

STUDIUM.AI: Datafying and Connecting the ‘Webs of Knowledge’ around the Premodern University of Leuven (1425-1797)

Yann Ryan¹, Margherita Fantoli¹, Yanne Broux¹, and Violet Soen¹

¹KU Leuven

This paper presents the design and implementation of STUDIUM.AI, a research infrastructure serving as a central hub for studying the academic ecosystem of the early modern University of Leuven (1425–1797). We begin by introducing the project’s core datasets—which document scholars, students, and books associated with the university—and outline the relational database architecture that interconnects them. A critical focus is our standardization pipeline for historical names and places, a preliminary step enabling robust data interlinking. Next, we address research data management challenges in this large-scale initiative, detailing our use of KU Leuven’s MANGO platform for storing active research data while adhering to FAIR principles (Findable, Accessible, Interoperable, Reusable). Finally, we demonstrate best practices for linking to external authority files, exemplified by our integration with the World Historical Gazetteer.

Keywords: Early Modern University of Leuven, Data interlinking, Name standardization, Research Data Management, FAIR Data, World Historical Gazetteer

1 Introduction

Founded almost 600 years ago, the early modern University of Leuven (1425-1797) turned into a thriving hub of students, professors, scholars, printers and publishers. These left us a vast array of traces and sources, be it their names noted down in the matriculation registers, the many extant student notebooks, or the impressive local production of handpress textbooks and treatises. Even if many of these historical sources perished in the fires of the first and second world war, in 2013 the UNESCO Memory of the World program recognized the remaining *Archives of the University of Leuven (1425-1797)* as *University Heritage of Global Significance*.¹ In the last decade, libraries and archives in Leuven, Louvain-la-Neuve and Brussels have digitized a

¹ <https://nieuws.kuleuven.be/en/content/2013/unesco-adds-university-archives-heritage-register>. Accessed 6 Oct. 2024.

significant portion of the surviving sources to valorize this exceptional intellectual and cultural heritage. At the same time, digitization sparked the creation of new metadata and search interfaces, whereas different research groups at KU Leuven started collecting prosopographical and book historical data from a DH-perspective.

By approaching the intellectual heritage and historical infrastructure of education of the early modern University of Leuven from different perspectives, the datasets bring about a new picture, yet one that can only be harvested across multiple platforms. This fragmentation of the data, both at the moment of its creation and through its digitized surrogates, prevents users from connecting information relating to the same actors and objects across the different datasets. For example, the famous humanist Erasmus of Rotterdam (?-1536) pops up in no less than 5 datasets, as he matriculated in 1518 and authored multiple treatises as well as textbooks. The Research Infrastructure STUDIUM.AI aims to bridge this gap by interconnecting this large array of materials, thus enabling joint searches across different platforms, and by enriching these data through linking and standardising information. footnote <https://studium-ai.org>. Accessed Oct. 6 2024. The Research infrastructure revolves around two main initiatives: on the one hand, the interlinking of several databases (on which this paper is based), and, on the other hand, the enrichment of the current resources by training an HTR-model aimed at transcribing the content in already digitized notebooks of students of the University in the premodern era (which will not be discussed here). Our paper discusses three main research questions relevant to the set-up of interlinking research infrastructures, namely:

- RQ1: Given the variety of the digitized sources and the objects of the databases concerned, what approach can be taken to interlink the data and valorize their thematic unity?
- RQ2: What workflow ensures the best transparency, reusability, and synchronicity of the data within this kind of interlinking infrastructure?
- RQ3: How can such a diverse infrastructure integrate permanent identifiers from external resources?

The paper is structured as follows: first, we outline the digital resources in the field of premodern university studies and the existing research and materials in the specific case of the early modern University (Section 2); second, we describe the ten datasets that we have currently integrated into the research infrastructure (Section 3). We then outline our current approach to interlinking the data (RQ1) (Section 4) and to ensure a stable workflow between the different partners in the project (RQ2) (Section 5). Finally, we illustrate the process and the advantages of integrating permanent identifiers to enrich our information (Section 6).

2 Literature Review

‘Datafying’ prosopographies of premodern universities is certainly not new. Inspired by the approach of serial sources pioneered by the French ‘School of the Annales’, statistical accounts of aspects of university life and recruitment were trending in the 1960s and 1970s, such as Jacques Verger’s work on the French universities, based on the *Suppliques* held by the Vatican Archives Verger (1970). As early as 1980, the digitization (entering the information in a database) of A. B. Emden’s *Biographical Registers* of

Oxford and Cambridge was carried out by the Oxford University Computing Service, a resource which was used in turn for writing new histories of the University (Aston, 1980).

Most recently, this digital prosopographical work has witnessed a resurgence using newer methods in the field of Digital Humanities. Premodern universities have been at the centre of several research projects.² For example, the *RAG-Repertorium Academicum Germanicum* (Deutsche Forschungsgemeinschaft 2011-9) set the tone to detect geospatial clusters in the late medieval population of c. 62,000 students in the medieval Holy Roman Empire (Gubler, Kaspar, 2022; Schwinges, 2020), including Louvain up to 1527 (and some cases up to 1550). The data are stored in a relational database that is browseable online following different facets, and dynamic visualization guides the user in data exploration. The database is currently developed using NodeGoat, which is also used to produce the visualizations. Secondly, the *SP-Studium Parisiense* prosopographical database aims at gathering information for ca 40,000 people belonging to the schools and universities of Paris between the 12th and 16th centuries (Genet et al., 2016).³ Another database in development is the *Universitas Magistrorum*, which seeks to record various attributes of the professors at the University of Prague between 1458 and 1622 (Synovcová Borovičková and Škudrnová, 2021). At the European level, the project *RETE-Repertorium Eruditorum Totius Europae* under the direction of David de la Croix gathers information about scholars active in European universities between 1200 and 1800. RETE aims to describe the evolution of the learned society of Europe in multiple respects, for instance by analyzing the network of universities resulting from the movements of scholars from one institution to another; or by devising a measure of the Human Capital Index based on the publication output of historical scholars.⁴ Similar projects, aiming at reconstructing the networks around historical universities, have been developing recently (Brizzi and Frijhoff, 2018). Such projects share specific goals and challenges, and the European Network Atelier Héloïse was set up to bridge the different “national” projects in the broader framework of the European Digital Academic History.⁵

Most of these initiatives could thrive through wider initiatives relating to the digitization of cultural heritage to make it more accessible. Projects in this vein have attracted large-scale funding in recent years, such as *Europeana*⁶ and the *Towards a National Collection* project in the UK.⁷ These kinds of projects have employed the use of metadata, and now, increasingly, the linking of records and entities through the use of linked open data to make records interoperable and searchable across different archives and repositories. For early modern history, many of these projects working on interlinking records have focused on correspondence data because of its relatively easily interoperable metadata (i.e. sender, recipient, dates and places of sending and receipt). The COST action project *Reassembling the Republic of Letters 1500 – 1800* and its associated volume Hotson and Wallnig (2019a) summarised many of the challenges and opportunities in modelling and linking correspondence data from different projects and corpora, a task also taken up by the *SKILLNET* project (Van Miert, 2022),

² A list can be found here: <https://heloise.hypotheses.org/base-de-donnees>. Accessed 6 Oct. 2024.

³ <http://studium.univ-paris1.fr/>.

⁴ The RETE database can be browsed here: <https://shiny-lidam.sipr.ucl.ac.be/scholars/>. Accessed 6 Oct. 2024.

⁵ <https://heloise.hypotheses.org/>. Accessed 6 Oct. 2024.

⁶ <https://www.nationalcollection.org.uk/>. Accessed 1 Oct. 2024.

⁷ <https://www.europeana.eu/en>. Accessed 1 Oct. 2024.

the LettersSampo project at the University of Helsinki, and *Early Modern Letters Online* (Hotson and Wallnig, 2019b).⁸ This has been paired along with wider thinking about the value of research infrastructure and its relation to cultural heritage, particularly as manifested in libraries (Waters, 2023).

