From Documents to Data:
Digital Technologies in the Study of Notarial Charters
- poster proposal -

Franziska Decker[1], Sandy Aoun[1], Giuseppe Consolo[1]
[1] University of Graz, Austria – {firstname}.{lastname}@uni-graz.at

ABSTRACT (ENGLISH)
Our contribution presents an ongoing research effort that investigates late medieval notarial charters using advanced computational methods. Drawing from the Monasterium.net[1] archive, we build a curated corpus of 14th–15th-century documents, primarily from Southern Italy, Austria, and Germany. We aim to identify and analyze notarial practices and documentary typologies through text mining, information extraction, and network analysis. A custom annotation scheme is being developed to deploy machine learning and rule-based models tailored to the historical and linguistic complexity of the sources. By extracting named entities and relationships and visualizing them, we explore the social networks and documentary formulas embedded in the charters. The work contributes to the development of digital diplomatics and supports the evolution of Monasterium.net as a platform for computational cultural heritage research.
Keywords: Digital Diplomatics; Information Extraction; Network Analysis; Late Medieval History

ABSTRACT (ITALIANO)
Dai documenti ai dati: le tecnologie digitali per lo studio delle carte notarili.

Il nostro contributo presenta un progetto in corso che esplora le carte notarili tardo-medievali attraverso l'impiego di metodologie digitali e computazionali avanzate. Il corpus selezionato è tratto dal portale Monasterium.net e comprende documenti del XIV e XV secolo, provenienti in particolare dall'Italia meridionale, dall'Austria e dalla Germania. L'obiettivo è analizzare le pratiche notarili e le tipologie documentarie mediante tecniche di text mining, NER (Named Entity Recognition) e network analysis. Per affrontare la complessità storica e linguistica delle fonti abbiamo sviluppato uno schema di annotazione personalizzato, adatto sia a modelli di apprendimento automatico che a modelli basati su regole. L'estrazione automatica di entità e relazioni, e la loro successiva visualizzazione, consente di ricostruire formule documentarie e reti sociali implicite nei documenti. Il progetto intende contribuire allo sviluppo della diplomatica digitale e alla trasformazione di Monasterium.net in una piattaforma sempre più adatta alla ricerca computazionale nel campo del patrimonio culturale.
Parole chiave: Diplomatica Digitale; Information Extraction; Network Analysis; Basso Medioevo

1. INTRODUCTION
The study of medieval notarial charters has long been a cornerstone for understanding administrative and social practices in pre-modern Europe (Schmoeckel & Schubert, 2009; Ostos Salcedo & Pardo Rodríguez, 1997; Trenchs, 1989; Schuler, 1976; Amelotti & Costamagna, 1975; Cencetti, 1964).
Diplomatics have long identified a well-structured normative and formal system in notarial writing characterized by a rigorous codification of documentary formulas. The evolution of the *instrumentum publicum* attested in the 12th century confirms the evidential value of the notarial document, conferring a central role in the juridical and administrative production to the notary (Zabbia, 2009).
Notarial charters are composed of highly formalized sections and free parts where the notary adapts the form to the needs of the individual act. This textual structure, studied in the work of Sabatini (1965; 1968), allows us to identify phenomena of linguistic micro-variation, which reveal significant differences in documentary practices depending on the geographical and institutional context (Greco, 2017). The analysis of documentary clauses allows identifying distinctive features of documentary typologies and notarial circles, providing valuable data for reconstructing professional networks and interactions between social actors (Capriolo, 2017).
In this context, our work aims to integrate computational methods with the diplomatists traditional approach to analyze late-medieval notarial documentation preserved on Monasterium.net (hereinafter called MOM-CA[2]). Through text mining and network analysis, we intend to identify correlations between

[1] *Monasterium.net,* https://www.monasterium.net (23/01/2025).
[2] The abbreviation stands for "Monasterium - Collaborate Archive".

the individuals involved in the charters' creation, textual formulas, and different documentary typologies, thus contributing to a deeper understanding of notarial practices in the fourteenth and fifteenth centuries. The research aligns with the themes of the AIUCD conference, which emphasizes computational approaches to cultural heritage, including text analysis, network visualization, and the application of cutting-edge digital tools.

