Building a virtual research environment to move from digital to distant Diplomatics (ERC project DiDip)

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich

Luger, Daniel
daniel.luger@uni-graz.at
Universität Graz, Österreich

Nicolaou, Anguelos
angelos.nicolaou@uni-graz.at
Universität Graz, Österreich

Kovacs, Tamas
tamas.kovacs@uni-graz.at
Universität Graz, Österreich

Atzenhofer-Baumgartner, Florian
florian.atzenhofer-baumgartner@uni-graz.at
Universität Graz, Österreich

Lamminger, Florian
florian.lamminger@uni-graz.at
Universität Graz, Österreich

Aoun, Sandy
sandy.aoun@uni-graz.at
Universität Graz, Österreich

Decker, Franziska
franziska.decker@uni-graz.at
Universität Graz, Österreich

The ERC Project "From Digital to Distant Diplomatics" (DiDip, https://didip.eu) attempts to build an innovative and sustainable (virtual) research infrastructure and environment (VRE) to facilitate large-scale analyses of historical documents. It will extend the Monasterium.net infrastructure, which is provided in an aging software (Bürgermeister et al. 2018). However, Monasterium.net is still the largest repository of digital representations of medieval and early modern charters. For the future use of this corpus, it is crucial to make data, in particular gold standard annotations, and methods as open as possible. We plan to combine traditional approaches to analyzing such charters with state-of-the-art computational methods and artificial intelligence. The data produced and the methods used will be available under open licenses (code repository https://github.com/Didip-eu).

The project addresses an unsolved problem in the domain of diplomatics, i.e., the historical auxiliary science, studying medieval and early modern single sheet legal documents: With pure human intellectual capacity, the empirical part of this research had to focus on local, regional, or chancery level in the face of overwhelming quantities of charters (Hlavacek 2006). While earlier approaches to applying digital methods to the field focussed on digital representation of individual descriptions (Ambrosio et al. 2014, Vogeler 2009, Bradley et al 2019), the large-scale "distant reading" approach has been scarce. This changed only recently: In the field of computer vision, Handwritten Text Recognition (HTR) has provided the first results in changing this (Hodel 2017). In addition, Computer Vision (CV) can provide quantified stylistic attributes of all graphical features of a charter. Leifert et al. 2020 and Christlein 2018 extracted graphical elements from charters (e.g., decorations, notarial signs). There are indications that CV can infer the date of a historical document (Cloppet 2017, Seuret 2021) and can classify the handwriting style from a paleographical perspective.

We conclude that a typical CV pipeline for analyzing a charter should consist of: layout analysis, HTR or word segmentation, and finally an analysis of non-text attributes such as style, material, seals etc. Most of these tasks utilize publicly available datasets that are focussed on manuscript books (Simistira 2016) that cannot encapsulate the diversity that is observed in large charter collections.

The most precious resource in such a pipeline is the diplomatist's time spent on annotating data. We drastically economize annotation effort by reformulating layout analysis as an object detection problem instead of the typical image segmentation approach. Indicatively, this allowed us to annotate the layout of 1175 charter images in a fraction of the time that would be needed normally. Figure 1 shows an example of this kind of annotation. Preliminary experiments demonstrate that this approach works well, e.g., it can detect seals with an accuracy above 95% when using a YOLOv5 (Jocher 2022) based model. With a 50% Intersection over Union (IoU) threshold this result is, of course, mainly usable for classification tasks, while segmentation will have to make use of approaches (Leipert et al. 2021). For tasks like HTR, writer identification, and layout analysis, we consider binarization as a useful step. We made experiments indicating that purely synthetic data could be used for the binarization step (Nicolaou 2022), yet comprehensive performance analysis on charters specifically would need manual annotation of the ground-truth pixel-by-pixel. Although the target CV pipeline will be under construction for a while, a few stand-alone methods required for such a pipeline have already been successfully developed and tested.