When narrowing the focus on the University of Leuven in the premodern era, compiling data on the academic community and book productions has been the focus of a broad array of scholarly initiatives for decades. Already in the beginning of the twentieth century, Louvain archivists and historians teamed up to edit the matriculation registers (Reusens et al., 1903-1990) as well as other source documents of the first four centuries of the university's existence.⁹

Recently, as we describe in more detail in Section 3, several projects have carried out systematic digitizations of (re)sources about the Old University. Projects can be grouped into two categories: the first one focused on the digitization of the sources, whether handwritten notes of Louvain students (as in the portal site of *Magister Dixit*¹⁰), thesis sheets they had published for defense (as in the digitized curated collection of the KU Leuven libraries¹¹), the published work of Louvain professors and scholars (as in the portal site of *Lovaniensia*¹²), or the digitization of the matriculation registers (and the additional compilation of an Excel list of names registered in these sources). The second type of dataset is related to research projects with a specific focus on data collection about research questions, such as the printed textbook production in Leuven, as in *Manuale Lovaniense* (Cammaerts, 2024)¹³, the student notes on the topic of logic as in *LLLogeia* (see for instance Geudens et al. (2020)), the persons and books at the Collegium Trilingue, as in *DaLet*¹⁴, or the Aristotelian Diagrams¹⁵, as in Leonardi's subsection on Louvain (see Demey (2024)). Besides the data creation, the project also prompted the beginning and completion of several PhD theses and research projects investigating the teaching activity at the early modern University (see for example the research projects on Logic by Geudens and Masolini (2016), the Louvain Trilingue within the *DaLet*-team, Feys et al. (2025) forthcoming, and the 'higher' Faculties of Canon Law, Civil Law and Theology in @Aulam, Soen et al. (Forthcoming)).

The novelty of the STUDIUM.AI infrastructure is twofold: on the one hand, we aim at matching databases and metadatasets of the two groups, to allow the users to jointly navigate across prosopographical information, and the information/digitized version of the documents which testify to the scholarly and student activity of the community around the premodern university. This part of the workflow is discussed in the rest of the paper. On the other hand, by exploiting HTR and NLP techniques (which are not discussed in this paper), we want to enrich the above resources in a semi-automated manner, by linking the information of the databases to the people and places mentioned in the documents.

⁸ Early Modern Letters Online. <http://emlo.bodleian.ox.ac.uk/home>. Accessed 7 Oct. 2024; LettersSampo. <https://lettersampo.demo.seco.cs.aalto.fi/en>. Accessed 5 Oct. 2024.

⁹ Especially the editions of the matriculation registers.

¹⁰ <https://www.kuleuven.be/lectio/research/MagisterDixit>. Accessed 6 Oct. 2024.

¹¹ https://kuleuven.limo.libis.be/discovery/collectionDiscovery?vid=32KUL_KUL:KULeuven&collectionId=81531994440001488.

¹² <https://www.lovaniensia.be>. Accessed 6 Oct. 2024.

¹³ <https://www.odis.be/hercules/search2.php?histid=1003367>, URL generated on 7 Oct. 2024.

¹⁴ <https://www.dalet.be/>. Accessed 6 Oct. 2024.

¹⁵ <https://leonardi.logicalgeometry.org>. Accessed 6 Oct. 2024.

3 Data

More recently, both GLAM institutions (such as KU Leuven Libraries and the State Archives) and academic research groups have invested significant resources in the digitization of sources and prosopographical information. The State Archives, through the impressive dedication of volunteers, digitized the university's matriculation records, where any person related to university (mostly students, but also professors and publishers) were registered, and also populated an Excel-list on the basis of the twentieth-century edition of Reusens, Wils and Schillings. A second, cleaned and enriched version (cross-checked again with the sources), was published as a joint effort between the State Archives, the interdisciplinary KU Leuven Research Institute LECTIO and an intern of the KU Leuven Advanced Master of Digital Humanities, Lydia Janssen (Janis et al., 2020). Within the KU Leuven, Magister Dixit, a joint project of KU Leuven Libraries and the Research Center LECTIO (now a KU Leuven Research Institute) has resulted in the—still ongoing—digitization of more than 500 manuscripts containing student notes of the early modern University, together with the creation of a metadata catalogue currently available on Zenodo (KU Leuven Libraries, 2024). Moreover, a few data-creation groups, used the platform ODIS – a digital spinoff of KADOC, Documentation and Research Centre on Religion, Culture and Society – to set up relational databases encoding information on academic prints in the Low Countries of the Early Modern Period, and on the actors (authors, publishers, printers and engravers) who contributed to the books. More precisely, they are *Manuale Lovaniense*,¹⁶ and *Scholars@OldLouvain*.¹⁷ KU Leuven Libraries has focused on printed materials, too. The *Collectio Academica Antiqua* (CAA) is a “reconstructed” collection, whose goal is to gather all printed books authored by members of the premodern University. The metadata, in the form of MARC21 XML, has been released as open data. The CAA currently offers the basis for a large-scale digitization project *Lovaniensia* whose volumes were recently OCRed via a partnership with Google Books.¹⁸ The ‘Theses from the Old University of Leuven’ collection is a set of 3,157 broadsheet academic dissertations, written by students of the Old University, kept and digitized at KU Leuven.¹⁹ At the intersection of manuscripts and prints, the DaLeT database records the students’ manually written notes on the side of printed books used at the *Collegium Trilingue*, an academic center founded in 1517 in Leuven where Ancient Greek, Latin and Hebrew were taught, with a curriculum inspired by Erasmus’ humanistic perspective (cf. (Papy, 2018)).²⁰ Finally, the *Leonardi.DB* linked data store preserves all Aristotelian diagrams published across time: this also includes diagrams preserved in students’ notes of the Old University.²¹ Table 1 offers a schematic overview of the materials available. Note that a ‘record’ in each case counts a different object: those from catalogues count individual collection items, such as a book or thesis broadside. Databases from datasets, such as the matriculation records count rows, either persons or others.

¹⁶ <https://www.odis.be/hercules/search2.php?histid=1003367>, URL generated on 7 Oct. 2024, see Cammaerts 2024.

¹⁷ The list is currently in final preparation by An Smets and Violet Soen, and will be deposited on KU Leuven RDR as well as made available through the project’s website: <https://www.kuleuven.be/lectio/research/aulam>.

¹⁸ <https://lovaniensia.be>. Accessed 6 October 2024.

¹⁹ https://kuleuven.limo.libis.be/discovery/collectionDiscovery?vid=32KUL_KUL:KULeuven&collectionId=81531994440001488&sortItemsBy=date_a.

²⁰ A partial export of the database can be found at: <https://zenodo.org/records/8305741>.

²¹ ‘Leonardi.DB’. <https://leonardi.logicalgeometry.org/>. Accessed 3 October 2024.

Table 1: Databases linked by the STUDIUM.AI infrastructure.

Database	Types of records	# records	Format
Matriculation books Louvain (MBL)	Matriculation records	143303	Excel spreadsheet
Manuale Lovaniense	Printed books	509	Relational DB/JSON exports
Scholars@OldLouvain	People	300	Relational DB/JSON exports
Magister Dixit	Manuscripts	599	MARC XML
Collectio Academica Antiqua	Printed books	2056	MARC XML
Lovaniensia	Printed books	2057	Dublin core Omeka export
Thesis sheets	Printed texts	3157	MARC XML
DaLeT	People + students notes	795	Filemaker relational DB
Leonardi.DB	Aristotelian diagrams	>6000	Linked Data
Lllogeia	Manuscripts	87	Filemaker relational DB

These data are each time gathered from particular source sets, which are historical (like in the case of the matriculation books) or artificial (like the CAA curated at the KU Leuven libraries' Special Collections divisions). This comes with some disadvantages: for the matriculation books, two books have been lost in the time, leaving lacunae for the years between 1569 and 1616 (a third volume was lost in the first World War, but as it had already been turned into a printed edition by Schilling and others, the information has survived). The thesis sheets are inevitably 'ephemeral' print which was meant as an invitation, and were often recycled afterwards as scrap paper, etc. The survival rate is very uncomplete and often depends on capricious finds in the archives (Soen and Fantoli, Forthcoming). Concerning the data contained in the 'artificial' CAA, recent research has shown that by the sheer amount of rare books collected, quantitative information deducted from the metadata corresponds significantly with information from 'gold-standard'-catalogues, like the Universal Short Title Catalogue (Scebba and Fantoli, Forthcoming). DaLeT, Leonardi.DB, LLlogeia and Manuale Lovaniense, however, are datasets made by researchers with the aim of exhaustivity, and as such, they are themselves gold-standard, as being entered and surveyed by experts in the field. The interlinking of a series of datasets (and getting them out of 'their silo') each times generates new persons, new names or name variants, as detailed below in 4.2. This not only creates larger data pools, but also helps to correct some of the earlier biases in scholarship and source records, such an almost exclusive focus on the propedeutic Faculty of Arts, which has now been flattened out in comparison with the 'higher' Faculties of Canon Law, Civil Law and Medicine (Soen and Ryan, Forthcoming).

