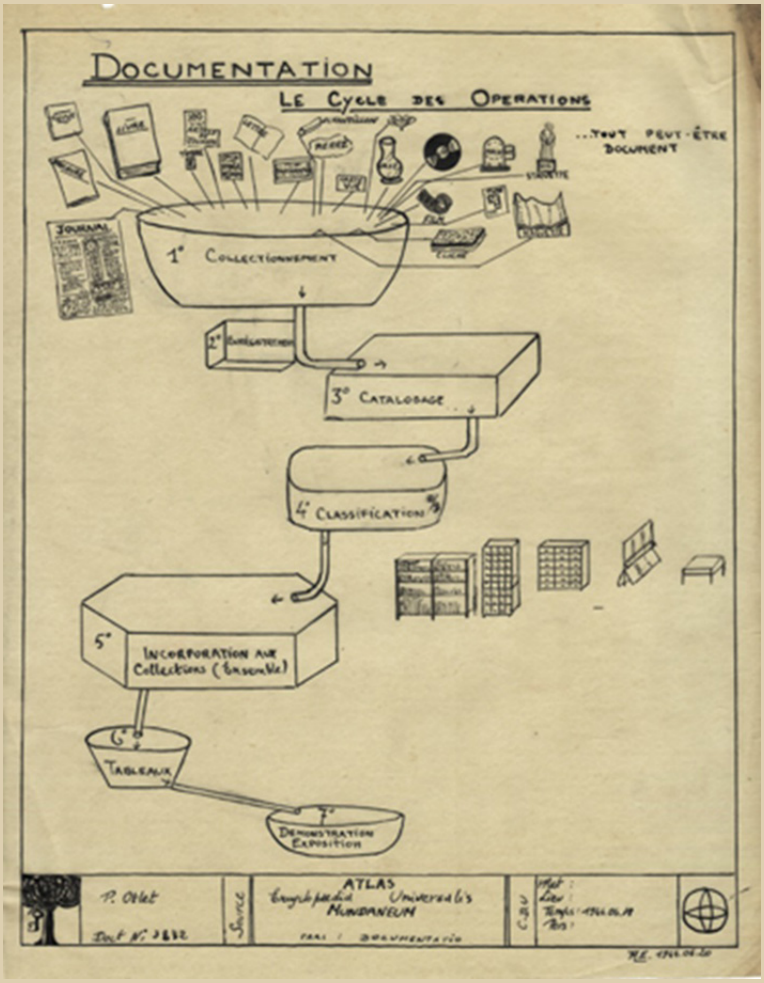
AIDAinformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

NUMERO 3-
ANNO 43
LUGLIO-DICEMBRE 2025



CACUCCI EDITORE
BARI

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 Tuscia*;
Laurence Favier, *Université Charles-de-Gaulle Lille 3*;
Madjid Ihadjadene, *Université Vincennes-Saint-Dénis 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 Chiaravalloti, *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.cacucciedi tore.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, Metadatazione delle tipologie di documenti clinici: un approccio alla standardizzazione mediante LOINC 73

Carla Beatriz Marques Felipe, Carla Maria Martellote 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

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-639-1
DOI 10.57574/596563911
pag. 9-50 (luglio-dicembre 2025)
Open access article licensed under CC-BY-NC-SA

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.

2025, The Author(s). This is an open access article, free of all copyright, that anyone can freely read, download, copy, distribute, print, search, or link to the full texts or use them for any other lawful purpose. This article is made available under a Creative Commons Attribution Share-Alike 4.0 International License. AIDAinformazioni is a journal of the Department of Culture, Education and Society, University of Calabria, Rende (CS), Italy, published by Cacucci Editore, Bari, Italy.

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 (Omame 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 (Bouzid 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 Holocaustic 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.

