-8163.

** Universidade Federal do Rio de Janeiro. Rio de Janeiro-RJ, Brasil. felipecarla12@gmail.com. ORCID: 0000-0001-5277-9165.

eral entities. However, a compilation of legislative proposals concerning women is already under consideration in the Chamber of Deputies (Ministério do Planejamento e Orçamento 2024), a body of the Brazilian Legislative Branch, preceding this plan.

Based on the *Cross-Cutting Women's Agenda*, how can legislative proposals concerning women currently under consideration in the Chamber of Deputies be represented and organized? The general objective of this research is to propose a classification system for legislative proposals under review that address women's issues, in alignment with the principles and dimensions outlined in the *Cross-Cutting Women's Agenda*. The specific objectives are: (1) to classify the legislative proposals introduced and under deliberation during the 56th Legislature; and (2) to assess the extent to which these proposals correspond to the thematic axes and priorities of the Cross-Cutting Women's Agenda.

The justification for this research is grounded in the principle that gender mainstreaming in public policy must acknowledge that inequalities between men and women – considering their diverse identities – permeate all areas of society and governance. This recognition requires concrete, cross-sectoral actions coordinated among various levels of government. Accordingly, it is essential to represent and organize parliamentary initiatives in order to assess whether they effectively address the diverse needs, realities, and experiences of women and girls, thereby contributing to the promotion of a more inclusive and equitable society.

2. Deconstructing Power and Reimagining Representation in Knowledge Organization

Knowledge Organization (KO) is the invisible architecture of our intellectual world. It is the intricate system of principles, schemes, and practices that governs how information is structured, described, and made accessible. From the Dewey Decimal system arranging books on a library shelf to the metadata tags ordering search results on the internet, KO systems are the foundational frameworks that shape how we navigate, encounter, and ultimately understand the vast universe of human knowledge (Zeng 2008; Hjørland 2016).

For much of its history, this field operated under a paradigm of universalism and neutrality, aspiring to create a single, objective map of all knowledge that could be universally applied. However, this seemingly benign mission of efficient retrieval has come under intense scrutiny. Scholars from critical, feminist, and postcolonial perspectives have compellingly argued that these traditional systems are anything but neutral. Instead, they are cultural artifacts, deeply embedded with the biases, hierarchies, and power structures of the dominant societies that created them. This critique reveals that the act of

classifying is never a simple act of ordering; it is an act of defining, of including and excluding, and of asserting a particular worldview as authoritative (Bowker and Star 2000).

A contemporary, socially responsible approach to KO must actively dismantle the epistemological assumptions of traditional universalism. By integrating the profound critiques of feminist theorists like Olson (1998; 2002; 2007), postcolonial thinkers like Spivak (2010), and the practical innovations of phenomenological and user-centric models, KO can transition from a discipline focused on mere document retrieval to one centered on epistemic justice and equitable representation. To achieve this, KO must move beyond rigid hierarchical structures, embrace the complexities of context and experience, and empower marginalized communities to participate in the construction of their own representation.

2.1. Cracks in the Traditional KO Paradigm

The traditional model of Knowledge Organization, exemplified by systems like the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC), was born from a 19th-century positivist ambition to create a comprehensive and objective order for the world's knowledge. The underlying assumption was that knowledge could be neatly divided into discrete, mutually exclusive categories arranged in a logical, hierarchical structure, mirroring a supposed natural order. This top-down, universalist approach promised efficiency and standardization, allowing any user in any library to navigate its collection using the same logical framework (Hjørland 2008).

The critical flaw in this paradigm, however, is its unexamined claim to neutrality. As Berman (1971) famously exposed in his *Prejudices and Antipathies*, library classification systems and subject headings are replete with biased, pejorative, and outdated terminology that marginalizes entire groups of people. His work demonstrated how subject headings often rendered women, LGBTQ+ individuals, and ethnic minorities invisible or pathologized them, using language that reflected and reinforced societal prejudice. For example, topics related to women were frequently subordinated under broader, male-centric headings, and non-Western belief systems were often categorized under *mythology* or *folklore*, while Christianity was treated as the normative standard for *religion*.

This critique reveals that classification is not a passive reflection of reality but an active construction of it. As Bowker and Star (1999) argue in *Sorting Things Out*, classification systems are powerful political and ethical tools that have profound real-world consequences. By creating categories and setting their boundaries, these systems create "visibility and invisibility," conferring legitimacy on certain ways of knowing while erasing others. The hierarchical

structure itself, a hallmark of traditional KO, inherently creates relationships of subordination and superordination, often mirroring colonial and patriarchal power dynamics. This structure struggles to accommodate interdisciplinary knowledge and complex, intersecting identities, forcing them into ill-fitting and reductive categories. The result is a system that, under the guise of objective order, perpetuates a specific, dominant worldview – typically that of the Western, white, male academic – and in doing so, commits a form of epistemic violence against those who fall outside its narrow parameters.

2.2. Weaving Webs of Situated Knowledge

It is precisely this illusion of neutrality and the rigidity of hierarchical structures that feminist KO scholars have sought to dismantle. Leading this charge, Olson (1998; 2007) provides a powerful framework for rethinking the ethical and political dimensions of knowledge organization from a feminist perspective. Olson argues that the very foundations of traditional KO – its universalism, its hierarchical logic, and its disavowal of context – are fundamentally patriarchal in nature. She proposes a radical alternative, not as a replacement system, but as a set of guiding principles for creating more inclusive and just organizational structures.

At the heart of Olson's critique is the rejection of a single, universal framework. Drawing on feminist standpoint theory, which posits that knowledge is always situated and that marginalized groups possess a unique and valuable perspective on the nature of power (Harding 2004), Olson argues that any claim to a universal "view from nowhere" is a disguised "view from the top." Instead of a single truth, she advocates for a system that can accommodate a multiplicity of truths and perspectives. This requires shifting from a hierarchical pyramid to a decentralized network or web. A web-like structure, unlike a rigid tree, allows for multiple points of entry, non-linear relationships, and the representation of complex, intersecting concepts. It reflects a world where categories are not mutually exclusive but porous and interconnected, better representing the messy reality of lived experience (Olson 1998; 2007).

This emphasis on context and experience is crucial. Olson insists that KO systems must move beyond organizing abstract concepts and instead engage with the concrete, embodied experiences of users, particularly women (Olson 1998; 2007). This aligns with Haraway's (1988) influential concept of "situated knowledges", which rejects the false dichotomy between relativism and objectivity, arguing for an embodied objectivity rooted in specific, partial perspectives. For KO, this means that the knowledge of communities, especially those historically silenced, must be treated as a valid and essential resource in the design of information systems. It necessitates a participatory approach where catalogers and system designers are not distant authorities but

facilitators who work in collaboration with knowledgeable communities. Finally, Olson (2002) demands that we recognize power as an inherent factor in knowledge, acknowledging that the “power to name” is a significant political act that shapes discourse and reality.

2.3. The postcolonial challenge

Parallel to the feminist critique, postcolonial theory offers a potent lens for understanding the limitations of Western-centric KO systems. Spivak's (2010) seminal essay, *Can the Subaltern Speak?*, while not about library science, resonates deeply with the central dilemmas of knowledge representation. Spivak introduces the “subaltern” to describe those who are socially, politically, and geographically outside the hegemonic power structures, rendered voiceless by colonial and neocolonial discourse. Her question challenges the very possibility of authentic representation, asking whether a person from a marginalized position can ever truly be heard or understood within the epistemological frameworks of the dominant culture.

In the context of KO, this translates to a critical question: “Can subaltern knowledge be represented without being distorted or co-opted by the very systems that have historically marginalized it?”. Spivak's work pushes beyond a simple call for inclusion. She challenges the liberal belief that the mere production of information or the provision of access automatically constitutes knowledge or empowerment (Spivak 2010).

Access to a library or a database is of limited value if its underlying organizational logic erases or misrepresents one's history, culture, and way of being. For the subaltern woman, who exists at the intersection of multiple forms of oppression – gender, race, class, and colonial history – this problem is compounded. As Crenshaw's (1991) theory of intersectionality powerfully demonstrates, systems of power are interlocking, and an analysis that focuses only on gender without considering race or class will fail to capture the specific nature of her experience.

Spivak's interventionist and contestatory style provides a model for an alternative KO practice – one that is not content with simply adding new categories but is committed to questioning the foundations of the system itself (Spivak 2010). This requires an approach that, as argued by scholars in the decolonizing knowledge movement, actively works to center knowledge systems, treating them not as objects of study but as valid, parallel frameworks for organizing the world. It means creating spaces for “epistemic disobedience,” challenging the hegemony of Western scientific and bibliographic standards and fostering a pluriversal, rather than a universal, approach to knowledge.

2.4. Synthesizing a new path forward

Considering these powerful critiques, a path forward requires synthesizing their insights with emerging models within KO itself. The work of thinkers like Vignaux (2000), Souza (2006), and Szostak (2024) provides practical and philosophical tools for building this new paradigm. Vignaux (2000) reminds us that classification is a fundamental cognitive act. In a world perceived as fragmented, humans constantly seek to create order by classifying, subdividing, and applying labels to understand phenomena. This impulse to name and categorize is not inherently oppressive; the challenge lies in how it is implemented.

This is where the principles outlined by Souza (2006) become essential. Souza argues that any effective classification process must be developed based on two fundamental reference points: the nature of the information being classified, and the specific characteristics and needs of the user community. This user-centric principle directly counters the top-down universalism of traditional KO and aligns perfectly with the feminist call for context, experience, and community involvement.

Building on this, Szostak (2024) advocates for a fundamental shift in the very purpose of KO. He argues that the field should embrace purposes beyond merely classifying documents for retrieval, moving toward a phenomenon-based approach that aims to reflect ontological reality. Instead of organizing books about women, such a system would seek to organize knowledge about the complex, multifaceted phenomenon of “womanhood” itself, drawing connections across disciplines like health, labor, politics, sociology, and art.

This approach, like the work of Gnoli et al. (2018) about the Integrative Levels Classification, favors a holistic, interdisciplinary model that can better capture the richness and complexity of the world. Such a system would be inherently more web-like and less hierarchical, allowing users to explore the myriad interconnected facets of a given phenomenon, providing a practical structure for realizing the theoretical goals outlined by Olson and other critical scholars.

3. Methodological Procedures

For conducting the research, the methodological procedures are outlined based on a mixed-methods approach, combining qualitative and quantitative dimensions with exploratory objectives, aiming to deepen the knowledge on the identified theme. As recommended by Creswell and Creswell (2018), this approach allows for a more complete and robust understanding of the phenomenon, integrating the breadth of numerical data with the depth of content analysis.

The investigation begins with bibliographic research to build a solid theoretical framework. This conceptual foundation is grounded in two main areas: Knowledge Representation and Organization (KRO), which explores how social categories are constructed and maintained in information systems, based on authors like Olson (2002) and Hjørland (2017); and gender studies, using the theoretical contribution of Scott (1986) to analyze how gender operates as a primary category of power in social and political relations.

The next stage consists of documentary research, which, according to Gil (2008), uses primary sources that have not yet undergone analytical treatment. The documentary corpus is composed of the *Cross-Cutting Women's Agenda* and the legislative proposals of the 56th Legislature, retrieved from the Chamber of Deputies' portal through the *Sileg-Tramitação* system. The collection was carried out on October 14, 2024, using the descriptor *mulher* (woman).

The quantitative dimension of the research is manifested in the mapping and categorization of the legislative proposals found. In contrast, the qualitative dimension focuses on the content analysis of the documents, seeking to understand the different representations of the term *mulher* and the underlying logics of the agendas and proposals, to unveil the knowledge structures that support them.

4. Results

After reviewing the *Cross-Cutting Women's Agenda*, which is part of the Brazilian federal government's 2024–2027 Multi-Year Plan (2024b), the classification of legislative proposals concerning women initiated and under review in the 56th Legislature of the Chamber of Deputies was organized into five dimensions (Fig. 1): (1) Living free from violence; (2) Economic autonomy; (3) Equality in decision-making and power; (4) Health and the environment; and (5) Institutional capacity.

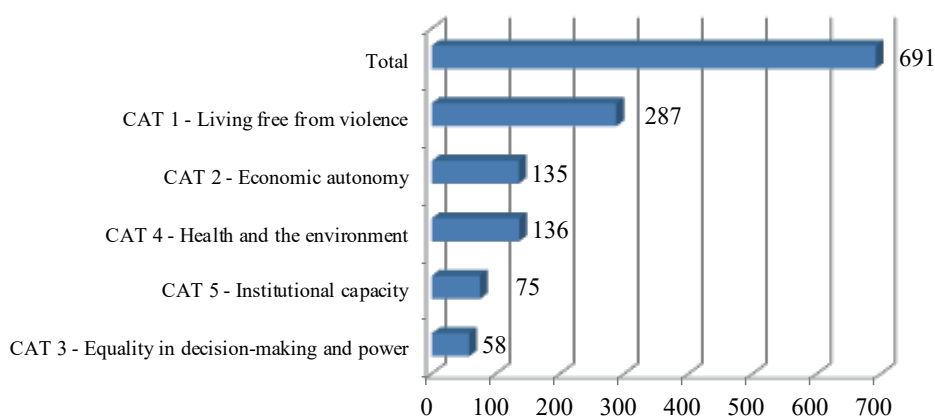


Figure 1. Legislative Proposals on Women.

Based on these five categories (CAT), 713 legislative proposals were classified, retrieved from the legislative system using the Portuguese descriptor *mulher* in titles or indexing terms. Of this total, 22 proposals were discarded: 5 (five) for addressing topics contrary to the *Cross-Cutting Women's Agenda*, and 17 for being related to honors and commemorative dates (Fig.1).

Proposals were considered contrary when they reinforced stereotypes, denied gender rights, or promoted setbacks in equality. The assessment was context-sensitive and aligned with the guiding principles of the *Cross-Cutting Women's Agenda (2024–2027)*.

Out of the 691 classified proposals, 58 are in *CAT 3 - Equality in decision-making and power*; 75 in *CAT 5 - Institutional capacity*; 136 in *CAT 4 - Health and the environment*; 135 in *CAT 2 - Economic autonomy*; and 287 in *CAT 1 - Living free from violence* (Fig.1).

CAT 1 – Living Free from Violence (287 Proposals - 41.5%).

This category overwhelmingly dominates the legislative agenda. It reflects a persistent political and societal recognition of gender-based violence as Brazil's most urgent women's issue. The large volume of bills suggests broad consensus and possibly public pressure, driven by high-profile femicide cases and alarming national statistics. However, the breadth of this category may also mask duplication and reactive policymaking rather than systemic reforms.

CAT 4 – Health and the Environment (136 Proposals - 19.7%)

This category shows strong representation, likely influenced by growing awareness of reproductive rights, mental health, and environmental justice as intersecting feminist issues. It may also encompass maternal health, access to healthcare in rural areas, and women's roles in environmental protection. The prominence of this dimension reveals an expanding policy vision that integrates social determinants of health with gender.

CAT 2 – Economic Autonomy (135 Proposals - 19.5%)

Legislative activity here suggests growing concern with women's access to employment, entrepreneurship, equal pay, and financial literacy. This near-equal representation alongside health points to a policy shift from protection to empowerment, aligning with international agendas.

CAT 5 – Institutional Capacity (75 Proposals - 10.9%)

Proposals in this category are likely addressing legal frameworks, data collection, gender mainstreaming in public institutions, and training of officials. Although smaller in number, their strategic significance is high. They underpin the successful implementation of all other categories, suggesting that while the quantity is lower, the quality or impact potential of these proposals may be foundational.

CAT 3 – Equality in Decision-Making and Power (58 Proposals - 8.4%)

This underrepresentation is concerning. Despite being central to dismantling structural inequality, this area is least legislated. It reveals the inertia in

shifting entrenched patriarchal political structures. Policies in this area likely involve quotas, leadership incentives, and political participation reforms. The low number may indicate legislative discomfort or resistance to redistributive changes in political power.

The legislative distribution across the five categories of the *Agenda Transversal das Mulheres* reveals a strong emphasis on combating violence, but significant underrepresentation in proposals promoting women's political empowerment. This pattern underscores a reactive rather than transformational legislative approach to gender equality. While Brazil is making strides in areas like health and economic rights, more deliberate action is needed to institutionalize gender perspectives and democratize power structures within governance.

That way, the imbalance across categories suggests that while immediate and visible threats to women's safety receive attention, transformative policies targeting systemic inequality, especially in politics, remain marginalized. Like this, the classification process enhances transparency and accountability, enabling civil society and researchers to track legislative alignment with national gender goals.

These results contribute to *Sustainable Legislative Infofeminism* (Viola 2023), a movement advocating for the enhancement and expansion of data access and information dissemination, based on technology, to develop knowledge about the evolving rights of women.

5. Conclusions

In conclusion, the evolution of Knowledge Organization from a technical discipline of ordering to a critically engaged field of social ethics marks a vital paradigm shift. The foundational critiques offered by feminist and postcolonial scholars have irrevocably demonstrated that traditional, universalist systems are inadequate and often harmful, perpetuating biases that silence and marginalize.

By embracing the principles articulated by Olson – of networked structures, situated knowledge, and the recognition of power – and heeding Spivak's challenge to create space for subaltern voices, KO can begin to redress these historical injustices. The future of a responsible and effective KO lies in synthesizing these critical insights with pragmatic, phenomenon-based, and user-centric models. This is not a quest for a new, perfect universal system, but rather an ongoing, collaborative commitment to building a pluriverse of knowledge systems that are flexible, contextual, and co-created with the communities they are meant to serve, ensuring that the power to name is shared by all.

Ensuring women's rights and gender equality depends on eradicating violence, promoting economic autonomy, and increasing female participation in positions of power. This transformation requires the collaboration of various government bodies and public policies based on data. Therefore, it is important to monitor future public policies that may arise from legislative proposals.

The transparency and accessibility of these results foster women's participation in the legislative process and promote awareness of and tracking legislative proposals in the Chamber of Deputies, thereby providing visibility to legislative data and information, with the aim of achieving the sustainability of women.

While the legislative content is Brazil-specific, the methodology – grounded in gender analysis and knowledge organization – is transferable. The five categories resonate with global frameworks like *Convention on the Elimination of All Forms of Discrimination Against Women* (United Nations 1979) and the *2030 Agenda* (United Nations 2015), allowing for adaptation to different national contexts.

This research also promotes citizen engagement and supports public policymaking. Systematizing legislative data with a gender perspective also counters disinformation and strengthens the democratic process through evidence-based monitoring.

References

- Berman, Sanford. 1971. *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People*. Scarecrow Press.
- Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press.
- Crenshaw, Kimberle W. 1991. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics (1989)." In *Feminist Legal Theory*, edited by Katharine T. Bartlett and Rosanne Kennedy, Routledge. <https://doi.org/10.4324/9780429500480-5>.
- Creswell, John W., and J. David Creswell. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th ed. Sage Publications Inc.
- Gil, Antonio Carlos. 2019. *Métodos e técnicas de pesquisa social*. 7th ed. GEN Atlas.

- Gnoli, Claudio, Andreas Ledl, Ziyoun Park, and Marcin Trzmielewski. 2018. "Phenomenon-based vs. disciplinary classification: possibilities for evaluating and for mapping." In *Challenges and opportunities for knowledge organization in the digital age: proceedings of the Fifteenth International ISKO Conference*, 9-11 July 2018, Porto, Portugal, edited by Fernanda Ribeiro and Maria Elisa Cerveira. Baden-Baden: Ergon, 653-62.
- Haraway, Donna. 1988. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14(3): 575-99. <https://doi.org/10.2307/3178066>.
- Harding, Sandra. 2004. *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies*. Routledge.
- Hjørland, Birger. 2008. "What Is Knowledge Organization (KO)?" *Knowledge Organization* 35(2/3): 86-101. <https://doi.org/10.5771/0943-7444-2008-2-3-86>.
- Hjørland, Birger. 2016. "Knowledge Organization." *Knowledge Organization* 43 (6): 475-84. <https://doi.org/10.5771/0943-7444-2016-6-475>.
- Hjørland, Birger. 2017. "Domain Analysis". *Knowledge Organization* 44 (6): 436-64. <https://doi.org/10.5771/0943-7444-2017-6-436>.
- Ministério do Planejamento e Orçamento. 2024. *Agenda Transversal das Mulheres*. Série Planejamento Nacional.
- Olson, Hope A. 1998. "Educação para catalogação é/como estudos femininos." *The Serials Librarian* 35 (1-2): 153-66. https://doi.org/10.1300/J123v35n01_09.
- Olson, Hope A. 2002. *The Power to Name: Locating the Limits of Subject Representation in Libraries*. Kluwer Academic Publishers.
- Olson, Hope A. 2007. "How we construct subjects: a feminist analysis." *Library Trends*, 56 (2): 509-41.
- Presidência da República. 2024. "Lei nº 14.802, de 10 de janeiro de 2024. Institui o Plano Plurianual da União para o período de 2024 a 2027."
- Scott, Joan W. 1986. "Gender: A Useful Category of Historical Analysis." *The American Historical Review* 91 (5): 1053-75. <https://doi.org/10.2307/1864376>.
- Souza, Rosali Fernandez de. 2006. "Organização e representação de áreas do Conhecimento em Ciência e Tecnologia: Princípios de agregação em grandes áreas segundo diferentes contextos de produção e uso de informação." *Encontros Bibli: Revista eletrônica de Biblioteconomia e Ciência da informação* 11 (1):27-41. <https://doi.org/10.5007/1518-2924.2006v11nesp1p27>.

- Spivak, Gayatri Chakravorty. 2010. "Can the Subaltern Speak?" In *Can the Subaltern Speak? Reflections on the History of an Idea*, edited by Rosalind C. Morris, Part 1. New York: Columbia University Press.
- Szostak, Rick. 2024. "Additional Purposes of Knowledge Organization." *Knowledge Organization* 51 (5): 371–75. <https://doi.org/10.5771/0943-7444-2024-5-371>.
- United Nations. 1979. *Convention on the Elimination of All Forms of Discrimination Against Women*. Treaty Series, vol. 1249. <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-elimination-all-forms-discrimination-against-women>.
- United Nations. 2015. *Transforming Our World: The 2030 Agenda for Sustainable Development*. A/RES/70/1. Accessed November 12, 2017. <https://sdgs.un.org/2030agenda>.
- Vignaux, Georges. 2000. *O Demónio da Classificação: Pensar, Organizar*. Instituto Piaget.
- Viola, Carla Maria Martellote. 2023. "Informações Legislativas e os Direitos em Construção das Mulheres Brasileiras: Proposta de Categorização Rumo à Agenda 2030." Tese de Doutorado, Universidade Federal do Rio de Janeiro e Instituto Brasileiro de Informação em Ciência e Tecnologia. <http://ridi.ibict.br/handle/123456789/1290>.
- Zeng, Márcia L. 2008. "Knowledge Organization Systems (KOS)." *Knowledge Organization* 35 (2-3): 160–82. <https://doi.org/10.5771/0943-7444-2008-2-3-160>.

Modellizzazione delle conoscenze: Thesaurus Nuovo soggettario e Wikidata, due strumenti a confronto

Uno studio di caso nel dominio della fotografia

Elena Cencetti*, Camillo Carlo Pellizzari di San Girolamo**, Elisabetta Viti*

Abstract: The paper aims to show the structural and functional similarities and differences between the *Nuovo soggettario* Thesaurus and the Wikidata knowledge base, providing an overview about their reciprocal interoperability, due to the compliance of both tools with the Findable, Accessible, Interoperable, and Reusable (FAIR) principles and Resource Description Framework (RDF) standard. As a case study, a group of terms in the Thesaurus relating to photography has been manually reconciled with the corresponding Wikidata items. The reconciliation process involved a review of the meaning and form of the terms and is described in detail with some significant examples. On the basis of this work, some general conclusions are drawn about the possibilities of reciprocal quantitative and qualitative improvement and a few possible future projects are outlined.

Keywords: Nuovo soggettario Thesaurus, Wikidata, Ontology, Thesaurus, Photography.

1. Introduzione

Questo contributo ha l'obiettivo di mettere in evidenza affinità e differenze strutturali e funzionali tra il Thesaurus *Nuovo soggettario* (da ora in poi ThNS) e la base di conoscenza libera Wikidata (da ora in poi WD), e di illustrare le modalità per la creazione di un'interoperabilità semantica e tecnica reciproca.

Dall'analisi emergerà come il confronto trasversale tra contesti e professionalità diverse possa essere proficuo e comportare benefici per i due strumenti, soprattutto sul versante dell'incremento e perfezionamento reciproco di dati strutturati e di collegamenti con database di biblioteche, archivi e musei.

* Biblioteca Nazionale Centrale di Firenze, Firenze, Italia. elena.cencetti@unifi.it. ORCID: 0009-0004-8731-6418. elisabetta.viti@cultura.gov.it. ORCID: 0000-0002-7627-4062.

** Scuola Normale Superiore, Pisa, Italia. camillo.pellizzaridisangirolamo@sns.it. ORCID: 0000-0003-2699-1693.

Il testo è stato scritto in totale collaborazione e accordo tra gli autori, tuttavia vanno ascritti a Elena Cencetti il paragrafo 5.1 e 6, a Camillo Pellizzari di San Girolamo i paragrafi 2 e 4.2 e 5.3 e a Elisabetta Viti i paragrafi 3 e 4.1 e 5.2.