2. METHODOLOGY

2.1 CORPUS CREATION

The documents for this study will be sourced from MOM-CA, a collaborative digital archive with over 600,000 entries of European charters, which are mainly but not exclusively from German, Italian, Austrian, Slovakian, Czech Republic, and Hungarian archives. The collection for our research will include documents from the 14th and 15th centuries, with a preliminary focus on German, Austrian, and Southern Italian charters.

Notarial charters suffer from underrepresentation in multiple ways. Firstly, their numbers are much lower in comparison to non-notarial sealed charters, especially in regions above the Alps (e.g. they constitute only 4% of the total archival material scrutinized by Weileder, 2019; see also Leipert et al., 2019). Furthermore, as for the late Middle Ages we are presented with a plethora of documents difficult to handle manually, the scholarly focus has rather been on regal and papal charters, leading to only few transcriptions of notarial charters in contrast to their vast amount. Additionally, if provided, many of the available charter abstracts focus only on the main points of a charter's content but lack information about its scribe, i.e. the notary, where applicable.

To build the corpus of our research study, our data undergoes multiple filtering steps. The first step consists of filtering the entire MOM-CA database by retaining all the charters that have an assigned date definitely pointing to the 14th and 15th century, and by excluding all the entries that have assigned dates certainly belonging to other centuries. Charters which have unclear information or no date value are also excluded and stored for potential manual control. The second step consists of filtering the charters belonging to the 14th and 15th century based on terms and keywords unique to notarial documents which can be found in their respective abstracts or full texts. Charters with existing annotations for the CEI-XML[3] tags `<cei:notariusDesc>`, `<cei:notariusSign>`, or `<cei:notariusSub>` are also preserved at this stage. Implementation-wise, both filtering steps make use of the Python library pandas. Finally, the last step consists of filtering the extracted late medieval charters by the language of their abstract or full text transcription. This is done by applying a language detection tool which has been developed within the authors' project[4] and has been fine-tuned on MOM-CA data.

Since this work is part of a larger project called 'From Digital to Distant Diplomatics' (DiDip), the examined corpus will undergo changes with progress in the research areas of other team members carrying out work for instance on handwritten text recognition, providing more full text charter transcriptions for the corpus. Also, object detection and classification tasks from the field of computer vision using visual clues on charters like seals and notary signs have already yielded promising results (Leipert et al. 2019; Nicolaou et al., 2023) and will enable the identification of further notarial material, especially when only image data exists. Lastly, while progressing in our research, new common textual formulas might become apparent which can broaden the adopted search terms (Brugnoli, 2011).

2.2 CORPUS ANNOTATION

To extract information from the charters, we plan to create ground truth data, which would enable us to build and evaluate computational models. Annotation guidelines employed in the context of extracting named entities from German charters abstracts are quite generic (Aoun et al. 2024), while the BeNASch[5] project, which targets challenges particular to historical German, focuses on complex entity nesting. Both approaches do not fully fit our data and research case study. Hence, we are developing a suitable custom annotation scheme, which defines relevant fine-grained named entity types and pertinent relationships between them. While toolkits such as INCePTION and Prodigy provide interfaces for semantic

[3] *CEI (Charters Encoding Initative),* https://www.cei.lmu.de/.
[4] https://huggingface.co/ERCDiDip/langdetect.
[5] https://dhbern.github.io/BeNASch/.

annotation, an application[6] which caters more to diplomatists needs is also being developed in the DiDip project.