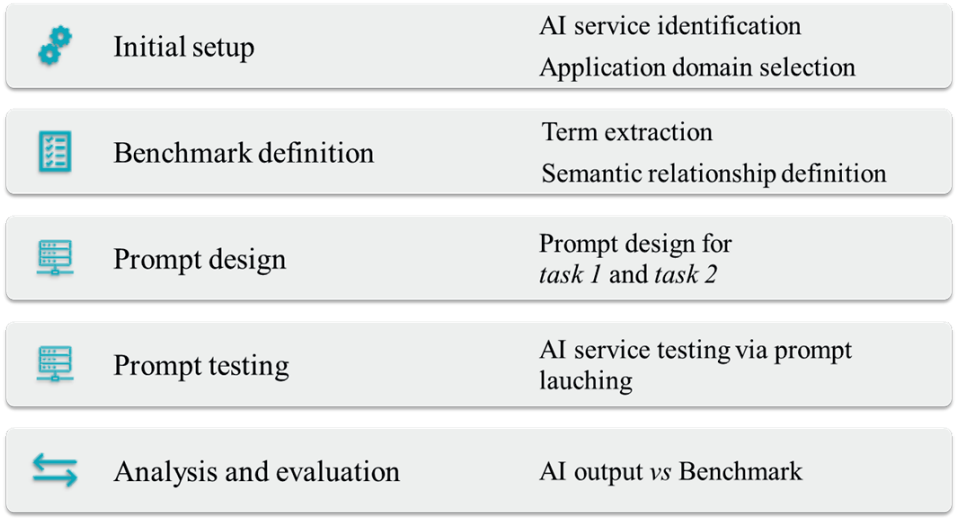


Figure 1. Overview of methodology.

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 of 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 = *Gualchiere* [*Fulling mills*] <*Strumenti* [*Tools*]> and t2 = *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 (EQ) → t2 if and only if t2 (EQ) → 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 *Calcàre* (*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 reasoning[8].

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 inexpert 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 (ICMATE-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.

Bouzid, Sara, and Loïs Piron. 2024. "Leveraging Generative AI in Short Document Indexing." *Elecronics* 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 Reykjavík 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 demoetnoantropologici*, 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.

Omame, 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, Osma, 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 “concerie” (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 “orefi-cerie” (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)* → *t1 (NT)*, where BT = Broader Term and NT = Narrower Term
- *t1 (EQ)* → *t2*, where EQ = Equivalent
- *t1 (RT)* → *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) → t2 (NT)" and "t2 (BT) → 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) → t2" if and only if "t2 (EQ) → t1"
- The RT relation is symmetric, so "t1 (RT) → t2" if and only if "t2 (RT) → 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)	Calzaturifici (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 (Materiali)	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	
Macinazione (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)	Minerali 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)	Mattonifici (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	
Combustibili (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 (Strumenti)	Crogiuoli (Strumenti)	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)	Falegnamerie (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 (Strumenti)	Fucine (Strumenti)	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	
Vetrerie (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	
Minerali 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 (Strumenti)	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	
Depurazione (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 (Materiali)	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 (Strumenti)	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	
Combustibili (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) → 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	
Mulini (Impianti Produttivi)	Molini (Impianti Produttivi)	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	
Cardatura (Attività)	Lana (Materiali)	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	
Pietre (Materiali)	Selce (Materiali)	t1 (BT) → t2 (NT)	1	1	√	1	√
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2		0		0	
		t1 (RT) → t2		0.3		0	
		None		0		0	
Pietre (Materiali)	Ossidiana (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	
Mulini (Impianti Produttivi)	Mulini ad acqua (Impianti Produttivi)	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	
Folloni (Strumenti)	Gualchiere (Strumenti)	t1 (BT) → t2 (NT)		0.9	x	1	x
		t2 (BT) → t1 (NT)		0		0	
		t1 (EQ) → t2	1	0		0	
		t1 (RT) → 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 (Strumenti)	Pane (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	
Calcare <Luoghi di produzione> (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	

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 Chiaravalloti, Maria Teresa Guaglianone, Elena Cardillo, Metadatazione delle tipologie di documenti clinici: un approccio alla standardizzazione mediante LOINC

Carla Beatriz Marques Felipe, Carla Maria Martellote 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 Chiaravalloti, 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 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

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