Building a virtual research environment to move from Digital to Distant Diplomatics

DiDip
From Digital to Distant Diplomatics

Georg Vogeler, Sandy Aoun, Florian Atzenhofer-Baumgartner, Franziska Decker, Florian Lamminger, Daniel Luger, Tamas Kovacs, Anguelos Nicolaou, Nicolas Renet
→{firstname.lastname}@uni-graz.at

Data & Infrastructure

Monasterium.net is the largest publicly available collection of digitized medieval charters:

- more than 650.000 from all over Europe, with a bias towards Central Europe (Germany, Italy, Austria, Slovakia, Czech Republic, and Hungary)
- managed by an international consortium, the „International Centre for Archival Research“ (ICARus) as a community effort of archives and research institutions
- provided by the „MOM-CA" (Monasterium Collaborative Archive) open source software (XQuery-based software package that is deployed in an eXist-db installation)
- includes a JavaScript-based graphical user interface to edit the XML of the charter descriptions („EditMOM3“)

Problems: heterogeneous descriptions, data bias, highly customized end-of-life software, legacy XML-CEI format, poorly documented public API

Task 1: Cleaning data
Tackle inconsistencies and redundancies across data that stem from acquisition history and changing processing conventions

Task 2: Adding data
Reduce regional bias by acquisition of further charters (100.000+ in the upcoming 4 years: France, Northwestern Europe, Scandinavia)

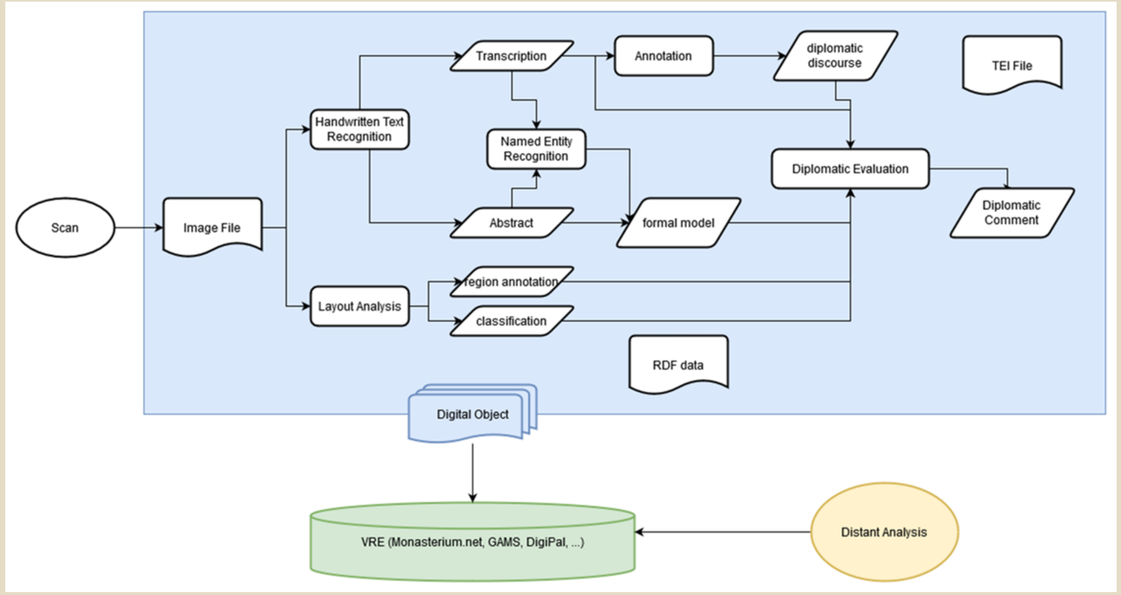
Task 3: Enriching data
Create representative gold standard in manually annotated subsets of 1000/5000 charters; pave way for automated metadata enrichment

Task 4: Processing data
Design and align data transformation flows to handle relational impedance; implement data version control; advance CEI towards TEI

Task 5: Publishing data
Make the data better available by enabling computational access to the data (publically described RESTful API, IIIF, W3C Linked Data Platform); provide consistent descriptions; support human selection (Information Retrieval)

Digital/Distant Diplomatics

Charters are of great importance and high value as sources for historical research because human community life is built on contracts and statutes, i.e., legal acts. They are studied in the research field of diplomatics. Traditional diplomatics usually focus on single documents or highly restricted groups of charters in terms of time or region. What can we establish as computational methods able to deal with the particular challenges presented in diplomatics?