4 Interlinking the datasets

The pivotal information that bridges all the datasets, both prosopographical and book historical, is the learned community revolving around the Old University of Leuven, i.e. the people. Book historical resources contain detailed information about the books' contributors, who were often employed at the University of Leuven. Furthermore, the Magister Dixit dataset lists the students taking the notes and the professors giving the classes, who are also recorded in the matriculation books and in the relational datasets such as DaLeT and Scholars@OldLouvain. Hence, in order to partially automate the matching across datasets, we identified a common core of prosopographical information (names, place of birth, place of death, and dates related to biographical and professional events) that we could extrapolate from the different datasets in order to match the attestations of individuals. The relational database hosting the 'common core' has been modelled from scratch and revolves around the distinction between the level of attestations of people in the source, and records of people connecting the different attestations, as described in detail in Broux (2024a).

As expected based on the different provenance and focus of the datasets, the information about people was recorded in very different ways. Some of the databases were developed by researchers (Manuale Lovaniense, Scholars@OldLouvain, DaLeT), who generally agreed on a set of (loose) standardization guidelines (standardization with the Latin name of the scholars/students, structured encoding of dates etc.). Others, such as the library metadata, were set up by expert librarians in endeavours spanning several years. In addition, priority was given to a faithful representation of the information printed in the books, which resulted in a low degree of standardization. To give only one example, the Abbot of Tongerlo from 1608–1629, who appears in the CAA metadata as a dedicatee, is listed as both 'Adriaan Stalpaerts' and 'Hadrianus Stalpaert' there.

We decided to start the standardization process from the matriculation records (MBL henceforth, for Matriculation Books Leuven), which reflect exactly what was transcribed in the critical edition of Reusens, Schillings et al. (1903), without any standardization procedure for the first and last names of students and professors. This choice was based on the observations that the MBL are by far the largest dataset available with over 143,000 entries spanning the four centuries of the Old University, and that they represent a valuable source of variants of given names and last names. In addition, it is highly unlikely that duplicate entries will be found for the same person since every university member was registered only once (and re-matriculation remained the exception to the rule). This means that there is an approximate 1:1 correspondence between records and individuals; in the bibliographical datasets, on the contrary, people may appear multiple times.

4.1 Standardizing names for matching people

This section provides a general overview of the process of standardizing names of persons, which is described in full in Broux (2024b). In the Excel file of the MBL, the student(s) were described with the fields 'Voornaam' (given name), 'Naam' (family name), 'Herkomst' (origin), and 'Bisdome' (diocese). The first step to ensure a proper standardization of names and the identification of individuals of the MBL consisted of categorizing the parts of the string identifying the full name of a matriculated person into several descriptors such as: first name, middle name, last name, patronym, function and origin. For instance, for the name 'Carolus Josephus Gummarus Bosmans

Lyranus’, the elements are classified as first name (Carolus), middle names (Josephus, Gummarus), last name (Bosmans), origin (Lyranus, from the Brabant city of Lier). In some cases, ambiguity is registered when a component of the name could be interpreted in a double sense, for instance, both as a patronym or a last name. In the enrollment by ‘Franciscus Jacobi Balduini Avesnensis’, the string ‘Balduini’ could plausibly be either the last name or the patronymic of the father Jacobus, i.e. Franciscus’ grandfather (see Broux (2024b) for an in-depth discussion of these cases of ambiguity). For the standardization and interlinking process, we kept given names and last names as separate categories for two reasons, even though several names can be used as both. On the one hand, from the onomastic perspective (since the infrastructure aims at grouping together variants of the same name under a same ID, i.e. at lemmatizing the attestations of names), the lack of distinction between first names and last would generate noise for the cases in which a first name and a last name have the same form. For example, all the variants of the last name ‘Frans’ (e.g. Franssen(s), Frandsen, Fransing) would be integrated as variants of the first name ‘Frans’, which does not correspond to the historical reality behind the data. On the other hand, since the identification of duplicate individuals in the datasets relies on the matching of the full name, this distinction avoids the conflation of, for instance, an individual named Frans Jacobs and one named Jacobus Frans (see Broux (2024b)). Hence, the next paragraphs detail the two paths taken for the standardization of first names and family names (the standardization of place names identifying the origin is discussed in detail in Section 6).

4.1.1 Standardizing family names

The MBL dataset contains 72,701 unique variants for family names and patronyms, for a total of 146,618 attestations. Their standardization relied mostly on the third edition of the dictionary of family names for Belgium and the North of France, prepared by Frans Debrabandere (*Woordenboek van de familienamen in België en Noord-Frankrijk*, hereafter referred to as WFN) that we converted into a relational database, stored within the Filemaker infrastructure.²² More precisely, we included all the variants of the family names listed by WFN, and then grouped them under the lemma proposed by WFN. Only 25% of the family names listed in the MBL (i.e. 18,274 out of the 72,701) had a precise match with one of the variants of WFN, hence we relied on a fuzzy matching approach to speed up the manual lemmatization of the remaining variants. To ensure an efficient matching, we developed a key-collision approach, or “fingerprint” approach, consisting of the simplification of the string to its essential components, both for the WFN variants and for the MBL attestations. For instance, vowels, articles, particles are removed from the string, and we clustered some interchangeable components (the details can be found in Broux (2024b)). In this way, the FileMaker interface allows team members to easily select the right WFN lemma for each MBL attestation. Once a lemma has been chosen for one MBL attestation, it is automatically assigned also to all the MBL attestations whose form is considered a variant of this

²² Debrabandere 2003. The author kindly shared a Word file of the 2014 version, available online (<https://www.cbgfamilienamen.nl/nfb/aanhangsels/wfb-voorwerk.pdf>. Accessed on 5 October 2024.). The conversion into a relational dataset was done by the STUDIUM.AI co-pi Mark Depauw. Since when we started the standardization, we were not available of the datasets for the standardization of Dutch names provided by NAMES project (<https://www.clariah.nl/projects/names-dutch-corpus-of-person-name-variants>), by Gerrit Bloothoof, and in order to avoid possible confusion coming from the implementation of two different standards, we did not use the NAMES lemmatization.

lemma by the WFN (see subsection 4.3 for a discussion of our “human in the loop” approach). As the example by Broux (2024b) illustrates, the last name Bachusius is lemmatized with the WFN last name Bachus. All the variants of this last name present in the MBL database (Bachus, Bachuijs, Bachuys, etc.) can then easily be assigned to the same lemma. To this day, 28, 869 unique variants have been processed, and they account for 66.5% of all the attestations of the MBL, while the remaining ones are each very rarely attested (once or twice) because we made the choice to first process the frequently attested ones. The standardization process resulted in the attributions of unique identifiers at the level of both variants and lemmas (resp. NamVar and Nam ID), and to every element of the original name (Descriptor ID).

4.1.2 Standardizing first names

Given names display much less variation. The 3,783 unique variants account for 170,100 attestations of given names. Because of the relatively small number of variants, we proceeded with a manual approach to the clustering, supported by the consultation of reference lists such as those found in the *Nederlandse Voornamenbank* of the Meertens Instituut,²³ the *Oxonian A Dictionary of First Names* (Hanks et al., 2006), the WFN and *Behind the Name: the Etymology and History of First Names*,²⁴ a website which collects names from various traditions and sources and has a system of manual verification for each new entry. Because each of these resources singularly only covered a small percentage of the MBL first names, and adopted different criteria in terms of grouping variants together, we did not systematically follow the lemmatization of one resource (as we did for the last names). However, the dataset downloaded from the website Behind the name, could be used to enrich our given name variants to include the gender automatically. At the end of the manual process of clustering, 2,271 distinct lemmas were identified for the given names, and received a unique identifier (as did the variants). In Figure 1, we summarize the workflow for standardizing the MBL names. Place names are discussed in Section 6.

4.2 Integrating other datasets

The benefit from of the standardisation and linking of names is widely recognized, something which is particularly important when working with digital networks, for example (Ahnert and Ahnert, 2023; Finegold et al., 2016). For the other nine datasets, a fairly uniform procedure was implemented to standardize the data, with some deviations depending on the particular data formats. There are a number of different approaches which can be taken to match. A fully automated record linkage pipeline, using software such as DeezyMatch, which uses a deep learning approach to fuzzy record linkage by generating word vectors which can be used for candidate matching and clustering, is one possible approach (Hosseini et al., 2020). Another example of an automated approach is the Entity Matching Tool developed as part of the European Holocaust Research Infrastructure consortium.²⁵ Another widely used automated approach is rules-based, for example matching various combinations of names and variants, sometimes taking into account spelling and name variations (Hill et al., 2019). This can include considering ‘edit distances’ or phonetic spelling to find fuzzy matches. However, these methods, particularly when they involve unsupervised or

²³ <https://nvb.meertens.knaw.nl/>. Accessed 6 Oct. 2024.

