Journal of Cultural Analytics

Scebba, Rossana. 2026. Bibliographic Metadata as Relational Data: A Cross-Disciplinary Methodological Reflection. *Journal of Cultural Analytics*, 11(1): pp. 1–34. DOI: https://doi.org/10.22148/jca.1097

Bibliographic Metadata as Relational Data: A Cross-Disciplinary Methodological Reflection

Rossana Scebba, IRES/LIDAM, Université Catholique de Louvain and Research Unit of Early Modern History, Katholieke Universiteit Leuven, rossana.scebba@kuleuven.be, rossana.scebba@uclouvain.be

In this paper, I reflect on the growing cross-disciplinary convergence of the use of bibliographic metadata as empirical material in digital humanities and computational history as well as in quantitative economic history. I discuss the main challenges involved in preprocessing metadata from library catalogs, comparing techniques from the abovementioned fields. A case study based on the *Collectio academica antiqua* of the Old University of Louvain serves to demonstrate how its metadata can be parsed, disambiguated, and reorganized into relational format, that is, into linked tabular datasets describing different aspects of the collection's records. Building on this foundation, I examine the use of bibliographic data as empirical material in historical network analysis, with particular attention to the assumptions underlying co-occurrence representations. I show that role-differentiated participation encoded in the original metadata is not naturally accommodated by flat network projections, and I argue that a multilayer network representation provides a coherent way to preserve this heterogeneity. I use this to assess how differences in participation across roles relate to the institutional context in which publications were produced. I also highlight both the analytical limits of bibliographic metadata when used in isolation and the gains that arise when relational representations derived from catalog data are integrated with complementary sources.

Journal of Cultural Analytics is a peer-reviewed open access journal published by the Center for Digital Humanities. © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.
OPEN ACCESS

Section 1: Introduction

Traditionally compiled by librarians for purposes of bibliographic control and collection management, metadata from union catalogues and curated library collections are increasingly being repurposed as empirical material in historical research both in the fields of digital humanities and computational history as well as in the field of quantitative social sciences. Digital humanities scholars have been at the forefront of this reappropriation. Building on the framework introduced by Lahti et al., "Bibliographic Data Science and the History of the Book," who coined the term *bibliographic data science* to describe the systematic use of catalog metadata for historical inquiry, a growing body of work has shown how information encoded in title pages, imprint statements, dedications, approbations, and physical descriptions can support large-scale analyses of intellectual and book history. Key contributions include studies drawing on large-scale book catalogs to examine cultural production, canon formation, book pricing and the circulation of classical and vernacular texts (see Lahti et al., "A Quantitative Study of History in the English Short-Title Catalogue," Tolonen et al., "A Quantitative Approach to Book-Printing in Sweden and Finland," Tolonen et al., "Examining the Early Modern Canon," Hill et al., "Reconstructing Intellectual Networks," Tiihonen and Tolonen, Fantoli et al.), as well as network analyses of print culture that trace influence through dedications, printers' relationships, and publishing communities (Gavin, Hill et al., "Communication and Idea Transmission Across Historical Communities," Ladd, Greteman, Gittel, Valleriani et al., Ryan and Tolonen, "Networks of Influence in Scottish Enlightenment Publishing," Ryan and Tolonen, "The Evolution of Scottish Enlightenment Publishing," Heßbrüggen-Walter). A parallel strand of scholarship draws on metadata of epistolary sources, as in Hotson and Wallnig on the Republic of Letters or Roller, "Tracing the Footsteps of Ideas" on the circulation of Reformation ideas through correspondence.

In parallel, researchers in the social sciences—particularly, though not exclusively, applied economic historians—have also begun to employ bibliographic data to address topics such as the decline of Islamic science and preindustrial city growth (Chaney, "Religion and the Rise and Fall of Islamic Science," Chaney, "Modern Library Holdings and Historic City Growth"), teacher-directed scientific progress in early modern England (Koschnick), the diffusion of ideas through railroad networks in nineteenth-century German-speaking regions (Chiopris), the relationship between limited academic career prospect and the rise of dissenting religious print (de Pleijt and Koschnick), and the role of the Republic of Letters in Britain's Industrial Revolution take-off (Cervellati et al.) This convergence around the same materials reveals both the expanding research potential of bibliographic data and the

persistence of distinct disciplinary conventions in how these data are employed. The digital humanities have often developed without full integration with the quantitative social sciences (Lemercier 273), which helps explain why the two traditions, despite similar empirical aims, have adopted different, though occasionally overlapping, methodological conventions.

The way the very same bibliographic data are mobilized in the two literatures reveals this divergence most clearly. While not all digital humanities or computational history projects using these materials adopt network methods, the network perspective remains one of the most common modelling frameworks paired with bibliographic metadata in the field. Although catalog data do not inherently encode relationships, they are often interpreted as networked systems by treating the co-occurrence of agents within the same publication event as evidence of collaboration or shared involvement. This approach has been productive for mapping and exploration, though its analytical development remains uneven. In this perspective, skepticism toward digital approaches among historians and humanists trained in textual or archival traditions is not without foundation (Gregory 2). Yet this limitation is increasingly being addressed more broadly, with Roller's "Theory-Driven Statistics for the Digital Humanities: Presenting Pitfalls and a Practical Guide by the Example of the Reformation," insisting that quantitative methods be tied to explicit research questions and historiographical theories that establish testable relationships between measurable concepts of interest.

By contrast, in quantitative economic history, recent trends shaped in part by methodological norms associated with the so-called "credibility revolution" (Angrist and Pischke) have reinforced expectations about explicit research design and the use of linked datasets (Cantoni and Yuchtman 216). Within this framework, bibliographic data are routinely employed in research designs that integrate multiple data sources to produce theoretically grounded interpretations of historical phenomena. Economic history thus offers a useful point of comparison for digital humanities and computational history because it faces similar preprocessing and interpretative challenges and shows how formal analytical modeling can expand the scope of the same bibliographic metadata corpora.

The paper builds on this cross-disciplinary comparison by examining how metadata from library catalogs can be de-structured, parsed, and reorganized in relational form and how the resulting data can support analytical modeling within a network framework. In doing so, it responds to calls for closer alignment between quantitative analysis and explicit theoretical frameworks (Roller, "Theory-Driven Statistics"). Related work in the digital humanities has pursued more explicitly analytical modeling

strategies, including multimodal combinations of network methods and text analysis (Hill et al., "Communication and Idea Transmission Across Historical Communities") and temporal modeling approaches (Roller, "Tracing the Footsteps of Ideas"). More broadly, the present study aligns with the commitment articulated by Lahti et al., "Best Practices in Bibliographic Data Science," to treat bibliographic metadata as a substantive source for historical inquiry.

Against this background, this paper is guided by three related research questions. First, how can bibliographic catalog records, originally designed for item-level description, be transformed into relational data structures suitable for longitudinal and network-based analysis? Second, what modeling assumptions are implicitly introduced when bibliographic metadata are interpreted as co-occurrence networks, and how do these assumptions shape the interpretation of relationships inferred from catalog data? Third, what analytical leverage is gained by preserving role-differentiated participation through a multilayer network representation compared with flat co-occurrence projections commonly used in digital history?

To address these questions, I focus on the metadata of the *Collectio academica antiqua*, the cultural heritage collection of the Old University of Louvain (1425–1797), and I examine how relational modeling choices affect the identification of systematic patterns in academic print production. The central claim of the paper is that bibliographic data encode role-differentiated forms of participation that are not naturally accommodated by standard, flat co-occurrence network representations. When such data are modeled as such, distinct forms of involvement in book production are implicitly collapsed. I argue that, if one adopts a network perspective, a multilayer representation offers the most coherent way, within a co-occurrence framework, to preserve this role differentiation instead of flattening it into an undifferentiated relation. This modeling choice is justified because it preserves distinctions encoded in the original sources and because these distinctions might correspond to historically meaningful differences in modes of participation to book production. I use my case study of the old academic collection of Louvain to address whether holding multiple production roles is associated with a higher likelihood that a publication involves university professors.

This analysis also makes clear that recovering features of Louvain's academic publishing ecosystem requires integrating bibliographic metadata with external sources beyond the catalog itself. However rich, bibliographic data *alone* offer limited leverage for addressing complex historical questions. Their analytical potential is substantially expanded when they are linked to complementary sources, such as prosopographical—the study of background characteristics—or institutional datasets.

The comparison with quantitative economic history reinforces this point: In that field, bibliographic data typically function as one component within a broader data infrastructure combining multiple sources.

The paper proceeds in five sections. Section 2 outlines the structure of metadata from union catalogs and library collections and discusses how key bibliographic elements can be identified, extracted, and transformed into a relation format. Section 3 details the conceptual and practical steps involved in preprocessing the resulting data for systematic quantitative analysis. Section 4 examines the assumptions involved in interpreting catalog metadata as network data and shows how a multilayer co-occurrence representation allows role-differentiated participation to be modeled and combined with external sources for analytical use. Section 5 concludes.

Section 2: What is in a Catalog? Anatomy of Bibliographic Metadata

Digital library catalogs are curated databases describing and indexing the holdings of one or more libraries. The structured information stored in these catalogs is known as bibliographic metadata. Metadata refer to attributes about catalog entries, rather than their full content, and support organization, classification, and discovery. For printed book collections, such metadata include elements such as title, publication date and place, names of agents involved in the creation and distribution, the language, and the physical description of each holding.