2. Cenni storici

Il progetto rientra nell'ambito di una collaborazione quasi ventennale tra la Biblioteca nazionale centrale di Firenze (da ora in poi BNCF) e Wikimedia Italia (da ora in poi WMI), istituzionalizzata nell'ottobre del 2013 con una lettera d'intenti, seguita da successivi rinnovi, e puntualizzata con un accordo del 2025 che definisce ampi obiettivi di cooperazione anche sul fronte dell'interoperabilità semantica e tecnica tra il ThNS e WD (Biblioteca Nazionale Centrale di Firenze and Wikimedia Italia 2025).

Una prima fase si è basata sull'assegnazione semiautomatica di equivalenti WD ai termini preferiti del ThNS in un campo dedicato (Equiv. Wikidata), laddove fosse già presente un *deeplink* a Wikipedia in italiano (itWP) nel campo Fonte. In particolare, tramite un'interrogazione dell'*endpoint* SPARQL di WD sono stati estratti dati strutturati ed è stato fatto un match con i termini del ThNS, con successiva verifica intellettuale dell'effettiva corrispondenza, risoluzione di incongruenze ed eventuale ricerca di un elemento WD semanticamente più coerente. Nel tempo l'interoperabilità è diventata reciproca, nel maggio 2013 è stata creata in WD la proprietà P508 (*identificativo Thesaurus BNCF*) e si è cominciato a importare i valori del template Thesaurus BNCF presenti nelle voci di itWP come valori della proprietà P508 nei corrispondenti elementi WD. L'attribuzione degli equivalenti WD ai termini del ThNS e viceversa è diventata ormai una prassi intellettuale consolidata a completamento dell'attività di allestimento del ThNS (Cencetti et al. 2025).

Una seconda fase, più sperimentale e pionieristica, si è focalizzata sul perfezionamento e incremento di un'ontologia relativa al dominio della fotografia in WD partendo dalla terminologia del ThNS come fonte di riferimento (cfr. par. 6).

Questa collaborazione su più fronti ha prodotto ottimi risultati e dall'inizio del 2024 le modifiche effettuate in WD dal Settore ricerche e strumenti di indicizzazione semantica della BNCF sono state circa 3.000; in particolare, sono stati creati circa 30 nuovi elementi e sono stati aggiunti circa 530 legami col ThNS; su WD è stata creata una pagina di coordinamento, costantemente aggiornata, per raccogliere materiali e bibliografia riguardo alla collaborazione (Wikimedia Foundation 2025a).

A giugno 2025 i link dal ThNS a WD sono 13.454 e da WD a ThNS sono 17.497 (per ulteriori statistiche si veda il par. 5.3).

3. Thesauri e ontologie

I *Knowledge Organization Systems* (KOSs), declinati in un'ampia gamma di risorse (thesauri, ontologie, classificazioni, soggettari, glossari, dizionari, enciclopedie, tassonomie, schemi di categorizzazione, *semantic network*, ecc.), sono

funzionali alla raccolta, organizzazione e recupero di informazioni provenienti da specifiche aree di studio. Infatti costituiscono lo strumento “ponte” tra le esigenze informative dell’utente e le risorse all’interno delle collezioni, e come *roadmap* sollecitano e guidano la ricerca attraverso un processo di continua scoperta, talvolta anche inaspettata (serendipità), creando connessioni inter- e transdisciplinari (Hodge 2000).

I thesauri e le ontologie sono sistemi altamente strutturati, basati su relazioni semantiche tra concetti e termini, e spesso vengono utilizzati come base di conoscenza reciproca per confronti e arricchimenti.

A differenza di un dizionario, in cui le voci d’entrata sono accompagnate da più definizioni, che riflettono un uso comune contestualizzandolo anche in ambiti disciplinari diversi, nei thesauri i concetti, ristretti nello scopo e nel significato, sono rappresentati dai termini in un rapporto biunivoco (secondo il principio di uniformità e univocità ogni concetto è espresso da un solo termine e un termine rappresenta uno solo concetto). La loro modellizzazione in una struttura sistematica sindetica fatta di relazioni gerarchiche, associative e di equivalenza, aiuta a esplorare e a comprendere l’organizzazione in categorie (es. Organismi, Persone, Oggetti, ecc.), a navigare nella struttura semantica di un termine per esplorare significati più generali, più specifici o associati, e a esaminare l’apparato sinonimico e le varianti grafiche di uno stesso concetto/termine.

Come viene specificato nello standard ISO 25964-2013 (International Organization for Standardization 2011; International Organization for Standardization 2013), rispetto ai thesauri che prevedono tre tipi di relazioni semantiche standardizzate, le ontologie offrono un’ampia ed esplicita sistematizzazione terminologica per guidare a una più appropriata rappresentazione semantica di un dominio concettuale. Infatti, sono caratterizzate da un alto livello di astrazione che garantisce una concettualizzazione dettagliata della conoscenza specifica, tramite metadati, fornendo una rappresentazione astratta della realtà di un determinato dominio attraverso l’organizzazione dei fenomeni (concetti, oggetti e altre entità) in una struttura semantica organizzata in classi (entità di dominio), istanze (oggetti attuali presenti nel dominio) e proprietà (attributi o qualità), funzioni e relazioni (che permettono di collegare le entità all’interno del dominio) e assiomi (regole logiche che restringono significati e l’uso di classi, istanze, funzioni e relazioni) (Gruber 1993). Un’ontologia definisce una semantica formalmente definita per una comprensione comune della struttura delle informazioni tra umani e/o agenti software: un vocabolario comune ai ricercatori, che devono condividere informazioni in campi di conoscenza delimitati e in ambienti applicativi altamente specializzati e, allo stesso tempo, significati condivisi e relazioni binarie concetti/oggetti interpretabili dalla macchina (Studer et al. 1998, 184).

I thesauri grazie alla propria struttura sindetica, al controllo del vocabolario, e alla disponibilità in formati e protocolli standard (come SKOS/RDF), possono essere reingegnerizzati e trasformati in ontologie (es. MeSH, AGROVOC, Cybersecurity Thesaurus). Infatti la loro semantica viene migliorata grazie all'applicazione delle regole del Web Ontology Language (OWL)¹ che garantiscono un maggiore livello di personalizzazione nella selezione dei link tra concetti permettendo così al sistema informativo di scambiare informazioni strutturate in un formato di esportazione comune e condiviso e, conseguentemente, rispondendo in modo più circostanziato e preciso alle esigenze informative dell'utente finale (Lanza 2022). Kless et al. nel loro saggio spiegano come la trasformazione non sia solo una conversione sintattica in un linguaggio formale o l'automazione di alcuni passaggi ma consista nel trattare un thesaurus come un modello di dati. Il thesaurus e l'ontologia offrono una chiara distinzione nel modo di gestire e rappresentare domini specialistici e nello strutturare ed esplicitare le relazioni tra concetti. La strategia è quindi quella di perfezionare le relazioni thesaurali paradigmatiche specificando le varie tipologie. Per esempio si potrebbe definire in modo più preciso il rapporto gerarchico BT (*Broader term*)/NT (*Narrower term*): genere/specie → sottoclasse di; tutto/parte → parte di; classe/individuo → istanza di. Per esempio, il concetto di “mobili” nel ThNS diventa una classe dell'ontologia e si relaziona con altre entità che ne costituiscono tipologie o parti, es. Mobili da giardino → sottoclasse di Mobili; Cassetti → parte di Mobili. Oppure chiarire e precisare la relazione reciproca iperonimi/iponimi, assegnando a una classe di concetti un certo numero di “individui” che rappresentano istanze cioè nomi specifici di concetti di dominio, es. Battaglia di Lepanto → istanza di Battaglie (Kless et al. 2012).

Anche le relazioni associative di un thesaurus possono essere trasformate in una relazione ontologica tra due concetti puntualizzando la specificazione formale dei concetti come classe dell'ontologia e restringendo le proprietà della classe stessa. Le *Object Properties* espresse in OWL rappresentano una correlazione binaria tra entità e descrivono classi più circoscritte. Per esempio, in un thesaurus concetti opposti sono legati da una relazione associativa e ciascuno di loro da una nota d'ambito per meglio specificare il significato del termine, cosa diversa in un'ontologia dove i due termini sono rappresentati usando `rdf:type owl:ObjectProperty` (sottoclasse della classe `RDF rdf:Property`) “*HasAsContrary*” che specifica esplicitamente la tipologia di correlazione tra i due concetti (Lanza 2022, 105–25).

In questo nostro studio, come verrà evidenziato e spiegato successivamente, abbiamo tentato di mettere a confronto la struttura dei due strumenti proprio

¹ OWL definisce costrutti avanzati per descrivere la semantica di risorse RDF, supportando l'aggiunta di vincoli, restrizioni e caratteristiche delle proprietà (transitiva, ecc.) che permettono di abilitare il ragionamento logico automatico nel web semantico.

per rilevare similitudini e differenze e per ipotizzare una sperimentazione futura di reingegnerizzazione del ThNS.

4. Caratteristiche dei due strumenti

4.1. Thesaurus Nuovo soggettario

Il ThNS, allestito e implementato dalla BNCF già dal 2006, è la componente principale dell'omonimo sistema impiegato per l'indicizzazione per soggetto e per la descrizione di risorse di varia natura (testuali, iconografiche, sonore, oggetti, ecc.) su supporto analogico e/o digitale non solo nell'ambito del Servizio bibliotecario nazionale (SBN) ma anche in contesti Galleries, Libraries, Archives, Museums (GLAM). Il ThNS ha un patrimonio terminologico multidisciplinare di 74.490 termini, che viene aggiornato ogni sei mesi (ultimo aggiornamento marzo 2025). È conforme allo standard ISO 25964-2013 e ai principi FAIR (ricercabilità, accessibilità, interoperabilità e riusabilità), è disponibile nella nuvola dei Linked Open Data (Cyganiak et al. 2025) e sulla piattaforma dati.cultura.gov.it del Ministero della Cultura (2025), e scaricabile e riusabile secondo la licenza CC BY 4.0.

Fin dal 2012 è stato impiegato anche come base di conoscenza per l'indicizzazione automatica di risorse digitali; ad oggi viene sperimentato nell'Infrastruttura e servizi digitali per il Patrimonio Culturale (I.PaC)² per l'attribuzione di parole chiave a testi e immagini digitalizzate, processati tramite sistemi di *machine learning*.

Grazie alla sua disponibilità in protocolli e formati aperti, come Simple Knowledge Organization System (SKOS)/RDF, MARC21 e ZThes, e all'interrogabilità tramite l'endpoint SPARQL, costituisce un "ponte semantico" con altri strumenti di indicizzazione, come Library of Congress Subject Headings (LCSH), Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU), Gemeinsame Normdatei (GND) ed Encabezamientos de materia de la Biblioteca Nacional de España (EMBNE), e database di istituti della cultura, con thesauri, ontologie, classificazioni, enciclopedie, dizionari, ecc. Si veda Figura 1 per un esempio.

² Ecosistema digitale del patrimonio culturale italiano che presenta un'infrastruttura informativa innovativa basata sulla rappresentazione unificata e accessibile della conoscenza attraverso un complesso sistema di servizi avanzati e funzioni, per la gestione e il dialogo di dati stratificati, eterogenei e appartenenti a domini diversi, fondati sia su modelli e schemi predefiniti (motori a regole e ontologie) sia su algoritmi di intelligenza artificiale.



Nuovo soggettario - Thesaurus



[Criteri](#) [Aiuto alla ricerca](#) [Sigle e simboli](#) [Fonti](#) [Novità](#) [Download](#)



[Torna alla ricerca](#)

Palladiotipie

GERARCHIA

Macrocategoria: **Categoria** Cose:Forme

Nota d'ambito Stampe fotografiche ottenute con un procedimento di stampa identico alla Platinotipia (il palladio sostituisce il platino)

Usato per [Palladiotipi](#), [Stampe al palladio](#), [Stampe palladiotipiche](#)

Termine apicale [Forme](#)

Termine più generale [Fotografie](#)

Termine associato [Palladiotipia](#), [Platinotipie](#)

Fonti [AAT](#): Palladium prints; FSR (voce: Palladio)

Equiv. in altri strumenti di indicizzazione

 LOC TGM: [Palladium prints](#)

 EMBNE: [Paladiotipos](#)

Equiv. Wikidata

 [Q131597990](#)

Proponente [BN](#)

Status del record Termine strutturato

Identificativo 76375

Notizie bibliografiche

- Catalogo della BNCF
→ Opere
- Stringhe di soggetto
- Catalogo SBN
→ Opere

 [Suggerimenti sul termine](#)

Figura 1. Il termine Palladiotipie (76375) del ThNS.

4.2. Wikidata

WD, uno dei progetti Wikimedia, è una base di conoscenza libera che raccoglie dati strutturati di tutte le entità (argomenti, concetti astratti, oggetti, persone, luoghi, eventi, ecc.) che rientrano nei suoi criteri di inclusione (Wikidata community 2025a), ossia principalmente tutte le entità che abbiano una voce in almeno una versione linguistica di Wikipedia e/o che siano chiaramente identificabili e descritte da fonti attendibili e pubblicamente consultabili.

Gli elementi, che rappresentano qualsiasi cosa esista nella conoscenza umana, in WD sono strutturati in ontologie di dominio, intese come concettualizzazioni formali (trattabili e processabili dalle macchine) che permettono di rappresentare un settore disciplinare specifico, individuandone categorie, entità e loro relazioni. Entità individuali e concetti sono legati da relazioni gerarchiche principalmente tramite la proprietà *sottoclasse di* (P279) che lega gli elementi alla tipologia semantica più ampia alla quale appartengono (cfr. par. 5.2). I dati di WD sono rilasciati in licenza CC0, sono disponibili in numerosi formati e interrogabili con un endpoint SPARQL.

WD ha oltre 116 milioni di elementi e 12.500 proprietà (giugno 2025) (Wikidata community 2025b e 2025c). Si veda la Figura 2 per un esempio.

palladiotipia (Q131597990)

stampa fotografica realizzata con il procedimento al palladio modifica

palladiotipie | palladiotipi | palladiotipo | stampe al palladio | stampa al palladio | stampe palladiotipiche | stampe palladiotipica

▼ In altre lingue

Lingua	Etichetta	Descrizione	Anche conosciuto come
italiano	palladiotipia	stampa fotografica realizzata con il procedimento al palladio	palladiotipie palladiotipi palladiotipo stampe al palladio stampa al palladio stampe palladiotipiche stampe palladiotipica
tedesco	Palladiotype	Fotodruck nach dem Palladiumverfahren	
greco	Nessuna etichetta definita	Nessuna descrizione definita	
inglese	palladium print	photographic print made by the palladium process	
spagnolo	paladiotipos	impresión fotográfica por paladio	
francese	palladiotype	impression de photos par le procédé au palladium	
latino	Nessuna etichetta definita	Nessuna descrizione definita	
predefinito per tutte le lingue	Nessuna etichetta definita	—	
portoghese	Nessuna etichetta definita	Nessuna descrizione definita	

Dichiarazioni

sottoclasse di

fotografia modifica

► 1 riferimento

stampa fotografica modifica

► 1 riferimento

+ aggiungi valore

metodo di fabbricazione

palladiotipia modifica

▼ 0 riferimenti

+ aggiungi riferimento

+ aggiungi valore

+ aggiungi dichiarazione

Identificativi

identificativo BNE (datos.BNE.es)	XX5176194	modifica
	soggetto indicato come	Paladiotipos
	▼ 0 riferimenti	+ aggiungi riferimento
	+ aggiungi valore	

identificativo AAT	300134677	modifica
	soggetto indicato come	palladium prints
	▼ 0 riferimenti	+ aggiungi riferimento
	+ aggiungi valore	

identificativo Thesaurus BNCF	76375	modifica
	soggetto indicato come	Palladiotipie
	▼ 0 riferimenti	+ aggiungi riferimento
	+ aggiungi valore	

identificativo Thesaurus for Graphic Materials	tgm007408	modifica
	soggetto indicato come	Palladium prints
	▼ 0 riferimenti	+ aggiungi riferimento
	+ aggiungi valore	

+ aggiungi dichiarazione

Figura 2. L'elemento palladiotipia (Q131597990) di WD.

5. Analisi e confronto dei due strumenti

La comparazione dei due KOS si è basata sull'applicazione di due metodi di analisi, deduttivo (dall'alto verso il basso) e induttivo (dal basso verso l'alto), facendo emergere riflessioni generali su:

1. natura, dimensione strutturale e funzionale dei due strumenti;
2. poliedricità della terminologia presa in esame;
3. area culturale, utenza e *mission* dei soggetti curatori.

5.1. Finalità, struttura e tipologia dei termini

Entrambi gli strumenti hanno come obiettivo comune il controllo del vocabolario cercando di evitare le ambiguità espressive proprie del linguaggio

naturale. Grazie alla loro struttura formale standardizzata, e all'uniformità/univocità del rapporto concetto/termine, si creano costellazioni di senso (attraverso rimandi verticali e orizzontali tra termini semanticamente più ampi, ristretti e associati) che rendono possibile anche l'identificazione di significati specifici afferenti a domini più tecnici. I gruppi di termini con legami logici e di significato guidano l'utente ad affinare la ricerca, a modificarla, ampliarla, restringerla e a scegliere ulteriori punti di accesso, in modo da facilitare la navigabilità e il recupero dei contenuti, ma anche l'accesso a classi di oggetti descritti in siti eterogenei che usano linguaggi differenti.

Per quanto riguarda la struttura i due strumenti si differenziano già per la specificità della loro natura.

Il ThNS ha una macrostruttura basata su un'organizzazione sistematica (classificazione dei fenomeni) fondata su: quattro macrocategorie; tredici categorie che definiscono la struttura classificatoria fondamentale dei fenomeni, suddividendo il vocabolario in classi omogenee e facilitando così la costruzione delle gerarchie; faccette, espresse da un solo termine o da etichette di nodo (locuzioni tra parentesi quadre) che rappresentano principi di suddivisione di una categoria secondo caratteristiche specifiche, consentendo così lo sviluppo ordinato delle gerarchie e la loro coerenza con i principi che regolano la costruzione dello strumento. Prevalentemente monogerarchico, come già accennato ha un patrimonio terminologico di 74.490 termini (*release* marzo 2025), di cui preferiti 34.470 con 540 poligerarchici (circa 1,5%).

WD, invece, ha un'ontologia dotata di un unico elemento apicale, *entità* (Q35120), e successivi livelli sottostanti, per esempio il 1° livello è costituito da 27 elementi, il 2° livello da 524 elementi e il 3° livello da 5.211 elementi; tali livelli sono stati individuati analizzando i valori di P279 (24 giugno 2025); anche WD ha un'ontologia prevalentemente monogerarchica³; non è previsto in generale l'uso di etichette di nodo.

Per quanto riguarda la tipologia dei termini, nel ThNS sono presenti concetti espressi da nomi comuni (sostantivi e sintagmi nominali, preposizionali e verbali) costituiti da una o più parole; sono assenti nomi propri, eccetto alcune categorie, per esempio formaggi, vini, feste, battaglie, ecc. I termini possono essere: preferiti e non preferiti (entrambi con proprio ID).

In WD sono presenti nomi comuni, nomi propri di persone, organizzazioni o enti, titoli di opere, luoghi, specifici popoli o gruppi etnici, nomi di specifiche marche, ecc. In ciascun elemento sono presenti nel *termbox* (ossia la tabella comprendente etichette, descrizioni e alias divise per lingua; le lin-

³ Attualmente (24 giugno 2025), considerando gli elementi dotati di P279, l'84,1% ne ha un solo valore, quasi il 14,8% due valori e l'1,1% tre o più valori. Considerando, invece, tutti gli elementi dotati di almeno una tra P31, P279, P361 e P171, il numero di valori totali per queste proprietà presenti in ciascun elemento è uno nell'87,1% dei casi, due nel 10,8% dei casi e tre o più nel 2,1% dei casi.

gue disponibili sono 624 al 24 giugno 2025) sia termini preferiti, usati come etichette, sia termini non preferiti, usati come alias; l'identificatore concerne quindi l'elemento coi suoi vari nomi, che non hanno ID propri.

Il ThNS è corredato da un ricco apparato di note, in particolare: note di definizione (in rari casi inserite per chiarire il significato di neologismi, tecnicismi e termini di prestito); note d'ambito (per specificare il significato, nei termini polisemici); note di orientamento (per guidare l'utente del Thesaurus a scegliere termini che possano avere altro significato).

In WD la definizione di un termine è inserita nel campo descrizione all'interno del *termbox* di ciascun elemento. La proprietà *distinto da* (P1889) ha una funzione corrispondente a quella della nota di orientamento del ThNS.

5.2. Relazioni semantiche

Entrambi gli strumenti garantiscono una sistematizzazione dei concetti modellando le informazioni tramite specifiche relazioni paradigmatiche che creano legami di significato o meglio metodi usati per collegare tra loro i concetti. Il ThNS si struttura su relazioni semantiche: gerarchiche di *Broader Term* (BT) e *Narrower Term* (NT), sinonimiche di *USE/Used For* (USE/UF), associative di *Related Term* (RT) conformi allo standard ISO 25964-2013 (International Organization for Standardization 2011, par. 10.2.4). In particolare il legame generico di BT sottintende tre tipologie specifiche di relazione menzionate nello standard ISO:

- genere-specie ossia BTG (*Broader Term Generic*);
- parte-tutto ossia BTP (*Broader Term Partitive*);
- esemplificativa ossia BTI (*Broader Term Instancial*).
- Le relazioni di BT hanno automaticamente una reciproca relazione di NT, e le relazioni RT sono automaticamente reciproche.
- WD è organizzata in base a relazioni simili a quelle succitate, ottenute tramite differenti e specifiche proprietà: le relazioni gerarchiche sono rese tramite quattro proprietà principali, che possono essere ricondotte alle diverse casistiche di BT (Tab. 1):
- la proprietà *sottoclasse di* (P279) stabilisce una relazione generica, fra un genere e le sue specie (BTG); una relazione di tipo generico è resa anche dalla proprietà *taxon di livello superiore* (P171), che lega il *taxon* di livello inferiore a quello superiore più vicino (è utilizzata però esclusivamente nel caso di elementi relativi a un sistema di classificazione tassonomico);
- la proprietà *parte di* (P361) stabilisce una relazione parte-tutto (o partitiva), fra un intero (tutto) e una sua parte (BTP);

- la proprietà *istanza di* (P31) stabilisce una relazione di tipo esemplificativo fra una classe espressa da un nome comune e un esempio individuale di questa classe (BTI).

Proprietà	Descrizione	Esempio
sottoclasse di (P279)	questo elemento è un sottoinsieme, una sottoclasse, un tipo di un altro elemento	mela (Q89) è sottoclasse di frutta (Q3314483)
taxon di livello superiore (P171)	il più vicino taxon del livello tassonomico superiore rispetto a quello in questione	Homo sapiens (Q15978631) taxon di livello superiore Homo (Q171283)
istanza di (P31)	questo elemento è un'istanza, un tipo di questa classe, categoria o gruppo di oggetti	Alessandro Manzoni (Q1064) è istanza di umano (Q5)
parte di (P361)	questo elemento è parte di quell'elemento; proprietà inversa di consiste di (P527)	ecosfera (Q3175157) è parte di biosfera (Q42762)
consiste di (P527)	elemento di cui è formato/composto/ costituito il soggetto il quale non è istanza o sottoinsieme del soggetto; proprietà inversa di parte di (P361)	sistema solare (Q544) consiste di Mercurio (Q308), Venere (Q313), ecc.
contiene elementi del tipo (P2670)	il soggetto consiste di istanze di questa classe (generalmente il soggetto non è una classe)	alfabeto greco (Q8216) contiene elementi del tipo lettera greca (Q19793459)
è una metaclassa per (P8225)	tutte le istanze del soggetto sono sotto-classi dell'oggetto	tipo di nave (Q2235308) è una metaclassa per nave (Q11446)

Tabella 1. Proprietà di WD corrispondenti a relazioni BT.

Non sono invece esplicitati nella struttura dei singoli elementi di WD i concetti più specifici rappresentati nel ThNS dai NT (termini più specifici); possono tuttavia essere ricavati come inversi delle succitate proprietà corrispondenti a BT e sono visibili attraverso il gadget Related items (Wikidata community 2025d) e il tool Wikidata Graph Builder (AngryLoki 2016). In WD al posto degli UF si usano gli *alias*: anziché creare relazioni tra un termine preferito e più termini non preferiti tramite UF, i termini non preferiti sono inseriti nell'elemento come alias in ciascuna lingua, mentre il termine preferito in ciascuna lingua è inserito come etichetta.

Le relazioni associative nel Thesaurus sono stabilite su base definitoria tra concetti semanticamente correlati quando cioè un concetto/termine è necessario per definire e descrivere l'altro. In WD tali relazioni sono rese da un vasto insieme di proprietà; le principali sono elencate nella seguente Tabella 2.