²⁴ <https://www.behindthename.com>. Accessed 6 Oct. 2024.

²⁵ <https://emt.ehri-project.eu>.

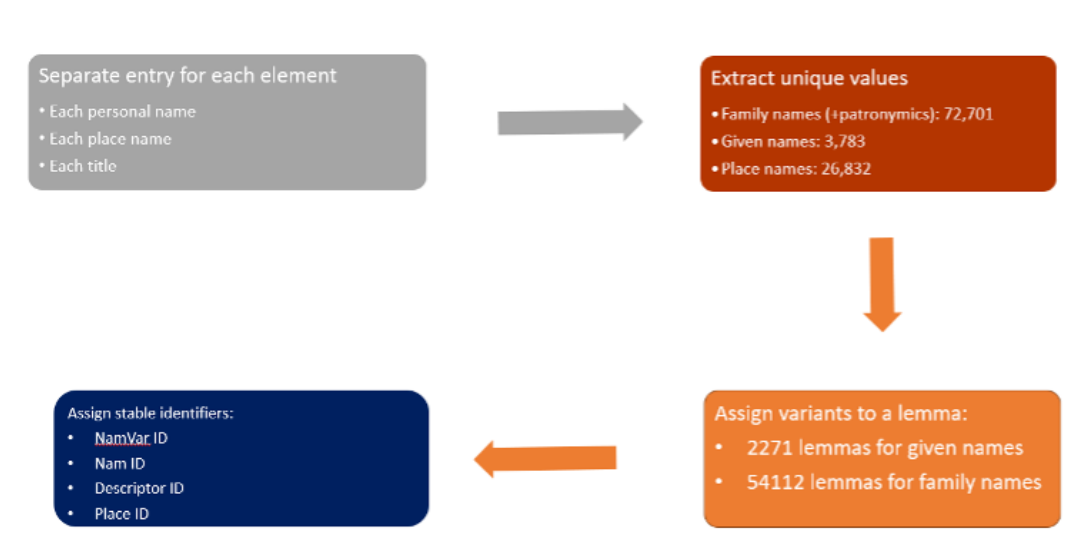


Figure 1: The essential steps for standardizing the MBL names.

semi-supervised clustering algorithms, are not necessarily suitable for a database meant for information retrieval, where naming incorrect matches is more detrimental than missing correct ones, in other words, precision is more important than recall.²⁶

To this end we implemented a rules-based approach without fuzzy matching, augmented with a ‘human in the loop’ workflow. Since these datasets generally start from the source record (i.e. a printed book, a thesis sheet, a bounded manuscript with student notes), the data about persons they contain is subsidiary and, in most cases, does not include a disambiguating person level, only a person-related-to-source, i.e. person attestation, level.²⁷ This did not require much additional work in most of the datasets, apart from standardizing the event date and the role the person had in relation to the event. The second step consisted of creating preliminary person records. As explained above, most of these datasets have already implemented a limited form of name standardization. This has the benefit that it is fairly straightforward to create an overarching person level for each dataset simply by grouping together those name strings with an exact match under a (temporary) unique ID (definitive IDs were only awarded upon import in the live STUDIUM.AI database). Since the standardization was not always carried out rigorously, the matching was done by using a concatenation of the name string + the date string of a person, e.g. ‘PhilippusvanDormael1611-1635’.²⁸ Only those attestations with an exact match were grouped together under a single person ID. If other attestations exist with the exact same name but a slightly deviating date (or no date at all), these were kept apart as distinct individuals to be checked during the final disambiguation round involving all datasets (as described in Subsection 4.3). The third step involved creating the Descrip-

²⁶ See for instance Shoilee et al. 2024.

²⁷ DaLet and Manuale Lovaniense are the two exceptions here. The former assigned its own Person IDs. The data of the latter is stored in ODIS, a research infrastructure used by various research projects where person disambiguation has already taken place; its person attestations are therefore linked to ODIS’ stable person IDs.

²⁸ Spaces were removed as they could potentially cause matches to be overlooked by creating meaningless variation (van Dormael is the same name as Vandormael) or by input errors such as double or trailing whitespaces.

Table 2: Number of new name variants and new names per dataset.

Database	New given name variants	New family name variants
Lovaniensia	421	1,211
CAA	332	1,049
ODIS	35	210
Magister Dixit	42	100
Thesis sheets	32	142
DaLet	106	154
LLogeia	[to be processed]	[to be processed]
Leonardi	[to be processed]	[to be processed]

Table 3: Number of new records per table in STUDIUM.AI's People section for each dataset. PS indicates unique persons, while person attestations their mentions in each database.

Dataset	PS	PSAtt	Descriptors
MBL	143,303	143,303	474,771
Lovaniensia	2,415	6,465	14,889
CAA	3,070	7,311	16,529
Manuale Lovaniense	326	1,377	2,927
Magister Dixit	620	1,605	3,793
Thesis sheets	6,538	5,305	21,296
DaLet	795	1,776	3,385
LLogeia	[to be processed]	[to be processed]	[to be processed]
Leonardi	[to be processed]	[to be processed]	[to be processed]
Total	157,067 (pre merge)	167,142	537,590

tor records. For this, the identification string listed for each person's attestation had to be broken down into distinct elements, mostly names, but sometimes also an origin or a suffix used for homonyms. As with the matriculation data, a distinction was made between first names, middle names, patronymics, and family names. The attestations of the first two categories were matched with the given name variants extracted from the matriculation records, while those of the last two categories were matched with the family name variants. In case of a match, the corresponding NamVar and Nam IDS were imported into the working file. Name variants without a match needed to be integrated into the existing setup. The Nam table, recording the lemmatized version of name variants, was sifted for a suitable candidate. If one was found, the new variant was linked to it; if not, a new standard name was created for it. This was repeated for every next dataset that was processed (see Table 2).

Once each descriptor was linked to the correct name or place ID, the dataset was ready for import. Table 3 gives an overview of the number of records in each table created from the files provided by the various participating projects.

The workflow for ingesting these datasets is described in Figure 2.

4.3 'Human in the loop' matching process

As well as this rules-based matching, the project is undertaking to match more ambiguous cases by relying on input from scholars. The human-in-the-loop approach, or semi-automated one, where automatic methods such as string matching are checked manually by experts, sometimes including 'crowd-sourcing', is a good solution where

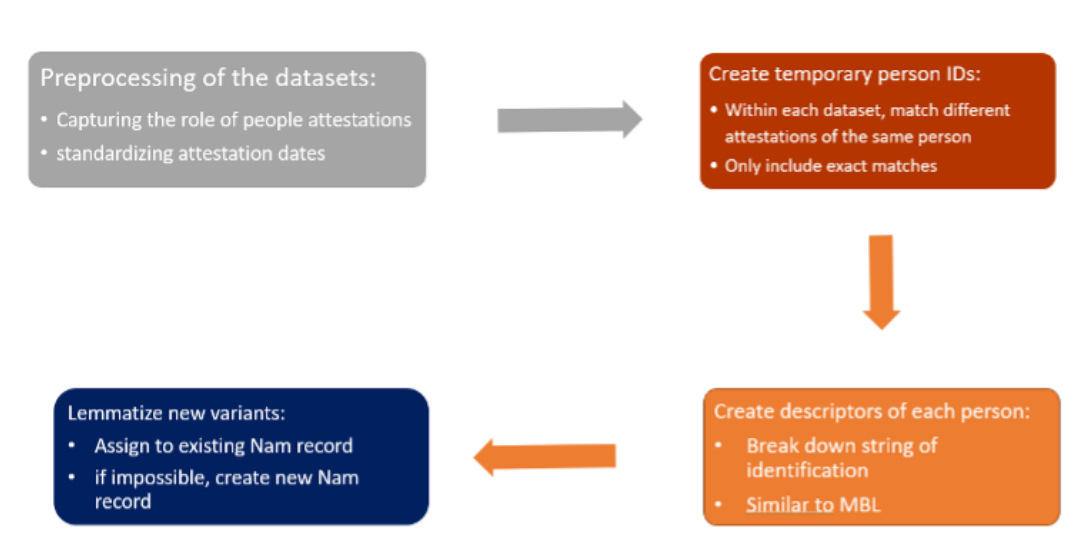


Figure 2: General workflow for digesting datasets other than the MBL.

precision is very important. This has the advantage of producing more accurate results while at the same time greatly speeding up a fully manual approach. Ruth and Sebastian Ahnert, for example, developed a tool called ‘The Disambiguation Engine’, used for the reconciliation of person records in Tudor and Stuart correspondence. This tool grouped together identical names and displayed them alongside other metadata, allowing a user to either mark multiple similar names as the same person, or mark identical names as separate people, based on the sources themselves. A similar tool, RECON, was developed for the database Early Modern Letters Online (Hyvönen et al., 2019). These tools enable those with historical expertise to benefit from the fast and scalable methods of data analysis, and vice versa. In both RECON and the Disambiguation Engine, a relatively straightforward approach to names was taken. Names were matched primarily on simple string matches, perhaps with some basic data cleaning such as removing punctuation or spaces.