Internally, library catalogs rely on two complementary layers: an encoding format and a content standard. The encoding determines how metadata is structured for storage and exchange, whereas the content standard governs what information is recorded and how it is expressed. Most libraries use MARC (MAchine-Readable Cataloging) as their encoding format. Developed at the Library of Congress by Henriette Avram and later formalized as MARC 21, this schema makes cataloging data legible to computers across libraries and associates text strings with numbered fields that are conventionally linked to specific bibliographic properties. Despite its limitations and its limited alignment with linked-data infrastructures (Tennant), MARC 21 continues to underpin most catalog systems. Even when library records are exported to relational tables or Extensible Markup Language (XML) formats such as MARCXML or Metadata Object Description Schema (MODS), they retain the underlying MARC 21 structure. Researchers must therefore parse fields, indicators, and subfields according to MARC conventions. Tools such as pymarc assist with extraction and conversion but still require familiarity with MARC 21 logic.[1]

[1] pymarc is a Python library for reading, writing, and parsing MARC21 records. For the documentation, see https://pypi.org/project/pymarc/.

While encoding formats define *where* information is stored, content standards determine *what* information is entered and how consistently. These standards shape decisions such as how titles are transcribed, how imprints are formatted, and how names, roles, and subject information are recorded. General collections typically follow Resource Description and Access (RDA), rare book collections often rely on more specialized descriptive standards such as Descriptive Cataloging of Rare Materials (DCRM), and archival holdings use standards such as Describing Archives: A Content Standard (DACS). As a result, records encoded in the same MARC 21 infrastructure may display distinct descriptive logics depending on the content standard applied. In other words, it is only in light of the cataloging conventions that one can interpret which MARC fields are actually available for analysis.

I work with the *Collectio academica antiqua*, a curated collection of early modern printed works associated with the Old University of Louvain, the first university founded in the historic Low Countries, in the Brabant region (present-day Belgium) in 1425, from which the contemporary sister universities of Katholieke Universiteit Leuven (KU Leuven) and Université Catholique de Louvain (UCLouvain) descend, following its abolition in 1797.[2] The collection gathers works linked to the university, its affiliated scholars, its illustrious alumni, and its institutional history. Although printing activity in Louvain is attested from 1473 onward, incunabula at the KU Leuven Libraries are held in a separate preservation collection, so the *Collectio academica antiqua* begins artificially in 1501 and so does not reflect the chronology of local printing (Scebba and Fantoli). The catalog records of the *Collectio academica antiqua* are encoded in MARC 21 and follow the RDA-based cataloging standard used at the KU Leuven Libraries, supplemented by internal guidelines. For early printed materials, catalogers also draw on the Short Title Catalogue Vlaanderen (STCV) rule set, which provides descriptive guidance for rare books (i.e., books printed before 1801).

Table 1 summarizes the MARC 21 fields most commonly encountered in early modern and rare book contexts. It generalizes from the data structure of the *Collectio academica antiqua* and from comparable early modern catalogs. Whether these fields become analytically useful depends on the research question as well as on the content standards and cataloging practices that shaped the metadata. This dependence is

[2] The institutional history is more complex: the original university, known as the Old University of Louvain, was suppressed under French revolutionary rule in 1797; then briefly refounded as the secular State University of Louvain in 1817 before being abolished again in 1835; and finally re-established as the Catholic University in 1834. In 1968, amid the linguistic tensions between Flemish and Francophone communities that have long shaped Belgian public life, it was split into the Dutch-language Katholieke Universiteit Leuven, remaining in Louvain (in the province of Flemish Brabant), and the French-language Université Catholique de Louvain, relocated to the newly built campus of Louvain-la-Neuve (in the province of Walloon Brabant).

evident when early modern catalogs are compared with contemporary circulating collections. For example, Petras et al. use metadata from modern library collections to construct time-period directories that allow users to navigate holdings by historical era and place. Their approach relies on subject strings containing explicit chronological and geographic subdivisions, which are common in contemporary catalogs but generally absent from early modern collections such as the *Collectio academica antiqua*.

This closer examination of the technical structure of catalog data underscores that working with bibliographic metadata requires an explicit understanding of their internal organization and their transformation into relational form. It also makes immediately clear that the extracted data often require harmonization. Because cataloging practices prioritize item-level description over cross-record consistency, records frequently lack internal coherence, particularly in the identification of persons, corporate entities, and place names (Padilla 20). Unless a catalog relies on robust authority control or controlled vocabularies, these elements tend to appear in multiple, nonaligned forms. As a result, they must be interpreted, cleaned, and restructured before they can serve as input for empirical modeling. The next section outlines the conceptual and practical decisions involved in converting catalog records into harmonized relational data.

Section 3: From Structured to Relational Data

Transforming catalog metadata into a form suitable for quantitative modeling involves a series of preprocessing steps. This stage raises questions that are increasingly shared across disciplines. Scholars in both digital humanities and social sciences working with historical data are confronting the same challenges: how to identify and reconcile references to historical persons, places, and subjects across historical data, including bibliographic metadata. Digital humanists have long grappled with these issues in prosopographies, bibliographies, and textual corpora (Ehrmann et al.), whereas social scientists are more recently engaging with similar problems (Arora et al.). Both traditions are converging toward workflows that combine domain knowledge with scalable, semi-automated tools in order to prepare historical data for analysis. In this context, extracting and parsing structured bibliographic metadata is a crucial step because converting MARC's heterogeneous field structure into an analytically usable form requires normalizing it into relational data, with entities separated into tables and linked through keys. In what follows, I show how these workflows unfold in practice, using the *Collectio academica antiqua* as a case study across three key stages: actors identification, spatial referencing of place names, and content classification.

Identifying Historical Actors: Disambiguation and Authority Alignment

Catalogs often record the names of both personal and corporate entities associated with a given holding, in line with the MARC 21 schema.[3] These entries may include additional identifying information such as life dates or floruit, numeration, pseudonyms, religious affiliation or noble titles, and places of activity, as well as the role under which the individual appears in the publication. One of the most challenging steps in working with bibliographic metadata is the identification and disambiguation of historical actors based on these name strings. This difficulty arises from variant spellings, inconsistent orthography, and uneven role attribution across entries. These challenges may be more or less pronounced depending on the uniformity of cataloging conventions and the degree of authority control implemented, and this may apply to curated collections and union catalogs alike. To construct meaningful relational data, it is essential to reconcile these inconsistencies and unambiguously identify historical individuals. The disambiguation step ensures that relationships in the data are constructed around unified entities instead of fragmented or duplicated name strings. This process relies on accompanying biographical attributes to distinguish homonyms and resolve pseudonyms. In the case of homonyms, these additional attributes can help differentiate distinct individuals with similar names, while in the case of pseudonyms or variant forms referring to the same person, shared dates or roles may allow for accurate consolidation even when string similarity alone is insufficient. Ideally, at the end of this reconciliation process, one should be able to assign a persistent identifier to each person or corporate entity.

In practice, two main strategies are typically employed: a bottom-up approach, based on string similarity and clustering, or a top-down strategy that aligns name strings with external authority files. Both techniques can be employed at different stages of the workflow and often both might benefit from a semi-automated approach. For instance, fuzzy string matching and clustering algorithms can be used to group likely name variants, which are then manually reviewed for validation. Similarly, a name search can be launched against a selected external authority file to identify potential matches, but the final assignment of identifiers should remain under human supervision.

In my pilot study using the *Collectio academica antiqua*, I implemented a structured pipeline to extract, standardize, and cluster personal and corporate names from MARC 21 records. I began by identifying the relevant MARC 21 tags containing names of interest, then used pymarc to extract their content into a flat, tabular format while retaining the original tag associated with each entry. I grouped all name strings from

[3] Corporate entities usually designate institutions such as religious orders, universities, academies, government bodies, or, in some cases, printing houses and workshops, which may be catalogued as such rather than as individual printers.

subfield $a into a single column, regardless of the MARC field they originated from. I applied the same principle to other associated attributes—such as numeration, life dates, and roles—by creating standardized columns based on subfield content, across different MARC tags.

Tag	LoC Definition	Content
Agents		
100	Main Entry–Personal Name	Personal name primarily responsible for the work (e.g., author)
110	Main Entry–Corporate Name	Institutional body primarily responsible for the work (e.g., university press)
600	Subject Added Entry–Personal Name	Person discussed or referenced in the work (e.g., biography subject)
610	Subject Added Entry–Corporate Name	Corporate entity discussed or referenced (e.g., religious order, university)
700	Added Entry–Personal Name	Additional individual associated with the work (e.g., editor, translator)
710	Added Entry–Corporate Name	Additional institutional body involved (e.g., publishers, sponsors)
Content		
245	Title Statement	Full title of the work, including subtitles
650	Subject Added Entry–Topical Term	Keywords or subject headings describing the topic
655	Index Term–Genre/Form	Material type or genre designation (e.g., ephemera, pamphlets, and other genre or form designations)
Paratext and imprint		
500	General Note	Miscellaneous notes, often paratextual (e.g., dedications, colophons)
260/264	Publication, Distribution, etc. (Imprint)	Publisher information, place, and date of publication
041	Language Code	Language(s) in which the work is written
Access information		
852	Location	Physical location and call number of the holding institution
856	Electronic Location and Access	URL or link to digital version or online resource

Table 1: The table lists selected MARC 21 fields with their Library of Congress definitions and brief content description.