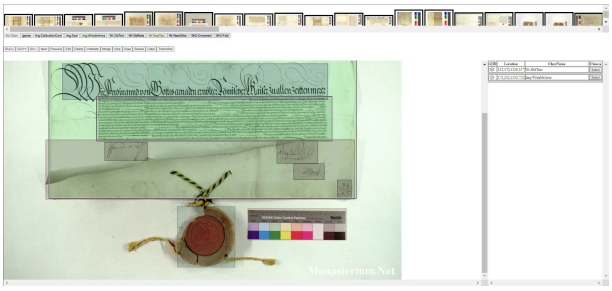

Figure 1: Image annotation example using FRAT (https://github.com/anguelos/frat)

A "distant reading" approach has been taken by Telihuan et al. 2012 and 2014, Perreaux 2021, Leclerq et al. 2021, who study statistical features of larger text corpora extracted from charters. We plan to generalize these approaches with the application of Natural Language Processing (NLP) as a custom, multi-level pipeline. It will resolve information retrieval from both HTR and human-produced data, i.e., address tasks like named entity recognition or relationship extraction, text reuse, and in particular formulaicity detection, but also reflect on the possibilities of named entity linking and text summarisation. We are currently working on laying the basis for this. The wide diversity of dialects and less-resourced languages (early vernacular) is one of the most significant analytic difficulties experienced in medieval and early modern charters. Additionally, the existing OCR of charter texts has insufficient quality for further NLP processing. We show the adoption of a multilingual generic system to tackle both problems by BERT (Devlin et al 2019) models. We create a domain-specific model currently under the pseudonym of "RatisBERT" for OCR post correction, that will be made available through the project's GitHub repositories. It will be based on human controlled data from Monasterium.net itself, charter corpora like ALIM, CDLM, DEEDS, Glessgen 2016, CAO, and enriched by non-charter specific gold standard corpora like the reference corpora for medieval and early modern German (REM, REF, REN including Rhenish) or the PalaFraFro V2-2. The system takes into consideration a variety of errors originating from HTR and various historical periods and linguistic regions, and provides an effective and automated post-correction approach. We use XLM-RoBERTa for language and variant detection. Through the second layer, the pipeline identifies named entities in the formulaic language of charters, thus forming a solid subset for the abstract generating task, which creates a condensed version of a document in English and other modern languages while preserving its essential information in the standardised format of the charter abstract.

The project is planning to integrate these solutions built on the collection of Monasterium.net with generic Digital Humanities (DH) tools through RESTful application programming interfaces (API) and provides access to its own methods through their own, thriving to set up the domain-specific diplomatics VRE as part of the growing DH API infrastructure.

Bibliographie

ALIM - Archivio della Latinità Italiana del Medioevo http://www.alim.dfll.univr.it/

Ambrosio, Antonella, Sébastien Barret, and Georg Vogeler (Eds.). *Digital Diplomatics: The Computer as a Tool for the Diplomatist?* Archiv für Diplomatik. Beiheft 14. Köln, Wien: Böhlau Verlag, 2014. https://www.degruyter.com/view/title/496882 .

Bradley, John, Dauvit Broun, Alice Rio, und Matthew Hammond. „Exploring a Model for the Semantics of Medieval Legal Charters". *International Journal of Humanities and Arts Computing* 13, Nr. 1–2 (10. Juni 2019): 136–54. https://doi.org/10.3366/ijhac.2017.0184 .

Bürgermeister, Martina, Schneider, Gerlinde, Makowski, Stephan, Jeller, Daniel, Bigalke, Jan, Theisen, Christian, und Vogeler, Georg. „‚Software Aging' in den DH: Kritik des reinen Forschungswillens". In *Kritik der digitalen Vernunft. DHd2018. Konferenzabstracts* . Köln: DHd, 2018: 308–11.

CAO - Corpus der altdeutschen Originalurkunden, elektronische Fassung ed. Kurt Gärter, Andrea Rapp. Trier, [2007]. https://tcdh01.uni-trier.de/cgi-bin/iCorpus/CorpusIndex.tcl .

Christlein, Vincent. „Automatic Detection of Illuminated Charters". In *Illuminierte Urkunden. Beiträge Aus Diplomatik, Kunstgeschichte Und Digital Humanities* , herausgegeben von Gabriele Bartz und Markus Gneiß. Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde, Beihefte 15. Wien: Böhlau, 2018: 45–51.

———. „Technical Tools for the Analysis of High Medieval Papal Charters". In *Papstgeschichte im Digitalen Zeitalter. Neue Zugangsweisen zu einer Kulturgeschichte Europas* , herausgegeben von Klaus Herbers. AfD Beiheft. Wien: Böhlau Verlag GmbH & Cie, 2018: 45–53.

CDLM - Codice Diplomatico della Lombardia Medievale. 2000-2022. https://www.lombardiabeniculturali.it/cdlm/

Cloppet, Florence, Véronique Eglin, Marlène Helias-Baron, Cuong Kieu, Nicole Vincent, und Dominique Stutzmann. „ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script". In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* , 2017: 1371–76. https://doi.org/10.1109/ICDAR.2017.224 .