2.3 CORPUS ANALYSIS

We intend to analyze our corpus by tackling the information extraction (IE) challenge of the computational linguistics discipline. Based on our designed annotation scheme, our initial objective would be to tackle the named entity recognition and relation extraction IE tasks. Different methodologies can be tried out in this context, ranging from a rule-based setup to highly intricate yet quite opaque deep learning-based models (Ehrmann et al., 2023).
The very nature of our data, diplomatics research questions, and singular annotation scheme motivates us to test the applicability of a rule based approach against that of a machine learning based one.
Computationally, a rule based approach would rely on the extraction of linguistic features and the usage of regular expressions; while in the machine learning realm, frameworks such as spaCy and Flair, which allow for training custom models, can be exploited.
At a more advanced stage, network analysis and data visualisation can be carried out by leveraging the Python library NetworkX as well as the Gephi open-source software.

2.4 ANTICIPATED CHALLENGES AND SOLUTIONS

Given the work's early stage, several challenges are anticipated. One major challenge for our project is based on the nature of the MOM-CA archive itself. Collecting documentary data from numerous European archives, said data and its quality are highly heterogeneous and barely standardized. Most of its records are accompanied by only minimal data because merely a unique identifier (the charter's archive signature) and an assigned date are sufficient for the upload to MOM-CA. While charter images and abstracts (*regesta*), as well as metadata can frequently be found as additional data, only a minority of the documents come with a transcription or edition of their full text or even annotations.
Even after filtering charters that already have transcriptions and abstracts, the next challenge in such data would be to disambiguate and link the named entities with a low error rate (Jeller, 2016).
The NER task itself is challenging too, as it is performed on historical languages (true even for the majority of the charter abstracts) with no standardized orthography or grammar, and only few to no out-of-the-box models are available for the specific historical languages. Also, many manual annotations will be necessary for potentially training models, which will require the development of custom annotation schemes.
This research's challenges pose a chance to bridge the gap between computational methods and traditional diplomatics. It benefits from the interdisciplinarity of the DiDip project members and presents an exemplary use case for the expected audience of the new generation of the MOM-CA platform, which is currently in development.

3. EXPECTED CONTRIBUTIONS

This work's purpose is to contribute to the relatively new field of digital diplomatics (Vogeler, 2004; Vogeler, 2007; Duranti, 2009; Ambrosio, 2020), as well as the enrichment of MOM-CA as a resource for computational text analysis and broader application in cultural heritage studies. By automatically extracting named entities and relevant relationships between them, and subsequently displaying them through the use of network analysis and data visualisation tools, we aim to understand the communities and contextual tendencies of individuals who chose to consult specific notaries.
Thus, this study serves as a practical use case highlighting the specific needs of diplomatists in regard to the functionalities of the next generation of MOM-CA. This ensures that the platform evolves in response to scholarly requirements.

ACKNOWLEDGEMENTS

The contributing authors are part of the ERC project "From Digital to Distant Diplomatics" (Grant No. 101019327), affiliated with the Department of Digital Humanities at the University of Graz. Tamás Kovács is developing the custom annotation tool to aid in the corpus annotation part of the presented work. He, along with the project leader Georg Vogeler and colleague Johannes Laroche are providing valuable

[6] https://github.com/kreeedit/ANNIE.

feedback on the research concept and annotation scheme. Florian Atzenhofer-Baumgartner has provided help in the corpus creation part.
The authors would like to thank the European Research Council for their generous funding, as well as the International Centre for Archival Research (ICARus) and the involved archives for providing their data on Monasterium.net.

REFERENCES

Ambrosio, A. (2020). La diplomatica e il digitale. Il fondo della Biblioteca della Società Napoletana di Storia Patria. Research Trends in Humanities Education & Philosophy, 1–15.

Amelotti, Mario, Costamagna, Giorgio (1975). Alle Origini Del Notariato Italiano. Studi Storici sul Notariato Italiano, 2.

Aoun, Sandy et al. (2024). Information Extraction from German Medieval Charters Abstracts. 19th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2024), Washington, D.C., August 6-10, 2024. https://doi.org/10.5281/zenodo.14671230.

Brugnoli, A. (2011). Insediamento, territorio e formule notarili: Una verifica (Verona, IX-XII secolo). Reti Medievali Rivista, 12/1. http://dx.doi.org/10.6092/1593-2214/310.