At this stage, I cleaned the strings to ensure consistency: I removed leading and trailing spaces, standardized the formatting of numeration and dates, and eliminated unnecessary punctuation. While reviewing the grouped values, I also checked for misplacements (e.g., numeration incorrectly stored in a name subfield) and reassigned them to the appropriate column when necessary.

I then processed the resulting dataframe in OpenRefine, which allows for semi-automated clustering with high user oversight.[4] I applied different clustering techniques and manually evaluated the suggested matches. I recommend working on a duplicate column and proceeding in successive passes: first clustering on the combination of surname, given name, numeration, and parsed life dates (possibly formatted as YYYY–YYYY or as YYYY in case only one date is available); then reapplying clustering without the dates; and lastly filtering by common life dates to identify remaining variants that earlier steps may have missed. After this bottom-up procedure, I adopted a top-down approach by matching the clustered names against an external authority file. Given that my case study mainly involved early modern persons active in Europe, the Consortium of European Research Libraries (CERL) Thesaurus proved particularly useful.[5] I relied on the dedicated Python library, cerl, which is a wrapper for the CERL Thesaurus API.[6] However, other reconciliation workflows are equally viable. Many thesauri, including Virtual International Authority File (VIAF), Gemeinsame Normdatai (GND), and Wikidata, offer dedicated reconciliation services within OpenRefine or provide public APIs, which allow users to build custom pipelines for automated queries. Crucially, querying authority files often returns more than one potential match per name. These results must be carefully reviewed and manually approved, ideally by inspecting supporting metadata such as life dates, places of activity, or roles. This step requiring human validation is essential to ensure accurate reconciliation, especially when dealing with common names or minimal contextual information.

External authority files offer three main advantages. First, they often have already resolved duplications across records by grouping known name variants. Second, they enrich the dataset with additional information such as birthplaces, deathplaces, or institutional affiliations, which can be repurposed in later stages of our own analysis.

---

[4] OpenRefine is an open-source data cleaning and transformation tool designed for working with messy or semi-structured data. For more, see OpenRefine, https://openrefine.org/.

[5] The CERL Thesaurus is maintained by the Consortium of European Research Libraries and aggregates historical name authorities for persons and corporate entities active in book production. For more on this, see https://data.cerl.org/thesaurus/.

[6] For more on the Python library cerl, see https://pypi.org/project/cerl/. The library was developed by Andreas Walker. A minor compatibility fix is required to use it with recent versions of urllib3.

Third, if the metadata needs to be merged with other datasets, records reconciled against the same thesaurus can be matched more reliably and seamlessly by relying on the same identifier. Of course, authority files come with some limitations, namely, they are inevitably incomplete, and many lesser-known historical figures remain unlisted. In such cases, these individuals will likely be unmatched in the thesaurus. Still, the majority of complex name variation tends to cluster around well-documented individuals, precisely those who are more likely to appear in authority files. In practice, researchers may need to consult multiple thesauri. To bring consistency to the resulting patchwork of matches, it becomes essential to generate a consistent internal identifier system, one that accounts for entities matched across authority files, those identified through clustering, and those that remain unmatched.

In the context of my pilot study on the *Collectio academica antiqua*, the disambiguation process began with 5,018 name strings extracted from MARC 21 records. I first applied a bottom-up procedure using OpenRefine's clustering functions, combined with biographical information such as life dates and numeration. I then adopted a top-down approach by matching these entities to external authority files. I ended up identifying a total of 4,192 distinct individuals. Among these identified individuals, 47.5% (2,033) were matched to the CERL Thesaurus and a further 5% (213) to other authority sources such as Wikidata or Onderzoekssteunpunt en Databank Intermediaire Structuren (ODIS). The remaining 2,032 (47.5%) had no external match and were retained as internally disambiguated entities.

Disambiguation remains one of the most time-consuming steps in preparing bibliographic metadata, yet it is essential. Without it, individual identities would not be consistently established, and relationships would rest on unreliable associations. Crucially, it is what gives the data meaning for relational study. The process can also create synergies: matching entities to authority files often yields additional information that proves useful even beyond the immediate project.

Geolocating Place Names

Bibliographic metadata is typically rich in geographic information because it often records the place of publication, manufacture, and distribution for each library holding. As with personal and corporate names, place names are typically transcribed as they appear on the title page. This practice, once again, prioritizes fidelity to the original source over internal consistency of the catalog. As a result, the same city may appear under multiple variants. These may reflect different languages, historical spellings, political jurisdictions, or typographical inconsistencies. To make this information usable for relational analysis, place names must be standardized and geolocated. This

task is not fundamentally different from disambiguating historical actors, but it is often somewhat easier due to the more stable nature of place identities. Nonetheless, caution is essential. Even with present-day place names, it is surprisingly easy to misidentify a location because many places can share the same name within or across countries. This issue may be less common for cities that are on average better known and thus more consistently recorded, like printing centers. Still, historical geolocation poses a challenge. The process often requires close attention to historical context and supporting metadata.

For harmonizing place names, best practice is to reconcile each string against a standardized entry in a historical urban gazetteer or geographic authority file. These resources contain a wide range of name variants, including multilingual and obsolete forms, which significantly facilitate disambiguation and geolocation. Among the most comprehensive options are Wikidata, GeoNames, the CERL Thesaurus, the RBMS/BSC Latin Place Names File, and the World Historical Gazetteer, all of which are particularly well suited to historical case studies.[7] In my pilot study of Louvain's *Collectio academica antiqua*, I processed 631 valid place-name strings, reconciled into 166 distinct locations. **Table 2** reports, for the ten most frequent printing locations, the number of distinct place-name variants and the corresponding number of imprints.

After harmonization, locations can be geographically referenced. Although coordinates typically refer to a single point rather than the full spatial extent of a city or region, they still enable meaningful spatial representation. For example, in Schich et al. and de la Croix and Scebba, coordinates are used to visualize city attractiveness based on the mobility of, respectively, notable individuals and medieval and pre-modern scholars affiliated to European academic institutions.

Once coordinates are available, they can be plotted on vectorized shapefiles or digital atlases, which represent geopolitical or cultural regions as spatial polygons. In historical analyses, it is particularly important to work with maps that reflect time-consistent boundaries. Using modern administrative boundaries for early periods can misrepresent political geography and bias spatial patterns. A wide array of historical vectorized digital atlases or shapefiles are available to support this kind of analysis.

---

[7] GeoNames is geographical database providing rich geospatial metadata for each location, including names in multiple languages and spatial coordinates. For more on the World Historical Gazetteer, see https://whgazetteer.org/. The World Historical Gazetteer is a curated index of historical place names with spatial and temporal metadata and can be used by uploading individual or group projects. For more on RBMS/BSC Latin Place Names File, see https://rbms.info/lpn/. The RBMS/BSC Latin Place Names File is a curated list of Latin place names and their modern equivalents, maintained by the Bibliographic Standards Committee (BSC) of the Rare Books and Manuscripts Section (RBMS) of the American Library Association.

Place name	Number of variants	Number of imprints
Louvain	64	1,143
Lyon	18	129
Antwerp	68	493
Ingolstadt	5	9
Mainz	5	17
Amsterdam	14	62
Brussels	30	176
Hanover	1	11
Rome	7	31
Cologne	16	215

Table 2: Top ten unified locations by number of imprints.

Both computational humanists and applied economic historians are increasingly contributing projects in this direction. However, the limited integration between disciplines means that humanists are often unaware of the spatial datasets used by economists, and vice versa. These resources may vary in geographic coverage and temporal resolution, where some provide century-scale snapshots, while others offer yearly updates with shifting polity boundaries. Below, I present a selection of the most commonly encountered and widely used resources across computational humanities and quantitative economic history.

- The Centennia Historical Atlas: Academic Research Edition offers high-resolution, time-sensitive boundaries of European polities between the late Middle Ages and the nineteenth century (Reed).
- The Cliopatria database records overlapping and successor polities from 3400 BCE to the present (Bennett et al.).
- EurAtlas provides digital cartographic reconstructions of European boundaries spanning two millennia.[8]
- Developed by economist Victor Gay and hosted by Harvard Dataverse, the Third Republic France Geographic Information System (TRF-GIS) maps French administrative constituencies spanning the period 1870–1940.[9]

[8] Developed by Christos and Marc-Antoine Nüssli. For more on EurAtlas, see https://euratlas.com/.

[9] For more on the Harvard Dataverse, see https://dataverse.harvard.edu/dataverse/TRF-GIS.

- The IPUMS Mosaic project, as the name suggests, provides a patchwork of harmonized historical census microdata and corresponding geographic boundary files for selected countries in the European area at different points in time.[10]
- The Historical GIS Collection at ETH Zurich, compiled by computational sociologist Ramona Roller, supplies static polygon boundaries for territories within the sixteenth-century Holy Roman Empire, along with attributes such as foundation dates and confessional status.[11]
- The project (Re)counting the Uncounted, led by historian and digital humanist Rombert Stapel, produced the Historical Atlas of the Low Countries, a detailed GIS dataset designed to anchor premodern population censuses to historically accurate administrative units.[12]

A comparative overview of these resources is provided in **Table 3**, which summarizes their temporal and geographic coverage, resolution, and accessibility. While the Centennia, Cliopatria, and EurAtlas shapefiles offer broad historical coverage—either pan-European or global—and span from antiquity to the modern era, they were primarily designed as general-purpose tools and are widely used in applied economic

Name	Coverage (Time)	Coverage (Geography)	Temporal Resolution	Open Access
Centennia Historical Atlas	ca. 1000–2000 CE	Europe	Variable, up to one-tenth of a year	No
Cliopatria	3400 BCE–present	Global	Variable	Yes
EurAtlas	1–2000 CE	Europe	100-year steps	No
TRF-GIS	1870–1940	France	Annual	Yes
IPUMS Mosaic	1770–2003	Europe and selected countries	30-year steps (Europe-wide); finer for some states	Yes
ETH Zurich HGIS	ca. 1500–1600	Holy Roman Empire	Static (single date)	Yes
Historical Atlas of the Low Countries	1350–1850	Low Countries	Variable (per census)	Yes

Table 3: The table compares selected historical shapefiles by time span, spatial scope, temporal resolution, and accessibility.