Proprietà	Descrizione	Esempio
aspetto di (P1269)	argomento di cui questo elemento è aspetto	clima di Miami (Q5133566) è aspetto di Miami (Q8652)
contrario di (P461)	l'elemento contrario a questo elemento; è una proprietà reciproca	inverno (Q1311) è contrario di estate (Q1313)
distinto da (P1889)	elemento distinto dal soggetto con il quale è spesso confuso; è una proprietà reciproca	telefotografia (Q117833345) è distinto da fototelegrafia (Q117833347)
considerato essere uguale a (P460)	il soggetto potrebbe essere uguale all'oggetto (per relazioni di quasi equivalenza, o di equivalenza contestata da alcune fonti); è una proprietà reciproca	Giacomo il Giusto (Q26925) è considerato essere uguale a Giacomo d'Alfeo (Q44047)
parzialmente coincidente con (P1382)	parzialmente parte di, ma non completamente parte di (P361)	trigonometria (Q8084) è parzialmente coincidente con geometria del triangolo (Q1257470)
causato da (P828)	il soggetto è effetto della seguente causa; proprietà inversa di è causa di (P1542)	malaria (Q12156) è causato da Plasmodium malariae (Q133969)
è causa di (P1542)	il soggetto è causa dell'avverarsi del seguente effetto; proprietà inversa di causato da (P828)	Plasmodium malariae (Q133969) è causa di malaria (Q12156)
uso (P366)	principale uso, funzione, ruolo o scopo del soggetto	libro (Q571) uso lettura (Q199657)
praticato da (P3095)	tipo di agenti che studiano questo argomento o lavorano in questo campo; proprietà inversa di campo di questa occupazione (P425)	filologia romanza (Q1277348) è praticata da filologo romanzo (Q2504617)
campo di questa occupazione (P425)	campo corrispondente ad una certa occupazione; proprietà inversa di praticato da (P3095)	filologo romanzo (Q2504617) campo di questa occupazione (P425) filologia romanza (Q1277348)
studia (P2578)	il soggetto è una scienza o dominio che studia l'oggetto; proprietà inversa di studiato da (P2579)	egittologia (Q145903) studia antico Egitto (Q11768)
studiato da (P2579)	il soggetto è studiato da questa scienza o dominio; proprietà inversa di studia (P2578)	antico Egitto (Q11768) è studiato da egittologia (Q145903)
prodotto o materiale prodotto (P1056)	materiale o prodotto da una azienda, ente o un processo di produzione	cianografia (Q372532) prodotto o materiale prodotto cianotipo (Q64029200)
metodo di fabbricazione (P2079)	metodo, processo utilizzato per realizzare, lavorare, un oggetto	biscotto (Q13270) metodo di fabbricazione cottura al forno (Q720398)

Tabella 2. Selezione di proprietà WD corrispondenti a relazioni RT.

Le relazioni stabilite tra due elementi attraverso queste proprietà non sono automaticamente reciproche: le relazioni inverse vanno create manualmente; esistono gadget che semplificano tale aggiunta manuale (ad esempio *Consistency check add* (Frettie 2025)).

5.3. Legami con altri strumenti

Nel ThNS ci sono dei limiti tecnici che non consentono di legare un concetto/termine a due concetti/termini di un altro strumento di organizzazione della conoscenza, oppure un concetto/termine di un altro strumento di indicizzazione a due concetti/termini del ThNS.

In WD, invece, sono possibili entrambe le opzioni. I casi di legami multipli possono essere individuati tramite le query che servono a verificare le violazioni dei vincoli di valore singolo (ossia i casi in cui ci si aspetta che un elemento si leghi a un solo ID di un certo database e invece si lega a due o più ID) e dei vincoli di valore unico (ossia i casi in cui ci si aspetta che un solo elemento si leghi a un dato ID di un certo database e invece due o più elementi si legano ad esso), e poi verificati per trovare eventuali errori⁴.

Nella seguente Tabella 3 si mostrano i legami reciproci tra WD e il ThNS, con quattro thesauri di biblioteche nazionali a cui entrambi linkano e con il thesaurus *Art & architecture thesaurus online* (AAT); i dati sono ricavati da query eseguite il 30 settembre 2025. In WD le proprietà usate per i quattro thesauri delle biblioteche nazionali sono le stesse usate anche per i rispettivi *authority file*; per evitare, quindi, che anche elementi relativi a singoli esseri umani/enti/luoghi vengano inclusi nei risultati, le query marcate con * vengono limitate agli elementi relativi a concetti, che sono dotati di P279, e quindi comparabili con ThNS.

n°	legami	da NS	Query	n°	legami	da WD	query
1	a WD	13454	https://tinyurl.com/4fpxxjr	12	a NS	17550	https://w.wiki/EZKd
2	a LCSH	17383	https://tinyurl.com/yjypw9h2	13	a LCSH*	27280	https://qlever.cs.uni-freiburg.de/wikidata/xBPm04
3	a LCSH + WD	9524	https://tinyurl.com/ytun8j9k	14	a LCSH* + NS	9933	https://w.wiki/EZKf
4	a RAMEAU	15792	https://tinyurl.com/4x56j7he	15	a RAMEAU*	12743	https://qlever.cs.uni-freiburg.de/wikidata/4C1Zmf
5	a RAMEAU + WD	8805	https://tinyurl.com/3vw3sfzt	16	a RAMEAU* + NS	8211	https://w.wiki/EZKg
6	a GND	9969	https://tinyurl.com/34ufju3u	17	a GND*	29457	https://qlever.cs.uni-freiburg.de/wikidata/wNghsp
7	a GND + WD	6356	https://tinyurl.com/mrxspepj	18	a GND* + NS	8815	https://w.wiki/EZKh

⁴ Sull'uso dei vincoli in Wikidata, cfr. (Wikimedia Foundation 2025b); le query per le violazioni di vincolo relative al ThNS sono consultabili in (Wikimedia Foundation 2025c).

8	a EMBNE	6634	https://tinyurl.com/4x8de4kb	19	a EMBNE*	4089	https://qllever.cs.uni-freiburg.de/wikidata/1ntpUi
9	a EMBNE + WD	4241	https://tinyurl.com/2s4j24da	20	a EMBNE* + NS	2636	https://w.wiki/EZKj
10	ad AAT	385	https://tinyurl.com/4bbdh8j9	21	ad AAT	25025	https://w.wiki/EZKw
11	ad AAT + WD	312	https://tinyurl.com/mwc95fzn	22	ad AAT + NS	5302	https://w.wiki/EZKx

Tabella 3. Relazioni tra WD, NS e altri thesauri; * indica che la query è limitata in Wikidata a elementi contenenti P279; i risultati di tutte le query in tabella sono disponibili su Zenodo (Pellizzari di San Girolamo 2025).

6. Caso di studio: la fotografia

6.1. Metodologia

La metodologia seguita per questo confronto si è basata sulla selezione di un corpus terminologico di dominio, particolarmente significativo e rappresentativo delle sue categorie concettuali, anche sulla base di suggerimenti di esperti: 160 termini relativi ad oggetti (es. Macchine fotografiche); attività, tecniche e relativi prodotti (es. Calotipia e Calotipi), persone (es. Fotografi) e strutture (es. Laboratori fotografici), organizzazioni (es. Archivi fotografici e Fototeche), materia e materiali (es. Sali d’argento, Carta al carbone)⁵.

La rosa terminologica si è basata sull’estrazione di terminologia di dominio già presente nel ThNS, sull’acquisizione di nuove proposte formulate o in fase di indicizzazione semantica corrente dalla Bibliografia nazionale italiana (BNI) e dagli enti che collaborano allo sviluppo del ThNS, o frutto dell’analisi e della revisione del *Glossario di terminologia fotografica* dell’Associazione Italiana Biblioteche (AIB) (Corti e Gioffredi Superbi 2004) e del *Thesaurus per materia-tecnica e materiale composito* dell’Istituto Centrale per il Catalogo e la Documentazione (ICCD).

I termini scelti sono stati poi sottoposti a un’attenta analisi semantica di natura intellettuale che ha preso in considerazione concetti, significanti, significati, struttura relazionale e mettendoli poi a confronto con i corrispondenti elementi di WD per appurarne corrispondenza formale e semantica.

Questo paragone, logicamente, ha portato alla riconsiderazione di scelte pregresse e a eventuali modifiche per migliorare la coerenza interna dei due strumenti. Le principali tipologie di modifiche sono state:

- incremento dell’apparato sinonimico (UF in ThNS e *alias* in WD);

⁵ I contenuti di questo paragrafo sono principalmente tratti dalla tesi di master di Elena Cencetti (Cencetti 2025b). La lista dei 160 termini di ambito fotografico selezionati è disponibile su Zenodo (Cencetti 2025a).

- puntualizzazione nella definizione dei concetti tramite l'introduzione o la modifica di note d'ambito, note di orientamento, di definizione nel ThNS e di descrizioni (principalmente in italiano e in inglese) in WD;
- revisione e miglioramento dei legami gerarchici (BT in ThNS e P279 *sottoclasse di* in WD);
- arricchimento di *link* ad altri strumenti di indicizzazione semantica di biblioteche nazionali straniere (LCSH, RAMEAU, GND, EMBNE), a thesauri specialistici come l'AAT del Getty Research Institute, a database di archivi e musei (es. Gallerie degli Uffizi, Museo Galileo, Museo Egizio di Torino, Museo nazionale della scienza e tecnologia Leonardo da Vinci, ecc.) e a voci di Wikipedia;
- creazione di nuovi elementi WD nel caso ci fosse una mancata corrispondenza con i termini del ThNS;
- valutazione ed eventuale introduzione di nuova terminologia nel ThNS;
- verifica/creazione di legami reciproci tra i concetti dei due KOS.

Nell'analisi comparata e nell'integrazione dei dati è stato necessario preservare e mantenere la logica strutturale e relazionale interna ai due strumenti. Per esempio in WD etichette, descrizioni, *alias*, dichiarazioni e identificativi devono fornire un quadro chiaro, logico, coerente e privo di contraddizioni del concetto espresso dall'elemento; infatti sono piuttosto frequenti casi di elementi "incoerenti", che richiedono un'attenta analisi per la risoluzione di eventuali incongruenze (cfr. paragrafo 6.2.2).

I primi interventi sono stati rintracciare in WD l'elemento corrispondente al termine del ThNS e modificarne adeguatamente il *termbox*:

- si è verificata la compilazione e la correttezza formale dell'etichetta in italiano e in inglese, e possibilmente anche in spagnolo, tedesco e francese, e delle relative descrizioni, sulla base di un confronto con i significati attribuiti dalle rispettive voci di Wikipedia (se esistenti) o dalle note di definizione e di ambito del ThNS e di AAT;
- sono stati inseriti, inoltre, tutti gli *alias* ritenuti pertinenti come ulteriori punti di accesso all'informazione: tra quelli in lingua italiana sono state aggiunte (se non già presenti) le forme sinonimiche che nel ThNS costituiscono UF con rinvio al termine preferito, così come tra quelli in lingua straniera, soprattutto in inglese, sono state aggiunte le varianti segnalate da AAT.

Per realizzare un'ontologia il più possibile salda e accurata, nella sezione Dichiarazioni, è stata posta particolare attenzione alla già descritta proprietà *sottoclasse di* (P279). Sono state eliminate sia le dichiarazioni errate (come nel caso di *procedimento carbro* (Q1035568) che, prima delle correzioni, risultava erroneamente sottoclasse di *stampa fotografica* (Q56055236), ma un processo non può logicamente essere sottoclasse di un oggetto), sia quelle ridondanti: nella costruzione dell'ontologia è infatti superfluo e, soprattutto, inefficace che

un elemento A sia contemporaneamente sottoclasse di B e di C se B è già a sua volta sottoclasse di C. Proprio in relazione a ciò, è stato necessario intervenire in casi come:

- *fotografia pubblicitaria* (Q1657850), indicata come sottoclasse sia di *fotografia commerciale* (Q12029910), sia di *fotografia professionale* (Q828872); quest'ultima dichiarazione è stata rimossa perché risultava superflua, in quanto *fotografia commerciale* (Q12029910) è già a sua volta sottoclasse di *fotografia professionale* (Q828872);
- stessa cosa per *ologramma* (Q3139490) che era sottoclasse sia di *positivo* (Q1413646), sia di *fotografia* (Q125191); *positivo* (Q1413646) è però già sottoclasse di *fotografia* (Q125191), quindi la seconda dichiarazione era ridondante.

In ogni caso, data l'importanza della relazione ontologica stabilita dalla proprietà P279 e per rendere più completa e autorevole l'informazione da essa fornita, si è ritenuto utile aggiungere, contestualmente, un riferimento per specificare la provenienza della dichiarazione riportata, in questo caso AAT (Fig. 3).



Figura 3. Riferimento ad AAT nella dichiarazione sottoclasse di (P279) dell'elemento WD ferrotipia (Q913381).

Per perfezionare e arricchire l'ontologia della fotografia sono state poi utilizzate altre specifiche proprietà come:

- *istanza di* (P31), che indica che un elemento è un'istanza, un tipo di una determinata classe, categoria o gruppo di oggetti: ad esempio *fotografia Kirlian* (Q1128987) è istanza di *tecnica fotografica* (Q1439691), oppure *fotografia a luce notturna* (Q1894442) è istanza di *genere fotografico* (Q3100808);
- *aspetto di* (P1269), che lega un elemento a un altro in quanto suo specifico contesto o argomento; ad esempio *fotografia spaziale* (Q81932940) è aspetto di *fotografia aerea* (Q191839);

- *parte di* (P361), che sottolinea che un dato elemento è parte o componente di un altro: ad esempio *sviluppo fotografico* (Q172839) è parte di *fotografia* (Q11633);
- *prodotto o materiale prodotto* (P1056) e *metodo di fabbricazione* (P2079), che collegano reciprocamente l'oggetto al processo che lo produce e viceversa: ad esempio *ambrotipia* (Q459832) ha come prodotto o materiale prodotto *ambrotipo* (Q67200337); viceversa, *ambrotipo* (Q67200337) ha come metodo di fabbricazione *ambrotipia* (Q459832);
- *praticato da* (P3095) e *campo di questa occupazione* (P425), che legano l'attività alla persona e alla professione che normalmente la pratica (e viceversa): ad esempio *fotografia scientifica* (Q2586218) è praticata da *fotografo scientifico* (Q21550957); viceversa, *fotografo scientifico* (Q21550957) ha come campo di questa occupazione *fotografia scientifica* (Q2586218);
- *distinto da* (P1889), che fa chiarezza tra concetti che potrebbero essere confusi l'uno con l'altro, in maniera analoga alla funzione svolta dalle note di orientamento nel ThNS: ad esempio *telefotografia* (Q117833345) è distinto da *fortelegrafia* (Q117833347), e viceversa;
- *contrario di* (P461), che collega tra loro concetti opposti: ad esempio *fotografia astratta* (Q333327) è contrario di *fotografia concreta* (Q1781772), e viceversa.

Passando invece alla sezione Identificativi, il processo di allineamento di un elemento di WD con il corrispondente termine del ThNS ha necessariamente compreso la verifica (o l'inserimento laddove assente) del relativo identificativo del ThNS, tramite la sopra ricordata proprietà *identificativo Thesaurus BNCF* (P508); oltre a questo, sono stati controllati e aggiunti i collegamenti con gli identificativi degli strumenti di indicizzazione semantica delle biblioteche nazionali straniere, mediante altre apposite proprietà:

- P244 collega l'elemento in questione alla LCSH;
- P268 conduce al RAMEAU;
- P227 porta al GND;
- P950 punta agli EMBNE.

Proprio per la specificità del dominio analizzato si è ritenuto inoltre opportuno e utile assegnare agli elementi WD anche il corrispondente identificativo di AAT, utilizzando la proprietà P1014.

Qualora fossero già presenti due o più identificativi AAT per lo stesso elemento WD, si è valutata la loro eventuale conferma o rimozione (in quest'ultimo caso spostandoli, se possibile, verso elementi più rispondenti). Per questo motivo, ad esempio, dall'elemento *fotografia di moda* (Q1146354) è stato rimosso il collegamento sia a *fashion photographs* (300226299), sia a *fashion photographers* (300386242), lasciando solamente il corretto link a *fashion photography* (300135798).

In tutti gli identificativi degli strumenti sopra elencati (ThNS, LCSH, RA-MEAU, GND, EMBNE, AAT), per maggiore completezza e comodità del lettore umano, è stata esplicitata la corrispondente forma verbale preferita, tramite il qualificatore *soggetto indicato come* (P1810) (Fig. 4).

astrofotografia (Q280855)

Identificativi

identificativo BNF 119383576 modifica
 soggetto indicato come Photographie astronomique
 ▶ 1 riferimento
 + aggiungi valore

identificativo GND 4122059-6 modifica
 soggetto indicato come Astrofotografie
 ▶ 1 riferimento
 + aggiungi valore

Figura 4. Riferimento di utilizzo del qualificatore soggetto indicato come (P1810) nell'elemento WD astrofotografia (Q280855).

Operazione facoltativa, ma utile, è stata anche quella di precisare il livello di equivalenza tra l'elemento e un identificativo esterno tramite il qualificatore *tipo di relazione* (P4390), che permette di indicare se si tratta, per esempio, di una corrispondenza esatta, parziale, ecc. (Fig. 5).

fotografia (Q11633)

identificativo Thesaurus BNCF 1526 modifica
 soggetto indicato come Fotografia
 tipo di relazione **corrispondenza esatta**
 nomi alternativi Attività fotografica
 Processo fotografico
 ▶ 1 riferimento
 + aggiungi valore

Figura 5. Riferimento di utilizzo del qualificatore tipo di relazione (P4390) nell'elemento WD fotografia (Q11633).

Le stesse procedure sono state seguite anche nei casi in cui, dopo adeguate verifiche che hanno confermato l'assenza dell'elemento ricercato, si è dovuto procedere a una creazione *ex novo*.

Tutte queste operazioni hanno portato all'elaborazione di interessanti riflessioni di confronto su differenze e affinità riguardo a natura, struttura e funzionalità dei due strumenti e a un perfezionamento e arricchimento reciproco sul piano terminologico.

6.2. Casi significativi

Nel corso delle operazioni di conciliazione tra il ThNS e WD, legate al progetto Ontologia della fotografia, sono emersi alcuni casi significativi e nello stesso tempo “insidiosi” che riportiamo di seguito a scopo esemplificativo.

6.2.1. Ricerca di elementi “nascosti” nell'ontologia

La prima criticità riscontrata riguarda tutti quei casi in cui non è immediatamente evidente l'esistenza o meno, nell'ontologia di WD, dell'elemento da collegare al termine del ThNS. È capitato spesso, infatti, che l'elemento ricercato fosse già presente nell'ontologia, ma non risultasse subito rintracciabile perché privo dell'etichetta in italiano; in molti casi, individuandone l'etichetta in inglese (o in tedesco, francese e spagnolo), è stato possibile procedere, dopo dovuti controlli e verifiche, a migliorare l'elemento nel modo sopra descritto. Nel peggiore dei casi, l'elemento può esistere ma avere solamente etichette in lingue non conosciute da chi opera, a volte addirittura in alfabeto non latino, che potrebbero quindi non comparire nei risultati di una semplice ricerca per stringa.

Prima di confermare con ragionevole sicurezza che l'elemento in questione non esiste, procedendo a quel punto a una creazione *ex novo*, può quindi essere utile scandagliare il succitato *Wikidata Graph Builder*, che consente di indagare gli alberi genealogici visualizzando graficamente le relazioni ontologiche, per individuare se il concetto sia già presente in altra forma all'interno delle catene gerarchiche dell'ontologia. Analoga operazione può essere svolta tramite l'utilizzo del *Wikidata Query Service* (WDQS), che consente di avere, tramite un'equivalente query SPARQL, una lista ordinata di risultati che rispondono ai requisiti appositamente impostati (es. “Tutte le entità che sono ricorsivamente sottoclasse di *fotografia* (Q11633)”, cioè del concetto di fotografia in quanto attività, con etichetta in italiano o in inglese, oppure “Tutte le entità che sono ricorsivamente sottoclasse di *fotografia* (Q125191)”, cioè del concetto di fotografia in quanto oggetto/risultato dell'omonima attività, con etichetta in italiano o in inglese); perciò, tramite *query* che ricercano ricorsivamente le sottoclassi è possibile ottenere elenchi di entità in cui provare a cercare, in maniera più lineare e talvolta più immediata e chiara, un determinato elemento “nascosto”.

6.2.2. *Casi di incongruenze interne agli elementi*

Altro problema spesso riscontrato riguarda il fatto che gli elementi concettuali possano non essere del tutto coerenti in sé stessi, ossia “conflati” (Pellizzari di San Girolamo 2024): può capitare infatti che si presentino incongruenze tra quanto affermato in *alias*, descrizioni (ad esempio, definizioni in lingua diversa veicolano, in certi casi, significati sensibilmente diversi), dichiarazioni, identificativi e *sitelink*, che possono talvolta erroneamente riguardare un concetto simile o in qualche modo collegato ma, comunque, differente.

Pertanto, è innanzitutto necessario comprendere quale sia il concetto che l'entità rappresenta, e per questo è fondamentale controllare i *sitelink* (in particolare i collegamenti alle voci di Wikipedia nelle varie lingue eventualmente presenti nell'elemento in questione). Anche tra questi collegamenti, talvolta, possono presentarsi delle incoerenze, in quanto le voci di alcune lingue possono essere riferite a un concetto diverso da quello cui rimandano altre; in questi casi è opportuno valutare il significato prevalente e stabilire il contenuto semantico rappresentato dall'elemento, procedendo poi a modificarlo e a migliorarlo (ad esempio, sistemando la proprietà P279, le descrizioni, gli *alias*, ecc.) e a rimuovere le parti individuate come non attinenti. Queste ultime, come già in precedenza sottolineato, andrebbero contestualmente anche spostate nell'elemento cui in realtà pertengono, in modo da evitare ulteriori errori o dubbi futuri, sempre con l'obiettivo di fare ordine nell'ontologia; qualora tale elemento non esista ancora, l'ideale sarebbe procedere alla sua creazione *ex novo*, assegnandogli poi i dati in questione. È quanto è stato fatto, ad esempio, nel caso di *procedimento alla gomma bicromatata* (Q519042), che presentava alcune incongruenze che creavano confusione, come l'erronea associazione all'identificativo AAT *gum bichromate prints* (300134949), relativo al risultato del processo e non al processo stesso, motivo per cui tale collegamento è stato rimosso e sostituito col corretto riferimento a *gum bichromate process* (300053501); al contempo, dal momento che non esisteva ancora un'entità relativa al risultato/oggetto, si è creato l'elemento *stampa fotografica alla gomma bicromatata* (Q130677822), inserendo poi l'identificativo AAT appena rimosso dall'altra entità.

6.2.3. *Compresenza di più elementi che sembrano coincidenti nel significato*

Altra criticità si riscontra nel momento in cui, andando a cercare nell'ontologia di WD un elemento da legare a un termine del ThNS si trovano due (o a volte più) elementi diversi che però sembrano rappresentare lo stesso concetto. A volte tali entità vanno semplicemente unite, in quanto rappresentano lo stesso concetto e non c'è nemmeno una fonte, tra quelle indicate, che le distingue: in linea di massima, infatti, in WD si tende a mantenere un elemen-

to solo se c'è almeno una fonte che lo differenzia dagli altri, mentre, in caso contrario, lo si unisce a quello più antico per data di creazione (che di solito è anche il più ricco di dati). Per questo motivo, l'elemento Q123360691 è stato unito a *fotografia stereoscopica* (Q17165350), andando a costituirne, quindi, un reindirizzamento.

A volte, invece, due (o più) elementi diversi che sembrano rappresentare lo stesso concetto paiono sovrapponibili solo alla prima impressione: in realtà hanno ragione di esistere distintamente perché legati a fonti differenti (per esempio, a diversi identificativi degli strumenti di indicizzazione stranieri o di AAT o, più frequentemente, a diverse voci di Wikipedia). Questo avviene perlopiù quando non tutte le Wikipedie nelle varie lingue sono allineate nell'avere una sola voce in relazione a un determinato concetto, e può magari capitare che una Wikipedia presenti due voci distinte che ne analizzano aspetti o sfumature anche lievemente differenti che in altre lingue non sussistono⁶; ad esse corrispondono, di conseguenza, due diversi elementi di WD, che sarebbero quindi sovrapponibili per tutte le fonti, eccetto che per quella particolare Wikipedia. Questo si verifica anche nel caso degli elementi *fotografia* (Q11633) e *registrazione fotografica* (Q1428014), rispettivamente collegati alle voci *Fotografie* e *Fotografische Aufnahme* di Wikipedia in tedesco. Solamente Wikipedia in tedesco propone questo grado di specificità, avendo due voci distinte laddove tutte le altre fonti ne presentano solo una. L'elemento *registrazione fotografica* (Q1428014), che apparentemente sembra un duplicato di *fotografia* (Q11633), in realtà ha quindi una sua ragione di esistere perché collegato alla relativa voce della Wikipedia tedesca, anche se questa risulta a tutti gli effetti essere l'unica fonte per quell'entità.

