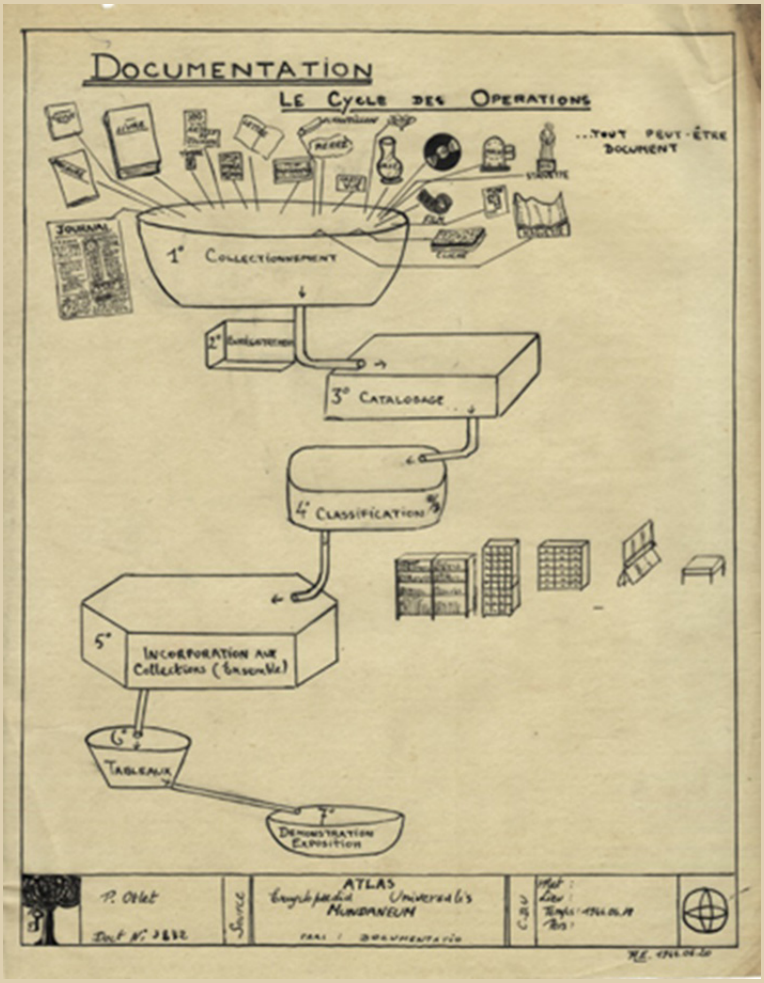
AIDAinformazioni

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.cacuccieditore.it e-mail: info@cacucci.it

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

Sommario

Giovanna Aracri, Antonietta Folino, Maria Teresa Guaglianone, Marco Manna, Identification of thesaurus relationships with the support of ChatGPT and Gemini 9

Alessandra Bonsignori, Terminologia giuridica e organizzazione della conoscenza 51

Maria Teresa Chiaravalloti, Maria Teresa Guaglianone, Elena Cardillo, 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 10.57574/596563917
pag. 141-162 (luglio-dicembre 2025)
Open access article licensed under CC-BY-NC-SA

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

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

Bianca La Manna*

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

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

1. Introduzione

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

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

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

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.