[10] For more on the IPUMS Mosaic project, see https://mosaic.ipums.org/historical-gis-datafiles.

[11] For more on the Historical GIS Collection at ETH Zurich, see https://www.research-collection.ethz.ch/handle/20.500.11850/472583.

[12] For more on *(Re)counting the Uncounted*, see https://datasets.iisg.amsterdam/dataverse/recountingtheuncounted.

history. By contrast, the other atlases are more narrowly focused, both in geographic scope and in time frame. **Figure 1** plots publication centers of the *Collectio academica antiqua*'s holdings over time, employing time-consistent and evolving boundaries.

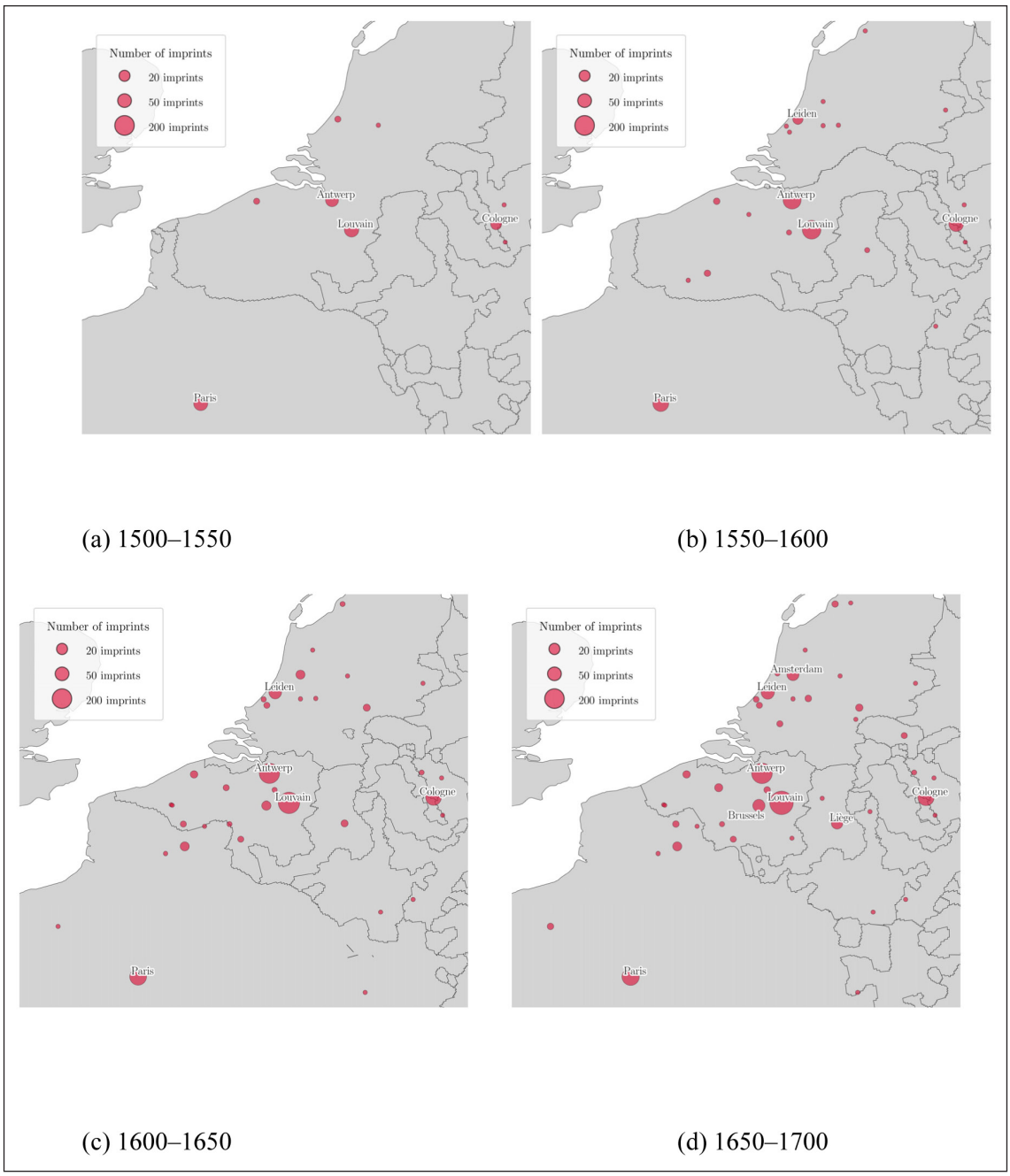

Figure 1: The maps show printing centers in the Low Countries and surroundings at fifty-year intervals. Bubble area is proportional to the number of printed holdings from the *Collectio academica antiqua* associated with each city. Historical boundaries are time-varying and drawn from the Centennia Historical Atlas.

Classifying Content: Assigning Topical Terms

Library catalogs typically include some indication of what each book is about. In the MARC 21 schema, this information tends to be stored in tags 650 and 655, which refer respectively to topical words and genre categories. However, this metadata might be incomplete, inconsistently applied, or skewed by cataloging practices. While a close reading approach might allow for a direct understanding of the subject of each book, this method is not feasible in large-scale, data-driven studies. In these cases, a systematic and ideally automated workflow for classifying the content of library holdings becomes essential. Most of these approaches draw on a key element, that is, textual input, most commonly the book title, which in early modern imprints is often lengthy and content-rich. Although full-text content could, in theory, offer even richer input, the focus here is on completing and standardizing existing catalog metadata to enable structured comparisons across records. Assigning accurate subject metadata is an essential preprocessing step for higher-level analyses tracking thematic trends over time, identifying disciplinary clusters, or studying the diffusion of ideas across communities.

Automated classification techniques broadly fall into two categories: supervised and unsupervised methods. Supervised methods rely on a labeled dataset, that is, a subset of records for which the correct classification (e.g., field or topic) is already known. Machine-learning models are trained to learn these associations and predict labels for unclassified data. An example of supervised topic modeling in the case of bibliographic metadata is introduced in Koschnick and it involves fine-tuning a pretrained transformer model such as DistilBERT (Sanh et al.) on a dataset of titles or abstracts paired with subject headings. The model learns to classify each input into one of the predefined topics, based on patterns in the text. Unsupervised methods, by contrast, do not require labeled input. Topic modeling techniques, such as Latent Dirichlet Allocation (LDA) (Blei et al.) attempt to infer hidden thematic structures by analyzing word co-occurrence patterns in a corpus. Similarly, algorithms such as k-means clustering can group texts based on similarity in their vector representations without any prior knowledge of categories. Applications of k-means clustering to bibliographic metadata include the analysis of genre-indicating subtitles in German literature (Gittel) and the definition of academic fields from publication titles (Curtis and de la Croix).

While unsupervised methods can help identify thematic patterns, they are often less appropriate when the goal is to complete or standardize existing subject metadata. In such cases, including the present paper, supervised approaches are favored because they directly learn from cataloged examples and preserve alignment with established classification schemes.

In my case study of the *Collectio academica antiqua*, I used a supervised approach. I identified a topical classification term for 38.4% of the holdings (1,450 out of 3,777 records). This subset served as the labeled data for training. I then proceeded to derive a higher-level classification by grouping subject headings in more general categories. To assign topics to the books in the *Collectio academica antiqua* that lacked subject metadata, I fine-tuned a DistilBERT model to classify titles based on the labeled records. Each title was mapped to one of three broad disciplinary categories—theology, humanities, or sciences—based on the existing catalog labels. Because the model was pretrained on English text, I first translated titles to ensure compatibility. The model was then trained over four epochs, that is, it went through the entire training dataset four times to refine its predictions.

To evaluate the model, I set aside 20% of the labeled data (290 titles) as a validation set. Model performance differs markedly across categories. As displayed in **Table 4**, precision, recall, and F1 scores are high for theology and humanities, the two dominant classes in the labeled data, indicating that the classifier learns stable and interpretable patterns for these categories. By contrast, the sciences category is severely underrepresented in the training set, and the model fails to predict this class in the validation data, resulting in zero recall for this category. As a consequence, overall accuracy is driven by the two majority classes, while macro-averaged performance metrics provide a more informative summary under class imbalance.

	Precision	Recall	F1-score	Support
Humanities	0.793	0.836	0.814	110
Sciences	0.000	0.000	0.000	6
Theology	0.897	0.897	0.897	174

Table 4: The table reports precision, recall, and F1 scores by class for the validation set of the *Collectio academica antiqua*. Support indicates the number of titles per category in the held-out validation data. Performance is strong for theology and humanities, the two majority classes, while the sciences category is severely underrepresented and not predicted by the model, reflecting the limits imposed by class imbalance rather than classifier instability.