6.2.4. Accortezze necessarie nell'attribuzione della proprietà sottoclasse di (P279)

Un'altra criticità spesso riscontrata e non sempre di immediata risoluzione riguarda la proprietà ontologica *sottoclasse di* (P279). Come precedentemente affermato, tale proprietà in WD esprime solo parzialmente il concetto thesaurale di BT, andando a coprire, in linea di massima, il solo ruolo del BTG. Capita spesso, quindi, che la relazione gerarchica tra due elementi/concetti espressa in WD dalla proprietà *sottoclasse di* (P279) non corrisponda a quella che lega due termini/concetti all'interno di un thesaurus, che può essere, infatti, non solo di tipo generico (BTG) ma anche partitivo (BTP) o esemplifi-

⁶ Questa situazione può essere legata sia all'attività di traduzione, cioè "dire la stessa cosa in un'altra lingua" – e quindi relativa al contesto linguistico che può determinare varie accezioni di significato, in quanto lingue diverse esprimono i concetti in maniera diversa – sia a un maggiore o minore grado di precisione e di approfondimento di un concetto applicato dalle singole persone che operano alla compilazione delle voci, che porta quindi a raggruppare più concetti in un'unica voce oppure a scinderli in voci distinte.

cativo (BTI). Per questo motivo, le relazioni di BT dei vocabolari controllati devono essere sempre usate con cautela come fonte di affermazione della proprietà P279 nell'ontologia di WD; non sempre, infatti, per tale riferimento ci si è potuti basare automaticamente sulla scelta del ThNS. In quest'ultimo, per esempio, *Sviluppo fotografico* (70749) ha BT *Fotografia* (1526) secondo una relazione gerarchica partitiva, resa in WD, come visto, non tanto dalla proprietà *sottoclasse di* (P279), quanto da *parte di* (P361), essendo lo sviluppo fotografico non un tipo di attività o processo fotografico, bensì una parte di esso; così, al corrispondente elemento *sviluppo fotografico* (Q172839) è stata inserita sia la dichiarazione “parte di *fotografia* (Q11633)”, sia “sottoclasse di *tecnica fotografica* (Q1439691)”.

6.2.5. “Ostacoli” alla creazione del legame con gli strumenti di indicizzazione stranieri

Un'altra questione cui si è dovuto talvolta far fronte riguarda quei casi in cui uno degli strumenti di indicizzazione stranieri con cui attivare il collegamento ha, spesso per motivi legati all'attività di soggettazione delle risorse, più concetti compresi in un solo termine; RAMEAU, ad esempio, ha quasi sempre un unico termine che esprime sia il significato di un'attività fotografica, sia quello del risultato da essa ottenuto, concetti che, invece, sia il ThNS sia l'ontologia di WD mantengono distinti. Stessa cosa capita molte volte anche in EMBNE. Si è dovuto perciò capire in che modo relazionare gli strumenti in casi come questi: considerato che il software del ThNS non consente di attribuire a due termini diversi lo stesso equivalente straniero, è stata presa la decisione di attivare il collegamento in questione solamente al termine corrispondente al concetto di “risultato dell'attività fotografica” (categoria *Forme*). L'ontologia di WD, invece, non esige questo tipo di scelta selettiva, permettendo di collegare al medesimo identificativo più elementi; salvo rare eccezioni, comunque, se lo stesso identificativo è associato a più di una entità, la piattaforma segnala una violazione del vincolo di valore unico, indice di un potenziale errore. È però possibile risolvere tale problema con l'aggiunta dell'apposito qualificatore *identificativo condiviso con* (P4070), che permette di segnalare la condivisione, puntualizzando in ciascuno dei due elementi a quale altro elemento è stato assegnato lo stesso identificativo (Fig. 6).

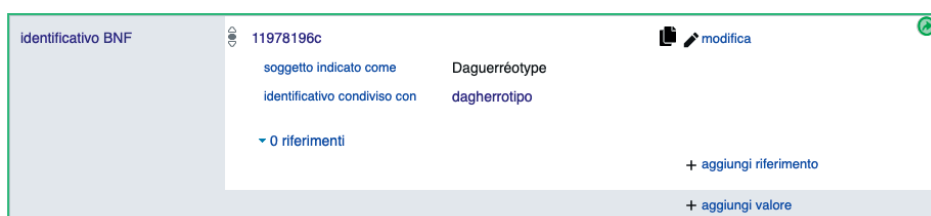


Figura 6. Identificativo RAMEAU collegato a dagherrotipia (Q179744), col qualificatore identificativo condiviso con (P4070); lo stesso qualificatore è presente anche nell'altro elemento dagherrotipo (Q64029175).

6.2.6. Casi particolari (o particolarmente insidiosi)

Il termine del ThNS *Macchine fotografiche reflex* (16398) non presentava, al momento dell'analisi per l'allineamento tra gli strumenti, il collegamento al relativo elemento di WD. In fase di ricerca dell'entità corrispondente, oltre al più generico *fotocamera reflex* (Q2858617), si è rilevata anche la presenza di *reflex a obiettivo singolo* (Q196353), anch'esso potenzialmente collegabile al termine del ThNS, in quanto in quest'ultimo *Reflex a obiettivo singolo* (77040), *Reflex monobiettivo* (77041), *Single-lens reflex* (77043) e *SLR <Single-lens reflex>* (77042) risultano tutti rinvii al preferito *Macchine fotografiche reflex* (16398).

Si è cercato, quindi, di approfondire l'indagine, analizzando le voci di Wikipedia e gli identificativi collegati a *fotocamera reflex* (Q2858617) e *reflex a obiettivo singolo* (Q196353): i due elementi di WD sono collegati rispettivamente alle distinte voci della Wikipedia in italiano *Fotocamera reflex* e *Reflex a obiettivo singolo*, e ai distinti identificativi AAT *reflex cameras* (300022654) e *single lens reflex cameras* (300022655). In WD, quindi, per i motivi sopra descritti, i due elementi hanno ragione di esistere distintamente; si è cercato, a questo punto, di capire se fosse opportuno attuare questa distinzione anche nel ThNS, andando quindi a staccare *Reflex a obiettivo singolo* (77040) – e, di conseguenza, anche gli altri UF *Reflex monobiettivo* (77041), *Single-lens reflex* (77043) e *SLR <Single-lens reflex>* (77042) – da *Macchine fotografiche reflex* (16398), arrivando perciò ad avere due distinti termini preferiti. Proprio per questo, oltre ad approfondire le ricerche su fonti e repertori specialistici, si è indagato anche quanto è stato fatto negli strumenti di indicizzazione stranieri⁷, cosa che ha portato, inoltre, a scoprire che sia LCSH sia RAMEAU hanno anche un termine relativo al concetto di “reflex biottica” (*Twin-lens cameras* (sh85139038) e *Appareils photographiques reflex biobjectifs* (124195199)). Anche WD ha un elemento (*reflex biottica* (Q209811)) relativo specificatamente

⁷ Al momento non esiste un equivalente EMBNE.

alla reflex biottica⁸; in ThNS, invece, al momento dell'analisi, non esisteva un concetto corrispondente e negli Online Public Access Catalog (OPAC) della BNCF e SBN c'era una esigua e datata garanzia bibliografica, quindi si sono presentate le seguenti opzioni di trattamento:

1) staccare *Reflex a obiettivo singolo* (77040), e di conseguenza gli altri UF sopra menzionati, da *Macchine fotografiche reflex* (16398) rendendoli entrambi preferiti e introducendo, come preferito anch'esso, un termine corrispondente al concetto di reflex biottica⁹;

2) mantenere come termine preferito solamente *Macchine fotografiche reflex* (16398) con UF *Reflex a obiettivo singolo* (77040), *Reflex monobiettivo* (77041), *Single-lens reflex* (77043) e *SLR <Single-lens reflex>* (77042) e aggiungere come ulteriori UF uno o più termini relativi al concetto di reflex biottica.

Dopo attenta valutazione della garanzia bibliografica sull'argomento (come detto scarsa e datata) e ulteriori ricerche sia su repertori specialistici sia su monografie e manuali presenti negli OPAC BNCF e SBN, si è deciso di adottare la seconda opzione: si è mantenuto come termine preferito solamente *Macchine fotografiche reflex* (16398), confermando i rinvii già presenti in ThNS e aggiungendo come ulteriore UF *Reflex biottiche* (77630) e *TLR <Twin-Lens Reflex>* (77631). Conseguentemente, è stato creato il reciproco collegamento tra *Macchine fotografiche reflex* (16398) (che in un primo momento era stato connesso all'elemento di WD *reflex a obiettivo singolo* (Q196353)) e il più generico *fotocamera reflex* (Q2858617).

Il termine del ThNS *Microfotografia* (27298) presentava già il link all'elemento di WD *microfotografia* (Q3274922), attribuito con la sopra descritta procedura di abbinamento semiautomatico. L'analisi dell'elemento ha confermato la corretta attribuzione dell'equivalente WD, anche se la dichiarazione "distinto da *fotomicrografia* (Q130416)" ha attirato l'attenzione e sollevato alcune riflessioni, in quanto in ThNS *Fotomicrografia* (76993) risultava come termine non preferito con rinvio a *Microfotografia* (27298). Dopo aver esteso e approfondito le ricerche sia sui repertori specialistici sia su manuali e monografie relative all'argomento, si è potuto constatare che le due tecniche sono solo apparentemente coincidenti, soprattutto per la confusione derivata dai nomi molto simili, ma di fatto utilizzano procedure e strumenti differenti, che conducono a risultati distinti. Alcuni strumenti e repertori schiacciano i due concetti (es. RAMEAU schiaccia *Microphotographie* su *Photomicrographie* (11979866m); AAT ugualmente schiaccia *photomicrography* su *microphotography* (300053550)), mentre LCSH ed EMBNE più correttamente

⁸ Esso è legato alla voce di Wikipedia in italiano *Reflex biottica* e al corrispettivo identificativo AAT *twin-lens reflex cameras* (300022657).

⁹ In questo caso, probabilmente, sia *Reflex a obiettivo singolo* (77040) sia l'altro termine relativo al concetto di reflex biottica sarebbero diventati NT di *Macchine fotografiche reflex* (16398).

distinguono i due concetti (rispettivamente *Microphotography* (sh85084884) / *Photomicrography* (sh85101386) e *Microfotografia* (XX525323) / *Fotomicrografia* (XX525338)), con tanto di note di orientamento reciproche. Anche la WebDewey(IT) distingue i due concetti, ponendo la fotomicrografia a 502.82 tra le tecniche e procedure ausiliari della scienza, mentre la microfotografia a 686.43 sotto la fotocopiatrice, tra le attività relative alla tecnologia. Si è ritenuto importante, quindi, anche per favorire il sempre maggiore utilizzo del ThNS da parte di realtà specialistiche, scindere le due tecniche, modificando innanzitutto la nota d'ambito di *Microfotografia* (27298) da "Fotografia di soggetti molto piccoli realizzata con l'ausilio di microscopi" (definizione errata in quanto si riferiva non tanto alla *Microfotografia*, quanto alla *Fotomicrografia*) a "Tecnica mediante la quale si riproduce fotograficamente un'immagine rimpicciolita di un oggetto relativamente grande". Si è poi trasformato il termine *Fotomicrografia* (76993) da non preferito a preferito, inserendo anche qui una precisa nota d'ambito: "Tecnica mediante la quale si riproducono immagini ingrandite di oggetti molto piccoli, generalmente per mezzo del microscopio"; entrambi i termini sono poi stati corredati di note di orientamento reciproche. Ovviamente, per registrare l'avvenuta modifica, il termine *Fotomicrografia* (76993) è stato opportunamente corredato dalla nota storica "Precedentemente in Thesaurus come termine non preferito, con rinvio a Microfotografia (fino a marzo 2025)". A questo punto si è potuto procedere all'attivazione dei giusti collegamenti con i corrispondenti elementi di WD: *Microfotografia* (27298) con *microfotografia* (Q3274922) e *Fotomicrografia* (76993) con *fotomicrografia* (Q130416).

7. Conclusioni e progetti futuri

La collaborazione tra ThNS e WD ha prodotto dei risultati proficui per entrambe le comunità, portando a un miglioramento dei due strumenti grazie a una puntualizzazione delle relazioni tra concetti/termini/entità e di conseguenza a una maggiore definizione della loro struttura sintetica.

Quale futuro ci aspetta? Desideriamo concludere questo contributo con alcune proposte e progetti che vorremmo realizzare prossimamente e che investono il piano semantico e strutturale del ThNS e di WD.

In primis, vorremmo proseguire l'analisi e la comparazione tra i due strumenti anche in altri domini, secondo la metodologia già sperimentata, per incrementare allineamenti semantici o per risolvere eventuali criticità di colloquio. Questo porterebbe a un maggiore sfruttamento delle potenzialità di WD nel *workflow* di allestimento delle voci del ThNS, anche come possibile fonte di nuova terminologia e di collegamenti con altri KOSs.

Inoltre sarebbe interessante poter sperimentare la reingegnerizzazione e ristrutturazione del ThNS in un formato ontologico in modo da offrire un più

vasto numero di relazioni semantiche definite con maggiore precisione: ad esempio, miglioramento delle gerarchie e dei criteri di assegnazione dei BT (distinti nelle loro tre tipologie) in casi potenzialmente poligerarchici; valutazione di eventuali RT per relazioni antonimiche e relazioni causa-effetto.

In questo processo di implementazione e sviluppo sarebbe auspicabile anche un miglioramento della struttura dell'ontologia WD basandosi su principi di raggruppamento (categorie e sottocategorie) e di divisione più rigorosi al fine di creare un modello concettuale più razionale, organizzato e organico.

Riferimenti bibliografici

- AngryLoki. 2016. *Wikidata Graph Builder*. GitHub Pages. <https://angryloki.github.io/wikidata-graph-builder/>.
- Biblioteca Nazionale Centrale di Firenze, e Wikimedia Italia. 2025. *Convenzione tra la Biblioteca Nazionale Centrale di Firenze e Wikimedia Italia*. Firenze: Biblioteca Nazionale Centrale di Firenze. 16 giugno 2025. <https://bncf.cultura.gov.it/wp-content/uploads/2019/11/20250616-Convenzione-BNCF-WMI-signed-signed.pdf>.
- Cencetti, Elena. 2025a. "Terminologia di ambito fotografico presente nel Thesaurus Nuovo soggettario." Zenodo, marzo 2. <https://doi.org/10.5281/ZENODO.14956464>.
- Cencetti, Elena. 2025b. "Thesaurus Nuovo soggettario e Wikidata: analisi di un corpus terminologico relativo all'ambito della fotografia." Tesi per il Master in Organizzazione e gestione degli archivi, catalogazione e meta-dazione di risorse manoscritte, stampate e digitali, Università degli Studi di Firenze.
- Cencetti, Elena, Camillo Carlo Pellizzari di San Girolamo, e Elisabetta Viti. 2025. "Termini, dati e collegamenti: 'conversazioni' tra il Thesaurus del Nuovo soggettario e Wikidata." *Imagines* 12: 85–94.
- Corti, Laura, e Fiorella Gioffredi Superbi. 2004. *Glossario di terminologia fotografica*. Roma: Associazione Italiana Biblioteche. <https://www.aib.it/aib/lis/lpi13eg.htm>.
- Cyganiak, Richard, Anja Jentzsch, Max Schmachtenberg, and Christian Bizer. 2025. "LOD Cloud – Linking Open Data". Lod-Cloud.net. Ultimo accesso 17 Novembre 2025. <https://lod-cloud.net/>.
- Frettie. 2025. *User:Frettie/consistency_check_add.js*. Ultimo aggiornamento 15 maggio 2025. Wikidata. https://www.wikidata.org/wiki/User:Frettie/consistency_check_add.js.

- Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5 (2): 199–220. <https://doi.org/10.1006/knac.1993.1008>.
- Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: Digital Library Federation, Council on Library and Information Resources.
- International Organization for Standardization (ISO). 2011. *ISO 25964-1:2011 – Information and documentation. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*. Geneva: ISO.
- International Organization for Standardization (ISO). 2013. *ISO 25964-2:2013 – Information and documentation. Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies*. Geneva: ISO.
- Kless, Daniel, Ludger Jansen, Jutta Lindenthal, and Jens Wiebensohn. 2012. "A method for re-engineering a thesaurus into an ontology." In *Frontiers in Artificial Intelligence and Applications*, edited by Giancarlo Guizzardi and Maureen Donnelly. Amsterdam: IOS Press. <https://doi.org/10.3233/978-1-61499-084-0-133>.
- Lanza, Claudia. 2022. *Semantic Control for the Cybersecurity Domain: Investigation on the Representativeness of a Domain-Specific Terminology Referring to Lexical Variation*. Boca Raton: CRC Press. <https://doi.org/10.1201/9781003281450>.
- Ministero della Cultura. 2025. "Descrizione dataset." *Dati della Cultura*. Ultimo accesso 17 novembre 2025. http://dati.cultura.gov.it/descrizione_dataset/.
- Pellizzari di San Girolamo, Camillo Carlo. 2024. "Conflations and Duplications in Wikidata Items: Causes, Detection, Solutions, and Issues." In *Proceedings of the Wikidata Workshop 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023)*, Athens, 2023. CEUR-WS. <https://ceur-ws.org/Vol-3640/paper4.pdf>.
- Pellizzari di San Girolamo, Camillo Carlo. 2025. "Links between Wikidata, the Thesaurus Nuovo soggettario and other thesauri." Zenodo, September 30. <https://doi.org/10.5281/zenodo.17235305>.
- Studer, Rudi, V. Richard Benjamins, and Dieter Fensel. 1998. "Knowledge Engineering: Principles and Methods." *Data & Knowledge Engineering* 25 (1–2): 161–97. [https://doi.org/10.1016/s0169-023x\(97\)00056-6](https://doi.org/10.1016/s0169-023x(97)00056-6).
- Wikidata community. 2025a. "Wikidata: Notability." *Wikidata*. Ultimo accesso 17 novembre 2025. <https://www.wikidata.org/wiki/Wikidata:Notability>.

- Wikidata community. 2025b. "Special:Statistics." *Wikidata*. Ultimo accesso 17 novembre 2025. <https://www.wikidata.org/wiki/Special:Statistics>.
- Wikidata community. 2025c. "Wikidata: List of properties." *Wikidata*. Ultimo accesso 17 novembre 2025. https://www.wikidata.org/wiki/Wikidata:List_of_properties.
- Wikidata community. 2025d. "MediaWiki:Gadgetrelateditems.js." *Wikidata*. Ultimo accesso 17 novembre 2025. <https://www.wikidata.org/wiki/MediaWiki:Gadgetrelateditems.js>.
- Wikimedia Foundation. 2025a. *Wikidata: Gruppo Wikidata per Musei, Archivi e Biblioteche/Nuovo soggettario*. https://www.wikidata.org/wiki/Wikidata:-Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/Nuovo_soggettario.
- Wikimedia Foundation. 2025b. *Help:Property Constraints Portal*. Wikidata. Ultimo accesso 18 novembre 2025. https://www.wikidata.org/wiki/Help%3AProperty_constraints_portal.
- Wikimedia Foundation. 2025c. *Property talk:P508*. Wikidata. Ultimo accesso 18 novembre 2025. https://www.wikidata.org/wiki/Property_talk%3AP508.

The transferability of brilliance

Pontish Yeramyán*, Alexander Murzaku*, Nick Uva*, Curt Anderson*

Abstract: Brilliance is the thinking that causes our most extraordinary performance. Enhancing one's brilliance and adopting brilliance from others is hard to do reliably. The transfer of brilliance from one person to another is often difficult; it most often happens in a family context where there is frequent and intense contact between brilliant people and someone with the desire to acquire that brilliance. In the search for what kind of thinking constitutes brilliance, we identify seven themes of brilliance. Our goal is to bring clarity to what brilliant thinking is and outline ways that brilliance can be documented, shared, and studied.

Keywords: Knowledge transfer, Knowledge management, Tacit knowledge, Human brilliance, Knowledge documentation.

1. Introduction

Brilliance is the exceptional intellectual capacity, creativity, and profound insight that drives groundbreaking ideas and innovations – an invaluable asset. Reliably increasing one's own brilliance and transferring brilliance to others is extraordinarily difficult. The successful transfer of brilliance from one person to another is rare, posing significant challenges in perpetuating exceptional thinking and innovation, whether it be from master to apprentice, from coach to trainee, or across generations. This rarity stems from the difficulties inherent in capturing and conveying this special form of tacit knowledge (*tacit* in the sense of Polanyi 1966), intuition, and deep cognitive processes that together constitute brilliance. Traditional mentorship and educational models often focus on explicit knowledge transfer, neglecting the nuanced thought patterns and experiential wisdom that underlie exceptional capabilities.

* Genius Institute at Gap International, Springfield, PA, USA. pontish.yeramyán@gapinternational.com. ORCID: 0009-0004-9218-5707; alex.murzaku@gapinternational.com. ORCID: 0009-0003-1939-6543; nick.uva@gapinternational.com. ORCID: 0009-0004-2818-3571; curt.anderson@gapinternational.com. ORCID: 0000-0002-8515-2948.

Based on thousands of interviews with high-performing individuals, we propose that brilliance is characterized by seven major themes, which we call Genius Universals. These themes are mindsets that a person embodies, and from these mindsets a person takes action in order to perform at their highest potential. Geniuses and other brilliant individuals are often unaware that these themes are part of their thinking, and it takes effort to distinguish and discover the nature and structure of one's own thinking.

We propose that a core aspect of brilliance is tacit knowledge about one's sense of self, purpose, and understanding of the world. Brilliance draws upon not just the individual's factual understanding of their area of expertise, but also includes a layer of thinking that is deeply ingrained and oftentimes inaccessible. As Einstein is quoted saying in a biography (Pais 2005, 462): «The essential of the being of a man of my type lies precisely in what he thinks and how he thinks, not on what he does or suffers.» This knowledge is a type of tacit knowledge. The Genius Universals, similarly, are themes not of explicit knowledge or “what he does,” but of tacit knowledge, or “how he thinks.”

It is widely recognized that tacit knowledge can be difficult to externalize, document, and transfer to other individuals (Nonaka 1994). While it is possible to transfer tacit knowledge, transfer is often unreliable and inconsistent, requires extensive amounts of time and personal contact between individuals, and cannot be done at scale within an organization. This difficulty is the critical impediment to the transfer of brilliance. If brilliance is the unique mindset of an individual which allows them to achieve exceptional results, successful transfer of this brilliance allows others to adopt and operate from this mindset themselves. One can learn not just techniques and processes from someone accomplished, but also the mindset that enabled them to do the things that led to them being successful. It also allows for the possibility of self-transfer, by which we mean the ability for someone to recall and recognize their own brilliance and apply it in new contexts.

Our goal is to bring clarity to what brilliant thinking is, and once it is identified what brilliance is, outline ways that brilliance can be documented, shared, and studied. The paper is organized as follows. First, we distinguish the concept of brilliant thinking as the type of thinking exhibited both by people recognized for their accomplishments and also by everyday people achieving at their highest level. Next, we propose what we call the Genius Universals, seven themes that recur frequently in the thinking of brilliant people, and which we claim constitute brilliant thinking. We establish that the Genius Universals can be identified in the language of historical and contemporary geniuses, but raise the problem that the brilliance of many thinkers is unreliably transmitted to their students, their employees, and other people. Following that, we argue that families are a case where brilliance is often successfully transferred, and identify the key features of the family environment that allow for transfer to be successful. We assimilate the

problem of transfer of brilliance to the problem of tacit knowledge documentation and transfer, and sketch an interview process for uncovering and documenting brilliance. Finally, we take stock and point out future goals for the scientific understanding of brilliance and its documentation and transfer.

2. On brilliance and the Genius Universals

Geniuses generate original and transformative ideas and achieve exceptional mastery in a specific domain, often reshaping how we understand or approach complex challenges. Whether it is a scientific breakthrough, an artistic masterpiece, or a simple yet profound insight, geniuses leave a lasting impact through their contributions. We call the mindset that enables the extraordinary performance of a genius “brilliance”¹.

When we discuss brilliance, we refer to a particular kind of thinking that allows for one to achieve exceptional results. For example, there are brilliant scientists, writers, painters, musicians, teachers, doctors, and so on, individuals who are not just highly competent in their area of expertise, but who stand out among their peers. Brilliance isn't limited to those who we broadly recognize as geniuses, but is a kind of thinking that everyone exhibits when they are working at their highest level. There are geniuses we recognize for their outstanding, widely known achievements, but brilliant thinking is also present in people who will never gain that kind of recognition. Although we use geniuses as examples of brilliance in this paper, our scope is wider than just geniuses; it is the entire phenomenon of brilliance.

Brilliance cannot be grounded in simple mastery of the facts and procedures of a simple discipline. If it were, we would recognize many skilled, technically competent people as being brilliant in their craft. Instead, we view brilliance as arising from a person's way of thinking. It is often difficult for one to describe the ways of thinking which constitute their own brilliance, much less do so systematically. Brilliance is also not necessarily “about” a specific discipline or competency, but is domain independent; many successful people are able to be successful in separate areas, suggesting that the mindset that makes them successful is sufficiently general purpose that it can be applied across different contexts.

What constitutes a brilliant mindset? Over the last twenty years, we have conducted thousands of interviews with businesspeople with the goal of understanding the mindsets that led to their accomplishments (whether personal or in business)². Domain knowledge has varied considerably across the people that we

¹ As a note, we will use the terms *brilliance* and *Genius* as near synonyms. *Brilliance* is used as a term for the overall category of thinking, while *Genius* is our technical term for this kind of thinking.

² We interviewed businesspeople because they are the client base of our organization. We assert that *Genius* is present in all areas of achievement, not only business.

have interviewed, but once we ask questions that probe the thought processes behind their accomplishments, patterns begin to emerge. We’ve identified seven themes that recur frequently enough that we believe them to be core aspects of brilliant thinking, regardless of industry, vocation, functional role, and experience. We call these themes “Genius Universals” (cf. “themes of brilliance” in Murzaku et al. 2024). These themes co-occur in a person’s brilliance; nobody’s brilliance is solely defined by a single one of these themes, but many or even all appear when we look at an individual’s brilliance. This implies that brilliance is multidimensional and cannot be described by a single property.³

The Genius Universals are briefly described in Table 1: Each Genius Universal is named by a cluster of four words or phrases that evoke the general semantic contour of the theme⁴. The accompanying descriptions are an attempt to distill the theme into natural language.