Capriolo, G. (2017). Pratiche redazionali nel Regno di Napoli in età aragonese: Realtà territoriali a confronto. Scrineum Rivista, 14, 501–518. https://doi.org/10.13128/Scrineum-21996.

Cencetti, G. (1964). Il notaio medievale italiano. Atti del XIII Congresso Nazionale del Notariato, Genova, IX–XXIII. http://www.rmoa.unina.it/id/eprint/5638.

Duranti, L. (2009). From Digital Diplomatics to Digital Records Forensics. Archivaria, 68, 39–66.

Ehrmann, Maud et al. (2023). Named entity recognition and classification in historical documents: A survey. ACM Computing Surveys, 56/2, 1–47. https://doi.org/10.1145/3604931.

Greco, P. (2017). La formula documentaria della defensio nelle carte notarili latine della Langobardia minor (IX secolo): Uno studio linguistico. Philologica Jassyensia, 13/2, 71–88. https://hdl.handle.net/11588/699461.

Jeller, Daniel (2016). Urkunden als Netzwerke. https://doi.org/10.58079/nmde.

Klie, Jan-Christoph et al. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico. Association for Computational Linguistics, 5–9.

Leipert, Martin et al. (2020), The Notary in the Haystack – Countering Class Imbalance in Document Processing with CNNs. Proceedings of the 14th IAPR International Workshop on Document Analysis Systems, Wuhan, China, July 26–29, 2020. Lecture Notes in Computer Science, 12116, 246–261. https://doi.org/10.48550/arXiv.2007.07943.

Nicolaou, Anguelos et al. (2023). Efficient Annotation of Medieval Charters. Document Analysis and Recognition. Proceedings of the ICDAR 2023 Workshops. San José, CA, USA, August 24–26, 2023. Vol. 1. Lecture Notes in Computer Science, 14193, 284–295. https://arxiv.org/abs/2306.14071.

Ostos Salcedo, Pilar, Pardo Rodríguez, María Luisa (1997). Estudios sobre el notariado europeo (siglos XIV–XV).

Sabatini, F. (1965). Esigenze di realismo e dislocazione morfologica in testi preromanzi. Rivista di cultura classica e medioevale, 7, 972–998.

Sabatini, F. (1968). Dalla «scripta latina rustica» alle «scriptae romanze». Studi Medievali, Serie III, 9, 320–358.

Schuler, Peter-Johannes (1976). Geschichte des südwestdeutschen Notariats. Von seinen Anfängen bis zur Reichsnotariatsordnung von 1512.

Schmoeckel, Mathias, Schubert, Werner (Eds.) (2009). Handbuch zur Geschichte des Notariats der europäischen Traditionen. Rheinische Schriften zur Rechtsgeschichte, 12.

Trenchs, José (Ed.) (1989). Notariado público y documento privado. De los orígenes al siglo XIV. Actas del VII Congreso Internacional de Diplomática.

Weileder, Magdalena (2019). Spätmittelalterliche Notariatsurkunden. Prokuratorien, beglaubigte Abschriften und Delegationsurkunden aus bayerischen und österreichischen Beständen. Archiv für Diplpmatik. Beihefte, 18.

Vogeler, Georg (2004). Ein Standard für die Digitalisierung mittelalterlicher Urkunden mit XML. Bericht von einem internationalen Workshop in München 5./6. April 2004. Archiv für Diplomatik, 50, 23–34.

Vogeler, Georg (2007). Digital Diplomatics – Digitale Diplomatik: Historical research with medieval charters in a digital world - Die historische Arbeit mit Urkunden in der digitalen Welt. International Conference, Munich, 28.2.-2.3.2007.

Zabbia, Marino (2009). Notai e modelli documentari: Note per la storia della lunga fortuna di una soluzione efficace. Circolazione di uomini e scambi culturali tra città (secoli XII-XIV). A. L. Trombetti Budriesi (Ed.). 23–40. https://hdl.handle.net/2318/86408.