The detailed onomastic modelling carried out on the STUDIUM.AI project means a more sophisticated approach can be taken into account, by matching on name variants as well as straightforward strings. On STUDIUM.AI, this was implemented using a tool which finds potential matches by comparing combinations of name variants, and provides them within an interface which allows a user to check and merge the records together if they are deemed true matches. The tool is built into the Filemaker database. It displays all potential matches, starting with those where both a given and family name are shared (with standardisation taken into account). A user can then choose to merge records together if they conclude that the person is indeed the same. This is helpful because in many cases, the names are still not exactly the same, and we may have different sets of dates (floruit dates in one database and dates of birth and death in another, say). In other cases, we can check the original files for additional information. This is particularly useful in the case of named individuals in texts. In many cases, further information about the people can be found by consulting the original manuscript files, and we can merge these matches together. The tool was implemented at a series of data ‘edit-a-thon’ events, where interested parties and experts sat together to check and discuss the results of the automated matches.

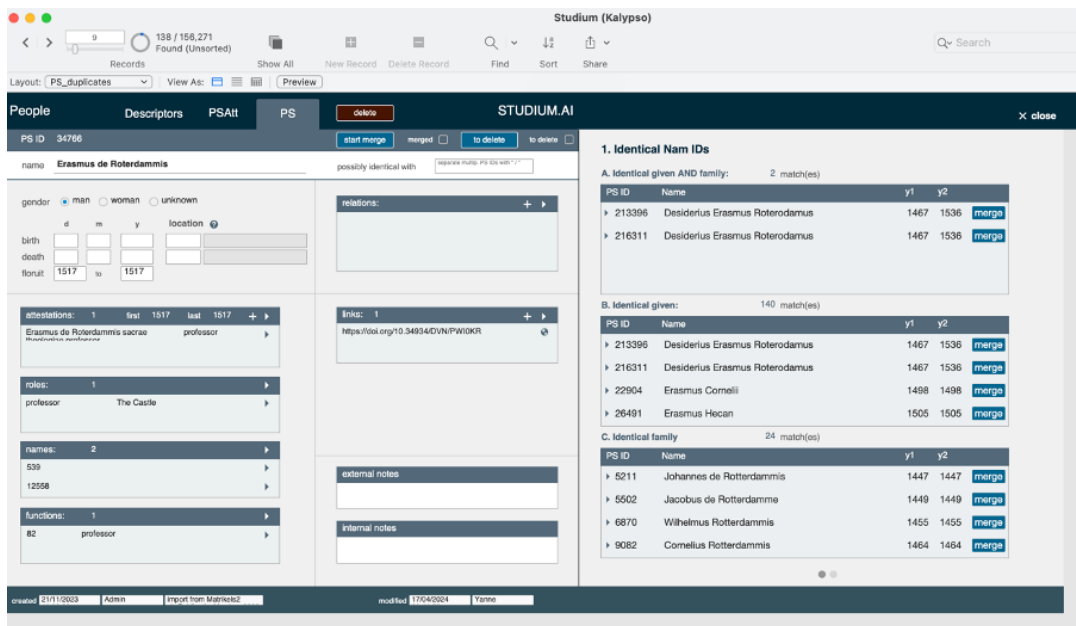


Figure 3: Screenshot of the matching tool. On the left, a matriculation record is displayed (matriculation of Erasmus Desiderius; on the right, the potential candidates, identified through name variants' similarities) for the Lovaniensia book data.

These events resulted in approximately 6,000 person records marked and merged with existing ones. A screenshot of the interface of the matching tool is provided in Figure 3.

This de-duplication process, carried out both within and across collections, forms the basis for the connections across the STUDIUM.AI datasets. 2,148 persons are found across more than one dataset, with the most frequent being the theologian Gommarus Huygens (1631–1702), who appears in 7 of the 10 datasets. The many traces of Huygens' connection to the old University—his matriculation record, monographs, printed lecture notes and appearances as a promotor in thesis sheets—are all now interlinked.

5 Working with the KU Leuven ManGO repository for active data

Our second research question relates to best practices regarding the data, to ensure the origin and cleaning of the data happens in a transparent and reproducible way, and to respect the intellectual property and provenance of the sources and their creators. Naturally, this linking process is only feasible if the underlying data workflow supports it. Because the motivation is that STUDIUM.AI functions as a linking resource rather than as a creator of data, it was important for the workflow to support an easy method to link to the underlying data sources. At the same time, we did have to carry out processing and, in the process of this, make changes to the data, for example by standardising names. In the front end (which will be developed in the course of 2025, hence the data are not shared yet), this means having IDs to the original records so they can be consulted. To properly record our changes or augmentation to the data, this meant that we developed a workflow which would ensure that the work we carried

out on the data was reproducible from the original partner files. As a project dealing with data from different partners and in different formats and states of readiness, there are a number of key challenges:

- How do we retain as much reproducibility as possible given that manual intervention by experts was unavoidable?
- How do we correctly link back to the original data and collections provided by the partners?
- How do we ensure that the correct usage rights are communicated alongside the different records, given that the data is coming from different circumstances with different approaches to copyright?

For the data workflow and management, we have divided the process into two: first, the active research data, and second the long-term storage. By active research data, we mean the files which are currently being worked on. In this paper, we primarily describe the processes relating to the ongoing active research data phase. For the active research data, we are using a service called ManGO provided by KU Leuven.²⁹ The system is based on the open source data management software iRODS, and can be accessed from an ordinary computer through a web-based interface, or through a number of programming clients. ManGO is designed to ensure that the data we use on STUDIUM.AI is managed according to FAIR principles (Findable, Accessible, Interoperable, and Reproducible). One key goal for FAIR data is to ensure it is ‘machine actionable’, meaning it can be easily worked on computationally without the need for manual intervention. In the context of a project like this, this might mean that at a later stage, we can easily recall and update the data used on the project with a new version without the need to re-process the data. *Findable*: Metadata for the original and processed datasets are described using the ManGO system and a standardized schema based on Dublin Core/DataCite standards³⁰, Keywords creator, publisher, geographic, and temporal coverage information are included to enhance findability. ManGO’s search function and Elasticsearch technology ensure easy retrieval of records. *Accessible*: Data is stored on centralized servers with a transparent authentication system via the ManGO platform, ensuring access to the most recent versions. Data and metadata are retrievable through the iRODS³¹ system, command line tools, and Python. *Interoperable*: Data is stored in open formats like .csv, .tsv, XML, and JSON, adhering to Dublin Core standards. Metadata and schemas are also in accessible, interoperable formats. Proprietary formats like .xlsx are converted to text-based formats (csv, tsv) for the project. *Reproducible*: Data processing is automated using scripts in Filemaker, R, or Python. Code, derived, and intermediate files are kept for reproducibility. Unique identifiers and references in Filemaker ensure FAIR data principles apply, allowing users to trace and reproduce steps with proper permissions.

A typical workflow is to import a MARC 21 XML file and convert it to a series of flat format dataframes, then extract and clean the relevant data fields. To give an example, this is the workflow used to import the data from one collection, a set of thesis sheets held by KU Leuven library.

²⁹ <https://rdm-docs.icts.kuleuven.be/mango/index.html>. Accessed 6 Oct 2024. ManGO also allows for fine-grained control over user permissions and access, meaning that both project members and partners can have access to and modify or change, if necessary, their own datasets.

³⁰ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>. Accessed 6 Oct. 2024.