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

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

2. Rappresentare e pubblicare gli atti notarili in LOD

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

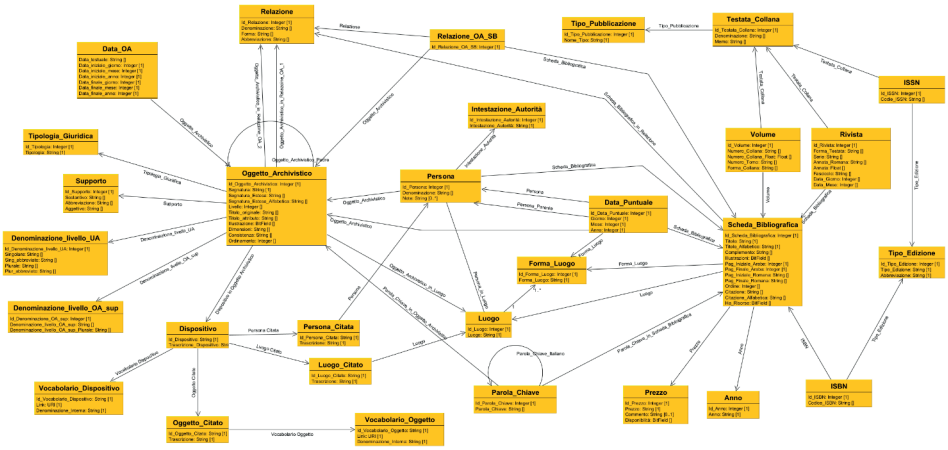


Figura 1. Elementi principali del diagramma E/R ricostruiti dalle classi e relazioni presenti nella base di dati relazionale.

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

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

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

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

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

- Bibliographic Framework (BIBFRAME): realizzata dalla Library of Congress (LOC 2016);
- FRBR-aligned Bibliographic Ontology (FaBIO) (Peroni e Shotton 2012) che fa parte della rete più ampia delle Semantic Publishing and Referencing Ontologies (SPAR Ontologies) (Peroni et al. 2018).

Altre ontologie fondamentali per la rappresentazione del patrimonio culturale e che includono nella loro descrizione entità e proprietà provenienti da vari settori della conoscenza sono: CIDOC Conceptual Reference Model (CIDOC-CRM) (CIDOC CRM Special Interest Group 2024) e Architettura della Conoscenza (ArCo) (Gangemi et al. 2023).

Il modello concettuale finale (Fig. 2) implementa le due entità principali già citate, ovvero oggetto archivistico (*oad:ArchivalResource*) e scheda bibliografica (*a-cat:CatalogueRecord*) e le relazioni tra di loro.

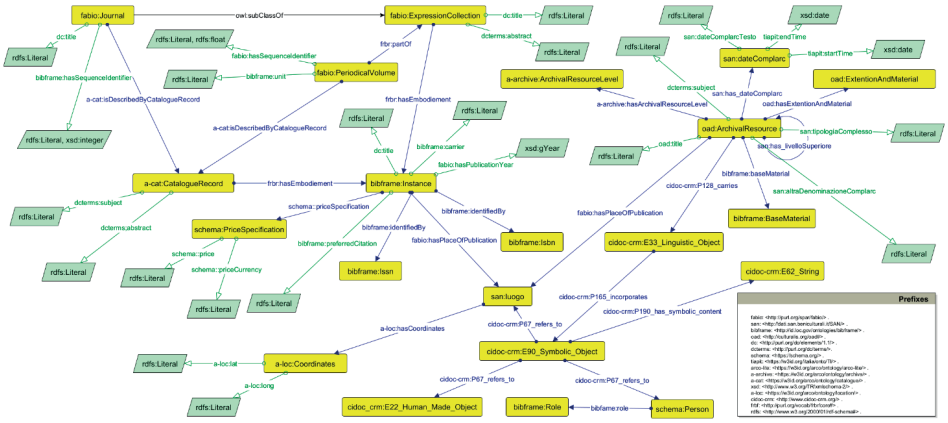


Figura 2. Modello a grafo creato con Graffoo (Peroni 2013).

Questa rappresentazione finale tiene conto della base di dati di riferimento, creata allo scopo di rappresentare e rendere accessibile una collezione nella sua interezza, attraverso le librerie digitali sopra citate, e all'interno della quale parte imprescindibile nella descrizione dell'oggetto archivistico è la sua relazione con il fondo e con le schede bibliografiche. Dal momento che questo progetto si propone di rappresentare nello specifico l'atto notarile, ci concentreremo sull'analisi della descrizione semantica dell'oggetto archivistico, di cui l'atto fa parte.

2.1. L'atto notarile in LOD

L'atto notarile oggetto della nostra ricerca corrisponde al documento che è parte di un oggetto archivistico più ampio, quale ad esempio un registro.