Prototypical workflow of computational methods (rectangles) applied to diplomatics research. (Draft by Sean M. Winslow)

We study large-scale developments in late medieval documentary practice by: *external features* (visual analysis) such as document layout; graphical/physical means of authentication; script style, and *internal features* (textual and semantic analysis) such as text reuse; text structure; lexicon; language style; roles of persons and institutions.

Computer Vision

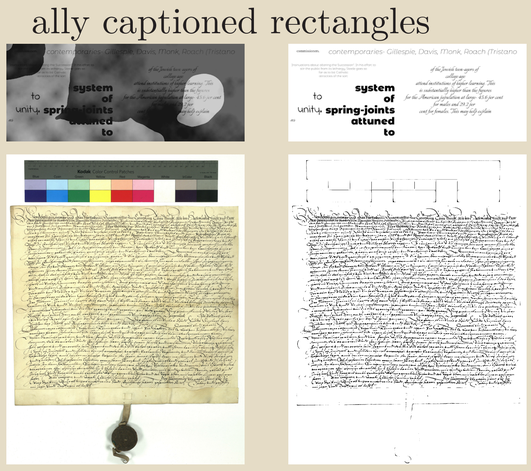
We hypothesise that charter style and form can indicate cultural interchange. Through distant viewing we intend to quantify charter similarity and answer questions such as if geographical proximity, institutional, or political proximity is a better predictor of stylistic similarity. Defining a charter prototype through quantitative analysis of their form allows us to see how the prototypical charter can evolve over time, geography, or institutions.

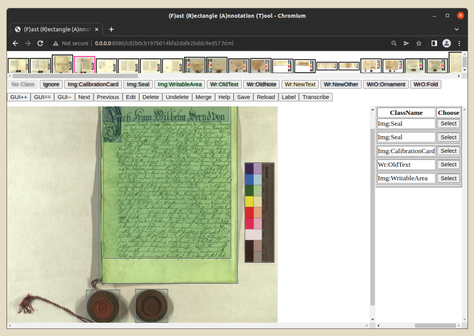
Planned ecosystem/pipeline:

- Binarization: UNets
- Layout analysis: YOLOv5
- Textline segmentation
- HTR: CTC RNN/Transformers
- Word segmentation: YOLOv5
- Word spotting: PHOCNet
- Texture (style) analysis: LBP

Data strategies:

- Realistic Augmentations
- Synthetic Self-supervision
- Groundtruth: Labeled, optionally captioned rectangles





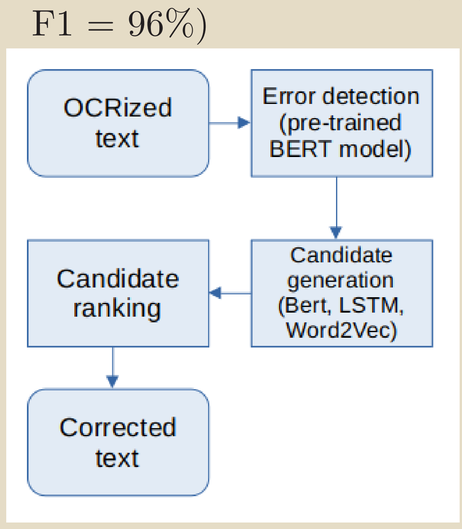
Natural Language Processing

Post-OCR-Correction
Predefined list of possible errors & manual error correction? Inefficient! It is better to find and fix typographical errors in texts (segmentation errors, misrecognition, or missing letters). Our approach uses:

- Fine-tuned Masked-Language model for error detection
- Statistical- & NN-based language models for candidate generation, that learn to link every vocabulary item to a continuous-valued feature vector
- Contextually, the most likely candidate replaces the detected error on the HTR-pipeline

Language Detection
Having difficulty distinguishing historical languages from those of similar eras due to structural & semantic similarities (e.g., different German dialects)? Our solution:

- Fine-tuned XLM-RoBERTa, BPE tokenizer, max. 512 tokens
- Weights for the loss function to handle class imbalances, macro-F1 = 96%)



Automatic Translation
The metadata of Monasterium.net contains at least 35 different languages. Our automatic translation uses Google Translate, Microsoft Translate, DeepL, and Yandex APIs, predefining the right tool for each language pair to get the best translation. We implement:

- Glossaries that helps translate historical names and idioms
- Experiments to build an AutoTrans model instead of API-based models to translate charter abstracts

CENTRE FOR INFORMATION-MODELLING AUSTRIAN CENTRE FOR DIGITAL HUMANITIES
UNI GRAZ
European Research Council
Established by the European Commission