Figure 2 illustrates this imbalance, reporting both raw counts and row-normalized percentages in order to make performance comparable across unevenly sized classes. Each row of the matrix represents the actual category of the entries from the validation set, while each column shows the predicted category assigned by the model. If the model worked perfectly, each class count would fall in the diagonal, where predicted labels match the actual ones, and the off-diagonal entries, which indicate misclassification,

would be empty. Most errors occurred between theology and humanities, which reflect their historical overlap. The sciences category is severely underrepresented in the training data, and the model fails to predict this class in the validation set, reflecting the limits imposed by class imbalance. Once the model was trained, I used it to assign topics to the remaining set of titles that lacked subject metadata. These predicted labels were then combined with the original metadata to generate a complete, standardized topical classification across the entire collection.

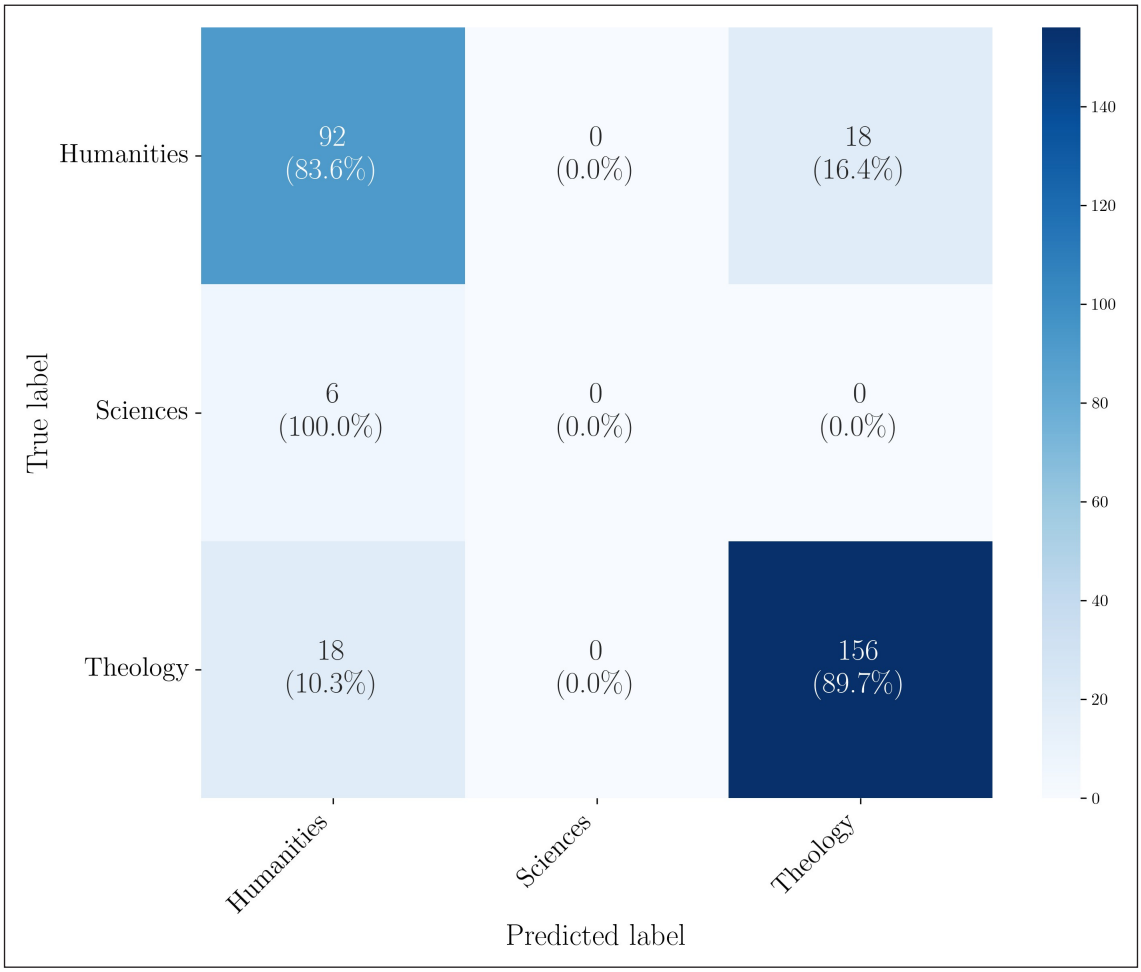

Figure 2: Confusion matrix for the validation set of the labeled holdings in the *Collectio academica antiqua*. Each row shows the true topic based on catalog metadata, and each column shows the model's predicted topic. Values along the diagonal (off-diagonal) indicate correct (incorrect) classifications.

Section 4: Modeling Catalog Data for Analysis

This section returns to the research questions outlined in the Introduction by examining how different modeling choices condition the interpretation of bibliographic metadata when represented as relational and network-based data. Having outlined the steps

required to transform bibliographic metadata into analyzable data, I now turn to their analytical modeling, focusing on network representations. First, I revisit the assumptions and limitations of interpreting catalog metadata as a graph. Then, drawing on a co-occurrence network derived from the *Collectio academica antiqua*, I show how the role attribute encoded in catalog metadata is not naturally accommodated by a single-layer projection and how a multilayer framework offers a coherent way to preserve this heterogeneity when adopting a co-occurrence representation. Lastly, I turn to quantitative economic history as a comparative analytical framework for the use of bibliographic metadata.

Adopting a Network Perspective

Catalog records are not inherently networked systems. However, they do encode structured associations among entities—authors linked to books, books linked to subjects, people linked to other people via shared roles in a publication, and so on. A network representation emerges only when we interpret these associations under a specific set of assumptions. For example, if two agents appear together on a title page, one might interpret that co-occurrence as a connection between them. Many studies in the computational humanities adopt this strategy. For instance, Gavin and Hill et al., "Reconstructing Intellectual Networks" on early English literary criticism and book trade, respectively; Greteman on the English print network from its origins through the eighteenth century; Ladd on early modern dedications; Valleriani et al. on networks of printers and publishers and their influence on the evolution of scientific knowledge; Ryan and Tolonen on Scottish Enlightenment publishing; and Heßbrüggen-Walter on interdisciplinarity in seventeenth-century German dissertations. All of these works use the co-occurrence of individuals in print artifacts as an interpretative lens to reconstruct intellectual, textual, or social communities. This raises two broader methodological questions: Under what interpretive assumptions can bibliographic metadata be treated as network data, and what are the limits of such representations? The *Collectio academica antiqua* provides a particularly fitting test for these questions because it documents a well-bounded institutional world (i.e., an academic print production orbiting around the university center of Louvain) where the logic of collaboration can be observed directly in the paratext and imprint data. This discussion directly addresses the second research question mentioned early on in the paper, by making explicit the interpretive assumptions under which bibliographic metadata can be treated as network data and by delineating the limits of co-occurrence-based representations.

First and foremost, library catalog data were never intended to capture explicit relationships between historical actors. Their purpose is providing bibliographic description and information management. Nonetheless, converting a collection's

metadata into a co-occurrence network means establishing links based on the structured associations present in the records (people, texts, places, and topics connected via shared bibliographic entries). This reinterpretation is not without pitfalls. A key risk is assuming that co-occurrence automatically implies a meaningful social relationship. Simply sharing a title page does not guarantee contemporaneity, collaboration, or even mutual awareness, especially in cases like posthumous editions or honorary dedications.

Yet, despite this ambiguity, the connections we derive from bibliographic metadata are grounded in historical evidence, that is, the sources themselves: title pages, imprint statements, colophons, and dedications. They are not arbitrary, ad hoc constructions imposed by the researcher. While the meaning of each connection may be open to interpretation, the very co-occurrence in the production of a shared work constitutes material evidence derived directly from the historical record (Hill et al., "Reconstructing Intellectual Networks"). At the same time, bibliographic metadata can sometimes support a stronger definition of network ties than simple co-occurrence. For instance, using bibliographic metadata from the *Sphaera* corpus, a curated collection of 359 astronomy and cosmology textbooks published between 1472 and 1650, Valleriani et al. reconstruct a network of book producers by defining "awareness relationships" between early modern printers and publishers. Such relationships arise when two similar editions are attributed to different producers, typically indicating imitation or participation in the same print run. Here, bibliographic fingerprint of editions and additional metadata on printers' and publishers' biographical timelines are used to establish, respectively, chains of editions and contemporaneity, yielding ties that can be interpreted more robustly than co-occurrence alone.

The Multilayer Approach

With these caveats in mind, bibliographic metadata have been used to construct various kinds of networks depending on the entities and relationships of interest. One of the most common projections is a person-to-person network where an edge represents two individuals' joint participation in the same publication. This co-occurrence network is frequently employed to study scholarly or intellectual communities. Even this straightforward representation involves important modeling choices. A first decision concerns an edge's directionality. Defining edges as directed or undirected depends on the interpretative aim. For example, cases reflecting unilateral acknowledgment like links from authors to dedicatees might call for directed edges, whereas undirected edges more accurately capture the symmetric nature of shared participation in other

contexts. A second choice concerns encoding some data as node attributes (e.g., relevant life dates, biographical information, and religious or institutional affiliation), or as edge attributes (e.g., the publication year and the number of shared works).