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

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

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

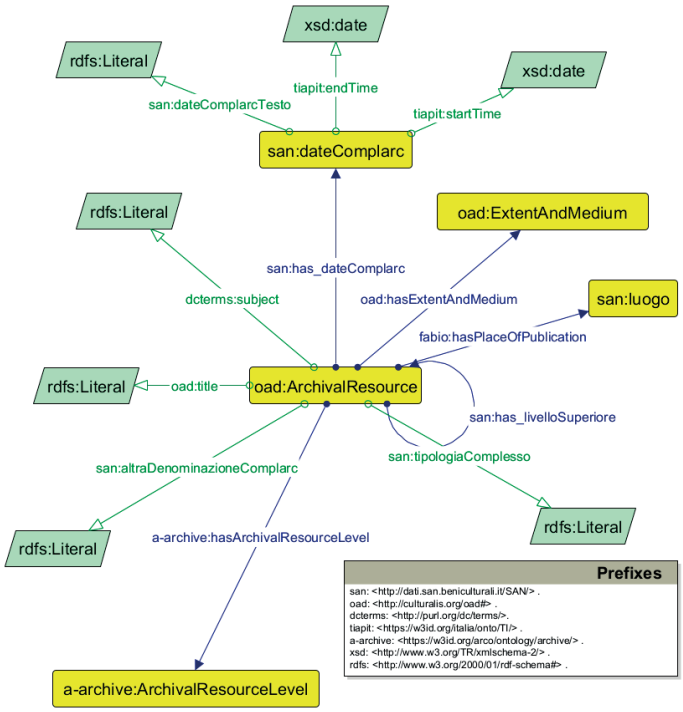


Figura 3. Descrizione dell'oggetto archivistico sulla base di ontologie esistenti, secondo la sintassi di Graffoo (in giallo le entità e i valori nominali).

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

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

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

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

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

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

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

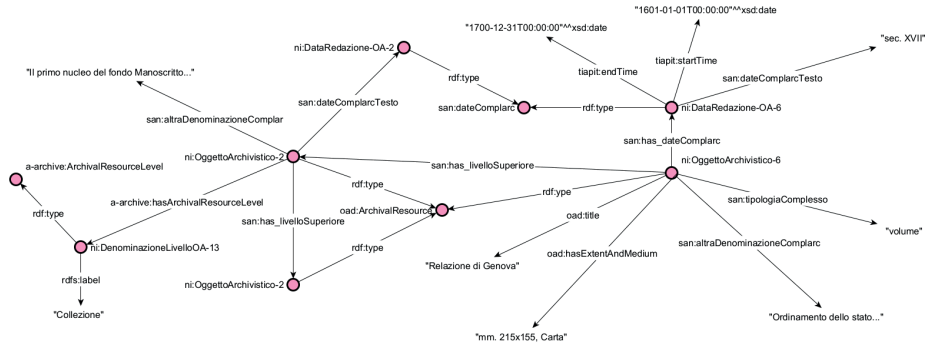


Figura 6. Rappresentazione visiva della stessa istanza di oggetto archivistico, "Relazione di Genova", realizzata tramite Graffoo: i punti rappresentano le istanze di entità, gli altri nodi rappresentano i valori letterali.

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

3. Il latino medievale nell'ambiente digitale

Realizzato il grafo di conoscenza, il secondo obiettivo è di arricchirlo con nuove entità estratte automaticamente. Questo passaggio ha come primo requisito necessario la disponibilità di testi, in questo caso di atti notarili, trascritti in formato digitale di testo semplice, in una quantità tale da essere rappresentativa della lingua. Uno degli obiettivi a lungo termine a cui il centro di studi sta lavorando è quello di realizzare un modello linguistico per la trascrizione automatica dei manoscritti, tramite l'utilizzo della piattaforma Transkribus (READ-COOP SCE 2023), che permetterebbe di ottenere i dati necessari per procedere con gli strumenti di NLP. Nella prospettiva di poter lavorare in futuro su questi dati, questa sezione si prefigge lo scopo di fornire la descrizione di uno studio preliminare volto ad analizzare i limiti e le possibilità dell'applicazione dell'NLP al latino medievale[2].

