STUDIUM.AI: datafying and connecting the 'webs of knowledge' around the premodern University of Leuven (1425-1797)

Introduction

This paper's goal is to discuss the process of interlinking an array of thematically related datasets developed in the past years by different research teams, using different tools and data formats. The work presented here falls under the umbrella of the STUDIUM.AI research infrastructure (RI), aiming to break down silos between multiple datasets concerning the University of Leuven in the premodern era, also known as the 'Old University' (1425-1797). From a data- and AI-driven perspective, it sets out to 'datafy' historical webs of knowledge around a premodern educational institution. With the help of the KU Leuven Core Facility Trismegistos+, the RI aims to ingest 160,000 records from 8 participating projects into a single relational database developed for data sharing and additional data entry by KU Leuven researchers.

In this paper we focus on linking prosopographical resources with bibliographical ones and discuss the complexity of preparing both kinds of resources in view of matching them. First, we introduce the datasets; then we discuss the preparation and standardization of the names of the people at the Old University of Leuven, the largest prosopographical resource; followed by the modeling of the book-historical datasets. At the end, we outline the scientific and logistical challenges inherent in this kind of work.

Section 1. Datasets of the premodern University

The interest in the Old University of Leuven is a long-standing one. The University is a rather unique case, and has been the target of various digitisation initiatives since the designation of its archives as UNESCO documentary heritage in 2013. Consequently, a number of databases have been produced which capture different angles of the historical evolution of the institution. The datasets participating in the STUDIUM.AI FWO medium-scale research infrastructure are both prosopographical (matriculation records, scholars of the Old University) and bibliographical (Magister Dixit, Manuale Lovaniense, Lovaniensia & Collectio Academica Antiqua (henceforth Caa), Lllogeia, Dalet, Leonardi.DB). Here, we aim to discuss the challenges and opportunities resulting from the work to combine these heterogenous sources into a single database. The goal of STUDIUM.AI is to develop an essential "common core" of information which allows us to 1) match the overlapping records and 2) link to the full records in the original resources. Figure 1 shows a simplified schema of the relational database of the RI. The full version of the database will be hosted on a dedicated server acquired for the project.

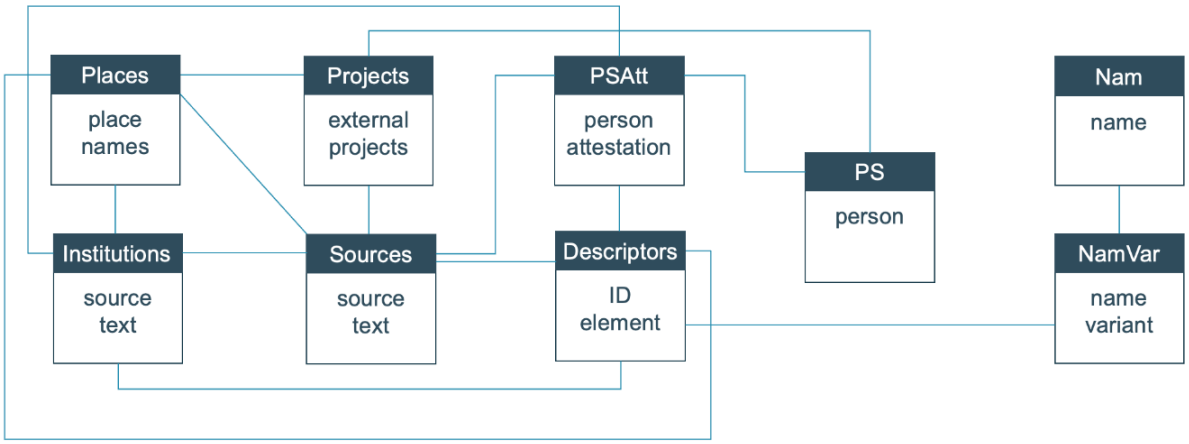

Figure 1: Relational schema of the different components of the Studium.AI RI resources

Section 2: Standardizing personal and place names

The integration of the largest dataset, the names in the university's matriculation registers, was preceded by a three-year coordinated effort between the State Archives and the STUDIUM.AI. Unlike the other datasets, where people are recorded with a standardized, often modern, version of their names, the entries of the matriculation records were transcribed verbatim. In the early modern period, there was no notion of a fixed spelling: names were often written down in a Latinized form representing what the scribe understood. Even common names appear in many forms: no less than 37 different spelling variants of the name 'Johannes' occur in the matriculation books. For example, in the matriculation records of 1560, Floris vander Haer, who studied at the Collegium Trilingue, was registered as 'Florentius ab Haer', while the Lovaniensia dataset lists his name as 'Floris Van der Haer'. Likewise, the author and professor Gommarus Huygens from Lier of the Caa is listed as 'Gummarus Huijgens Lyranus' in the 1673 matriculation register. Without preprocessing, automated interlinking would have been impossible.