³¹ <https://irods.org/>

1. Data received by the library in MARC21 XML format. The filename is versioned and a read-only copy is uploaded to the ManGO repository.
2. To convert MARC21 into something useable by Filemaker, the XML was first processed into a series of dataframes (one for each tag) using R.
3. The relevant information on people, places, and sources was extracted from these tables.
4. The data was processed according to the Filemaker STUDIUM.AI data model, including joining the names to the existing Filemaker names variants so they could be standardized.
5. For this particular dataset, there was a piece of information not found within the semi-structured data form on the MARC21 XML. Person names had been extracted and standardized from information on the title pages, and entered in the MARC 700 field. However, the names on the title pages also included, in most cases, information on the origin of the person, and this had not been extracted alongside the rest of the name. It was information that could assist in linking records as well as historically valuable so we decided to retain it.
6. To do so, the names were extracted from the title statements using regular expressions (these origins were often found in a standardized format such as *ex Lovanii* or *Chimacensis*, or from Louvain or Chimay).
7. The result of the regex matching was a messy and incomplete list of origin places. This list of non-corrected and non-standardised place names was exported as a .csv and uploaded to a shared drive.
8. Place names checked. In some cases the regex rules gave incorrect place names, but where they were correct, they were linked to existing Filemaker place records. This was done manually by entering the correct Filemaker record IDs for places into a cloud-based spreadsheet application.
9. Updated spreadsheets were re-imported to R and used to populate the Filemaker tables before upload.
10. Final set of tables exported from R and imported into Filemaker.

Each dataset provided by partners is kept in a separate collection. The metadata schema can be updated and there is a metadata versioning system. We are using a modified version of the Dublin Core schema, including title, description, collection, a partner url, a unique identifier for the project component, as well as various standardized metadata fields such as dates and geographic location. The original datasets are kept in read-only folders so they can be accessed but not changed. Supplementary and intermediate datasets are kept in a 'work' folder, for example those necessary to do matching. Scripts used to process the raw data files are also kept here. The last sub collection contains the final files used to export the data to the Filemaker database. As far as possible, this workflow means that everything is reproducible. There are some places where this 'chain' of reproducibility has to be broken – for instance the only way to clean some of the data, for example determining and correcting place names used as part of person descriptors, is to do this manually, using a UI such as a spreadsheet. As much as possible was automated using scripts and regular expressions. However, even

with complex regex rules, the resulting dataset needed to be checked and corrected by hand by multiple people with expertise. This meant exporting a semi-processed dataset, checking manually using a shared spreadsheet, and then importing the corrected data into the final data files. Keeping these files along with the scripts to show how they were created and then used retains as much of the reproducibility as possible. However, we believe that in many cases this kind of semi-automatic workflow is the only realistic one and the aim should be to make it as easy as possible to recreate or understand which decisions were automated and which were manual.

Another key principle was that we are linking datasets together rather than providing new data. In order to preserve this principle, we needed to ensure that that our database referenced the original files sent by data partners. ManGO allowed us to have stable files with version numbers which could be referenced in an import field in the Filemaker database. Therefore, original data files provided by data partners are given unique and version names, and these names are attached to any data import to the Filemaker system. Additionally, these file identifiers can be referenced in derived data and scripts, so that it is always possible to tell exactly which version of a file the data has come from. Particularly key is having a system for doing updates over time, which a good metadata schema and versioning system allows us to do. Naturally, having good documentation (of both the original files plus any derived versions) is vital. Lastly, it was important to stick to the principle of linking rather than changing data so that we can retain the credibility of the original data sent from partners but make it easier to access for the kinds of needs of researchers and digital humanities practitioners. This means that while the data is not changed, we do augment it, for example by adding authority files to place names and giving standardized versions.

6 Integrating Authority files

As well as linking across the project's datasets through names of individual persons, we link to external resources and datasets by including different types of authority files. These 'external links' are both to the original partner datasets, where possible, and authority resources such as VIAF and geographical references in Wikidata. This brings with it the potential to make the data more interoperable in the future, as well as enhance the accuracy of the data concerned. Links are stored in a one-to-many way using a links table, which stores related URIs, authority files, and IDs. First of all, STUDIUM.AI stores a range of links and URIs from the participating datasets: e.g. the entries for 'people' in the DaLeT project, or the stable URLs for 'persons' and 'publications' used by ODIS. For printed texts cataloged by KU Leuven Libraries, the stable URLs to the KU Leuven LIMO catalogue are also included for the relevant printed texts and rare books. Whenever the MARC21 data included links to Short Title Catalogues for rare books, mostly the Universal Short-title Catalogue (USTC),³² and the Short Title Catalogue Flanders (STCV),³³ the links were extracted as to enhance interoperability. Authority files for geographical entities were handled in a slightly more complex way, which is outlined in some more detail below. The datasets contain multiple sources of geographical information (about 'places'), recording different spatial characteristics and features, related to different parts of the metadata. Most of the 160,000 names in the matriculation records contain a toponym used as part of a name descriptor, e.g. 'Jacobus Philippus Debruyne Lovaniensis', recording the place

³² <https://www.ustc.ac.uk/>. Accessed 6 Oct. 2024.

³³ <https://www.stcv.be>. Accessed 8 June 2024.

of origin (rather than of birth). In the same vein, the thesis sheets use a very similar kind of descriptor in the names of the defendants (students) and their promotors (professors). A second source of geographical data is found in bibliographic records: here, the place of publication is almost always listed on the title page or colophon of printed works, and subsequently recorded in the MARC21 metadata. Consequently, there is a complex, interesting, and interlinked spatial network surrounding these records, from the origin places of the students matriculating or graduating to the places of publication for works associated with the university. This is a valuable source of information that can be used for research in spatial history and spatial humanities, ranging from map visualisations to spatial analysis, particularly in the context of book history (Black et al., 2021; Elliott and Gillies, 2009; Lahti et al., 2019). In total there are about 33,000 unique strings used as place descriptors in the STUDIUM.AI dataset. In the first step, these were grouped together under place names in much the same manner as with family and given names. The place Leuven, for instance, is represented in the dataset as the following non-standardised strings, coming from attestations of people and sources: *Leuven; Liuvensis; Loewensis; Lovaiensis; Lovanienij; Lovanienis; Lovanieniss; Lovaniensis; Lovaniensos; Lovanio; Lovaniuensis; Loviensis; Lowaniensis; Lowensis; a Lovanio; civis Lovaniensis; de Lona [Lovanio]; de Louvanio; de Lovania; de Lovanie; de Lovanio; de Lovaniolo; de Lovanis; de lovanio; in Castria Lovaniensis; natus Lovanii; opidi Lovaniensis; oriundus Lovanio; prop Lovanium; prope Lovaniam; prope Lovanium*. During this phase, these 33,000 unique strings collapsed to about 9,000 place records. To geocode the dataset we used a database and resource called the World Historical Gazetteer (Grossner and Mostern, 2021). The WHG is conceived as software enabling the connection of specialist collections of place names. Using the WHG allows us to incorporate this inherent conceptual ‘fuzziness’ and represent the changing aspects of historical place names. The WHG contains an interface for geocoding. We exported a dataset containing all the various variations of the place names. The WHG checks them against a database containing geographic entities from Wikidata and Geonames, and the Getty Thesaurus of Geographic Names – a total of 13 million place names. These are then manually verified using a user interface. The user interface displays all the information about the attestation, including all name variants, as well as coordinates if these had already been added to the record by the project. Alongside the information all found possible matches from the WHG index are displayed, using fuzzy matching across all the variants. Users checking the records can say whether each is a match or not, given the information at hand. Once finished, this augmented data can be downloaded and integrated within STUDIUM.AI.

So far, this process has allowed us to add 2,180 new geometries and 6,702 links (e.g Geonames, Wikidata, Getty Thesaurus of Geographic Names, BNF, and Library of Congress URIs). These primarily come from (the modern borders of) Belgium (1893 unique places found so far), the Netherlands (502), France (222), Germany (144), Italy (36), Ireland (31), Spain (17), Luxembourg (15), and Turkey (15). On the one hand—if we restrict the maps to the matriculation records alone—these first visualisations confirm that Leuven recruited mainly in the Hinterland (a radius of about 150 kilometers), yet on the other hand, they now clearly show how the Revolt and the emergence of the Dutch Republic condensed the recruiting era to Belgium and a part of Northern France, in other words, that territories which were still under the Catholic Habsburg regime in the seventeenth century.