Genius Universal	Description
POSSIBILITY / CREATION / CHOICE / FREEDOM	Freedom to imagine the unimaginable and openness to what could be; boldness to create a new reality
PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY	Personal drive and greater reason for being, rooted in the ambition to positively impact others
EXTREMENESS / INTENSITY / OBSESSION / ABOVE AND BEYOND	Drive to go above and beyond, channeling intense focus and dedication to deliver peak performance
CURIOSITY / DISCOVERY / LEARNING / GROWING	Appetite for learning, growth, and expansion; The need to deepen knowledge and push boundaries of what is known
AT-STAKENESS / CONVICTION / RESOLVE / PERSEVERANCE	Drive within oneself connected to the joy or pain of what may be gained or lost; the will to persist in the face of challenges
JOY / EXCITEMENT / EXHILARATION/SPIRIT	Passion and excitement and experiencing joy and energy; fostering an environment of contagious enthusiasm

³ Hjørland (2008) distinguishes the “facet-analytic” approach to knowledge organization. We can take the Genius Universals as facets of Genius that can be used for the organization and categorization of Genius language. Looking further down the road, the Genius Universals could be formalized with some combination of tools from formal ontology (e.g. General Formal Ontology, Herre (2007, 2013)), formal linguistic semantics (Heim and Kratzer 1997), or computational linguistics (WordNet, Miller et al. 1990; FrameNet, Baker et al. 1998), a possibility we are currently exploring.

⁴ One way of thinking about this is that the four words/phrases of a Genius Universal describe a fuzzy concept with a center and a periphery, perhaps a prototype (Rosch 1975; Rosch and Mervis 1975), a graded concept (Kay and McDaniel 1978), or a conceptual space (Gärdenfors 2000). See also Bianchini, Giusti and Gnoli (2017) who use a conceptually similar idea (the APUPA bell curve) in knowledge organization.

TOGETHERNESS/RELATIONSHIP/ CONNECTION/LOVE	Creating deep, meaningful connections with people and what can be accomplished as a collective and cohesive whole
---	---

Table 1. Descriptions of Genius Universals.

The Genius Universals stem from our work with clients and the interviews that we have conducted, but we find that each Genius Universal can be connected to independently motivated performance-related constructs in the literature. We won't explore these connections in detail here, but Table 2 provides an initial offer for how our Genius Universals might link to existing work.

Genius Universal	Connection to literature
POSSIBILITY / CREATION / CHOICE / FREEDOM	Divergent thinking, Guilford (1950, 1967)
PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY	PERMA model, Seligman (2011) Drive, Pink (2009)
EXTREMENESS / INTENSITY / OBSESSION / ABOVE AND BEYOND	Grit, Duckworth et al. (2007)
CURIOSITY / DISCOVERY / LEARNING / GROWING	Curiosity as motivation, Kashdan & Steger (2007) Growth mindset, Dweck (2006)
AT-STAKENESS / CONVICTION / RESOLVE / PERSEVERANCE	Grit, Duckworth et al. (2007)
JOY / EXCITEMENT / EXHILARATION / SPIRIT	Broaden and Build Theory, Fredrickson (1998, 2001)
TOGETHERNESS / RELATIONSHIP / CONNECTION / LOVE	Belongingness Hypothesis, Baumeister & Leary (1995)

Table 2. Tentative connections between Genius Universals and existing literature on performance.

Brilliant or Genius thinking is thus a kind of mindset that creates the possibility of performance way beyond the norm. The Genius Universals are part of the language of an individual – the things they say out loud to others, what they write, and even what they say to themselves – and they provide an access point to that person's brilliance. Our aim in understanding the Genius Universals is to precisely characterize the mindset behind brilliance. When that mindset can be made salient for someone, they can more readily adopt that mindset and bring themselves to a new level. In addition, by studying brilliant thinking and the Genius Universals, we can create ways for Genius thinking to be made explicit, transferable, and more easily recalled. These notions of a person acquiring and using Genius thinking, either the thinking of others or their own thinking, we call "transfer of brilliance" and "transfer of Genius."

3. Genius and Genius transfer

3.1. The Genius of geniuses

When seeking access to brilliant thinking, it is common to look to figures from history who are recognized for their brilliance. The legacy of a person whose thinking produced exceptional outcomes can be a great inspiration. Some figures are so strongly associated with brilliance that invocation of their name alone may serve as shorthand for intelligence or innovation, as in the educational children's series *Little Einsteins* (after physicist Albert Einstein), or the automaker Tesla (after 19th and 20th century inventor Nikola Tesla). People are often inspired by the brilliance and extraordinary achievements of others. Being inspired, though, is not the same as thinking in the way that caused that achievement. Learning to think like another brilliant person is not easy, and even in the closet mentorships, brilliant thinking does not transfer from mentor to mentee.

Leonardo da Vinci is perhaps our prototypical example of a genius, even more than five hundred years after his death. In his lifetime, he was widely recognized by his contemporaries as an exceptional talent, not just for his skill as an artist, but also for his wide-ranging studies in disparate academic disciplines. Today, he is recognized not only for his paintings, but also his studies of human anatomy, his engineering drawings, and his drawings of the natural world.

A significant amount of time and a large percentage of da Vinci's writings were devoted to investigating and explaining the concepts of "seeing" and "distinguishing". His work can perhaps best be summarized as a journey toward *saper vedere* "knowing how to see" (Cassirer 1963). It's telling that he was so preoccupied with the importance of seeing and distinguishing, that the word *occhio* "eye" appears in more than nine hundred folios across his writings, where he explains how to observe and to distinguish the world. In his studies, da Vinci observed, in minute detail and with obsession, how birds or insects fly, the movement of their wings, their anatomy, and even the air currents around them. He studied them, he drew them, he tried for years to recreate that miracle of nature, and he recorded this in his Codex Atlanticus. Da Vinci was very much about not just observing the natural world, but also understanding his own brilliance for observation and distinguishing crucial differences that had not been noted before.

Because of his extensive writings, da Vinci's brilliance is documented better than many other historical figures. Table 3 provides a sample of quotes from da Vinci, along with the Genius Universal that the quote appears to indicate, establishing the Genius Universals as part of da Vinci's brilliance.

Quote	Genius Universal ⁵
Obstacles cannot bend me. Every obstacle yields to effort. Not to leave the furrow. He who fixes his course by a star changes not.	AT-STAKENESS / CONVICTION / RESOLVE / PERSEVERANCE
Iron rusts from disuse, stagnant water loses its purity and in cold weather becomes frozen; even so does inaction sap the vigour of the mind!	EXTREMENESS / INTENSITY / OBSESSION / ABOVE AND BEYOND
As a well-spent day brings happy sleep, so life well used brings happy death.	PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY
The natural desire of good men is knowledge.	CURIOSITY / DISCOVERY / LEARNING / GROWING
May I be deprived of movement ere I weary of being useful. Movement will fail sooner than usefulness. Death rather than weariness. I never weary in being useful. I am not tired of serving others.	PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY

Table 3. Leonardo da Vinci quotes and Genius Universals.

Next we consider Steve Jobs, co-founder of Apple, who is often considered to be a modern example of a genius. Jobs positioned himself at the intersection of art, science, and technology, and his vision changed the computing and electronics industries repeatedly in the 1980s, 1990s, and 2000s. His accomplishments include developing the first Macintosh computer with Steve Wozniak, the formation of Pixar, and the development of the iPod, the iPhone, and the iPad during his second tenure as Apple CEO. More than a dozen biographies of Jobs have been written, and there are more than seventy interviews with him recorded and made publicly available.

Although many interviews and biographies of him are not conducted with the goal of discovering Jobs' brilliance in the sense that we develop in this paper, we can look at what he said throughout his life and infer something about his mindset, his vision for the role of computing in society. When we do so, we notice several of the Genius Universals present in what he says. Table 4 gives samples of Jobs' language and the Genius Universals present.

⁵ All Leonardo da Vinci quotes are excerpted from *the Notebooks of Leonardo da Vinci* (MacCurdy 1955).

Quote	Genius Universal	Source
Stay hungry. Stay foolish.	CURIOSITY / DISCOVERY / LEARNING / GROWING	Stanford Commencement Speech, 2005
I'm convinced that the only thing that kept me going was that I loved what I did.	JOY / EXCITEMENT / EXHILARATION / SPIRIT	Stanford Commencement Speech, 2005
Great things in business are never done by one person. They're done by a team of people.	TOGETHERNESS / RELATIONSHIP / CONNECTION / LOVE	<i>60 Minutes</i> interview, 2003
We attract a different type of person [at Apple]... Someone who really wants to get in a little over his head and make a little dent in the universe.	PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY	<i>Playboy</i> interview, 1985
We're gambling on our vision, and we would rather do that than make "me too" products. Let some other companies do that. For us, it's always the next dream.	POSSIBILITY / CREATION / CHOICE / FREEDOM	Interview during the release of the first Macintosh computer, 1984

Table 4. Steve Jobs quotes and Genius Universals.

Now let us turn to Marie Curie, a scientist whose genius shaped physics and chemistry. Curie's work led to groundbreaking discoveries in radioactivity, a term she herself coined. She became the first woman to win a Nobel Prize, the only person to ever win Nobel Prizes in two different scientific fields (Physics and Chemistry), and her research led to the isolation of the elements polonium and radium. Curie's genius lay not only in her technical mastery and innovative experimental methods but also in her unwavering commitment to the pursuit of knowledge, even in challenging and dangerous conditions. Biographies and collections of her letters reveal her determination and intellect. Table 5 provides examples of Curie's statements and establishes Genius Universals such as PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY and JOY / EXCITEMENT / EXHILARATION / SPIRIT as part of her brilliance.

Quote	Genius Universal	Source
I am among those who think that science has great beauty. A scientist in his laboratory... is also a child placed before natural phenomena which impress him like a fairy tale.	JOY / EXCITEMENT / EXHILARATION / SPIRIT	<i>Madame Curie, 1938</i>
Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something, and that this thing, at whatever cost, must be attained.	AT-STAKENESS / CONVICTION / RESOLVE / PERSEVERANCE	<i>Madame Curie, 1938</i>
Radium is not to enrich anyone. It is an element; it belongs to all people.	PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY	<i>Madame Curie, 1938</i>
All my life through, the new sights of Nature made me rejoice like a child.	JOY / EXCITEMENT / EXHILARATION / SPIRIT	<i>Pierre Curie, 1923</i>
We passed our days at the laboratory, often eating a simple student's lunch there... We lived in a preoccupation as complete as that of a dream.	EXTREMENESS / INTENSITY / OBSESSION / ABOVE AND BEYOND	<i>Pierre Curie, 1923</i>

Table 5. Marie Curie quotes and Genius Universals.

Brilliant people possess an internalized network of insights, heuristics, and mental models that guide their thinking and problem-solving processes. These cognitive frameworks have been honed over years of practice, reflection, and iterative learning. But when the same people attempt to convey their expertise, they often focus on procedures, formulas, and observable behaviors without examining the underlying thought processes that drive their brilliance. Brilliance is thus often not successfully transferred from person to person, like from teacher to student or mentor to mentee, and one does not sufficiently recognize their own brilliance. In the next section, we linger on this point that transfer of brilliance is difficult and rare, but point out where it does seem to happen most often and most fluently: in families.

3.2. The difficulty of Genius transfer

Having developed a few examples of brilliance, let's consider another aspect of brilliance. We want to turn an eye towards not just the brilliance of an individual, but how that brilliance is transferred to other individuals and the difficulty in doing so. To begin, let's go back to Leonardo da Vinci, who is our "model" genius. Striking about Leonardo da Vinci is that although he had his own students, these students have nowhere near the recognition that he has. Though his students were competent painters in their own right, they never seem to have inherited his unique way of looking at the world. Also consider Steve Jobs. Jobs revolutionized the computer hardware, software, film, and personal electronic industries through the companies he led and influenced. We can see his brilliance in his language. Since his death in 2011, while Apple has performed well financially, it has not yet innovated in the way it did with Jobs as its leader. The question to ask is why brilliance does not automatically transfer from one person to another, even when it would be advantageous.

But it is not the case that transfer of brilliance never happens. A crucial observation is that brilliance is often transferred in families. There are many cases of family businesses where the business is handed down from parents to children across generations. So too are there examples of creative dynasties, families where many members contribute to the same art, such as painting, sculpting, artisan crafts, music, theater, or film.

An example of brilliance transferred within a family is the Mellerio family in France. Members of the family have been excelling in goldsmithing for fifteen generations, since 1515 (Mellerio n.d.). The family places itself between art, technology, and science, and constantly gets inspiration from the Italian renaissance. The youngest Mellerio says, «since I was a child, I understood the pride that the family took in running it, generation after generation» acknowledging the mindset required to keep the business alive for over 400 years. He also remarks, «I grew up with my brothers constantly hearing conversations about jewelry: my parents discussing their ideas in front of us and with us. I always projected taking part in the writing of the next chapters of Mellerio's legacy» (Natural Diamond Council 2025). showing how the brilliance of his parents was ever present in his environment, and how it helped him adopt the mindset required to carry on his family's legacy, an embodiment of the PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY Genius Universal.

Another goldsmithing dynasty, lasting more than a hundred years, is the Buccellati family in Italy. The latest representative is Lucrezia, who designs iPhone and iPad covers bejeweled with sunburst diamonds inspired by the sketches of Leonardo da Vinci. She says, «I feel I have a very different aesthetic than my father and grandfather, but we are all linked together by our DNA, which is very identifiable in our designs» (Magnifissance 2025). Lucrezia is

saying that, although the aesthetics are different between her, her father, and her grandfather, there is a mindset about what it means to be artistic that she acquired from her father, and he from his father.

The Coppola family is another example of brilliance transfer through generations. This is a family that has been excelling in entertainment for at least three generations. Well-known among the general public is American filmmaker Francis Ford Coppola, and his children Sofia Coppola and Roman Coppola, also filmmakers. Less well-known is that actors Nicolas Cage and Jason Schwartzman are Francis Ford Coppola's nephews, through his brother and sister, respectively. Not only that, but Francis Ford's sister Talia Shire (the mother of Jason Schwartzman) is an actress, his grand-daughter Gia Coppola a film director and screen writer, his own father Carmine Coppola a composer, pianist, flautist, and songwriter, and his mother Italia a lyricist, to name a few other members of the family. The Coppola family is a clear example of where a mindset around performing and entertaining has been handed down through multiple generations.

Of course, there is transfer of brilliance out of family contexts as well. There have been many successful mentor-mentee, master-apprentice, and teacher-student relationships where the student not only gains skills and techniques from their teacher, but also a way of thinking about their work and the world. Academia is one place where this happens often. To give an example, the doctoral advisor of American linguist Noam Chomsky was Zellig Harris, who was also an influential figure in linguistics. Chomsky's students in turn went on to become important figures in linguistics, including Barbara Partee, John R. Ross, Ray Jackendoff, Joan Bresnan, Ivan Sag, and George Lakoff, to name but a few.

Since transfer can happen outside of families, what we do not want to say is that familial relationships are important for transfer. Rather, our position is that families provide a kind of environment where transfer of brilliance can happen more easily than in other kinds of environments. In a family environment, there is frequent, intense, and long-term exposure between family members (intra- and inter-generationally) in the context of their work, and this constant, continuous exposure to brilliance allows brilliance to transfer more easily than less frequent, less intense, or short-term exposure.

We see in these case studies that da Vinci and Jobs were able to extensively document and transfer their technical knowledge and expertise, preserved in the form of books, treatises, paintings, videos, and notes. They were less successful at transferring their brilliance. In contrast, families like the Melleries, the Buccellatis, and the Coppolas are able to transfer their brilliance from generation to generation. When these families describe their journeys, their language embodies themes of the Genius Universals such as PURPOSE / CONTRIBUTION / WORLD IMPACT / LEGACY. There is something unmistak-

ably different about transfer in the context of a family compared to other kinds of teacher-student dynamics, and we hypothesize that it has to do with the frequency, duration, and intensity of familial relationships in the context of brilliant achievements. In a family, children are continuously exposed to the brilliance of the previous generation (parents, but also aunts, uncles, and grandparents) and see first hand how that knowledge is used. In addition, children can immediately ask questions about why something is done, and repeat their inquiries until they are satisfied with the answer. Finally, in some cases, there is also a sense of family legacy and carrying on a tradition. These factors may be present in other teacher-student dynamics, but they appear more frequently in families and seem to be a crucial aspect of the environment that leads to brilliance transfer.

Brilliance transfer then presents a different set of challenges from the transfer of technical knowledge, which can be documented, or passed down through apprenticeships. Brilliance is not simply a collection of facts or skills. It also encompasses things that are harder to verbalize and are deeply embedded in an individual's experiences. This individual, inexpressible knowledge is the aspect of brilliance which is most difficult to transfer, and it is best understood through the theory of "tacit knowledge" (Polanyi 1966).

4. Tacit knowledge and tacit knowledge transfer

4.1. Brilliance as tacit knowledge

By laying out the Genius Universals, we can get a glimpse of the landscape of brilliance, and an individual can begin to attend to their own brilliance by noticing the Genius Universals. To provide a more complete account of brilliance, including the difficulty of transfer of brilliance, we propose that the knowledge and understanding on which brilliant thinking is founded comprises both explicit knowledge and tacit knowledge (Polanyi 1966; Collins 2010). Explicit knowledge is what we normally think of as knowledge: it is knowledge of procedures, processes, facts, formulas, instructions, and other things that we are conscious of and which can be easily articulated. In contrast, tacit knowledge is that knowledge which cannot be easily articulated; it is often personal and context-dependent, and encompasses know-how like riding a bike or recognizing a face, as well as mindsets and ways of thinking. It is often rooted in unconscious thought and experience, rather than explicit rules. Tacit knowledge is the foundation on which one can make connections within explicit knowledge, have new insights, and achieve exceptional innovations.

In distinguishing brilliance from other types of thinking, we stress that brilliance is not the same thing as tacit knowledge. Tacit knowledge and explicit

knowledge are types of knowledge. Brilliance, on the other hand, is a way of thinking that produces exceptional results. When thinking in a brilliant way, a person uses their knowledge, both explicit and tacit, to achieve their goals. Tacit knowledge is difficult to access, and in cases where one may not even be aware that they have it, the Genius Universals can help provide access to this tacit knowledge. The Genius Universals are indicative of Genius thinking.

4.2. The difficulty of tacit knowledge transfer

Tacit knowledge is acknowledged to be critical for exceptional performance (for instance, see Nelson and Winter 1982; Hadjimichael and Tsoukas 2019). As such, the concept of tacit knowledge has been extensively adopted in the field of organizational knowledge management. In the context of this field, tacit knowledge is one of the most crucial resources an organization can possess (Grant 1996; Sobol and Lei 1994), and effective management of it is a critical challenge to overcome (Park and Gabbard 2018; Faulconbridge 2007; Aoki 2010). As previously mentioned, tacit knowledge is more difficult to access and transfer than explicit knowledge, because it cannot be readily expressed directly in language. The successful transfer of tacit knowledge is essential for any organization which aims to leverage its benefits, and many methodologies have been proposed to facilitate this transfer.

The SECI model of knowledge transfer (Nonaka 1994; Nonaka and Takeuchi 1995), distinguishes four kinds of knowledge transfer, shown in Table 6, defined by the initial and final state of the knowledge. Socialization, where tacit knowledge remains tacit as it is transferred, can be accomplished through repetition, guidance, and observation. This requires extended in-person interaction, like the kind seen in families and apprenticeships. The Mellerio and Buccellati families are examples of successful tacit knowledge transfer through socialization. Their continued innovation and excellence across generations is grounded in the tight and dense social bonds of family, and extended periods of shared practice characteristic of the relationship between master and apprentice craftsmen. Externalization in contrast is a method of tacit knowledge transfer where the tacit knowledge is first expressed in language. Externalization is easier to scale within an organization than socialization, because documentation can be shared more readily than the time and attention of a single person. But tacit knowledge is difficult to externalize by nature, and development of externalization methods is an ongoing area of study, which we provide a few examples of below.

Transfer type	Knowledge source state	Knowledge recipient state
---------------	------------------------	---------------------------

Socialization	Tacit	Tacit
Externalization	Tacit	Explicit
Combination	Explicit	Explicit
Internalization	Explicit	Tacit

Table 6. Types of knowledge transfer in the SECI model (Nonaka 1994).

Mascitelli (2000) argues that the ability of an organization to achieve breakthrough innovation is dependent on the creation and sharing of tacit knowledge. To cultivate an environment within the organization which allows for individuals to develop tacit knowledge around their work, he proposes that managers engage in storytelling practices centered around narratives of innovation. These narratives enable the transfer to employees of tacit knowledge connected to the strategic vision and guiding principles of the organization. Mascitelli also says that managers should also create and reinforce unique identities for teams. The characteristics of the social identity of a team, and the degree to which members are aligned with this identity, play a significant role in the ability of the team to successfully innovate (Archer et al. 1998; Scott 1997). Project goals and requirements should also be specified only enough to adequately provide the direction of the project during early stages, in order to allow enough ambiguity to foster divergent thinking and spontaneous leaps of intuition (Leonard and Strauss 1997).

Ambrosini and Bowman (2001) propose a method for the transfer of tacit knowledge within an organization. They employ structured interviews to create a kind of cognitive map (Eden et al. 1981) to guide participants to externalize and structure their tacit knowledge. Cognitive maps represent thinking as a graph consisting of nodes which represent mental objects such as ideas and beliefs, and edges which represent the cognitive connections between those objects. They employ the causal map (Huff 1990), a class of cognitive map which places concepts in relation to one another with edges representing causal relationships. Another example is Mitchell, Harvey, and Wood (2021), who explore tacit knowledge transfer in the context of a researcher's ability to locate, quantify, and justify research impact. They assess common academic practices based on their capacity to facilitate tacit knowledge transfer, both through socialization and through externalization, and recommend that care is taken in selecting context-appropriate practices to maximize transfer. And Goffin and Koners (2011) focus on the product development cycle, where they recommend that the use of metaphor and storytelling be encouraged in discussions around these issues, as they are particularly effective at capturing and transferring this tacit knowledge.

What underlies all these methods is the necessity of documentation. Documentation of knowledge should include detailed accounts of reasoning, decision-making steps, and problem-solving strategies. In addition, documentation should describe not just successful pathways but also dead-ends and failures that contributed to learning. By capturing this depth of information, documentation provides the next generations with a more holistic understanding of brilliance. It opens a window into the minds of experts, allowing learners to internalize not just what was done, but how and why it was done. Documentation allows for that knowledge to be shared and internalized by others in a scalable way. This scalability is not possible with socialization-based methods of tacit knowledge transfer. In the next section, we sketch a process for documenting the brilliance of a person.

5. The Genius Inquiry

Interviews are a familiar process for documenting knowledge. The kind of interview that most people are familiar with is the journalistic interview, which seeks to extract a narrative about a matter of public interest, whether it be details and interpretations of current events, or stories from the life of a public figure (Roulston 2012). Many journalists and media personalities, such as Oprah Winfrey, Barbara Walters, and Larry King, have gained fame primarily due to their skill at conducting journalistic interviews (Platt 2002). Journalistic interviews, though, are not an ideal method for tacit knowledge documentation and transfer. The focus on narrative prioritizes explicit knowledge. The interview is an attempt to get at the pre-existing knowledge and thinking of the interviewee. It does not elicit new connections in their thinking which previously they had not been able to externalize.

Some types of interviews are better suited for the externalization of tacit knowledge. The motivational interview, used in psychology, asks interviewees questions in order to increase their motivation towards goals, and assist them in changing their behaviors (Miller and Rollnick 2012). The cognitive map interview asks interviewees about the choices they made which caused a certain outcome, and sequentially inquires backward about the choices which led to those choices (Eden et al. 1981).

In an effort to capture and document brilliance, we chose to use a structured interview process aimed at externalizing the tacit knowledge component of a person's brilliance. We call this process an inquiry. Every inquiry is conducted with the intent of discovering an individual's brilliance by asking how and why questions. By asking how and why questions repeatedly, the interviewer can go increasingly deeper into the thinking of the person being interviewed. The successive questioning by the interviewer encourages the interviewee to go beyond their most accessible conscious thinking and explore

tacit knowledge related to their purpose, sense of who they are as a person and what is meaningful to them, and how they relate to other people and the world. These are the fundamental aspects of their Genius.

An inquiry could be set up in many different ways. One method of setting up the inquiry is to focus on the interviewee's previous success factors, and elicit the thinking that contributed to them. This reveals the mindset. An inquiry could also be set up to start from exploring the Genius Universals. For instance, an interviewer could ask the interviewee to imagine a time when they embodied a specific Genius Universal like POSSIBILITY / CREATION / CHOICE / FREEDOM, and ask questions about what that was like: When did you embody that Genius Universal, how did it feel in the moment, and so on. What is key in an inquiry, however the interviewer starts the conversation, is that the interviewer asks questions that are intended to get increasingly deeper into the thinking of the person being interviewed. Subsequent Genius Inquiries often reveal additional facets that were unexplored. In our experience, though, many of the same themes and even the exact same phrasing in language occurs in repeated inquiries, which supports our belief that the Genius Inquiry process reveals something essential about a person's brilliance.