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

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

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

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

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

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

---

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

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

3.1. Il corpus

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

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

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

Il corpus, in numeri, consiste di:

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

3.2. Strumenti esistenti per il latino

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

- Classical Language Toolkit (CLTK): una libreria Python che prende spunto dai principali programmi di NLP (Johnson et al. 2021, 20; Johnson 2021), tra cui spaCy (Honnibal et al. 2023), per fornire un approccio modulare per l'elaborazione dei linguaggi premoderni che vada incontro alle loro necessità[4]. Il codice è disponibile su Github (Johnson 2025), sotto la licenza MIT;
- spaCy (LatinCy): una serie di operazioni spaCy per l'NLP applicato al latino (Burns 2023). Si tratta di uno strumento *general purpose*, ovvero non specializzato in nessuna fase specifica della lingua latina.

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

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

3.3. NLP applicata al latino medievale

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

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

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

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

3.3.1. Preelaborazione dei dati

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

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

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

Relativamente alle versioni adottate, per l'intero progetto è stato utilizzato Python 3.9, pip 24.0. Per quanto riguarda le librerie, la versione di CLTK è 1.3.0, compatibile con spaCy versione 3.7.4[5].

3.3.2. Estrazione automatica delle entità

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

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

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

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

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

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

Opzione 2: personalizzare un modello già allenato

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

Opzione 3: approccio basato su regole

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

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

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

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

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

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

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

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

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

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

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

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

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

3.4. LiLa

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

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

3.5. Quanteda

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

L'obiettivo di questa sperimentazione è stato quello di provare a classificare in maniera automatica gli atti notarili in base alla loro tipologia, allo scopo di inserire questa ulteriore informazione semantica all'interno del modello descrittivo e, di conseguenza, nella base di dati semantica e a grafo. Definite 15 categorie[6] e le parole chiave che ne caratterizzavano ciascuna, è stato creato un dizionario e, tramite la libreria *quanteda.textstats*, è stato possibile assegnare in maniera automatica ad ogni atto una o più categorie. Questa assegnazione viene effettuata sulla corrispondenza esatta della stringa della parola chiave contro il testo dell'atto. La sperimentazione è stata effettuata sempre sui 163 atti di Sambuceto. 124 atti su 156 sono stati classificati correttamente (7 non erano classificabili in quanto incompleti o frammentari).

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

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

4. Conclusioni

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

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

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

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

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

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

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

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

Ringraziamenti

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

Riferimenti bibliografici

Balard, Michel. 1983. *Notai genovesi in oltremare: atti rogati a Cipro da Lamberto di Sambuceto (11 ottobre 1296 – 23 giugno 1299)*. Istituto di Medievistica.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, et al. 2018. "quanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Muller, e Akitaka Matsuo. 2021. "quanteda: Quantitative Analysis of Textual Data." https://quanteda.io/index.html.

Boros, Emanuela, Veronica Romero, Martin Maarand, et al. 2020. "A Comparison of Sequential and Combined Approaches for Named Entity Recognition in a Corpus of Handwritten Medieval Charters." In *17th International Conference on Frontiers in Handwriting Recognition (ICFHR), 8-10 settembre, Dortmund, Germania*, edited by Juan Guerrero. Washington, DC: IEEE Computer Society Conference Publishing Services, 79–84. https://doi.org/10.1109/ICFHR2020.2020.00025.

Burns, Patrick J. 2023. "LatinCy: Synthetic Trained Pipelines for Latin NLP. Consultato il 5 novembre 2025. https://arxiv.org/abs/2305.04365v1.

Chastang, Pierre, Sergio Torres Aguilar, and Xavier Tannier. 2021. "A Named Entity Recognition Model for Medieval Latin Charters." *Digital Humanities Quarterly*, 15 (4).

CIDOC CRM Special Interest Group. 2024. *Volume A: Definition of the CIDOC Conceptual Reference Model version 7.1.3.* https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.3.pdf.

