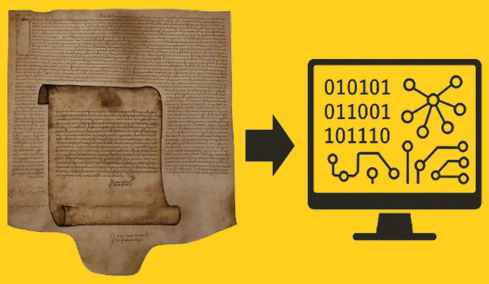
AICUD 2025, Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale

DiDip From Digital to Distant Diplomatics
www.didip.eu
@DiDip_ERC



From Documents to Data

Digital Technologies in the Study of Notarial Charters

11 – 06 – 2025

inôm
We work for tomorrow
UNI GRAZ
Department of Digital Humanities

Introduction

This work investigates 14th–15th century notarial charters available in the digital archive Monasterium.net ("MOM-CA"), with a focus on material from Southern Italy, Austria, and Germany.

Notarial charters are highly formalized legal documents that reveal standardized linguistic formulas, structured textual architecture, and complex professional practices. Their value lies in the wealth of information they contain about social actors, institutional settings, and documentary cultures.

We aim to extract and analyze these features using digital methods tailored to historical sources — namely named entity recognition, relation extraction, and network analysis — with the support of a custom annotation scheme. Documents are filtered, annotated, and computationally modeled to highlight recurrent formulas, documentary typologies, and professional networks.

By combining diplomatic expertise with computational tools, this work contributes to the advancement of Digital Diplomatics and facilitates the transformation of Monasterium.net into a robust platform for large-scale historical data analysis and cultural heritage research.

```python
def extract_context(text):
    matches = []
    words = text.split()
    for match in autoritate_pattern.finditer(text):
        start_idx = max(0, match.start())
        end_idx = match.end()

        # Find the positions of words around the match
        start_word_idx = len(text[:start_idx].split())  # Count words before the match
        end_word_idx = len(text[:end_idx].split())  # Count words till the end of match

        # Get 10 words before and 10 words after the match
        context_start_idx = max(0, start_word_idx - 10)
        context_end_idx = min(len(words), end_word_idx + 10)

        context = ' '.join(words[context_start_idx:context_end_idx])

        # Check for nearby 'notar' or 'notai'
        if re.search(notar_pattern, context):
            matches.append(context.strip())
```

Filtering for Authorisation Formulas in Notarial Charters

Corpus Creation, Annotation and Analysis

To **construct our corpus**, the Monasterium.net database is subjected to a sequence of filtering operations:

1. Selecting charters with assigned **dates** in the **14th or 15th century**
2. Retaining charters whose data includes pertinent **terms** for 'notary' and the ones that are tagged with **CEI-XML elements** <cei:notariusDesc>, <cei:notariusSign>, or <cei:notariusSub>
3. Selecting charters written in **desired languages** using a language detection tool (https://huggingface.co/ERCDiDip/langdetect) fine-tuned on monasterium.net data

To **annotate our corpus**, we developed a custom annotation scheme tailored to the semantic richness and structural complexity of notarial charters. This scheme defines specific categories of **Named Entities** (such as Persons, Locations, Dates, Organizations, and Commodities), as well as custom **Relations** between them, designed to reflect the logic of diplomatic sources. For instance, the relation notary-of links a person to their professional role, while issued-at, descendant-of, and sold-to express location, genealogy, and transactional roles respectively. Annotation is carried out using **ANNIE**, a web-based interface developed within the DiDip project. The toolkit supports layered tagging and custom relations, including compound expressions (e.g., *iudex ad contractus*) through advanced linking such as function_expansion. It is designed to handle the peculiarities of diplomatic syntax, with a growing dictionary of entity types and semantic relations optimized for the needs of diplomatists.

For **corpus analysis**, we aim to assess rule-based and machine learning-based techniques for information extraction:

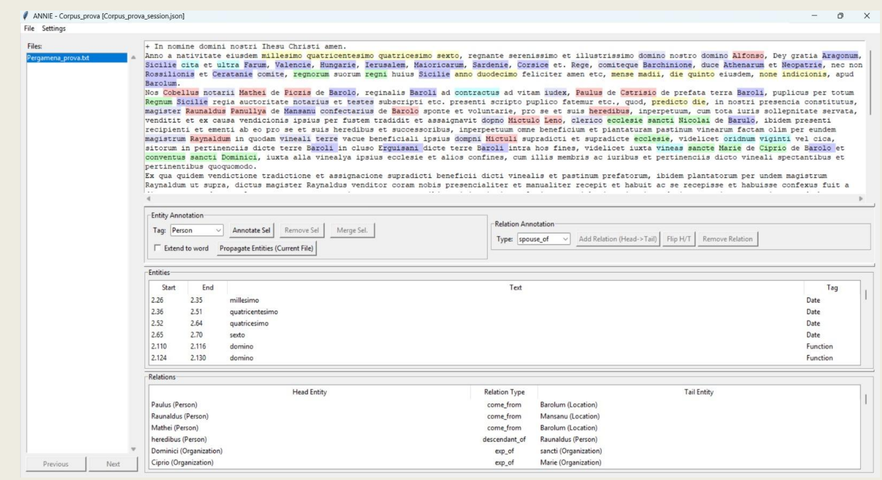
- Linguistic features and regular expressions would be used to implement a **rule-based system**
- spaCy and Flair frameworks would be used to train custom **machine learning models** on our annotated data

Network Analysis of the extracted data would then be performed to explore connections between individuals, locations, and types of documents. NetworkX and Gephi software would be employed to map notarial networks and detect clusters in professional practice.

As part of a complementary approach currently under development, we also want to **identify patterns** in the use of documentary formulas by training a diplomatics-informed model to **detect recurrent textual structures** in Latin notarial corpora.

BCB, "Pergamene", 26r

ANNIE Toolkit - https://github.com/kreeedit/ANNIE



Annotation Process

Challenges and Outlook

Working with late medieval notarial documents presents several challenges, both technical and conceptual. The MOM-CA archive is **highly heterogeneous**: while some documents include full transcriptions and metadata, others provide only minimal information such as a date or archive signature. This variability requires careful filtering, frequent manual intervention, and pre-processing during the corpus construction phase.

Another major difficulty lies in the **semantic annotation of historical texts**, which are often multilingual, inconsistent in spelling, and rich in complex expressions. No off-the-shelf NLP models exist for this context, which entails that manual annotation and iterative evaluation are crucial for developing effective extraction strategies.

Despite these obstacles, this work opens significant opportunities. By identifying recurring documentary formulas and mapping the social-professional networks behind notarial production, we contribute not only to the study of Digital Diplomatics, but also to the design of tools and workflows that are replicable across domains.

Looking ahead, our approach will support the integration of annotated diplomatic data into larger infrastructures such as Monasterium.net, offering a bridge between archival sources and computational analysis. This allows for a sustainable and extensible digital ecosystem for the study of legal and administrative documentation in pre-modern Europe.

Franziska Decker, Sandy Aoun, Giuseppe Consolo
Department of Digital Humanities, University of Graz, Austria
{firstname}.{lastname}@uni-graz.at

ASSOCIAZIONE PER L'INFORMATICA UMANISTICA E LA CULTURA DIGITALE