In effect, the inquiry process is a way of compressing the socialization process that we hypothesize is found in families. Within the family, children are in direct contact with the Genius of others. They can see that Genius in action, they receive explanations of the thought process driving the actions from their parents and grandparents, and they can immediately and continuously ask their own questions to elucidate and clarify what their parents and grandparents are thinking. The frequency, duration, and intensity of learning in a family environment amounts to a kind of continuous inquiry on the part of a child. Our inquiry process takes the most essential aspects of a child's continuous inquiry – the exploration of purpose and the motivating principles underlying what their parents do – and makes it into a process that can be conducted in a short period of time to uncover crucial facets of someone's Genius.

Brilliance can be documented in many ways. We have found success with a visual representation of the language from the interview. The interviewer selects sections from the participant's spoken language and arranges them into an inverted tree structure, similar in some ways to the causal maps employed by Ambrosini and Bowman (2001) in their method of documenting tacit knowledge⁶. The arrangement of language from the tree root to the leaves represents a continuum between purposeful and operational mindsets; language near the root of the tree is language revealing mindsets about the interviewee's understanding of the world and themselves, and the language nearer the leaves

⁶ Formally, the tree structure we use is a directed acyclic graph. Each node is language selected from the inquiry that shows the individual's brilliance, and edges between nodes create a hierarchy of thinking.

reveals thinking which drives and immediately guides action. The language in the tree is selected so that the mindset behind the language might be freely applied in other contexts. In other words, our visual representation is designed to document Genius thinking in such a way that it can be applied anywhere.

This representation is a tool for transfer of Genius. The interviewee receives a copy of the tree for their own use; they have a visual representation of their own thinking, accessible via language, and can use that language in the future to try to trigger specific mindsets that helped make them successful. The tree can also be shared with others. For instance, a supervisor might share their tree with an employee, so that the employee can understand their supervisor's mindset better, and even transfer that mindset to themselves by studying the language in the tree. If the tree is annotated so that each example of language is coded for a Genius Universal or Genius Universals, the content of the thinking can be studied. The ability to represent Genius in language makes recalling brilliance for oneself, studying brilliance, and sharing brilliance with others possible.

6. Looking ahead

The Genius Inquiry process, the Genius Universals, and the visual representation of Genius are closer to the beginning rather than the end of our journey to understand brilliance. As we discussed in Murzaku et al. (2024), we developed a novel corpus linguistic annotation process for the structured interviews that we discussed here, the Genius Inquiry. The goal of this annotation process is to allow us to study specific linguistic characteristics of the language of brilliance, including syntax, semantics, pragmatics, prosody and intonation, and body language. The research on these linguistic characteristics is still ongoing. This work will allow us to document and analyze the language of brilliance with automated tools and artificial intelligence, to massively increase access to understanding and transfer of brilliance. It will also allow us to continue to improve methods for internalization of tacit knowledge, allowing organizations to leverage the benefits of brilliance more effectively.

We have also pointed out that socialization in families is like a kind of continuous Genius Inquiry, and noted that both socialization in general and the Genius Inquiry process are limited by available time and resources. We raise the idea that AI could be used to conduct Genius Inquiries at larger scale and with greater frequency; an AI could be trained in the Genius Inquiry process so that a broader base of individuals within an organization could be interviewed, or to create an app that could do fast and regular check-ins with a person to capture what they are saying at that moment and the Genius thinking that it indicates.

7. Conclusion

In this paper, we have distinguished the notion of brilliance (which we also call Genius), a general purpose mindset that enables someone to perform at their highest potential. Brilliance is exhibited in people recognized as geniuses, but also in anyone who is achieving something exceptional whether or not they are publicly recognized for their achievements. What the nature of brilliant thinking is has remained elusive, but we propose seven themes characteristic of brilliant thinking, the Genius Universals. These themes are present in the thinking of brilliant figures throughout history, independent of their area of expertise.

Brilliance is difficult to transfer from person to person. The reason for this is that brilliance is a type of tacit knowledge, which is recognized to be difficult to externalize and transfer. The context in which brilliance transfer occurs most frequently is in the context of a family, where children are in direct contact with the brilliance of their caregivers, a type of socialization method of transfer. We have outlined a method for the elicitation and documentation of language that exhibits the Genius Universals. This allows for one to study and enhance their own brilliance, and for brilliance to be more readily shared with other people. Future directions include automating this process with AI, and studying the language of Genius using methods from linguistics. Our work allows for people to document and transfer their brilliance so that everyone can perform to their maximum potential and pass their unique way of thinking to others.

References

- Ambrosini, Véronique, and Cliff Bowman. 2001. "Tacit Knowledge: Some Suggestions for Operationalization." *Journal of Management Studies* 38 (6): 811–29. <https://doi.org/10.1111/1467-6486.00260>.
- Aoki, Masahiko. 2010. *Corporations in Evolving Diversity: Cognition, Governance, and Institutions*. Clarendon Lectures in Management Studies. Oxford; New York: Oxford University Press.
- Archer, Norman P., Fereidoun Ghazemzadeh, Penelope Jones, and Judith Jordan. 1998. "Knowledge Orientations and Team Effectiveness." *International Journal of Technology Management* 16 (1/2/3): 151–61. <https://doi.org/10.1504/IJTM.1998.002651>.
- Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. "The Berkeley FrameNet Project." *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* 1: 86–90. <https://doi.org/10.3115/980845.980860>.

- Baumeister, Roy F., and Mark R. Leary. 1995. "The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation." *Psychological Bulletin* 117 (3): 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>.
- Bianchini, Carlo, Luca Giusti, and Claudio Gnoli. 2017. "The APUPA Bell Curve: Ranganathan's Visual Pattern for Knowledge Organization." *Les cahiers du numérique* 13: 49–68. <https://doi.org/10.3166/LCN.13.1.49-68>.
- Cassirer, Ernst. 1963. *Individuum und Kosmos in der Philosophie der Renaissance [The Individual and the Cosmos in Renaissance Philosophy]*. Translated by Mario Domandi. New York: Harper & Row.
- Collins, Harry M. 2010. *Tacit and Explicit Knowledge*. Chicago; London: The University of Chicago Press.
- Curie, Eve. 1938. *Madame Curie*. New York: Doubleday.
- Curie, Marie. 1923. *Pierre Curie*. New York: MacMillan.
- Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92 (6): 1087–1101. <https://doi.org/10.1037/0022-3514.92.6.1087>.
- Dweck, Carol S. 2006. *Mindset: The New Psychology of Success*. New York: Random House.
- Eden, Colin, Sue Jones, David Sims, and Tim Smithin. 1981. "The Inter-subjectivity of Issues and Issues of Intersubjectivity." *Journal of Management Studies* 18 (1): 37–47. <https://doi.org/10.1111/j.1467-6486.1981.tb00090.x>.
- Faulconbridge, James R. 2007. "Relational Networks of Knowledge Production in Transnational Law Firms." *Geoforum* 38 (5): 925–40. <https://doi.org/10.1016/j.geoforum.2006.12.006>.
- Fredrickson, Barbara L. 1998. "What Good Are Positive Emotions?" *Review of General Psychology* 2 (3): 300–19. <https://doi.org/10.1037/1089-2680.2.3.300>.
- Fredrickson, Barbara L. 2001. "The Role of Positive Emotions in Positive Psychology: The Broaden-and-Build Theory of Positive Emotions." *American Psychologist* 56: 218–26. <https://doi.org/10.1037/0003-066x.56.3.218>.
- Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge: MIT Press.
- Goffin, Keith, and Ursula Koners. 2011. "Tacit Knowledge, Lessons Learnt, and New Product Development." *Journal of Product Innovation Management* 28 (2): 300–18. <https://doi.org/10.1111/j.1540-5885.2010.00798.x>.

- Grant, Robert M. 1996. "Toward a Knowledge-based Theory of the Firm." *Strategic Management Journal* 17 (S2): 109–22. <https://doi.org/10.1002/smj.4250171110>.
- Gubrium, Jaber F., ed. 2012. *The Sage Handbook of Interview Research: The Complexity of the Craft*. 2nd ed. Thousand Oaks: SAGE.
- Guilford, J. P. 1950. "Creativity." *American Psychologist* 5 (9): 444–54. <https://doi.org/10.1037/h0063487>.
- Guilford, J. P. 1967. "Creativity: Yesterday, Today and Tomorrow." *The Journal of Creative Behavior* 1 (1): 3–14. <https://doi.org/10.1002/j.2162-6057.1967.tb00002.x>.
- Hadjimichael, Demetris, and Haridimos Tsoukas. 2019. "Toward a Better Understanding of Tacit Knowledge in Organizations: Taking Stock and Moving Forward." *The Academy of Management Annals* 13 (2): 672–703. <https://doi.org/10.5465/annals.2017.0084>.
- Heim, Irene, and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Malden, MA: Wiley-Blackwell.
- Herre, Heinrich. 2010. "General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling." *Theory and Applications of Ontology: Computer Applications*, edited by Roberto Poli, Michael Healy, and Achilles Kameas, 297–345. Berlin: Springer.
- Herre, Heinrich. 2013. "Formal Ontology and the Foundational of Knowledge Organization." *Knowledge Organization* 40 (5): 332–39. <https://doi.org/10.5771/0943-7444-2013-5-332>.
- Hjørland, Birger. 2008. "What is Knowledge Organization (KO)?" *Knowledge Organization* 35 (2-3): 86–101.
- Huff, Anne Sigismund, ed. 1990. *Mapping Strategic Thought*. New York: Wiley.
- Kashdan, Todd B., and Michael F. Steger. 2007. "Curiosity and Pathways to Well-Being and Meaning in Life: Traits, States, and Everyday Behaviors." *Motivation and Emotion* 31 (3): 159–73. <https://doi.org/10.1007/s11031-007-9068-7>.
- Kay, Paul, and Chad K. McDaniel. 1978. "The Linguistic Significance of the Meanings of Basic Color Terms." *Language* 54 (3): 610–646. <https://doi.org/10.1353/lan.1978.0035>.
- Leonard, Dorothy A., and Susaan Straus. 1997. "Putting Your Company's Whole Brain to Work." *Harvard Business Review* 75 (4): 110–122.
- Magnifissance. Timeless Beauty and Wisdom. 2025. "Interview with Lucrezia Buccellati." <https://magnifissance.com/style/jewellery-watch/buccellati/>.

- Mascitelli, Ronald. 2000. "From Experience: Harnessing Tacit Knowledge to Achieve Breakthrough Innovation." *Journal of Product Innovation Management* 17 (3): 179–93. <https://doi.org/10.1111/1540-5885.1730179>.
- Mellerio. Depuis 1613. n.d. "Mellerio over time." Last accessed November 17, 2025. <https://www.mellerio.fr/pages/au-fil-du-temps>.
- Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. "Introduction to WordNet: An On-Line Lexical Database." *International Journal of Lexicography* 3 (4): 235–44. <https://doi.org/10.1093/ijl/3.4.235>.
- Miller, William R., and Stephen Rollnick. 2013. *Motivational Interviewing: Helping People Change*. 3rd ed. Applications of Motivational Interviewing. New York, NY: Guilford Press.
- Mitchell, Vincent-Wayne, William S. Harvey, and Geoffrey Wood. 2022. "Where Does All the 'Know How' Go? The Role of Tacit Knowledge in Research Impact." *Higher Education Research & Development* 41 (5): 1664–78. <https://doi.org/10.1080/07294360.2021.1937066>.
- Murzaku, Alexander, Pontish Yeramyan, Curt Anderson, et al. 2024. "Discovering and documenting brilliance: A novel multimodal annotation method." *AIDAinformazioni* 42 (1–2): 117–44. <https://doi.org/10.57574/596540766>.
- Natural Diamond Council. 2025. "Natural Diamond." <https://www.naturaldiamonds.com/historic-diamonds/meet-the-26-year-old-upholding-mellerios-diamond-legacy/>.
- Nelson, Richard R., and Sidney G. Winter. 2004. *An Evolutionary Theory of Economic Change*. Digitally reprinted. Cambridge, Mass.: The Belknap Press of Harvard Univ. Press.
- Nonaka, Ikujiro. 1994. "A Dynamic Theory of Organizational Knowledge Creation." *Organization Science* 5 (1): 14–37. <https://doi.org/10.1287/orsc.5.1.14>.
- Nonaka, Ikujiro, and Hirotaka Takeuchi. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. 1st ed. Oxford: Oxford University Press, Incorporated.
- Pais, Abraham. 2005. *Subtle Is the Lord: The Science and the Life of Albert Einstein*. Oxford: Oxford University Press.
- Park, Jongsoon, and Joseph L. Gabbard. 2018. "Factors That Affect Scientists' Knowledge Sharing Behavior in Health and Life Sciences Research Communities: Differences between Explicit and Implicit Knowledge." *Computers in Human Behavior* 78: 326–35. <https://doi.org/10.1016/j.chb.2017.09.017>.

- Pink, Daniel H. 2011. *Drive: The Surprising Truth About What Motivates Us*. New York, NY, US: Riverhead Books.
- Platt, Jennifer. 2001. "The History of the Interview." In *Handbook of Interview Research*, edited by Jaber Gubrium and James Holstein, 33–54. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. <https://doi.org/10.4135/9781412973588.n4>.
- Polanyi, Michael. 1966. *The Tacit Dimension*. Garden City, NY: Doubleday & Company, Inc.
- Rosch, Eleanor. 1975. "Cognitive Representations of Semantic Categories." *Journal of Experimental Psychology: General* 104 (3): 192–233. <https://doi.org/10.1037/0096-3445.104.3.192>.
- Rosch, Eleanor, and Carolyn B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7 (4): 573–605. [https://doi.org/10.1016/0010-0285\(75\)90024-9](https://doi.org/10.1016/0010-0285(75)90024-9).
- Roulston, Kathryn. 2012. "The Pedagogy of Interviewing." In *The Sage Handbook of Interview Research: The Complexity of the Craft*, edited by Jaber F. Gubrium, James A. Holstein, Amir B. Marvasti, and Karyn D. McKinney, 2nd ed. Thousand Oaks: SAGE.
- Scott, Susanne G. 1997. "Social Identification Effects in Product and Process Development Teams." *Journal of Engineering and Technology Management* 14 (2): 97–127. [https://doi.org/10.1016/S0923-4748\(97\)00007-6](https://doi.org/10.1016/S0923-4748(97)00007-6).
- Seligman, Martin E. P. 2011. *Flourish: A Visionary New Understanding of Happiness and Well-Being*. Flourish: A Visionary New Understanding of Happiness and Well-Being. New York, NY, US: Free Press.
- Sobol, Marion G., and David Lei. 1994. "Environment, Manufacturing Technology, and Embedded Knowledge." *International Journal of Human Factors in Manufacturing* 4 (2): 167–89. <https://doi.org/10.1002/hfm.4530040205>.

ISKO Italy in perspective:

An analysis of the 2017, 2019, 2021 and 2023 meetings from the ontological, epistemological and sociological dimensions of the domains

Ana Flávia Dias Zammataro*, Ana Cristina de Albuquerque*, Giulia Crippa**, Lucia Sardo**, Carlo Bianchini***, Maria Lígia Triques*

Abstract: This study examines four ISKO Italy meetings to determine which domain dimensions are most prominent in the presented works. The characterization of the domains is understood in terms of ontological, epistemological, and sociological dimensions, as outlined by Hjørland and Hartel (2003). Therefore, the goal of this article is to analyze the works presented in the ISKO Italy chapters through these dimensions. The guiding question is: how can the domains' dimensions, as presented by Hjørland and Hartel (2003), be understood in the themes discussed in the ISKO Italy chapters? The research uses qualitative, bibliographic, and documentary methods. Along with content analysis procedures, this approach allows for categorizing the works according to the domain dimensions. The results show a growing focus on themes related to the reality of researchers involved in strengthening theoretical, methodological, and applied studies in Knowledge Organization.

Keywords: Ontological dimension, Epistemological dimension, Sociological dimension, ISKO Italy.

1. Introduction

Knowledge Organization (KO), within the field of Information Science, shows that studies and research on its foundations, including its theoretical and epistemological assumptions, are reflected in the quality and scope of its processes, products, and tools. These foundations can be demonstrated through research that investigates a domain, considering, as (Hjørland and Hartel 2003) suggest, the ontological, epistemological, and sociological dimensions that define it, as well as the history of KO research itself, includ-

* Universidade Estadual de Londrina (UEL), Londrina, PR, Brasil. ana.flavia.zammataro@uel.br ORCID: 0000-0001-6520-3229; albuana@uel.br ORCID: 0000-0003-3506-0479; mlgia.triques@uel.br ORCID: 0000-0002-0450-7138.

** University of Bologna (Unibo), Department of Cultural Heritage, Ravenna, Italy. giulia.crippa2@unibo.it ORCID: 0000-0002-6711-3144; lucia.sardo@unibo.it ORCID: 0000-0001-6480-759X.

*** University of Pavia (Unipv), Department of Musicology and Cultural Heritage, Cremona, Italy. carlo.bianchini@unipv.it ORCID: 0000-0002-6635-6371.

ing the founding of the International Society for Knowledge Organization (ISKO) in Frankfurt, Germany, in 1989 (Dahlberg 1995).

The founding of ISKO was essential for establishing KO as a scientific discipline, granting worldwide institutional recognition to discussions of its investigative aspects. It is therefore recognized as: «a society in the field of ‘Organization and Representation of Knowledge’ responsible for the main actions surrounding its necessary consolidation and scientific construction» (Alves and Oliveira 2016, 104).

This article focuses on the institutionalization of KO’s scientific foundations through the creation of ISKO, especially ISKO Italy. It presents a guiding question: how can the domains’ dimensions offered by Hjørland and Hartel (2003) be seen in the themes discussed in the chapters of ISKO Italy?

The objective was to analyze the works presented in the chapters of ISKO Italy, in meetings held in 2017, 2019, 2021 and 2023, in light of the ontological, epistemological and sociological dimensions of the domains, elaborated by Hjørland and Hartel (2003). This time frame was chosen to address more recent meetings, seeking in the works presented updated discussions within the scope of Knowledge Organization.

According to Hjørland and Hartel (2003, 239), domains can be described by three dimensions, or «three types of theories or concepts» ontological, epistemological, and sociological. The ontological dimension involves describing elements of reality as they are organized, including explaining areas, fields, and objects, as well as the relationships between these domains, instruments, and systems. This organization, such as thesauri and classification systems, relates to a more applied aspect of KO studies, aiming to contribute to the development of these tools by examining terms, concepts, and areas of study.

The epistemological dimension pertains to the “study of knowledge” (Hjørland and Hartel 2003, 240), encompassing paradigms, traditions, and theoretical schools that can support Knowledge Organization. The sociological dimension, which is key to domain analysis, involves how society organizes knowledge through disciplines, professions, discursive and epistemic communities, and hobbies (Hjørland and Hartel 2003), as well as the relationship between users and KOS.

Thus, this research is based on Bufrem and Cervantes’s (2021) justification for theoretical studies that reflect on their own production. The authors explain that the development of a science manifests through scientific output; that is, via publications and the dissemination of scientific knowledge in various spaces, the scientific community gains access to knowledge that develops collectively and legitimizes itself. Observing one’s own field and its publications enables analysis and evaluation of the content produced by scientists, as well as trends, methods, and theoretical influences.

In this way, it is also justified by the possibility of understanding the fundamental aspects of Knowledge Organization and, more importantly, of reflecting on the ontological, epistemological and sociological dimensions that define a domain, as proposed by Hjørland and Hartel (2003).

The methodological procedures involve qualitative bibliographic research. Data collection and identification of the works were conducted on the ISKO Italy website, which provided the following information: the 8th meeting took place in 2017 in Bologna with 10 works; the 9th meeting was held in 2019 in Florence with 9 works; the 10th meeting occurred in 2021 in Cremona with 7 works; and the 11th meeting was held in 2023 in Rome with 17 works.

A total of 43 works were collected, which, through Content Analysis based on Bardin (1977), allowed for the identification and categorization of themes focused on the ontological, epistemological and sociological dimensions proposed by Hjørland and Hartel (2003).

Therefore, the author's inferences were based on the characteristics of these domains and contents from the works presented at these meetings. The features and key points of the analysis indicate a trend toward consolidating themes centered on the context and reality of researchers and professionals involved in strengthening theoretical, methodological, and applied studies in Knowledge Organization.

2. The Knowledge Organization: Foundations and Research Trajectory

Knowledge Organization (KO) has reshaped the visibility of the field of Information Science through studies and research that draw on discussions surrounding its foundations, theoretical, and epistemological assumptions, which, in turn, reflect the quality and scope of its processes, products, and instruments. These foundations can be highlighted through the study of the field's historical aspects, which include its relationships with the history of books and writing—that is, with Library Science and Documentation (Pinho 2009), and through its own research trajectory, including the creation of the ISKO, as mentioned, in 1989 (Dahlberg 1995).

When discussing Knowledge Organization, Hjørland (2008) highlights the importance of his studies being based on assumptions that contribute to the organization of documents, works and concepts, given that it is an area that deals with the analysis of the nature and quality of processes and systems aimed at the organization of knowledge.

The author also defines two approaches to the organization of knowledge, one more focused on the intellectual organization of knowledge, and therefore

more restricted, and the other, more comprehensive, called the social organization of knowledge:

As an example, Hjørland (2003) mentions the use of the periodic table or the zoological taxonomy of Biology, whose intellectually organized knowledge constitutes the basis for indexing systems. The other approach [...] refers to the organization of professions and disciplines, dividing knowledge according to the area of activity and research. (Zammataro 2023, 33).

This means that the intellectual knowledge organization involves scientific theories and concepts, while the social knowledge organization uses these theories and concepts to establish processes, tools, and systems applicable to everyday life (Pinho 2009). In other words, the social knowledge organization is based on sociological models that explain the divisions of professions and, at the same time, the different approaches of scientific disciplines, which directly impacts the quality of knowledge organization processes and systems (Hjørland 2003).

The development of instruments, processes and products based on theoretical and epistemological bases contributed to the recognition of Knowledge Organization as a scientific discipline, as well as to its international institutionalization through the creation of ISKO, with Ingetrout Dahlberg as one of its main founders (Pinho 2009) of this historical trajectory.

2.1. ISKO's historical trajectory and the Italian chapter

ISKO's historical trajectory is directly related to the foundations of Knowledge Organization. Therefore, bringing ISKO's history into the discussion also means highlighting the history of Knowledge Organization. Martins and Moraes (2015) reinforce the relationship between Knowledge Organization and Library Science and Documentation, explaining, based on Pinho (2009), that American librarian Henry Evelyn Bliss was the first to formally use the term "knowledge organization" in works published in 1929 and 1933, respectively, *The organization of knowledge and the system of sciences* and *The Knowledge Organization in Libraries*. Bliss's publications gained notoriety and contributed to the widespread use of the term "knowledge organization" in information studies (Martins and Moraes 2015). His work even influenced the name given to the institution that would later dedicate itself to research related to knowledge organization: the International Society for Knowledge Organization (ISKO).

With Ingetrout Dahlberg as one of the main creators of the society, ISKO contributed to giving KO a scientific and disciplinary character and expanding the studies that until then were carried out within the scope of the Society for Classification, also founded by Dahlberg, in 1977:

The International Society for Knowledge Organization (ISKO) was founded

in 1989 by Ingetraut Dahlberg with 9 colleagues, splitting from the existing Gesellschaft für Klassifikation. It is devoted to theory and practice of organizing knowledge, especially by intellectually-developed knowledge organization systems such as classification systems and thesauri. Research themes include general classification, interoperability, domain analysis, facet analysis, epistemology and information retrieval. (Szostak and Ohly 2020)

Thus, ISKO came to encompass phenomena that went beyond mathematical or statistical approaches, by bringing philosophical and conceptual perspectives to the debates, fostering research on conceptual, methodological and applied aspects of the organization of knowledge, and fundamental for the scientific recognition of the area, thus strengthening relations between researchers around the world:

«The establishment of ISKO and the scientific investment of researchers in the theories of this domain of study in the field of Information Science therefore contributed to the advent of a research area and, consequently, of a discipline that promotes the research, development and applications of conceptual knowledge organization systems, as well as the study of philosophical and semantic aspects related to conceptual organization and the dynamics of knowledge» (Santos, Neves and Souza 2019, 101).

On its official website, ISKO explains itself as an «international academic society dedicated to the theory and practice of knowledge organization» (ISKO n.d.). The website also promotes one of its main scientific communication channels, the journal *Knowledge Organization*, with eight regular publications per year.

Within the scope of its activities, there is also collaboration and the promotion of partnerships between the institution and several organizations, such as the United Nations Educational, Scientific and Cultural Organization (UNESCO), the International Federation of Library Associations and Institutions (IFLA), the American Society for Information Science and Technology (ASIS&T), in addition to dialogue with the Classification Research Group (Silva 2017). The website's homepage also provides information about ISKO chapters as an important means of giving the society global reach.

The chapters play an important role in the visibility of ISKO among the scientific community and currently number 15, namely: Brazil; Canada/United States; China; France; Germanic (comprising Germany, Austria and Switzerland); India; Iran; Italy; Netherlands (comprising Belgium, the Netherlands and Luxembourg); Maghreb (comprising Algeria, Morocco and Tunisia); Poland; Singapore; Iberian (comprising Portugal and Spain); United Kingdom; Africa (ISKO n.d.).

Within this framework, there is the ISKO Italy chapter, active since 2004, dedicated to developing research projects and organizing meetings and study days among ISKO members and between its members and individuals from

other institutions. Furthermore, it seeks to strengthen contacts between researchers wishing to develop theoretical or applied studies in Knowledge Organization (ISKO n.d.). Its objectives include collaborating: «with professional associations, universities, and research institutions interested in knowledge organization, including AIDA: Italian Association for Advanced Documentation and GRIS: AIB Subject Indexing Research Group» (ISKO n.d.).