To successfully map the overlap of individuals between the participating datasets, all related variants therefore need to be assigned to a common, standardized name. This is also the case for place names often added to a person's identification as origin designations. These standardized names can then be used to match the entries in the different datasets.

To the best of our knowledge, there is yet no comprehensive digital tool that collects all the information needed for this standardization. The most exhaustive collections available online (e.g. https://www.cbgfamilienamen.nl/nfb/index and https://nvb.meertens.knaw.nl for Dutch family and given names respectively) focus only on present-day variants. An exception is the NAMES project (https://www.clariah.nl/nl/projecten/names-dutch-corpus-of-person-name-variants) which has standardised name variants from nineteenth-century sources. We have obtained a digitized version of the dictionary of family names by Debrabandere 2003, where several historical variants are listed for many family names. We could thus automatically assign some 19,000 of the c.73,000 family name variants extracted from our datasets to a standard name in Debrabandere. The remaining pool is being processed semi-automatically using a matching process based on 'simplified strings' of the name variants where vowels are left out and certain consonants that are interchangeable in Dutch (e.g. 't' and 'd') are converted to a standard consonant. The specifics will be shared during the presentation. Currently (January 2024), 5,000 additional variants have thus been assigned. For the small pool of given names, the c.4,000 variants were grouped manually, consulting onomastic tools such as Debrabandere 2003 and the extensive collection shared on https://www.behindthename.com. As a result, the original dataset has been successfully imported and reworked within the infrastructure.

Section 3: Books at the Old University

High-quality, structured bibliographical data is becoming central to the work of some book historians and those interested in digitised book collections (Lahti et al. 2019; Tolonen et al. 2022). To date, we have ingested the bibliographical data from the datasets Manuale Lovaniense (576 records), the Caa (3,796 records) and Lovaniensia (942 records). Data management proved crucial for this step. Metadata created within KU Leuven library catalogues are frequently updated, and therefore a careful administration of the information is necessary to refer to the original records and to handle updates within STUDIUM.AI. Our database is designed to be citable and reproducible from both the data management and end-user perspectives. All ingested source files (and any associated scripts for data cleaning and extraction) are versioned and stored in an active research repository managed by KU Leuven (ManGO), and these files are referenced in the STUDIUM.AI database as metadata. Later in the project a web-based front-end and API

will be developed, and these will allow for dynamic citations, meaning that it will be possible to make a citation which includes both a query and a specific version of the database, following the DataCite guidelines. This ensures that analysis done using the STUDIUM database will be reproducible even if the database itself is updated.

