Review, renewal and relaunch of the web platform for Itinera Nova – a citizen science project of the City Archives Leuven and the Cologne Center for eHumanities

Sviatoslav Drach, Benedikte Löbbert, Claes Neuefeind (CCeH), Hadewijch Masure (Stadsarchief Leuven), Laurens Bastijns (Stad Leuven)

Introduction

The Itinera Nova project aims to digitize and transcribe the aldermen's registers and account books of the City of Leuven to make them publicly accessible. The registers and account books comprise a total of over 1,300,000 pages from the period 1362-1795. Following a citizen science paradigm, the project involves a community of volunteers that collaboratively work with a digital research platform, developed by the Cologne Center for eHumanities (CCeH) in close cooperation with the City Archives Leuven (see https://www.itineranova.be).

The focus of this contribution lies on the relaunch of the digital platform in June 2024. The platform was completely revised according to current standards and best practices in Digital Humanities. Extended search functionalities were developed for a better user experience, and the platform was embedded in the City of Leuven website for brand familiarity and swift access. The internal structure of the application was updated and new features were added to the transcription workflow, e.g. the integration of Handwritten Text Recognition (HTR).

About the project

The heart of the Itinera Nova project is a community of around 50 volunteers coordinated by a community manager at the City Archives Leuven. The volunteers are involved in all steps of the digitization process: the processing pipeline ranges from the creation of digital images and their archiving to the creation of metadata and full text transcriptions. For the transcription process, the Itinera Nova platform implements an editing and moderation system for revision and quality management. The digital images are made available to the volunteers via the digital work platform, where they first are indexed (i.e. each act is identified and enriched with metadata), and then transcribed, double checked, and published via the project website. Since the start of the project in 2009, about 150,000 acts have been indexed and around 95,000 were transcribed and published by the volunteers.

The development of such an extensive collection could hardly be conceived without their commitment within the framework of a citizen science project. The "club life" cultivated at the City Archives over the long term includes various events such as meetings in small groups, workshops and events in the Archives, at which the volunteers discuss the current project tasks, exchange ideas on specific topics, jointly develop their skills, clarify open questions or simply come together.

The Itinera Nova platform

Apart from community management, another major challenge lies in creating an appropriate digital infrastructure to support the volunteers. The project aims to provide high-quality scientific data in accordance with the FAIR principles (Wilkinson et al., 2016) and is therefore based on established data standards such as TEI-XML (https://tei-c.org). However, this poses a major hurdle for volunteers without the relevant technical expertise. Thanks to the easy-to-use editor which requires no knowledge of TEI-XML, volunteers are enabled to create transcriptions and to annotate them. The metadata currently only includes the most important information such as the name, date and language of the files. More comprehensive metadata was deliberately omitted in order to reduce the complexity for the editors. In addition, transcribed texts can include a fixed set of inline annotations such as abbreviation, superscript, unclear, deleted, new line and new page. More advanced annotations like person or place names were left out so far, a corresponding extension for inline annotations is currently being developed.

The Itinera Nova project relies on a project-specific, highly integrated solution that incorporates various technologies, frameworks and tools. Originally developed as an XRX web application (XForms, REST, XQuery) based on the XML database eXist (https://exist-db.org), the technical architecture of the application and its visual appearance were extensively revised in 2023. The aim of the current revision was to separate the technical components more clearly and thus achieve better maintainability, expandability and the option to integrate external software (Drach et al., 2021). This was achieved through a stronger separation of frontend and backend, modularization of software components and the implementation of an API interface (a RESTXQ API based on the XML-database BaseX, see https://docs.basex.org/wiki/RESTXQ), which was designed according to the OpenAPI specification (https://www.openapis.org). In addition, the data models were revised to allow for the annotation of persons and toponyms and to integrate results from automatic text recognition.

An important part of the revision is the integration of a HTR-workflow based on Transkribus (https://readcoop.eu/transkribus/) to accelerate the digitization process. Based on the existing manual transcriptions, several HTR-models were created for different time periods, mainly for the 15th century. These models show very good results in our preliminary tests, with character error rates of up to 3% (even though this period shows lots of abbreviations and multilingual content). However, since the registers and account books cover a large time span (more than 400 years), during which both the writing style and orthography have changed, new training data or models must be produced for various time periods (ideally for every few decades). After being processed with Transkribus, the texts are integrated directly into the transcription workflow as an optional starting point for editing. This way, HTR generated texts are already searchable before they are revised, but over time, can be reviewed by volunteers.

Conclusion and outlook

In our talk, we present the fully revised Itinera Nova platform and its new features, including extended search modes and the integration of HTR into the transcription workflow. Besides that, we give an outlook on further extensions of the platform, which are currently under

development. These include the annotation of toponyms (supported by NER-procedures), a map component to display the toponym information (on old and new maps), and further work on the account books, where HTR offers many opportunities (Bigalke et al., 2022).

Literature

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.

Drach, S., Sahle., P. und Neuefeind, C. (2021) „Anforderungen an Tools und Plattformen für kleinere und größere Transkriptionsprojekte anhand von zwei Anwendungsbeispielen.“ In: Partizipative Transkriptionsprojekte in Museen, Archiven und Bibliotheken – ein praktischer Erfahrungsaustausch. Hg. von Diana Stört, Franziska Schuster und Anita Hermannstädter. Oktober 2021. Berlin 2022, S. 55-60. doi: 10.7479/szm4-fs62.

Bigalke, J., Drach, S. und Neuefeind, C. (2022) „Semi-automatische Erschließung von Rechnungsbüchern am Beispiel des Stadtarchivs Leuven“. DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2022), Zenodo. doi: 10.5281/zenodo.6327963.