There is also a discussion group that ISKO members can join, the purpose of which is to «strengthen the network of contacts among those interested in these topics» (ISKO n.d.). The group disseminates the results of conferences and other events, both Italian and foreign, related to Knowledge Organization. The meetings held by ISKO Italy to date are listed in Table 1.

Meeting place	Year
University of Milan - Bicocca, Milan	2005
University of Milan - Bicocca, Milan	2006
University of Milan - Bicocca, Milan	2007
University of Turin, Turin	2009
Ca' Foscari University of Venice, Venice	2011
National Central Library of Florence, Florence	2013
University of Bologna, Department of Mathematics, Bologna	2015
University of Bologna, Department of Mathematics, Bologna	2017
National Central Library of Florence, Florence	2019
Department of Music and Cultural Heritage of the University of Pavia, Cremona campus (<i>online event</i>)	2021
Look beyond conference, Sapienza University of Rome	2023
Department of Culture, Education and Society, University of Calabria, Rende, Cosenza	2025

Table 1. Meetings promoted by ISKO Italy. Source: Elaborated by the authors based on (ISKO n.d.)

ISKO Italy chapter has big importance in the dissemination of research developed within the scope of the Knowledge Organization, both national as internationally, contributing to broaden discussions in the area.

3. The domain and its dimensions

The studies of the dimensions that can characterize a domain, proposed by Hjørland and Hartel (2003), were encouraged by the existence of the pragmatic approach called Domain Analysis (DA), presented for the first time in the field of Information Science by Hjørland and Albreschtsen (1995), with

the aim of better understanding the phenomena related to information in its respective domain and in its relationship with discursive communities.

DA brings a socio-epistemological perspective to Information Science and Knowledge Organization research, emerging as an alternative approach to the physical and cognitive paradigms that prevailed until the late 1980s. The approach contributed to the emergence of the relationship between discourse communities and their distinct domains, standing out as the only one to consider epistemological issues in the field of Knowledge Organization (Hjørland 2008). Thus, research involving domains within the scope of Knowledge Organization, according to Hjørland and Albreschtsen (1995), is considered «the best way to understand information, using discourse communities and their role in society» (Freitas and Albuquerque 2017, 2).

Barité (2015) explains that DA aims to represent knowledge by defining the conceptual structure of a field. Its application requires identifying domain-specific problems through quantitative, qualitative, or both surveys. The author adds that these methodologies serve both to build and to review and evaluate knowledge organization systems and structures.

Hjørland and Albreschtsen (1995) did not present an exact definition of the concept of domain; however, the literature provides some ways to understand it. According to Mai (2005, 605), a domain can be «an area of expertise, a corpus of literature, or a group of people working together in an organization». Amorim and Café (2017, 82) explain that a domain: «Can be delimited by a group of users, a discipline, a company, or a broad field of knowledge, endowed with informational needs and constituted by paradigms, traditions, and schools that define their theories». Hjørland (2017) emphasizes that domains establish relationships with the purpose of the research and with the researcher themselves, being modified with each new point of view, given their mutable nature.

Aiming at a better understanding of the domains under study and their implications for KO, Hjørland and Hartel (2003) proposed a characterization based on three dimensions: ontological, epistemological and sociological.

The authors state that the ontological dimension refers to elements of reality, seeking to explain how it is structured by defining terms involving areas, fields of knowledge, objects of study, problems, and domains. Thus, ontology is closely related to metaphysics, raising questions about the nature of the objects studied.

As one of the orientations involving the phenomenon of classification in contemporary times, Pombo (1998) highlights, based on Diemer (1974), the ontological orientation, also called by the author as *classification of beings*. The ontological orientation, according to Pombo (1998), refers to the way in which beings and phenomena are classified in the sciences, referring from the studies developed by Aristotle, to more recent research, whose domains de-

pend directly on the classification of beings, of human or natural phenomena, «such as biology, geology, cosmology, anthropology or psychological typology» (Pombo 1998, 2).

Aristotle, for example, in his work *Organon*, established 10 categories to represent elements of reality, seeking to guide the division made by the human mind. According to Burke (2003): «Aristotle had exposed a system of 10 general categories (substance, quantity, quality, relation, place, time, position, condition, action and passion)».

Aristotle's categorization greatly influenced studies developed over the centuries, reflecting how human and natural elements and phenomena could be organized and structured. Hjørland and Hartel (2003) highlight, in this sense, the theory of chemical elements as a major advance for humanity, since it used ontology to model chemical knowledge in a structured manner, listing the elements, their properties, and reactions. Thus, according to the authors, when ontological issues are addressed within the domains, questions such as, «areas, domains, elements, fields, types, objects, problems, subjects [...], topics» can be addressed (Hjørland and Hartel 2003, 239-240). These studies, by contributing to the understanding of how reality is structured, enable the development of instruments, processes and organizational systems, such as classification systems, thesauri and taxonomies, thus relating to a more applied sense of KO studies.

Because it involves the elaboration of theories and the study of paradigms, philosophies and schools of thought, the epistemological dimension, in the conception of Hjørland and Hartel (2003), is considered of extreme importance in Information Science, referring to: «the study of knowledge and the way in which knowledge is obtained, for example, the roles of observation, theoretical analysis, languages, traditions [...] and values in the production of knowledge» (Hjørland and Hartel 2003, 240).

Epistemological studies involve a reflection on knowledge, focusing on questions about how we acquire it, what the limits of human knowledge are, and how we construct it, including supporting scientific arguments and theories.

Considering the traditional approach, epistemology is the study of knowledge or gnosiology. On the other hand, there is a specific approach that presents epistemology as the critical study of the principles, hypotheses, and knowledge production of various sciences, looking more closely at the cognitive structure of scientific knowledge, values, and objectives. More than that, epistemology is concerned with the characteristics of scientific knowledge, its delimitations, and methodological processes in each domain (Araújo 2012 *apud* Araújo 2024, 141).

The epistemological dimension enunciates fundamental questions in the Knowledge Organization, among which are: «what are the necessary and suf-

ficient conditions of knowledge? What are its resources? What is its structure? And what are its limits? How do we understand the concept of justification? What makes beliefs justified, justified?» (Araújo 2024, 145). Thus, when considered in the context of Knowledge Organization, epistemological questions can underpin their theories and methods, referring to how to represent, organize and retrieve information (Hjørland 2002), concerning «approaches, metatheories, movements, paradigms, philosophies [...], regimes [...], schools [...], systems [...], traditions [...], trends» (Hjørland and Hartel 2003, 242). These questions contribute to elucidating problems related to the domain in question, helping to theoretically and methodologically support studies of KO for the development of Knowledge Organization System (KOS), aiming to meet the domains and their specificities.

Hjørland and Hartel (2003) describe four epistemological schools that can be appropriated by the epistemological dimension of the domains: empiricist, rationalist, historicist and pragmatist, whose meanings are expressed in Table 2.

Empiricist	Rationalist	Historicist	Pragmatist
Characterized by what is observable and can be experienced through experience. It therefore values observation and experimentation.	It is characterized by the idea that true knowledge must derive from reason and not from sensory experience. Universal truths, in this sense, can be reached through logical and deductive thinking.	Knowledge is conditioned by the historical context, valuing the understanding of contextualized knowledge and rejecting absolute and universal truths.	In pragmatism, knowledge is valid when applicable to reality, producing tangible and concrete effects. In this school of thought, knowledge is dynamic and constantly changing.

Table 2. Epistemological schools that can characterize the domain statements. Elaborated by the authors based on Hjørland and Hartel (2003).

These schools of thought, as highlighted by the authors, can elucidate some epistemological positions that can be adopted in the field of Information Science, influencing the development of instruments, processes and products of Knowledge Organization.

The sociological dimension, more comprehensive, refers to the way in which knowledge is organized, represented, acquired, sought and used by society and therefore involves concepts such as specialties, discursive and epistemic communities, disciplines, and even characteristic segments of everyday social life, such as hobbies (Hjørland and Hartel 2003).

For Hjørland and Albrechtsen (1995), the sociological dimension is central to Domain Analysis, since knowledge domains and discursive communities

are part of the social division of labor. Thus, it is assumed that knowledge is constructed collectively, within social, professional, or academic communities, and that domains can be understood from organized social contexts, with their own values, objectives, and shared practices that will shape the production, organization, and use of information.

Hjørland and Hartel (2003) highlight the importance of social interactions in the epistemological construction of domains and, to this extent, the interrelationship of these three dimensions: «We can say that an ontological principle constitutes an intellectual way of organizing knowledge, while a sociological principle constitutes a social way of organizing knowledge» (Hjørland and Hartel 2003, 243). Thus: «the study of a given domain necessarily involves the use of these three dimensions so that it can be thought and organized, and, in the scope of Information Science and Knowledge Organization, the combination of these three dimensions is fundamental» (Zammataro 2023, 127).

The complex interaction between the three dimensions of the domains, characterized by Hjørland and Hartel (2003) in the studies of Information Science and Knowledge Organization, is reinforced.

4. Analysis and discussion of results: the dimensions of the domains in the ISKO Italy themes

According to the methodology, the presentations were read in full, allowing for an analysis that highlights the main points, with emphasis mainly on the topics covered.

The analysis of the 43 papers presented at the meetings studied falls within the pre-analysis stage of content analysis (Bardin 1977), followed by the development of hypotheses based on this stage. Subsequently, exploration was carried out by identifying attributes that allowed us to identify the dimensions of the domains that prevail in these papers for subsequent categorization and inference.

Finally, the results were processed, which allowed the following categories to be determined: ontological dimension; epistemological dimension and sociological dimension and their respective inferences within the scope of the meetings.

It is noteworthy that, due to the relational nature of the proposed dimensions, some presentations were identified in more than one domain.

4.1. Regarding the ontological dimension

Table 3 presents an overview of the themes, followed by their analyses categorized in the ontological dimension.

MEETINGS	TITLE AND AUTHORS	ANALYSIS OF PRESENTATIONS
8th ISKO Italy, 2017, Bologna	Knowledge Organization for Digital Infrastructures and Application of Dahlberg's ICC in a LOD Environment, by Ernesto William De Luca	The presentation demonstrated how the <i>Information system Coding Classification</i> (ICC), created by Ingtraut Dahlberg, is applied in a Linked Open Data (LOD) environment at the Georg Eckert Institute (GEI), an educational media institute located in Leibniz, Germany. In his talk, De Luca demonstrates the challenges of ensuring that book searches, for example, are effective and meet informational needs.
8th ISKO Italy, 2017, Bologna	Creating CC Numbers for Literary Works from LOD, by Carlo Bianchini	The work demonstrated the functioning of <i>Colon Classification</i> (CC) from the creation of class numbers in a faceted classification scheme and how it relates to Linked Open Data (LOD) when indexing, by subject, in a faceted classification scheme.
8th ISKO Italy, 2017, Bologna	APUPA: Twilight in the Sources of <i>Marketing</i> and Knowledge Organization, by Luca Giusti	The presentation dealt with the APUPA model, created by Ranganathan, presenting a trajectory of the studies carried out by the Indian and how APUPA works.
8th ISKO Italy, 2017, Bologna	Representing Subjects in the IFLA LRM Model, by Pino Buizza	The representation of subjects was addressed based on the standard proposed by the Library Reference Model (LRM) of IFLA (International Federation of Library Associations) and Institutions).
8th ISKO Italy, 2017, Bologna	Representing CDD classes with computationally cheap codes, by Mauro Bertani	The work proposed the use of a classification system that employs knowledge of the faceted classification and the Dewey Decimal Classification, using a Dewey number as a single facet.
8th ISKO Italy, 2017, Bologna	Developing SciGator Facets, by Claudio Gnoli	SciGator research interface (http://scigator.unipv.it/) works, demonstrating how it is based on the CDD, and that a new version called <i>SciGator Facets</i> was developed based on <i>Colon Classification</i> (http://scigator.unipv.it/cc/).
9th ISKO Italy, 2019, Florence	Classification and <i>linked data</i> to the library: the case of the CDU, by Aida Slavic	The speaker addressed the UDC and the use of linked data within libraries, showing how they work, and explained how <i>linked data</i> allows access to information based on identification with other library catalogs or other records and how it makes it possible to link data within a bibliographic description such as author's name, title and publisher, for example, in addition to enabling free access to each of these described data, relating them semantically to resources on the <i>Web</i> .
9th ISKO Italy, 2019, Florence	Knowledge Organization for Digital Infrastructures and Application of Dahlberg's ICC in a LOD Environment, by Ernesto William De Luca	The presentation demonstrated how the <i>Information system Coding Classification</i> (ICC), created by Ingtraut Dahlberg, is applied in a Linked Open Data (LOD) environment at the Georg Eckert Institute (GEI), an educational media institute located in Leibniz, Germany. In his talk, De Luca demonstrates the challenges of ensuring that book searches, for example, are effective and meet informational needs.
9th ISKO Italy, 2019, Florence	Creating CC Numbers for Literary Works from LOD, by Carlo Bianchini	The work demonstrated the functioning of <i>Colon Classification</i> (CC) from the creation of class numbers in a faceted classification scheme and how it relates to Linked Open Data (LOD) when indexing, by subject, in a faceted classification scheme.
9th ISKO Italy, 2019, Florence	APUPA: Twilight in the Sources of <i>Marketing</i> and Knowledge Organization, by Luca Giusti	The presentation dealt with the APUPA model, created by Ranganathan, presenting a trajectory of the studies carried out by the Indian and how APUPA works.

9th ISKO Italy, 2019, Florence	Representing Subjects in the IFLA LRM Model, by Pino Buizza	The representation of subjects was addressed based on the standard proposed by the Library Reference Model (LRM) of IFLA (International Federation of Library Associations) and Institutions).
9th ISKO Italy, 2019, Florence	Representing CDD classes with computationally cheap codes, by Mauro Bertani	The work proposed the use of a classification system that employs knowledge of the faceted classification and the Dewey Decimal Classification, using a Dewey number as a single facet.
9th ISKO Italy, 2019, Florence	Developing SciGator Facets, by Claudio Gnoli	SciGator research interface (http://scigator.unipv.it/) works, demonstrating how it is based on the CDD, and that a new version called <i>SciGator Facets</i> was developed based on <i>Colon Classification</i> (http://scigator.unipv.it/cc/) .
10th ISKO Italy, 2021, Cremona (online)	Organizing Knowledge about Allergies, by Marcin Trzmielewski	Trzmielewski presented part of his doctoral research, seeking to demonstrate what the structure of a knowledge organization system (KOS) would be like based on the needs of allergists, in order to monitor and improve their work practices.
10th ISKO Italy, 2021, Cremona (online)	Updating the SciGator interface, by Giorgia Lottici	SciGator search system was discussed, explaining how the system has been updated over the years with the aim of improving searches in the University of Pavia library by users.
10th ISKO Italy, 2021, Cremona (online)	Automatic Classification with Linked Open Data, by Carlo Bianchini	It dealt with the operation of a tool for the automated classification of authors and literary works through a multifaceted classification and using Linked Open Data.
11th ISKO Italy, 2023, Rome	Indexing of Non-Textual Materials, by Athena Salaba	It addressed the indexing of non-textual materials, such as images, videos, and music, and the challenges of indexing these types of records, addressing the complexity of indexing and its inability to capture the full meaning of a work, concluding that it is possible to seek more precise and ethical ways of representing these contents to facilitate access and understanding by users.
11th ISKO Italy, 2023, Rome	Product Image Content Representation Model in Electronic Environment (MRCIP), by Gislene Rodrigues & Célia Dias	The authors presented a proposed conceptual model that can help brands and small entrepreneurs effectively represent product images used in e-commerce environments, favoring digital marketing and consumer purchasing decisions.
11th ISKO Italy, 2023, Rome	Broni's Curtain: Organizing Knowledge Beyond Books, by Claudio Gnoli	The author discusses knowledge organization (KO) theoretically, highlighting its potential for indexing diverse objects beyond books, such as works of art and cultural elements that convey information. He analyzes the curtain of the 19th-century Broni Theater, proposing the use of phenomenon-based classifications, better suited to the interdisciplinary and multimedia nature of these objects.
11th ISKO Italy, 2023, Rome	Michela Grossi, Roberta Schiavone, Beyond the Score: Indexing Noted Music for Performance	The presentation investigated the informative potential of semantic indexing of musical resources focused on performance, understood as execution and access to music, highlighting the Nuovo Soggettario as an appropriate tool and suggesting the enrichment of the Thesaurus with musical terms based on collaboration between librarians and specialists.
11th ISKO Italy, 2023, Rome	Strings and their semantics: the treatment of subjects in entity modeling processes, by Tiziana Possemato	The author analyzed the transition from traditional cataloging to entity modeling, highlighting the need for a conceptual shift in information processing and management. The paper also discusses the challenges of using owl:sameAs to connect identities across different databases and proposes, as a possible solution, the construction of "semantic chains" to capture the meaning of terms across heterogeneous sources.

11th ISKO Italy, 2023, Rome	Subject indexing in the ICCU Alphabetica portal, by Valentina Atturo	The study analyzed subject indexing on the Alphabetica portal, highlighting the use of a concept map that presents a list of terms related to the keyword search; a diagram centered on a term (initially the first in the list); and a detailed list with the number of results and the digital consistency of the selected term. The terms in the diagram, extracted from the BNCF thesaurus, include the central term, related terms, and specific terms, whose relevance is indicated by the number of associated records.
11th ISKO Italy, 2023, Rome	Automatic subject indexing: testing the Annif software for the Italian language, by Lorenzo Gobbo	The paper presented the results of training Annif, the National Library of Finland's automatic indexing tool, adapted for Italian. The author explained that to improve its performance, it is necessary to create a complete version of the SKOS vocabulary in Italian, and that the widespread adoption of automated systems and an initial collaborative effort would favor interoperability across different contexts and frameworks.
11th ISKO Italy, 2023, Rome	'Cataloging' experts by competencies: the Digital University project, by Vincenzo Maltese	The author presented the experience of the University of Trento in the Digital University project, which resulted in a system to index and locate experts by their skills, using a controlled vocabulary with around 3,000 concepts in Italian and English and managed according to Ranganathan's analytical-synthetic approach.
11th ISKO Italy, 2023, Rome	Bibliography: Open Data and BIBFRAME, by Andrea Marchitelli (EBSCO)	The work focused on BiblioGraph, an EBSCO linked data technology that transforms library catalogs into web-accessible resources. Based on BIBFRAME, it connects people, topics, works, and publishers, facilitating content discovery. The tool also offers automatic reports that monitor catalog usage via the web.
11th ISKO Italy, 2023, Rome	Author tagging: potential and limits of keywords in professional scientific production, by Alessandra Moi, Andrea Marchitelli and Carlo Bianchini	The study analyzed changes in subject indexing in libraries with the advent of digital technologies and the web, highlighting the replacement of controlled systems with free practices, such as author-specific keyword selection. The focus is on the impacts of this freedom on information retrieval, especially in scientific publications, and the challenges it poses to the effectiveness of this process.
11th ISKO Italy, 2023, Rome	Subject indexing of photographic images: proposals and models, by Margherita Naim	The presentation addressed the subject-specific indexing of photographic images, highlighting that, despite its importance, this topic is little explored in the specialized literature. The central idea is that indexing photographs requires specific criteria, different from those used for texts, and must follow controlled standards, such as vocabularies and thesauruses.
11th ISKO Italy, 2023, Rome	Between Textuality and Iconology: Insights on Subject Indexing in Maps, by Anna Lucarelli and Elisabetta Viti	The presentation addressed map indexing in libraries, highlighting the increased interest in these materials with digitization and the need for standardization in information description and retrieval. The authors explain that thematic indexing requires attention to aspects such as visual emphasis, different media, and toponyms. To ensure consistency, the use of controlled vocabularies, such as Nuovo Soggettario (New Soggettario), is essential.

11th ISKO Italy, 2023, Rome	Navigating an ocean of content: subject indexing of sound and audiovisual resources in ICBSA, by Antonella Fischetti	The paper addressed the transformations and challenges in cataloging and semantically describing sound and audiovisual materials, especially in the context of technological advancement and digitalization. The author highlighted that, in the digital context, semantic indexing has become an indispensable tool for quality access to sound and audiovisual materials, challenging professionals to develop increasingly precise and technically advanced descriptions.
11th ISKO Italy, 2023, Rome	Indexing of subjects of normative acts, by Ignazio Pirronitto	The presentation addressed the importance of subject-specific indexing of normative acts, such as laws and resolutions, emphasizing that, despite not being books, these documents should be included in knowledge organization systems to ensure their effective retrieval. The Nuovo Soggettario was discussed, along with the ISO 5963:1985 standard, to ensure standardized descriptive methods, thus avoiding ambiguities.
11th ISKO Italy, 2023, Rome	From the local thesaurus to the new subject list: subject indexing of photographs and postcards at the Gambalunga Civic Library in Rimini, by Mara Sorrentino	The presentation addressed the Gambalunga Civic Library's work in modernizing its photography collection, which boasts over 1.5 million images, highlighting the adoption of standards such as Nuovo Soggettario to improve the organization, access, and interoperability of its content. In short, the Library combines preservation and standardization to expand the use and sharing of its collection in national and international contexts.
11th ISKO Italy, 2023, Rome	Search and navigation features of the new Biblioteche Oggi Trends platform, by the Biblioteche Oggi Trends editorial team	It dealt with the periodical "Biblioteche Oggi Trends", which since 2021, has adopted a hybrid format, combining the traditional printed version with an improved digital edition, with the aim of providing a more interactive reading, with multimedia content and hyperlinks, in addition to expanding the freedom of authors in the use of unconventional sources.

Table 3. Categorization of ISKO Italy's themes in the ontological dimension. Elaborated by the authors based on survey data (2025).

The ontological dimension considers the nature of the objects of study within a knowledge domain. Presentations in this category demonstrate themes focused on the identification, definition, and development of processes, instruments, and products central to the development of a domain. The themes cover thesauri, classification systems, indexing, the use of platforms, and how concepts relate to each other in this universe. They consider different knowledge registers, such as music, videos, and photographs, and refer to the role of users, grounded in ontologies specific to the knowledge domain.

Thus, ontological commitment refers to the principle that scientific knowledge requires rigor and constant investigation of the very objects it studies. This dimension demonstrates that research objects play a decisive role in defining and delimiting a domain and that, if well observed and processed, they have repercussions on the conceptual configuration of these domains. Therefore, a rigid, pre-established ontology cannot be conceived for all domains since this ontology is constructed historically and dynamically (Hjørland 2004).

To visualize the themes, a word cloud was created with the words that emerged from the analysis (Fig. 1), enabling a co-understanding of the orientation of the ontological dimension.



Figure 1. Word cloud from analysis in the ontological dimension. Prepared by the authors based on research data (2025).

It is inferred that the ontological dimension enables the understanding that relates the delimitation of areas of knowledge, with categories, types of objects, research themes, topics and other elements that structure the field of knowledge in question (Hjørland and Hartel 2003), thus allowing the analysis of these elements and reflection on the domain itself and consequently, its ontological dimension.

4.2. Regarding the epistemological dimension

Table 4 presents an overview of the themes, followed by their analyses categorized in the epistemological dimension.

MEETINGS	TITLE AND AUTHORS	ANALYSIS OF PRESENTATIONS
8th ISKO Italy, 2017, Bologna	Mathematics, Cybernetics and Ranganathan's APUPA model, by Luca Giusti	Ranganathan 's academic career, including the research that led to the development of the APUPA model. It addressed the meaning of the acronym from the perspective of an information retrieval system and the trajectory of Ranganthan 's studies from a mathematical and philosophical perspective, demonstrating the relationship between his thinking and that of previous scholars.
8th ISKO Italy, 2017, Bologna	Considerations introductory on the return of FAIR Data Management, by Paola Rescigno	The lecture addressed how <i>FAIR Data Management</i> , a system that manages research data, has occupied more space in commercial contexts than in the field of Library Science and Information Science.

9th ISKO Italy, 2019, Florence	Remembering the work of Ingtraut Dahlberg, by H.-Peter Ohly	Work dedicated to the life and work of Dahlberg , highlighting his contributions to the development of classification systems.
9th ISKO Italy, 2019, Florence	APUPA: Twilight in the Sources of <i>Marketing</i> and Knowledge Organization, by Luca Giusti	The presentation dealt with the APUPA model, created by Ranganathan, presenting a trajectory of the studies carried out by the Indian and how APUPA works.
9th ISKO Italy, 2019, Florence	Developing ILC Classes for Mathematical Entities, by Daniele Pascal Morelli	It dealt with the relationship between the field of mathematical knowledge and the Organization of Knowledge, seeking to demonstrate how Mathematics can use theories and methods of the Knowledge Organization and, at the same time, how KO also uses Mathematics to develop its instruments and systems of knowledge organization, notably regarding the ILC system (<i>Integrative Levels Classification</i>).
10th ISKO Italy, 2021, Cremona (online)	Information Pyramids and Knowledge Organization, by Riccardo Ridi	The lecture discussed the concept of Knowledge Organization relating it to the perspective of the information pyramid DIK (data, information, knowledge) and DIKAS (data, information, knowledge, consciousness, self-knowledge).
10th ISKO Italy, 2021, Cremona (online)	Ahistorical-bibliographic paradigm for Library Science and Information Science, by Maria Teresa Biagetti	The author discussed her initial studies to define a historical-bibliographic paradigm for Information Science and Knowledge Organization, using as a basis an analysis of the European bibliographic heritage.
10th ISKO Italy, 2021, Cremona (online)	Explaining Ranganathan with Lattice Theory, by Mauro Bertani	The author presented a proposal for using Ranganathan's theory based on lattice theory, which is based on the idea that the same data can be perceived from different perspectives and that acts of knowledge determine the transition from one knowledge structure to another.
11th ISKO Italy, 2023, Rome	Broni's Curtain: Organizing Knowledge Beyond Books, by Claudio Gnoli	The author discusses knowledge organization (KO) theoretically, highlighting its potential for indexing diverse objects beyond books, such as works of art and cultural elements that convey information. He analyzes the curtain of the 19th-century Broni Theater, proposing the use of phenomenon-based classifications, better suited to the interdisciplinary and multimedia nature of these objects.