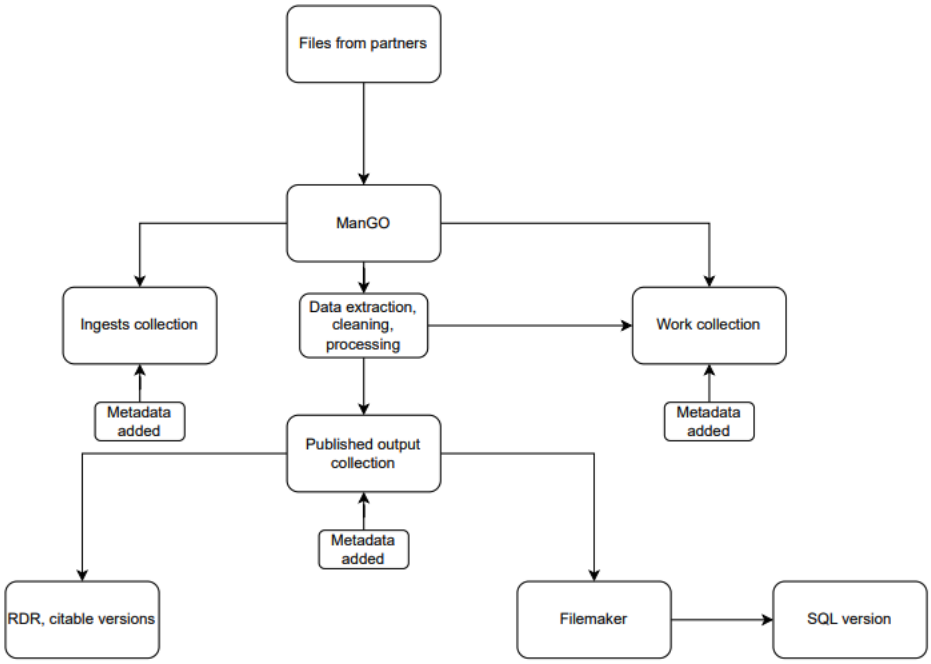

Figure 2: Data management workflow for bibliographical and prosopographical resources developed at KU Leuven

Based on this solid workflow, we proceeded to the extraction of information. The partner data comes in a variety of file formats and metadata standards, including flat .csv files with Dublin core metadata, JSON files, and MARC XML. In each case, we identified a 'common core' set of basic metadata found across each. Regarding the book sources, there are numerous challenges related to the data curation, for example with identifying places from a mix of historical and standardized descriptions, extracting correct dates, recording partial or uncertain information, and recording information in a simplified format where it has come from more complicated schemas e.g. XML or JSON. The MARC tags for publication places are in some cases standardised and in others the information from the imprint is transcribed verbatim, necessitating cleaning. Each unique string used as a place name description was exported to a separate file and manually matched either to existing places from the Matriculation Registers, or to Geonames and Wikidata authority files. E.g. for Caa, 673 separate strings are used in the place name field, which when cleaned resulted in 152 distinct places. During the presentation we report on the metadata mapping and the procedure developed to effectively extract the information.

Conclusions: Challenges and Hopes

The work on the STUDIUM.AI RI started with high hopes of breaking silos, but the way is paved with challenges that require constant adjustment of the work towards shared solutions: the complex collaborations with the source partners to enable transparent referencing to the original resources; conceiving a common core that can accommodate very diverse datasets; the divergent skill sets (historical and digital expertise) needed for the intense data cleaning that is crucial for data harmonization; and

finding and keeping the technical expertise required for server management (in order to host the backend and frontend databases).

The benefits of these efforts, however, emerge clearly from what is presented above. Data management practices are also improved for the source partners, ensuring proper referencing of the data. The information is enriched and made more findable via standardization and cleaning. The collaboration among different partners offers a clearer view of the shortcomings and strengths of both the original projects and the STUDIUM.AI common core. With this presentation we hope to enrich the discussion on 'breaking silos' with the experience developed within the project.

Works and databases cited

"Citation of Dynamic Datasets." *Datacite*, https://datacite-metadata-schema.readthedocs.io/en/4.5/guidance/dynamic-datasets/.

DaLeT. https://www.dalet.be/. Accessed 30 Jan. 2024.

Jannis, Suzanne, et al. *Belgian State Archives / State Archives of Leuven / Rijksarchief Leuven - Databank van Personen Ingeschreven in de Matrikels van de Oude Universiteit Leuven, 1426-1797*. Social Sciences and Digital Humanities Archive – SODHA, 2020. *DOI.org (Datacite)*, https://doi.org/10.34934/DVN/PWI0KR.

Lahti, Leo, et al. "Bibliographic Data Science and the History of the Book (c. 1500–1800)." *Cataloging & Classification Quarterly*, vol. 57, no. 1, Jan. 2019, pp. 5–23. *DOI.org (Crossref)*, https://doi.org/10.1080/01639374.2018.1543747.

Logical Geometry - Leonardi. https://logicalgeometry.org/leonardi/. Accessed 30 Jan. 2024.
Lovaniensia. https://lovaniensia.be/. Accessed 30 Jan. 2024.

Magister Dixit Project. https://www.kuleuven.be/lectio/research/MagisterDixit. Accessed 30 Jan. 2024.

Tolonen, Mikko, et al. "The Anatomy of Eighteenth Century Collections Online (ECCO)." *Eighteenth-Century Studies*, vol. 56, no. 1, Sept. 2022, pp. 95–123. *DOI.org (Crossref)*, https://doi.org/10.1353/ecs.2022.0060.