The WHG output format is called Linked Place Format – an extension of the GEOJSON format, which allows temporal scoping of place records, meaning a record can

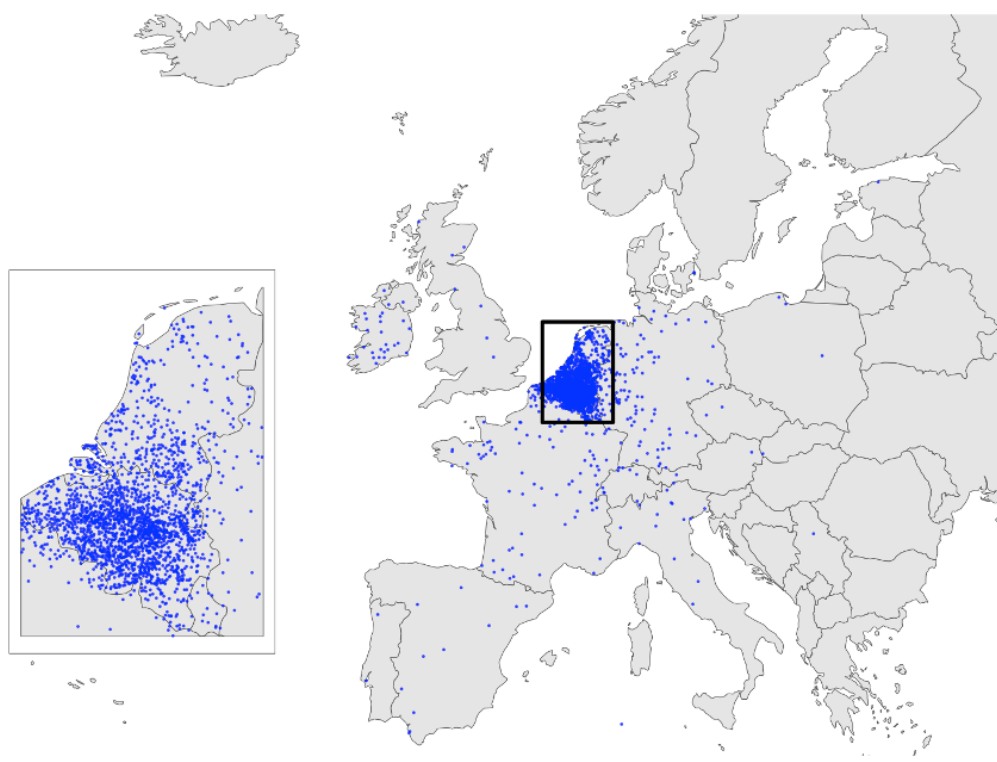


Figure 4: A map of the currently geocoded places from the STUDIUM.AI database, projected on a current map.

associated with both a place on earth (a geometry) and a set of dates.³⁴ Our final gazetteer will be published through the WHG, and we can use the published version to generate unique place URIs which will be referred to by the project. In this way, we do not have to rely on a single digital resource but can represent the fuzziness and historicity of the place names found within the datasets. The finished dataset can be later accessioned to the WHG ‘union index’, which collects all publicly-published attestations of places under the same record, with information on the source, temporal scope, various names, and so forth.. We will be able in this way to contribute in terms of the names and variants referring to existing places, which is of onomastic and historic interest.

7 Conclusions

STUDIUM.AI started from a modest idea to integrate the wide expertise in Digital Humanities to link existing datasets together, both those generated by GLAM-institutions and those dressed by research teams or individual researchers from different disciplines. This intent to ‘break silos’ in ongoing research on the early modern history of the University of Leuven (1425-1797) evolved into a shared endeavor to build a research infrastructure across 8 teams and 4 faculties of the current KU Leuven. The research infrastructure enables collaborative research in the future by a shared back-end of linked data in a common core. In this paper, we have discussed the infrastructure under three different perspectives. First, we have shown how the task of interlinking

³⁴ See also <https://github.com/LinkedPasts/linked-places-format>. Accessed 6. Oct. 2024.

data benefits from semi-automated approaches, that, while preserving the accuracy of the identified matches, accelerate the process. Moreover, our project integrated existing reference resources (such as onomastic dictionaries) into a digital workflow. This strengthens the reliability of the final output. Second, we have discussed our efforts in terms of data-management, in order to make our infrastructure sustainable to the (inevitable) evolutions both of the source datasets and the staff involved. Finally, we described the effort of enriching the data thanks to the integration of external authority files, which paves the way to further analysis and visualizations. Along with this project to build a common core of research data, currently a new HTR model is being developed to build a full-text corpus of the handwritten student notes, which will yield, through NER and NEL, additional data.

The standardization of data (be it names of persons or places, as discussed in this paper) might be a time-consuming process, but the added value is clear. The linking of data bears the potential to visualise networks across single datasets, here 'breaking the silos' of formerly separate (meta)data creation. For example, students are now being connected to professors (through student notes or defended theses), and in turn, these can be connected to printers and publishers and so forth. The linked data also help to identify new brokers or 'bridging figures' between the academic and the book world of early modern Leuven, individuals who might otherwise not show up in the single data sets. This standardisation also brings the possibility of linking across other university datasets and even national collections, though others have discussed the conceptual problems with standardising certain categories of information, such as roles and degrees, across different datasets and traditions (Rubio Muñoz 2022). As it is, the current data enrichment by integrating geographical information helps to build out a whole new set of research questions about the worldwide spread of the networks in and especially beyond Leuven. As such, the research infrastructure can help answer questions about places, such as printed outputs and matriculation, identifying temporal-spatial clusters across datasets, looking at spatial patterns in recruitment, etc. This new interpretation of Louvain as a hub in evolving temporal-spatial clusters of the transfer of knowledge can help to reassess the historiography of the early modern university from a DH-perspective.

References

- Ruth Ahnert and Sebastian E. Ahnert. Introduction: Tudor Letters in the Digital Age. In *Tudor Networks of Power*, pages 3–26. Oxford University Press Oxford, 1 edition, October 2023. ISBN 9780198858973 9780191891595. doi: 10.1093/oso/9780198858973.003.0001. URL <https://academic.oup.com/book/51646/chapter/419632240>.
- T. H. Aston. THE MEDIEVAL ALUMNI OF THE UNIVERSITY OF CAMBRIDGE. *Past and Present*, 86(1):9–86, 1980. ISSN 0031-2746, 1477-464X. doi: 10.1093/past/86.1.9. URL <https://academic.oup.com/past/article-lookup/doi/10.1093/past/86.1.9>.
- Fiona A. Black, Jennifer M. Grek Martin, and Bertrum H. MacDonald. Geographic Information Systems and Book History. In *Oxford Research Encyclopedia of Literature*. Oxford University Press, August 2021. ISBN 9780190201098. doi: 10.1093/acrefore/9780190201098.013.1151. URL <https://oxfordre.com/literature/view/10.1093/acrefore/9780190201098.001.0001/acrefore-9780190201098-e-1151>.

- Gian Paolo Brizzi and Willem Frijhoff. *Digital academic history: studi sulle popolazioni accademiche in Europa*. Studi e ricerche sull'università. Società editrice Il mulino, Bologna, 2018. ISBN 9788815275523.
- Yanne Broux. *STUDIUM.AI Filemaker database: manual. Version 3.0 (April 2024)*. 2024a. URL <https://lirias.kuleuven.be/4199372&lang=en>.
- Yanne Broux. *Standardizing names and disambiguating people in the STUDIUM.AI project A description of the process and current state of affairs (April 30 2024)*. 2024b. URL <https://lirias.kuleuven.be/4199373&lang=en>.
- Dieter Cammaerts. *Manuale Lovaniense. Een sociaal-economische en typografische studie van het gedrukte academische handbook in de vroegmoderne Leuvense Universiteit (1474-1650)*. PhD thesis, KU Leuven, Leuven, 2024.
- Lorenz Demey. *Leonardi.DB: Leuven Ontology for Aristotelian Diagrams Database*, 2024. URL <https://leonardi.logicalgeometry.org/>.
- Tom Elliott and Sean Gillies. Digital Geography and Classics. *Digital Humanities Quarterly*, 3(1), 2009. ISSN 1938-4122. URL <https://www.digitalhumanities.org/dhq/vol/3/1/000031/000031.html>.
- Xander Feys, Maxime Maleux, Andy Peetermans, and Raf Van Rooy, editors. *Student Notes from Latin Europe (1400–1750). A Research Companion*. Leuven University Press, Leuven, 2025.
- Michael Finegold, Jessica Otis, Cosma Shalizi, Daniel Shore, Lawrence Wang, and Christopher Warren. Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *Digital Humanities Quarterly*, 10(3), 2016. doi: 10.17613/M6B020. URL <https://works.hcommons.org/records/61327-m7s62>.
- Jean-Philippe Genet, Hicham Idabal, Thierry Kouamé, Stéphane Lamassé, Claire Priol, and Anne Tournieroux. General Introduction to the Studium Project. *Medieval People*, 31(1), January 2016. ISSN 2690-8182. URL <https://scholarworks.wmich.edu/medpros/vol31/iss1/9>.
- Christophe Geudens and Serena Masolini. Teaching Aristotle at the Louvain Faculty of Arts, 1425-1500. General regulations and handwritten testimonies. *Rivista di Filosofia Neo-Scolastica*, 108(4):813–844, 2016.
- Christophe Geudens, Jan Papy, and Lorenz Demey. *Louvain Theories of Topical Logic (c. 1450-1533). A Reassessment of the Traditionalist Thesis*. PhD thesis, KU Leuven, 2020.
- Karl Grossner and Ruth Mostern. Linked Places in World Historical Gazetteer. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, pages 40–43, Beijing China, November 2021. ACM. ISBN 9781450391023. doi: 10.1145/3486187.3490203. URL <https://dl.acm.org/doi/10.1145/3486187.3490203>.
- Gubler, Kaspar. Forschungsdaten vernetzen, harmonisieren und auswerten. Methodik und Umsetzung am Beispiel einer prosopographischen Datenbank mit rund 200.000 Studenten europäischer Universitäten (1200–1800). pages 127–145. Verlag Julius Klinkhardt, 2022. ISBN 978-3-7815-5952-3. doi: 10.25656/01:24857. URL https://www.pedocs.de/frontdoor.php?source_opus=24857.