However, there is one critical dimension that resists both types of attribute encoding: the role of each person within the book. This is because roles are not fixed properties of individuals. Rather, a node may be an author in a publication and a dedicatee in another. Nor do they characterize the tie itself, because each tie reflects individual involvement in a shared work, not the work as a whole. For example, the connection between a printer and an author cannot be reduced to either an attribute of the person or an attribute of their tie.[13] Basically, individuals' roles in each publication are asymmetric and two-way, which disqualifies both node and edge attributes as adequate carriers of this information. Capturing these asymmetries is essential for understanding early modern print economies, where the same person could occupy different positions across publications.

Hence, I argue that the most coherent solution is to model the co-occurrence network as a multilayer graph. Following the definition offered by Kivelä et al., a multilayer network extends beyond nodes and edges to include *layers* as core components, allowing each node to appear in multiple layers and enabling edges to connect any pair of node-layer instances. Additionally, multiple *aspects*, or features—that is, different sets of layers—can be modeled on top of each other (209). In practice, an aspect defines a distinct dimension of variation in the data, such as the type of activity, the temporal slice, or the institutional context, so that multiple aspects can coexist within one unified, multilayer structure.

In the present case, I only focus on a single aspect, the co-occurrence dimension in the catalog, where each layer corresponds to a distinct role and where nodes may participate in one or more role-layers depending on their involvement in each publication. This framework allows for both inter-layer edges (i.e., connections stemming from shared works in different roles) and intra-layer edges (i.e., ties between individuals sharing the same publication *and* role). This representation makes it possible to measure how role multiplicity structures collaboration and to distinguish between within-role cohesion and cross-role integration.

[13] An exception arises when focusing solely on a specific mode of co-occurrence, such as author–dedicatee pairings (see Ladd).

Figure 3 offers a visual example of what a multilayer co-occurrence network can look like.[14] I built a person-to-person co-occurrence network from the bibliographic metadata of the *Collectio academica antiqua*. The layers are defined from the roles recorded in the metadata, which I grouped into broader categories as detailed in **Table 6** in Appendix A.1.[15] The figure illustrates the methodological shift from descriptive catalog lists to an explicitly relational view, where each role layer isolates one channel of participation. Although the full network comprises several thousand individuals and ten role layers, the figure focuses on the fifty most active nodes and four main layers to improve legibility. In this representation, layers are the loci where nodes appear

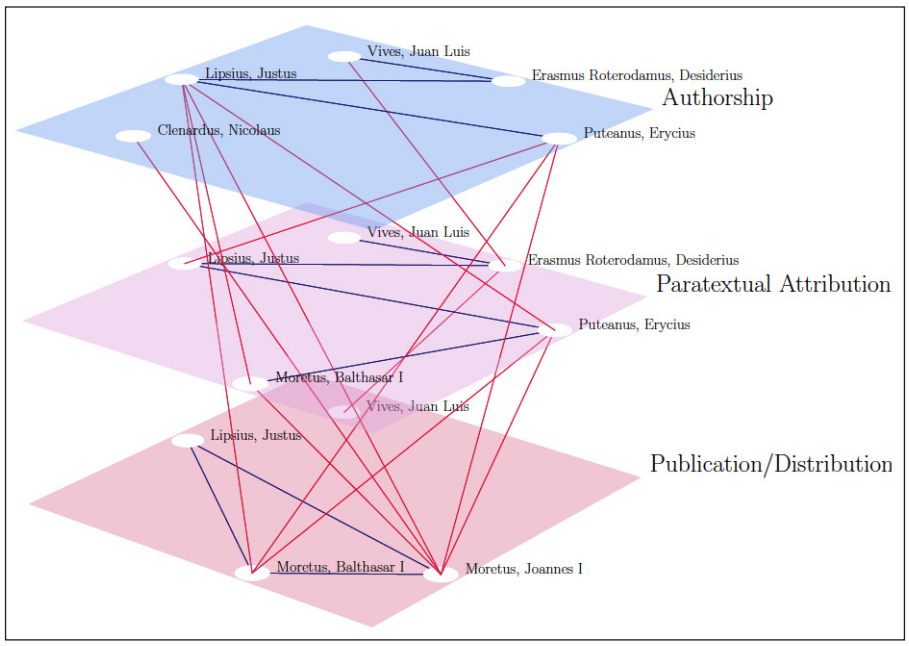

Figure 3: Multilayer person-to-person co-occurrence network constructed from the bibliographic metadata of the *Collectio academica antiqua*. Each layer isolates one type of role-based connection. Nodes represent individuals (only those with highest degree are shown for legibility). Intra-layer edges (blue lines) and inter-layer edges (red lines) connect individuals who co-occur in at least one book, respectively in the same or in different roles. The plots were generated with pymnet (Nurmi et al.).

[14] Multilayer networks are not to be confused with multiplex networks. As discussed in See Kivelä et al., the latter are a special case of the former in which inter-layer links *only* connect identical nodes across layers (diagonal coupling), and layers are often—though not necessarily—node-aligned (i.e., all nodes exist in every layer). In the present case, we refer specifically to a multilayer network, as the very nature of co-occurrence entails inter-layer edges between *different* nodes, and nodes are not guaranteed to appear in every layer.

[15] This grouping is a practical necessity for visualization and interpretive clarity, yet it is not without consequences. Aggregating fine-grained roles into broader categories reduces the apparent number of ties and slightly alters degree distributions, since individuals active under multiple specific designations (e.g., “printer” and “publisher”) are counted as one.

when a given role emerges in at least one shared publication. Thus, if an individual has different roles in different works, they appear in multiple layers. The layer classification preserves the co-occurrence principle: Edges are still defined by co-occurrence in publications. Dashed lines denote inter-layer edges and solid lines denote intra-layer edges.

Thinking of co-occurrence in multilayer terms allows for several modes of analysis. One can examine each layer separately to study the network of collaborations within a single role. This is particularly useful when intra-layer ties dominate or when the aim is to identify role-specific structural patterns. In practice, within-role co-occurrence in early modern data tends to be sparse. In the *Collectio academica antiqua* network, fewer than one-quarter of the ties link individuals who share the same role in a publication; more than 75% of connections are between people in different roles (inter-layer links; see **Table 8** in the Appendix). This is consistent with the low density observed in each individual layer (**Table 7**), which in turn reflects the historical reality that collaborative roles, like for instance co-authorship, were relatively rare in early modern print culture. If we were to apply this method to modern scientific publications, we would expect to see a much denser authorship co-occurrence layer, whereas the early modern dataset under analysis shows that authors typically connect to others through complementary roles (e.g., author to printer, author to dedicatee, etc.) rather than directly to other authors. This dominance of inter-layer ties underscores how production and intellectual labor were intertwined: Most connections bridge material and scholarly functions rather than occur within a single occupational sphere.

Second, the layers can also be collapsed into a single projection when applying classical metrics such as centrality or modularity, though doing so sacrifices the interpretive nuance that multilayer distinctions offer.

Third, and most distinctively, one can analyze patterns that involve the interplay between layers, something not possible in a single-layer projection. In this case, the layers linked to material production (printers, publishers, and booksellers) provide a controlled setting to study how the combination of roles across publications relates to the university ecosystem from which much of the corpus originated.[16] Focusing on the production side is particularly revealing because the *Collectio academica antiqua* primarily contains academic works—dissertations, textbooks, and reprints of classical

[16] Unlike in the general multilayer representation, in the production-side analysis I rely on the original, fine-grained role designations instead of the aggregated macro-categories. The distinction between "printer," "publisher," and "bookseller" is historically meaningful in the early modern context and captures variations in economic function that would be obscured by broader grouping.

authors—commissioned for or produced within the orbit of the Old University of Louvain. In such a context, holding multiple production roles can be read as an indicator of alignment with the university's printing economy, where a small number of trusted workshops frequently combined several functions under one roof. Analyzing these overlaps therefore allows one to test whether functional integration within print workshops corresponded to closer institutional or intellectual integration with the university world itself.

To examine this mechanism empirically, I traced the involvement of academic actors across the corpus. The Studium.AI research infrastructure is developing a comprehensive list of the students enrolled at the Old University of Louvain, but academic scholars can already be identified through the *Repertorium Eruditorum Totius Europae* (RETE) project, a pan-European prosopographical database documenting the university careers of medieval and early modern professors.[17]

Through the disambiguation process described in Section 3's Identifying Historical Actors, I linked each person in the *Collectio academica antiqua* to a unique CERL and Wikidata identifier. This, combined with a manual search, allows the matching of individuals appearing in the collection to scholars from the RETE database. Given RETE's pan-European scope, the linkage allowed me to identify which books involve professors in general and which specifically involve professors affiliated with Louvain. The integration with RETE data adds a socio-institutional dimension to the bibliographic co-occurrence network: Some nodes now carry information on whether the person held a university position, thus allowing the relationship between print production and academia to be tested empirically. I find that around three-fourths of the books of the *Collectio academica antiqua* see the participation of a university professor, specifically Louvain professors in most of the cases but also professors active at other European universities (around 7%). These figures confirm that although the collection is institutionally rooted in Louvain, its intellectual network extended beyond the city. After all, the transregional dimension of the book trade in the Low Countries is a well-established fact in the literature (De Ridder et al.).