Cyganiak, Richard, David Hyland-Wood, and Markus Lanthaler. 2014. "RDF 1.1 Concepts and Abstract Syntax. W3C Proposed Recommendation." Consultato il 5 novembre 2025. https://www.w3.org/TR/rdf11-concepts.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. "Named Entity Recognition and Classification on Historical Documents: A Survey." *ACM Computing Surveys* 56 (2): 1-47. https://doi.org/10.1145/3604931.

Erdmann, Alexander, David Joseph Wrisley, Benjamin Allen, et al. 2019. "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2-7 giugno 2019 Minneapolis, Minnesota, edited by Jill Burstein et al. Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2223–34.

Feliciati, Pierluigi. 2021. "Archives in a Graph: the Records in Contexts Ontology within the framework of standards and practices of Archival Description." *JLIS.It*, 12 (1): 92–101. https://doi.org/10.4403/jlis.it-12675.

Gangemi, Aldo, Margherita Porena, Maria Chiara Frangipane, Maria Letizia Mancinelli, and Valentina Presutti. 2023. *ArCo Ontology (ArCo network) version 2.0.* http://wit.istc.cnr.it/arco/lode/extract?lang=en&url=https://raw.githubusercontent.com/ICCD-MiBACT/ArCo/master/ArCo-release/ontologie/arco/arco.owl.

García-González, Herminio, Elena Albarrán Fernández, Jose Labra Gayo, and Miguel Calleja Puerta. 2020. "Converting Asturian Notaries Public deeds to Linked Data using TEI and ShExML." In *Workshop on Humanities in the Semantic Web (WHiSe) Proceedings 2 giugno 2020 Heraklion, Greece,* edited by Alessandro Adamou, Enrico Daga, and Albert Meroño-Peñuela. CEUR Workshop Proceedings. Aachen, Germany: CEUR-WS, 41-46. https://ceur-ws.org/Vol-2695/.

ICA/EGAD (International Council on Archives Expert Group on Archival Description). 2024. *International Council on Archives Records in Contexts Ontology (ICAR RiC-O) version 1.0.2*. Parigi: ICA (International Council on Archives).

ICAR - Centro MAAS.2015. Archived January 19, 2022 at https://web.archive.org/web/20220119183102/http:/dati.san.beniculturali.it/lode/aggiornato.htm#d4e1643.

ICAR - Centro MAAS.2015. Archived January 19, 2022 at https://web.archive.org/web/20220119183102/http:/dati.san.beniculturali.it/lode/aggiornato.htm#d4e1781.

ISO (International Organization for Standardization). 2019. *Date and time – Representations for information interchange – Part 1: Basic rules*. ISO 8601-1:2019. Parigi: ISO.

Istituto Centrale per gli Archivi, e Centro Metologie e Applicazioni per Archivi Storici. 2015. "SAN Ontologia version 1.1". http://dati.san.beniculturali.it/lode/aggiornato.htm

Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna e Regesta.exe. 2018. "OAD Vocabulary Specification version 1.3." Consultato il 5 novembre 2025. http://culturalis.org/oad/.

Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna.2025. "OAD - Ontology for Archival Description" Archived October 17, 2025 at https://web.archive.org/web/20250820013801/http:/culturalis.org/oad/#n.

Johnson, Kyle P. 2021. "CLTK: website." Consultato il 5 novembre 2025. http://cltk.org/.

Johnson, Kyle P. 2025. "CLTK: github repository." Consultato il 5 novembre 2025. https://github.com/cltk/cltk.

Johnson, Kyle P., Patrick Burns, John Stewart, Todd Cook, Clément Besnier, and William Mattingl. 2021. "The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations 1-6 agosto 2021 Bangkok, Thailand (online)*, edited by Heng Ji, Jong C. Park, and Rui Xia. Association for Computational Linguistics, 20-29. https://doi.org/10.18653/v1%2F2021.acl-demo.3.

Kleehammer, Michael. 2024. "pyodbc 5.2.0." Modificato il 17 ottobre 2025. https://pypi.org/project/pyodbc/.

La Manna, Bianca. 2025. "Grafo di conoscenza di Notariorum Itinera serializzato in formato Turtle (1.0)." Zenodo. https://doi.org/10.5281/zenodo.17339380.