- Patrick Hanks, Kate Hardcastle, and Flavia Hodges. *A Dictionary of First Names*. Oxford Reference. OUP Oxford, Oxford, 2nd ed edition, 2006. ISBN 9780198610601.
- Mark J. Hill, Ville Vaara, Tanja Säily, Leo Lahti, and Mikko Tolonen. Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material. In Costanza Navarretta, Manex Agirrezabal, and Bente Maegaard, editors, *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, volume 2364 of *CEUR Workshop Proceedings*, pages 201–219, Copenhagen, Denmark, March 2019. CEUR. URL https://ceur-ws.org/Vol-2364/#19_paper.
- Kasra Hosseini, Federico Nanni, and Mariona Coll Ardanuy. DeezyMatch: A Flexible Deep Learning Approach to Fuzzy String Matching. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 62–69, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.9. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.9>.
- Howard Hotson and Thomas Wallnig, editors. *Reassembling the Republic of Letters in the Digital Age: Standards, Systems, Scholarship*. Göttingen University Press, Göttingen, 2019a. ISBN 9783863954031. doi: 10.17875/gup2019-1146. URL <http://resolver.sub.uni-goettingen.de/purl?univerlag-isbn-978-3-86395-403-1>.
- Howard Hotson and Thomas Wallnig, editors. *Reassembling the Republic of Letters in the Digital Age: Standards, Systems, Scholarship*. Göttingen University Press, Göttingen, 2019b. ISBN 9783863954031. doi: 10.17875/gup2019-1146. URL <http://resolver.sub.uni-goettingen.de/purl?univerlag-isbn-978-3-86395-403-1>.
- Eero Hyvönen, Ruth Ahnert, Sebastian Ahnert, Jouni Tuominen, Eetu Mäkelä, Miranda Lewis, and Gertjan Filanrski. Reconciling metadata. In Howard Hotson and Thomas Wallnig, editors, *Reassembling the Republic of Letters in the Digital Age*, pages 223–236. Univ.-Verl. Göttingen, Germany, 2019.
- Suzanne Jannis, David Fondu, Lydia Janssen, Marc Carnier, and Valerie Vrancken. Belgian State Archives / State Archives of Leuven / Rijksarchief Leuven - Databank van personen ingeschreven in de matrikels van de Oude Universiteit Leuven, 1426-1797, 2020. URL <https://www.sodha.be/citation?persistentId=doi:10.34934/DVN/PWIOKR>.
- KU Leuven Libraries. KULEuvenDigitalisering/Magister-Dixit-Collection-Dataset: 202303 metadataset update, October 2024. URL <https://zenodo.org/records/13887967>.
- Leo Lahti, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. Bibliographic Data Science and the History of the Book (c. 1500–1800). *Cataloging & Classification Quarterly*, 57(1):5–23, January 2019. ISSN 0163-9374, 1544-4554. doi: 10.1080/01639374.2018.1543747. URL <https://www.tandfonline.com/doi/full/10.1080/01639374.2018.1543747>.
- Jan Papy. *The Leuven Collegium Trilingue 1517-1797: Erasmus, humanist educational practice and the new language institute Latin, Greek, Hebrew*. Peeters, Leuven Paris Bristol (Conn.), 2018. ISBN 9789042936225.

- Edmond-Henri-Joseph Reusens, Arnold Hubert Schillings, and Joseph Wils. *Matricule de l'Université de Louvain*. Académie Royale De Belgique. Commission Royale D'histoire. Kiessling, Bruxelles, 1903-1990.
- Rossana Scebba and Margherita Fantoli. Integrating Library and Prosopographical Data in the Early Modern Publication Network of the University of Louvain (1501-1797). In Violet Soen, Wouter Druwé, Wim François, and Ralph Dekoninck, editors, *Students, Scholars and their Books at the University of Louvain (1425-1797)*, number 17 in LECTIO Series. Brepols, Turnhout, Forthcoming.
- Rainer C. Schwinges. Gelehrte von Heidelberg und anderswo: Einblicke in die Datenbank des Repertorium Academicum Germanicum (RAG). In Heike Hawicks and Ingo Runde, editors, *Universitätsmatrikeln im deutschen Südwesten: Bestände, Erschließung und digitale Präsentation: Beiträge zur Tagung im Universitätsarchiv Heidelberg am 16. und 17. Mai 2019*, number Band 9 in Heidelberger Schriften zur Universitätsgeschichte, pages 275–308. Universitätsverlag Winter, Heidelberg, 2020. ISBN 9783825347260. OCLC: on1232446933.
- Violet Soen and Margherita Fantoli. Computing Women's Centrality in the Book Trade: Widow Printers in the University Town of Douai (1559-1659). In Lieke van Deinsen and Alice C. Montoya, editors, *Recovering Women's Book Culture and Literary History. (Middle Ages – 1830): New Digital Approaches*. Brill, Leiden, Forthcoming.
- Violet Soen and Yann Ryan. Who-is-Who in STUDIUM.AI? Towards New Metrics about Students, Scholars and Printers at the Early Modern University of Louvain (1425-1797). In Violet Soen, Wouter Druwé, Wim François, and Ralph Dekoninck, editors, *Students, Scholars and their Books at the University of Louvain (1425-1797)*, number 17 in LECTIO Series. Brepols, Turnhout, Forthcoming.
- Violet Soen, Wouter Druwé, Wim François, and Ralph Dekoninck, editors. *Students, Scholars and Their Books at the Early Modern University of Louvain (1425-1797)*. Number 17 in LECTIO Series. Brepols, Turnhout, Forthcoming.
- Jana Synovcová Borovičková and Jaroslava Škudrnová. Prosopographical Databases in the Context of Modern Research in the History of Universities – Universitas Magistrorum (1358–1622) Database. *AUC HISTORIA UNIVERSITATIS CAROLINAE PRAGENSIS*, 60(1):189–205, March 2021. ISSN 2336-5730, 0323-0562. doi: 10.14712/23365730.2020.26. URL <http://www.karolinum.cz/doi/10.14712/23365730.2020.26>.
- Dirk Van Miert. Social Networking in the Republic of Knowledge. *History of Humanities*, 7(2):313–329, September 2022. ISSN 2379-3163, 2379-3171. doi: 10.1086/721316. URL <https://www.journals.uchicago.edu/doi/10.1086/721316>.
- Jacques Verger. Le recrutement géographique des Universités françaises au début du XVe siècle d'après les Suppliques de 1403. *Mélanges d'archéologie et d'histoire*, 82(2):855–902, 1970. ISSN 0223-4874. doi: 10.3406/mefr.1970.7616. URL https://www.persee.fr/doc/mefr_0223-4874_1970_num_82_2_7616.
- Donald J. Waters. The emerging digital infrastructure for research in the humanities. *International Journal on Digital Libraries*, 24(2):87–102, June 2023. ISSN 1432-5012, 1432-1300. doi: 10.1007/s00799-022-00332-3. URL <https://link.springer.com/10.1007/s00799-022-00332-3>.