I then tested whether production agents who held multiple roles, for instance those appearing as both printer and bookseller in different publications, are more frequently associated with books linked to professors. The test is based on a simple comparison of proportions. For each book, I classified the involved production agents as single-role or multi-role and recorded whether the book involves at least one professor. I computed the share of professor-linked books among multi-role and single-role agents and

[17] For a provisional interface of the project, see https://studium-ai.org/. For more on the RETE project, see https://ojs.uclouvain.be/index.php/RETE/about.

assessed whether the difference is larger than expected by chance using a difference-in-proportions (z) test.[18]

Results in **Table 5** show that books involving multi-role production agents are significantly more likely to involve professors. The share of professor-linked books is about 83% for multi-role agents, compared with 73% for single-role agents, a difference of 9 percentage points ($z = 6.96$, $p = 3.35\times10^{-12}$). The association is even stronger for books involving professors affiliated with Louvain, where the corresponding shares are 77% and 63%, yielding a difference of 14 percentage points ($z = 9.53$, $p = 1.61\times10^{-21}$). These patterns illustrate how preserving role differentiation in a multilayer co-occurrence representation makes visible systematic associations between production structure and academic involvement that would be obscured in a flattened network view. In this sense, the case study demonstrates how modeling choices are not merely technical, but substantively shape the kinds of institutional and historical patterns that can be recovered from catalog data.

Group	$\text{Share}_{\text{profs}}$	$\text{Share}_{\text{Louvain}}$	Observations
Single-role	0.734	0.631	1,589
Multi-role	0.826	0.770	2,403
Difference (multi minus single)	0.092	0.139	
z statistic	6.96	9.53	
p value	3.35×10^{-12}	1.61×10^{-21}	

Table 5: The table reports association of multi-role production agents with professors.

The same reasoning applies when considering how many distinct layers each person appears in, sometimes called the actor's "multiplexity," or the number of role dimensions in their activity. In the *Collectio academica antiqua* network, only about 12% of individuals participate in more than one role layer (see **Table 9** in the Appendix). Those who do are often individuals who also served in another capacity (e.g., editors of other work, authors who also wrote dedications, etc.). **Table 10** in the Appendix shows the share of multi-role individuals in each layer. Unsurprisingly, roles like authorship and paratextual attributions account for a large fraction of the multiplex individuals, whereas very few censors, translators, or illustrators in this dataset also took on additional roles. The prevalence of paratextual connections (e.g., authors who share a printer also often share an editor or a contributor) is partly a reflection of

[18] The test compares proportions across two independent groups using a large-sample normal approximation to the difference in means of a binary outcome. Given the sample sizes involved, the approximation is appropriate.

how thoroughly this particular catalog records those secondary roles. By contrast, the censorship role appears underrepresented in the network. Historically, from the 1520s onward, works printed in places like Louvain often included an approbation or censor's note (Cammaerts 241), but in the data derived from the *Collectio academica antiqua*, such connections are few. This likely indicates that many approbations were not captured in the catalog metadata, not that censorship was absent. A comparison with a dataset where censorship information is fully recorded could yield a denser censorship layer and alter the overall network structure. This example highlights how the depth and focus of cataloging practices (which information was recorded or omitted) directly influence the networks we can extract. In this perspective, a comparative analysis using this analytical lens could substantiate how differences in cataloging practices shape resulting network structures.

Insights from Quantitative Economic History

Some studies in history and literary scholarship that use bibliographic metadata often combine meticulous preprocessing pipelines with network visualizations and descriptive commentary, typically focused on centrality measures. These exercises undoubtedly elevate our understanding of the collections and catalogues used as material, yet, as Lemercier and Zalc observe, such work frequently remains at the stage of mapping and visual speculation, without fully mobilizing networks as formal analytical tools. The patterns observed in Section 4.1 (e.g., the sparsity of within-role links or the rarity of multi-role connections) highlight both the potential and the limitations of co-occurrence networks derived solely from catalog data. On one hand, such networks can reveal hidden structures in the catalog or collection studied; on the other, relying on networked bibliographic data exclusively, without complementary data or further analytical techniques, may limit the types of historical questions we can answer. To overcome this limitation, bibliographic network studies can draw on frameworks that have demonstrated how relational data can be embedded in explicitly articulated analytical designs, most notably, quantitative economic history.

In recent years, bibliographic metadata have begun to feature in applied economic history, often alongside other sources. These studies go beyond descriptive and exploratory scopes and use a range of empirical tools beyond network analysis. Importantly, bibliographic metadata are not treated as econometric data but instead as surviving historical evidence, interpreted and combined with other sources within composite datasets tailored to specific historical questions, for example, building on library catalogues, biographical sources, and manuscript data. Chaney, "Religion and the Rise and Fall of Islamic Science", and "Modern Library Holdings Historic City Growth" constructs a georeferenced dataset of authors from which he derives metrics

such as author counts, the share of authors engaged in scientific topics, and number of authors' deaths. He uses this resource to address two distinct questions, respectively, that of the decline of scientific output in the medieval Islamic world and that of the enhancement of preindustrial city growth estimates.

Chiopris analyzes how spatial connections, namely the introduction of the railroad network in nineteenth-century German-speaking regions, shaped the creation and diffusion of ideas. Using comprehensive library catalogues from the German Collective Library Consortium, the author measures the emergence and spread of new ideas, broadly defined as novel publication topics, new words, or new combinations of existing concepts derived from enriched catalogue metadata.

Koschnick links author-level publication data from the English Short Title Catalogue (ESTC) to collegiate records from Oxford and Cambridge Universities to examine the diffusion of ideas between teachers and students from 1600 to 1800. In another study, de Pleijt and Koschnick combine ESTC data with information on scholars' socioeconomic backgrounds, and using sentiment analysis on the titles of printed works, they find that limited career prospects for less advantaged academics coincided with a rise in dissenting religious publications.

Finally, Cervellati et al. use metadata from the epistolary union catalogue Early Modern Letters Online (EMLO) to study the role of the Republic of Letters in explaining why Britain experienced a sustained economic take-off at the time of the Industrial Revolution. These contributions do not leverage networks as an analytical lens, but they do use bibliographic metadata as part of broader quantitative frameworks to address questions about scholarly output, institutional change, and idea diffusion. What they also show is that bibliographic metadata can sustain formal empirical strategies once linked with complementary datasets, to the end of the pursuit of a specific research question.

The disciplinary silo between quantitative economic history and digital or computational history has left analytical strategies developed in the former largely unexplored in the latter. From this perspective, large-scale patterns revealed by bibliographic metadata can be connected to deeper historical questions when analyzed with the appropriate combination of domain knowledge and formal methods. A crucial step in this direction involves enriching catalog metadata with external sources, based on the research question at hand. The multilayer model proposed in Section 4 might operationalize this integration as follows. The bibliographic co-occurrence network can be conceived as an initial aspect or feature, that is, a first set of layers, onto which additional layers sets could be stacked to encode relations for the same entities drawn from complementary sources. This approach facilitates the incorporation of external knowledge into digital history analyses and lay the groundwork for applying more sophisticated network techniques, like, for example, peer-effects models.

Section 5: Conclusions

In an effort to provide perspective on research materials and methods from neighboring fields, this paper has highlighted a growing convergence of computational history and quantitative economic history around shared sources and tools, most notably bibliographic metadata and network analysis.

Using a role-differentiated co-occurrence network representation of the *Collectio academica antiqua* of the Old University of Louvain, I demonstrated how different involvement types in book production (authorship, publication and distribution, dedication, censorship, and so on) can be modeled without being flattened into a single type of tie. The interplay of inter- and intra-layer connections reveals structural patterns that reflect both historical publishing practices and cataloging norms. Yet the exploration of my case study also shows that when bibliographic metadata is used in isolation and interpreted solely through network visualizations, the analytical power remains limited.

This underlines two distinct but complementary points. First, when bibliographic metadata are modeled as networks, analytical coherence depends on preserving the internal structure of the records rather than flattening heterogeneous forms of participation into undifferentiated ties. Second, even when such structure is preserved, bibliographic metadata alone offer limited leverage for addressing complex historical questions. Their analytical potential is fully realized only when relational representations derived from catalog data are linked to complementary sources that provide institutional, social, or biographical context. Together, these points respond to the research questions posed at the outset by clarifying both the conditions under which bibliographic metadata can be meaningfully modeled as relational data and the limits of such modeling when used in isolation.

In this sense, the contribution of the present article is threefold: a methodologically reflective case study, a role-aware relational modeling strategy, and a cross-field perspective that situates bibliographic metadata at the intersection of humanities and social science research.

Appendix

Mapping of Roles

This subsection presents the grouping strategy used to translate detailed role descriptors into broader layer categories for multilayer network construction. The table clarifies how fine-grained roles, such as "writer of preface," "censor," "engraver," and others, are aggregated into interpretable layers like Paratextual Attributions, Censorship, and Engraving/Illustration. The grouping reflects a deliberate analytical choice designed to balance representational fidelity with visual and conceptual clarity. Other researchers may opt for alternate classifications according to their research aims.

Layer	Granular roles
Authorship	author; author original work; dubious author
Paratextual Attributions	writer of preface; compiler; adapter; readapted by; author in quotations or text abstracts; narrator; commentator for written text; author of introduction, etc.; corrector; compiler of index; eulogist; collaborator; contributor
Translation	translator
Publication/Distribution	printer; editor; bookseller; publisher; auctioneer; patron
Dissertant	dissertant
Promotor	praeses; thesis advisor
Censorship	approbation; censor Engraving/Illustration engraver; illustrator; woodcutter; artist
Dedication	dedicatee; honoree; depicted; addressee
Unknown	other; responsible party; collection from; missing or unspecified role

Table 6: The table describes the mapping of granular roles to broader layer categories.