Library of Congress. 2016. *BIBFRAME 2 Ontology, version 2.6.0*. https://id.loc.gov/ontologies/bibframe.html.

Mambrini, Francesco, and Marco Carlo Passarotti. 2023. "The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms." *Journal of Open Humanities Data* 9 (28): 1–5. https://doi.org/10.5334/johd.145.

Montani, Ines, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. "explosion/spaCy: v3.7.2: Fixes for APIs and requirements (v3.7.2). Zenodo". https://zenodo.org/records/10009823.

Notarium Itinera. n.d.a. "Notarium Itinera." Consultato il 5 novembre 2025. https://notariorumitinera.eu/.

Notariorum Itinera. n.d.b. "Digital library archivistica." Consultato il 5 novembre 2025. https://notariorumitinera.eu/Digital_Library_Archivistica.aspx.

Notariorum Itinera. n.d.c. "Digital Library bibliografica." Consultato il 5 novembre 2025. https://notariorumitinera.eu/Digital_Library_Bibliografica.aspx.

Peroni, Silvio. 2013. "Graffoo." https://essepuntato.it/graffoo/specification/.

Peroni, Silvio, and David Shotton. 2012. "FaBiO and CiTO: ontologies for describing bibliographic resources and citations." *Journal of Web Semantics* 17: 33-43. https://doi.org/10.1016/j.websem.2012.08.001.

Peroni, Silvio, and David Shotton. 2018. "The SPAR Ontologies." In *The Semantic Web - ISWC 2018 8-12 ottobre 2018 Monterey, California*, edited by D. Vrandeˇci´c et al. Switzerland: Springer International Publishing AG, 119–36. https://doi.org/10.1007/978-3-030-00668-6_8.

Pilgrim, Mark, and Zope Foundation. 2025. "roman 5.0." Consultato il 18 luglio 2025. https://pypi.org/project/roman/.

R Foundation. 2025. "The R Project for Statistical Computing." Modificato il 13 giugno 2025. https://www.r-project.org/.

RDFLib Team. 2025. "rdflib 7.1.4." Consultato il 5 novembre 2025. https://rdflib.readthedocs.io/en/stable/.

READ-COOP SCE. 2023. "Transkribus website." Consultato il 5 novembre 2025. https://www.transkribus.org/it.

Roberts, Molly, Brandon Stewart, and Dustin Tingley. 2018. "stm. An R Package for the Structural Topic Model." Modificato il 31 luglio 2018. https://www.structuraltopicmodel.com/.

Russo, Alessandro, Giorgia Lodi, and Margherita Porena. 2023. Archived November, 9 2025 at https://web.archive.org/web/20250327055217/http://wit.istc.cnr.it/arco/lode/extract?lang=en&url=https://raw.githubusercontent.com/ICCD-MiBACT/ArCo/master/ArCo-release/ontologie/archive/archive.owl#d4e834.

Ruzzin, Valentina, ed. 2020. *Antonio de Inghibertis de Castro (Genova, 1330-1346)*. Genova: Società ligure di storia patria.

Selivanov, Dmitry. 2016. "text2vec." Consultato il 5 novembre 2025. https://text2vec.org/.

Sequeda, Juan, Syed Ahmed Tirmizi, Oscar Corcho, and Daniel P. Miranker. 2011. "Survey of Directly Mapping SQL Databases to the Semantic Web." *The Knowledge Engineering Review* 26 (4): 445–86. https://doi.org/10.1017/S0269888911000208.

spaCy. n.d. "Training Pipelines & Models." Consultato il 5 novembre 2025. https://spacy.io/usage/training.

Università Cattolica del Sacro Cuore. n.d. "LiLa: Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin." Consultato il 5 novembre 2025. https://lila-erc.eu/#page-top.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*, 3 (1). https://doi.org/10.1038/sdata.2016.18.

Won, Miguel, Patricia Murrieta-Flores, and Bruno Martins. 2018. "Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora." *Frontiers in Digital Humanities* 5. https://doi.org/10.3389/fdigh.2018.00002.

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

9 791259 656391 9 770112 100950