Table 4. Categorization of ISKO Italy's themes in the epistemological dimension. Elaborated by the authors based on survey data (2025).

The epistemological dimension considers the modes of production, theoretical validation, and how knowledge is structured within a domain. Thus, it discusses the theories, schools of thought, methods, paradigms, and models used to reflect on knowledge. Different domains may adopt different theoretical strategies, but the theoretical foundations that structure them must be clearly defined.

The presentations allocated in this category demonstrate themes focused on the discussion of authors such as Ranganathan and Dahlberg and their respective theories, such as classification systems and Faceted Classification, as well as philosophical perspectives in relation to Knowledge Organization and Information Science.

For Hjørland and Hartel (2003), epistemology represents an important axis, as epistemological conceptions directly influence the theories that structure the domains, the terms used, the way in which the objects of study are identified and chosen, the methodologies adopted to conduct practices within the domain and how the procedures to validate and communicate the knowledge produced work.

To visualize the themes, a word cloud was created with the words that emerged from the analysis (Fig. 2), enabling an understanding of the orientation of the epistemological dimension.



Figure 2. Word cloud from the analysis in the epistemological dimension. Prepared by the authors based on research data (2025).

It can be inferred that the epistemological dimension refers to both knowledge and the methods adopted to achieve and understand it and can also be seen as the accumulation of scientists' collective experience. The epistemological dimension of a domain is therefore manifested in the actions of professionals when they pay attention to aspects such as theoretical approaches, meta-theories, currents of thought, paradigms, philosophies, schools of thought, traditions, trends, and perspectives that underlie a specific field of knowledge (Hjørland and Hartel 2003).

4.3. Regarding the sociological dimension

Table 5 presents an overview of the themes, followed by their analyses categorized in the sociological dimension.

MEETINGS	TITLE AND AUTHORS	ANALYSIS OF PRESENTATIONS
8th ISKO Italy, 2017, Bologna	Recovery put gender and form according to the Library of Congress and IFLA, by Anna Lucarelli	The retrieval models based on document format and genre applied by the US Library of Congress and also by the International Federation of Library Associations and Institutions (IFLA) were discussed, and how these models can facilitate access and retrieval of works by users.
10th ISKO Italy, 2021, Cremona (online)	Organizing Knowledge about Allergies, by Marcin Trzmielewski	Trzmielewski presented part of his doctoral research, seeking to demonstrate what the structure of a knowledge organization system (KOS) would be like based on the needs of allergists, in order to monitor and improve their work practices.
10th ISKO Italy, 2021, Cremona (online)	Updating the SciGator interface, by Giorgia Lottici	SciGator search system was discussed, explaining how the system has been updated over the years with the aim of improving searches in the University of Pavia library by users.
10th ISKO Italy, 2021, Cremona (online)	Cognitive Aspects of Human Research, by Marcia Bates	She discussed the user experience when using search systems, highlighting the importance of thinking about classification and indexing systems considering the cognitive aspects and the needs of human beings in the construction of these systems.
11th ISKO Italy, 2023, Rome	Indexing of Non-Textual Materials, by Athena Salaba	It addressed the indexing of non-textual materials, such as images, videos, and music, and the challenges of indexing these types of records, addressing the complexity of indexing and its inability to capture the full meaning of a work, concluding that it is possible to seek more precise and ethical ways of representing these contents to facilitate access and understanding by users.
11th ISKO Italy, 2023, Rome	Product Image Content Representation Model in Electronic Environment (MRCIP), by Gislene Rodrigues & Célia Dias	The authors presented a proposed conceptual model that can help brands and small entrepreneurs effectively represent product images used in e-commerce environments, favoring digital marketing and consumer purchasing decisions.
11th ISKO Italy, 2023, Rome	Classifying fashion's heterogeneous cultural heritage, by Elisa Tosi Brandi	The presentation addressed the preservation of sources linked to fashion heritage, highlighting that good cataloging practices in public institutions can guide this field, related to Archival Science, Library Science, and Museology, seeking to raise awareness among specialists about the importance of managing and classifying this diverse cultural heritage.
11th ISKO Italy, 2023, Rome	Author tagging: potential and limits of keywords in professional scientific production, by Alessandra Moi, Andrea Marchitelli and Carlo Bianchini	The study analyzed changes in subject indexing in libraries with the advent of digital technologies and the web, highlighting the replacement of controlled systems with free practices, such as author-specific keyword selection. The focus is on the impacts of this freedom on information retrieval, especially in scientific publications, and the challenges it poses to the effectiveness of this process.

11th ISKO Italy, 2023, Rome	Indexing of subjects of normative acts, by Ignazio Pirronitto	The presentation addressed the importance of subject-specific indexing of normative acts, such as laws and resolutions, emphasizing that, despite not being books, these documents should be included in knowledge organization systems to ensure their effective retrieval. The Nuovo Soggettario (New Soggettario) was discussed, along with the ISO 5936:198 standard, to ensure standardized descriptive methods, thus avoiding ambiguities.
11th ISKO Italy, 2023, Rome	Search and navigation features of the new Biblioteche Oggi Trends platform, by the Biblioteche Oggi Trends editorial team	It dealt with the periodical “Biblioteche Oggi Trends”, which since 2021, has adopted a hybrid format, combining the traditional printed version with an improved digital edition, with the aim of providing a more interactive reading, with multimedia content and hyperlinks, in addition to expanding the freedom of authors in the use of unconventional sources.

Table 5. Categorization of ISKO Italy’s themes in the sociological dimension. Prepared by the authors based on survey data (2025).

The sociological dimension presents itself as an important part in understanding how knowledge is shared, as it complements the previous dimensions and mainly refers to how subjects constitute their discourses based on ontological and epistemological studies, that is, it deals with how social structures and discursive communities participate in the production and construction of knowledge.

It reflects the influences of social practices, institutions, culture, politics, and economics on the domain and how, based on the dynamics of social formations, domains are structured in the conception of social subjects, researchers, institutions, and traditions.

In the sociological dimension, it is possible to perceive the notion of discursive community, as it is the group of organized people who share language, objectives and methods related to a domain.

The presentations allocated in this category demonstrate themes always permeated by the figure of the user, with intersections with digital systems, decision-making, information search and retrieval processes, indexing, information needs in specific domains and work practices.

To visualize the themes, a word cloud was created with the words that emerged from the analysis (Fig. 3), enabling an understanding of the orientation of the sociological dimension.



Figure 3. Word cloud from the analysis in the sociological dimension. Prepared by the authors based on research data (2025).

It can be inferred that the sociological dimension is central to the approach and understanding of the concepts presented by Hjørland and Hartel (2003), as it is from this dimension that knowledge is recognized as socially constructed and situated in a context.

4.4. Discussion

The analysis and mapping of the dimensions in the works made it possible to categorize each meeting, based on the dimensions of the domains of Hjørland and Hartel (2003) and the respective inference, from the categorization together with the dimensions that were present in these meetings, as a result of the research.

Furthermore, a trend toward research focused on the ontological dimension of domains was observed, with topics focusing on discussions of knowledge organization systems, the structuring of concepts to represent domains, and their application in diverse informational contexts. The epistemological dimension ranked second among the domain dimensions that prevailed in these presentations, with papers discussing theoretical and methodological aspects of Knowledge Organization. Some topics also highlighted a tangency between two domain dimensions, demonstrating the possibilities for discussion that consider applicable aspects of Knowledge Organization, together with its theories and the surrounding sociocultural context.

It is important to note the possibility of identifying all three dimensions in each of these studies. However, one or two dimensions were identified as having the most prominent characteristics. In the case of studies in which two dimensions were identified, it was decided to always mention the most

prominent dimension first and the other second. The quantification of the dimensions in each meeting is explained in Table 6.

MEETINGS	Ontological	Epistemological	Sociological	Epistemological and Ontological	Ontological and Sociological	Epistemological and Sociological	Ontological and Epistemological
8th ISKO Italy, 2017	7	1	0	1	1	0	0
9th ISKO Italy, 2019	6	2	0	1	0	0	0
10th ISKO Italy, 2021	1	3	1	0	2	0	0
11th ISKO Italy, 2023	9	0	0	1	5	1	1

Table 6. Quantification of the predominant dimensions in ISKO Italy meetings. Based on: Survey data (2025).

The categorization of the four meetings results in the identification of the predominance of the ontological dimension of the domains, indicating the presence of research focused on the functioning of knowledge organization tools and systems, their applicability and efficiency in information units.

The characteristics and profiles of the authors are considered, as well as the relationship between ISKO Italy’s programs and the interests and needs of its members, as they are also expressed in the conferences and works presented.

The identification of a second research trend, involving the ontological and sociological dimensions of the domains, also allows us to connect these results to studies dedicated to investigating approaches that consider a multidimensional view of Knowledge Organization, such as Golub (2014), who articulates four elements: *phenomenon*, *perspectives*, *documents*, *collections*.

In a complementary way, it is then observed that it is possible to discuss the results obtained in light of these layers (or levels) that can be identified in the field of knowledge organization (Golub 2014). The author discusses how the *phenomenon*, the different disciplinary/cultural *perspectives*, the *documentary* representation, and the context of *collections* impact access to and retrieval of information.

In dialogue with this perspective, Gnoli (2016) offers a more explicit discussion of the dimensions by observing how this structure places Knowledge Organization Systems (KOS) in a broader context within the dynamics of recorded knowledge, articulating three fundamental components in the domains of knowledge: The object, the intellectual form, and the medium of the documents, which refer respectively to the objects of knowledge, the theoretical ways in which they are approached, and the ways in which they are used by society.

Thus, it is understood that works in which the ontological dimension predominates focus on the nature of objects of knowledge, engaging with the *phenomenon* in Golub (2014), which is that which exists in the world and which will later be represented, such as concepts, objects, processes, and events.

In works where the epistemological dimension predominates, the influence of how each discipline produces its own way of understanding the phenomenon and how the materialization of knowledge occurs can be observed. In Golub (2014), these elements can be considered together with the dimensions of *perspectives* and *documents*, involving how knowledge is theorized, categorized, and structured, and also how documents are vehicles of knowledge, shaped by production practices, patterns, and languages, in which their form of presentation influences their interpretation.

In relation to works where the sociological dimension predominates, the highlighted elements can be considered in relation to how they are understood by communities, given that these discursive and epistemic communities change throughout history and reflect values, interests, and disputes. In Golub (2014), a *collection* is the organized set of documents within an institutional context, reflecting institutional policies, cultural priorities, group interests, and professional traditions. Furthermore, it is possible to consider that each collection establishes its own organizational schemes and metadata standards, defining the logic of how knowledge can be retrieved, involving the epistemological dimension.

Therefore, this correlation helps in understanding how these layers intertwine as phenomena (ontological layer), perspectives and documentary forms (epistemological layer), and institutional collections and practices (sociological layer) combine for the organization and representation of knowledge.

This reinforces, in addition to the multidimensional nature of Knowledge Organization, the possibilities of expanding the themes of the works presented beyond the ontological dimension of the domains, that is, giving equal emphasis to the theoretical and epistemological foundations of Knowledge Organization in research that considers the user, society, its cultural and political aspects, and its needs as the focus.

5. Conclusions

This study sought to contribute to the deepening of the debates that underpin Knowledge Organization, based on the discussion that promotes an analysis of the themes presented in the meetings studied here at ISKO Italy. The prevalence of research with characteristics in the ontological dimension was highlighted, with studies involving instruments and systems of knowledge organization, their functioning and effectiveness in libraries, museums and

archives, therefore categorized in the ontological dimension of the domains of Hjørland and Hartel (2003).

The scenario allows for discussions on prevalent themes, casting a critical look on ISKO itself as a space for stimulating scientific research, in order to make research visible in other dimensions, especially in the sociological dimension of the domains, of low incidence.

Studies like these are considered important strategies for perceiving how an area behaves in a given research context, thus creating a guiding point for understanding and reflecting on the Knowledge Organization as a whole in its discursive community.

References

- Alves, Bruno H., and Ely F. T. de Oliveira. 2016. "O desenvolvimento do domínio da "Organização do Conhecimento" no contexto da Ciência da Informação a partir da ISKO-Brasil." *Brazilian Journal of Information Studies: Research Trends* 10 (2):103–08. <https://doi.org/10.36311/1981-1640.2016.v10n2.12.p103>.
- Amorim, Igor S., and Lígia M. Café 2017. "Agenciamento e análise de domínio: um encontro possível." *Informação & Sociedade* 27 (2): 75-88. <https://periodicos.ufpb.br/index.php/ies/article/view/30790>
- Araújo, Paula C. de. 2024. "Epistemologia: Um Conceito em Análise no Domínio da Organização do Conhecimento." In *Estudos críticos em organização do conhecimento*. Oficina Universitária, edited by Carlos C. de Almeida, Rosa San Segundo, Daniel Martínez-Ávila. São Paulo: Cultura Acadêmica: 139-64. <https://doi.org/10.36311/2024.978-65-5954-485-1.p139-164>.
- Bardin, Laurence. 1977. *Análise de conteúdo*. Lisboa: Edições.
- Barité, Mario. 2015. *Diccionario de Organización del Conocimiento: clasificación, indización, terminología*. 6th ed. Montevideo. CSIC.
- Bufrem, Leilah S., and Brigida M. N. Cervantes. 2021. "A Pesquisa em Ciência da Informação: dimensões epistemológica, política e metodológica." *Informação & Informação*, 26 (4): i-vii. <https://doi.org/10.5433/1981-8920.2021v26n4pi-vii>.
- Burke, Peter. 2003. *Uma história social do conhecimento: de Gutenberg a Diderot*. Translated by Plínio Dentzien. Rio de Janeiro. Jorge Zahar.

- Dahlberg, Ingetraut. 1995. "Current trends in Knowledge Organization." In *Organización del conocimiento en sistemas de información y documentación*, edited by Marco Garcia, Francisco J., Proc.1st ISKO-Spain Conference, Madrid, 4-5 Nov.1993. Universidad de Zaragoza, Zaragoza (1): 7-26. <https://dialnet.unirioja.es/servlet/articulo?codigo=2341306>.
- Diemer, Alwin. 1974. "L'Ordre (Classification) Universel des Savoirs comme Problème de Philosophie et d'Organisation." In *Conceptual basis of the Classification of Knowledge: Proceedings of the Ottawa Conference on the Conceptual Basis of the Classification of Knowledge, October 1 st to 5 th, 1971*, edited by J. A. Wojciechowski, Pullach/Munchen: Verlag Dokumentation, 144-60.
- Freitas, Lidiane M., and Ana C. de Albuquerque. 2017. "As abordagens da análise de domínio como aporte metodológico para a classificação arquivística." *Tendências da Pesquisa Brasileira em Ciência da Informação* (10) 2. <https://revistas.ancib.org/tpbci/article/view/424>.
- Gnoli, Claudio. 2016. "Classifying Phenomena Part 1: Dimensions." *Knowledge Organization*, [S.l.], 43 (6) : 403-15. <https://doi.org/10.5771/0943-7444-2016-6-403>.
- Golub, Koraljka. 2014. *Subject Access to Information: an Interdisciplinary Approach*. Bloomsbury Academic.
- Hjørland, Birger. 2002. "Domain analysis in information Science. Eleven approaches – traditional as well as innovative." *Journal of Documentation*, 58 (4) : 422-62. <https://doi.org/10.1108/00220410210431136>.
- Hjørland, Birger. 2003. "Fundamentals of Knowledge Organization." *Knowledge Organization*, 30 (2): 87-111. <https://www.imrpress.com/journal/KO/30/2/10.5771/0943-7444-2003-2-8>.
- Hjørland, Birger. 2004. "Domain analysis: a socio-cognitive orientation for Information Science research." *Bulletin of the American Society for Information Science and Technology* 30 (3): 17-21. <https://doi.org/10.1002/bult.312>.
- Hjørland, Birger. 2008. "What is Knowledge Organization (KO)?" *Knowledge Organization* 35 (2-3):86-101. <https://doi.org/10.5771/0943-7444-2008-2-3-86>.
- Hjørland, Birger. 2017. "Domain analysis." *Knowledge Organization* 44 (6):436-64. <https://doi.org/10.5771/0943-7444-2017-6-436>. Also available in ISKO Encyclopedia of Knowledge Organization, edited by Birger Hjørland and Claudio Gnoli, https://www.isko.org/cyclo/domain_analysis.

- Hjørland, Birger, and Hanne Albrechtsen. 1995. "Toward a new horizon in Information Science: domain-analysis." *Journal of the American Society for Information Science* 46 (6): 400-25. [https://doi.org/10.1002/\(SICI\)1097-4571\(199507\)46:6%3C400::AID-ASI2%3E3.0.CO;2-Y](https://doi.org/10.1002/(SICI)1097-4571(199507)46:6%3C400::AID-ASI2%3E3.0.CO;2-Y).
- Hjørland, Birger, and Jenna Hartel. 2003. "Afterword: ontological, epistemological and sociological dimensions of domains." *Knowledge Organization* 30 (3/4): 239-45. <https://www.imrpess.com/journal/ko/30/3-4/10.5771/0943-7444-2003-3-4-239>.
- ISKO "International Society for Knowledge Organization." Accessed November 16, 2025. <https://www.isko.org/>.
- ISKO International. "Isko Italy." Last modified September 25, 2025. <http://www.iskoi.org/documenti.htm>.
- ISKO Italy. 2017. "8th meeting Isko Italy." Last modified January 31, 2018. <http://www.iskoi.org/doc/bologna17.htm>.
- ISKO Italy. 2019. "9th meeting Isko Italy." Last modified May 8, 2019. <http://www.iskoi.org/doc/firenze19.htm>.
- ISKO Italy. 2021. "10th meeting Isko Italy." Last modified October 7, 2025. <http://www.iskoi.org/doc/cremona21.htm>.
- ISKO Italy. 2023. "11th meeting Isko Italy." Last modified September 4, 2024. <https://www.aib.it/notizie/look-beyond/>.
- Mai, Jens-Erik. 2005. "Analysis in indexing: document and domain centered approaches." *Information Processing & Management* 41 (3): 599-611. <https://doi.org/10.1016/j.ipm.2003.12.004>.
- Martins, Gracy K., and João B. E. de Moraes. 2015. "Organização e Representação do Conhecimento: institucionalização como disciplina científica no âmbito da Ciência da Informação." In *Annals of XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB)*, 16(01). João Pessoa: UFPB. <https://brapci.inf.br/v/187277>.
- Pinho, Fábio A. 2009. *Fundamentos da organização e representação do conhecimento*. Recife: Editora Universitária da UFPE.
- Pombo, Olga. 1998. "Da classificação dos seres à classificação dos saberes." *Leituras. Revista da Biblioteca Nacional de Lisboa, Ciência e Técnica Vitivinícola* (2): 19-33. <https://bibliotextos.wordpress.com/wp-content/uploads/2011/09/das-classificac3a7c3a3o-dos-seres-c3a0s-classificac3a7c3b5es-dos-saberes.pdf>.

- Sales, Rodrigo de. 2015. “A Organização do Conhecimento na comunidade científica brasileira da Ciência da Informação: uma investigação no âmbito da ISKO-Brasil.” In *Annals of XII Congreso ISKO Espanha e II Congreso ISKO Espanha-Portugal: Organización del conocimiento para sistemas de información abiertos*. Murcia: Universidad de Murcia. https://iskoiberico.org/wp-content/uploads/2015/11/cf078-85_sales.pdf
- Santos, Raimunda F. de, Dulce A.B. Neves, and Edivanio D. Souza. 2019. “A Organização do Conhecimento como domínio de estudo da Ciência da Informação: uma reflexão a partir dos aspectos epistemológicos.” In *Organização do Conhecimento responsável: promovendo sociedades democráticas e inclusivas*, edited by Thiago Henrique Bragato Barros, Natalia Bolfarini Tognoli, Proceedings of V ISKO Brasil, 95-104. <https://cip.brapci.inf.br/download/123253>.
- Silva, Alessandra R. da. 2017. “A dimensão discursiva da Organização do Conhecimento na Ciência da Informação Brasileira.” PhD diss., Universidade de Brasília, Unb, Brasília. <http://repositorio.unb.br/handle/10482/31312>.
- Szostak, Rick, and Peter Ohly. 2020. “The International Society for Knowledge Organization (ISKO).” In *Encyclopedia of Knowledge Organization* (corsivo), edited by B. Hjørland and Claudio Gnoli. Last modified September 14, 2020. <https://www.isko.org/cyclo/isko>.
- Zammataro, Ana F. D. 2023. “A Organização e Representação do Conhecimento em uma perspectiva decolonial: análise dos conceitos do Relatório Figueiredo a partir da classificação facetada.” PhD diss., Universidade Estadual de Londrina, Londrina, <https://repositorio.uel.br/items/476416e7-fcb6-4cf8-a2df-0269747badf4>.

AIDaInformazioni

Rivista semestrale di Scienze dell'Informazione

Anno 43

N. 3-4 – luglio-dicembre 2025

Contributi

GIOVANNA ARACRI, ANTONIETTA FOLINO, MARIA TERESA GUAGLIANONE, MARCO MANNA, Identification of thesaurus relationships with the support of ChatGPT and Gemini

ALESSANDRA BONSIGNORI, Terminologia giuridica e organizzazione della conoscenza

MARIA TERESA CHIARAVALLI, MARIA TERESA GUAGLIANONE, ELENA CARDILLO, Metadattazione delle tipologie di documenti clinici: un approccio alla standardizzazione mediante LOINC

CARLA BEATRIZ MARQUES FELIPE, CARLA MARIA MARTELLIOTE VIOLA, RAIMUNDA FERNANDA DOS SANTOS, Domain Analysis on Data Management

CLAUDIO GNOLI, EMANUELA VALENZANO, Che faccetta sei?

MANUELA GRILLO, Strumenti di convergenza tra archivi e biblioteche: i thesauri

BIANCA LA MANNA, Rappresentare gli atti notarili medievali in Linked Open Data (LOD)

CLAUDIA LANZA, L'univocità e la variabilità terminologica a supporto della rappresentatività delle risorse semantiche di organizzazione e rappresentazione della conoscenza specialistica

LARISSA MORAES MARTINS, CARLOS EDUARDO DA SILVA CARVALHO, ANA CRISTINA DE ALBUQUERQUE, MARIO BARITÉ, Possibilities for using cultural warrant to complement literary work

BIANCA THEREZINHA CARVALHO PANISSET, LUCIA MARIA VELLOSO DE OLIVEIRA, LETICIA SOUZA DA COSTA SAMPAIO, CAMILLA CAMPOI DE SOBRAL, Identification of types of documents and knowledge organization: transparency and access in records management

SOYOUNG YOON, CLAUDIO GNOLI, ZIYOUNG PARK, Applying Integrative Levels Classification to Korean KOS Registry

EMANUELA NICOLE DONATO, GRAZIA SERRATORE, MARIA TAVERNITI, MARIA TERESA CHIARAVALLI, L'evoluzione della struttura documentale nelle cartelle cliniche psichiatriche dell'archivio storico del manicomio di Girifalco (1881-1978)

LORENZA TALARICO, Manuscripts in the Digital Age

MARIA LÍGIA TRIQUES, ANANDA FERNANDA DE JESUS, MARIO BARITÉ, JOSÉ EDUARDO SANTAREM SEGUNDO, ANA CRISTINA DE ALBUQUERQUE, Vocabularies in the digital context

MARCIN TRZMIELEWSKI, Context-Driven Use of Automatic Tools in Medical Ontology Design: A Case Study in Allergology

CARLA MARIA MARTELLIOTE VIOLA, CARLA BEATRIZ MARQUES FELIPE, Exploring the relationships between the Cross-Cutting Women's Agenda - Multi-Year Plan 2024-2027 and the Knowledge Organization of legislative proposals under processing on women in Brazil

ELENA CENCETTI, CAMILLO CARLO PELLIZZARI DI SAN GIROLAMO, ELISABETTA VITI, Modellizzazione delle conoscenze: Thesaurus Nuovo soggettario e Wikidata, due strumenti a confronto

PONTISH YERAMYAN, ALEXANDER MURZAKU, NICK UVA, CURT ANDERSON, The transferability of brilliance

ANA FLÁVIA DIAS ZAMMATARO, ANA CRISTINA DE ALBUQUERQUE, GIULIA CRIPPA, LUCIA SARDO, CARLO BIANCHINI, MARIA LÍGIA TRIQUES, ISKO Italy in perspective



mundaneum

In copertina

Disegno di Paul Otlet, Collections Mundaneum, centre d'Archives, Mons (Belgique).

ISBN 979-12-5965-639-1

ISSN 1121-0095



9 791259 656391



9 770112 100950