CorA-ReN - Reference Corpus of Middle High German (1050–1350) https://www.linguistics.rub.de/rem/access/index.en.html

DEEDS - Documents of Early England Data set. https://deeds.library.utoronto.ca/

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, und Kristina Toutanova. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv, 24. Mai 2019. http://arxiv.org/abs/1810.04805 .

PALAFRAFRO-V2-2. In *palafra* , 2016. https://palafra.github.io/fr/texts.html .

Glessgen, Martin (ed.). *Documents linguistiques galloromans. Édition électronique* , 3eme ed. 2016 https://www.rose.uzh.ch/docling/

Gml-Mis - Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650). January 6, 2021. https://www.fdr.uni-hamburg.de/record/9195

Hlavácek, Ivan. „Das Problem der Masse: Das Spätmittelalter". *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 52 (2006): 371–93.

Hodel, Tobias. „Sending 15th-Century Missives through Algorithms: Testing and Evaluating HTR with 2,200 Documents". In *IMC Leeds 2017 Paper, 11th July* , 2017. https://solascriptum.wordpress.com/2017/07/11/imc-leeds-paper-sending-15th-century-missives-through-algorithms-testing-and-evaluating-htr-with-2200-documents/ .

Jocher, Glenn, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, u. a. *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference* (Version v6.1). Zenodo, 2022. https://doi.org/10.5281/ZENODO.3908559 .

Leipert, Martin, Georg Vogeler, Mathias Seuret, Andreas Maier, und Vincent Christlein. „The Notary in the Haystack – Countering Class Imbalance in Document Processing with CNNs". In *Document Analysis Systems* , herausgegeben von Xiang Bai, Dimosthenis Karatzas, und Daniel Lopresti. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 246–61. https://doi.org/10.1007/978-3-030-57058-3_18 .

Nicolaou, Anguelos, Vincent Christlein, Edgar Riba, Jian Shi, Georg Vogeler, und Mathias Seuret. „TorMentor: Deterministic Dynamic-Path, Data Augmentations With Fractals", 2022: 2707–11. https://openaccess.thecvf.com/content/CVPR2022W/ECV/html/Nicolaou_TorMentor_Deterministic_Dynamic-Path_Data_Augmentations_With_Fractals_CVPRW_2022_paper.html .

Perreaux, Nicolas. „Possibilities, Challenges and Limits of a European Charters Corpus (Cartae Europae Medii Aevi - CEMA)". *arXiv:2105.00932 [cs]* , 21. April 2021. http://arxiv.org/abs/2105.00932 .

ReF : „Referenzkorpus Frühneuhochdeutsch", [2018]. https://www.ruhr-uni-bochum.de/wegera/ref/ .

REM : Klein, Thomas; Wegera, Klaus-Peter; Dipper, Stefanie; Wich-Reif, Claudia (2016). Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0, https://www.linguistics.ruhr-uni-bochum.de/rem/. ISLRN 332-536-136-099-5.

REN : Schätzlein, Frank. „Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)". https://www.slm.uni-hamburg.de/ren.html .

Seuret, Mathias, Anguelos Nicolaou, Dalia Rodríguez-Salas, Nikolaus Weichselbaumer, Dominique Stutzmann, Martin Mayr, Andreas Maier, und Vincent Christlein. „ICDAR 2021 Competition on Historical Document Classification". In *Document Analysis and Recognition – ICDAR 2021* , ed. by Josep Lladós, Daniel Lopresti, und Seiichi Uchida, Cham: Springer International Publishing, 2021: 618–34.. https://doi.org/10.1007/978-3-030-86337-1_41 .

Simistira, Foteini, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, und Rolf Ingold. „DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts". In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* . 2016: 471–76. https://doi.org/10.1109/ICFHR.2016.0093 .

Tilahun, Gelila, Andrey Feuerverger, und Michael Gervers. „Dating medieval English charters". *The Annals of Applied Statistics* 6, Nr. 4 (Dezember 2012): 1615–40. https://doi.org/10.1214/12-AOAS566 .

Tilahun, Gelila, Michael Gervers, und Andrey Feuerverger. „Statistical Methods for Applying Chronology to Undated English Medieval Documents". In *Digital Diplomatics* , ed. by Antonella Ambrosio, Sébastien Barret, und Georg Vogeler. Köln: Böhlau Verlag, 2014: 211–24. https://doi.org/10.7788/boehlau.9783412217020.211 .

XLM-RoBERTa https://huggingface.co/docs/transformers/model_doc/xlm-roberta .

Vogeler, Georg (ed.). *Digitale Diplomatik* . Köln: Böhlau Verlag, 2009.