Per-layer Network Analysis

Layer	Nodes	Intra-layer edges	Density
Authorship	467	1,103	0.0101
Censorship	172	718	0.0488
Dedication	376	1,120	0.0159
Dissertant	42	50	0.0581
Engraving/Illustration	71	68	0.0274
Paratextual Attributions	419	1,574	0.0180

(Contd.)

Layer	Nodes	Intra-layer edges	Density
Promotor	4	2	0.3333
Publication/Distribution	948	1,200	0.0027
Translation	35	31	0.0521
Unknown	277	670	0.0175

Table 7: The table reports per-layer counts of nodes, intra-layer edges, and density.

Metric	Value
Total intra-layer edges	6,536
Total inter-layer edges	20,487
Total edges overall	27,023
Total unique nodes across all layers	4,119
Total node-layer combinations	4,905

Table 8: The table reports global multilayer network statistics, including node counts, layer combinations, and the number of intra- and inter-layer edges.

Inter-layer Network Analysis

Min	Q1	Median	Mean	Q3	Max	Single role	Multiple roles
1	1.0	1.0	1.2	1.0	7	3,595 (87.3%)	524 (12.7%)

Table 9: The table reports descriptive statistics of degree of multiplexity in the *Collectio academica antiqua* network.

Layer	Share
Authorship	0.618
Unknown	0.529
Paratextual Attributions	0.471
Publication/Distribution	0.328
Dedication	0.313
Censorship	0.095
Translation	0.088
Promotor	0.034
Engraving/Illustration	0.017
Dissertant	0.006

Table 10: The table reports the shares of multiplex nodes present in each layer.

Data Availability

All scripts and data supporting this article are available through the *Journal of Cultural Analytics* Dataverse repository.

Acknowledgements

The author acknowledges the support of the Global PhD Partnerships between KU Leuven and UCLouvain. The author thanks the two anonymous referees and the editor, as well as Margherita Fantoli, Violet Soen, Dirk van Miert, and Mikko Tolonen for their constructive feedback and comments.

Competing Interests

The author has no competing interests to declare.

Works Cited

Angrist, Joshua D, and Jorn-Steffen Pischke. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *The Journal of Economic Perspectives*, vol. 24, no. 2, 2010, pp. 3–30, DOI: 10.1257/jep.24.2.3.

Arora, Abhishek, Emily Silcock, Leander Heldring, and Melissa Dell. "Contrastive Entity Coreference and Disambiguation for Historical Texts." 2024, arXiv:2406.15576.

Avram, Henriette. "MARC: Its History and Implications." Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock number LC-1.2:M18/16), 1975.

Bennett, James S., Erin Mutch, Andrew Tollefson, Ed Chalstrey, Majid Benam, Enrico Cioni, Jenny Reddish, et al. "Cliopatria: A Geospatial Database of World-Wide Political Entities from 3400 BCE to 2024 CE." *Scientific Data*, vol. 12, no. 247, 2025, DOI: 10.1038/s41597-025-04516-9.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993–1022, DOI: 10.5555/944919.944937.

Cammaerts, Dieter. "Manuale Lovaniense: Een sociaaleconomische en typografische studie van het gedrukte handboek aan de vroegmoderne Leuvense universiteit (1474–1650)." 2024. KU Leuven, PhD dissertation.

Cantoni, Davide, and Noam Yuchtman. "Historical Natural Experiments: Bridging Economics and Economic History." *The Handbook of Historical Economics*, Elsevier, 2021, pp. 213–41.

Cervellati, Matteo, Sara Lazzaroni, Gianni Marciante, and Paolo Masella."The Rise of the Knowledge Economy: Republic of Letters and Communication Infrastructures in Early Modern England." Working paper, 2025.

Chaney, Eric. "Modern Library Holdings and Historic City Growth." Working paper, 2024.

Chaney, Eric. "Religion and the Rise and Fall of Islamic Science." Working paper, 2023.

Chiopris, Caterina. "The Diffusion of Ideas." Working paper, 2024.

Curtis, Matthew, and David de la Croix. "Seeds of Knowledge: Premodern Scholarship, Academic Fields, and European Growth." 2025, SSRN Working Paper, DOI: 10.2139/ssrn.5078307.

de la Croix, David, and Rossana Scebba. "Geolocalization and the Birth-to-Death Distance." *Repertorium Eruditorum Totius Europae*, vol. 14, 2024, pp. 37–42, DOI: 10.14428/rete.v14i0/Locations.

de Pleijt, Alexandra, and Julius Koschnick. "Alienated Intellectuals? Exploring the Political Consequences of the Educational Revolution in Early Modern England." Working paper, 2025.

De Ridder, Bram, Violet Soen, Werner Thomas, and Sophie Verreyken. *Transregional Territories: Crossing Borders in the Early Modern Low Countries and Beyond*. Brepols Publishers, 2020.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. "Named Entity Recognition and Classification in Historical Documents: A Survey." *ACM Computing Surveys*, vol. 56, no. 2, 2023, pp. 1–47, DOI: 10.1145/3604931.

Fantoli, Margherita, Jukka Suomela, Toon Van Hal, Mark Depauw, Lari Virkki, and Mikko Tolonen. "Quantifying the Presence of Ancient Greek and Latin Classics in Early Modern Britain." *Journal of Cultural Analytics*, vol. 10, no. 1, 2025, DOI: 10.22148/001c.128008.

Gavin, Michael. "Historical Text Networks: The Sociology of Early English Criticism." *Eighteenth-Century Studies*, vol. 50, no. 1, 2016, pp. 53–80, https://www.jstor.org/stable/43956564.

Gay, Victor. "Mapping the Third Republic: A Geographic Information System of France (1870–1940)." *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 54, no. 4, 2021, pp. 189–207, DOI: 10.1080/01615440.2021.1937421.

Gittel, Benjamin. "An Institutional Perspective on Genres: Generic Subtitles in German Literature from 1500–2020." *Journal of Cultural Analytics*, vol. 6, no. 1, 2021, pp. 1–38, DOI: 10.22148/001c.22086.

Gregory, Ian. "Challenges and Opportunities for Digital History." *Frontiers in Digital Humanities*, vol. 1, 2014, pp. 1–2, DOI: 10.3389/fdigh.2014.00001.

Greteman, Blaine. *Networking Print in Shakespeare's England: Influence, Agency, and Revolutionary Change*. Stanford UP, 2021.

Heßbrüggen-Walter, Stefan. "Interdisciplinarity in the Seventeenth Century? A Co-occurrence Analysis of Early Modern German Dissertation Titles." *Synthese*, vol. 203, no. 2, 2024, p. 67, DOI: 10.1007/s11229-024-04494-2.

Hill, Mark J., Ville Vaara, and Mikko Tolonen. "Communication and Idea Transmission Across Historical Communities: A Quantitative Analysis of Early Modern Nonconformist Networks." *Huntington Library Quarterly*, vol. 86, no. 2, 2023, pp. 377–407, DOI: 10.1353/hlq.2023.a936422.

Hill, Mark J., Ville Vaara, Tanja Säily, Leo Lahti, and Mikko Tolonen. "Reconstructing Intellectual Networks: From the ESTC's Bibliographic Metadata to Historical Material." *Proceedings of the Digital Humanities in the Nordic Countries*, 2019.

Hotson, Howard, and Thomas Wallnig, editors. *Reassembling the Republic of Letters in the Digital Age*. Göttingen UP, 2019.

Kivelä Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. "Multilayer Networks." *Journal of Complex Networks*, vol. 2, no. 3, 2014, pp. 203–71, DOI: 10.1093/comnet/cnu016.

Koschnick, Julius. "Teacher-Directed Scientific Change: The Case of the English Scientific Revolution." EHES Working Paper Series, no. 274, 2025.

Ladd, John R. "Imaginative Networks: Tracing Connections Among Early Modern Book Dedications." *Journal of Cultural Analytics*, vol. 6, no. 1, 2021, DOI: 10.22148/001c.21993.

Lahti, Leo, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. "Bibliographic Data Science and the History of the Book (c. 1500–1800)." *Cataloging and Classification Quarterly*, vol. 57, no. 1, 2019, pp. 5–23, DOI: 10.1080/01639374.2018.1543747.

Lahti, Leo, Niko Ilomäki, and Mikko Tolonen. "A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470–1800." *LIBER Quarterly*, vol. 25, no. 2, 2015, pp. 87–116, DOI: 10.18352/lq.10112.

Lahti, Leo, Ville Vaara, Jani Marjanen, and Mikko Tolonen. "Best Practices in Bibliographic Data Science." *Proceedings of the Research Data and Humanities (RDHum) 2019 Conference: Data, Methods and Tools*. Edited by Juhani Harri Jantunen, S. Brunni, N. Kunnas, S. Palviainen, and K. Västi. Studia Humaniora Ouluensia, U of Oulu, 2019, pp. 57–65.

Lemercier, Claire. "A History Without the Social Sciences?" Translated by Angela Krieger. *Annales: Histoire, Sciences Sociales*, vol. 70, no. 2, 2015, pp. 271–83, DOI: 10.1017/S2398568200001163.

Lemercier, Claire, and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction*. Translated by Arthur Goldhammer. U of Virginia P, 2019.

Nurmi, Tarmo, Arash Badie-Modiri, Corinna Coupette, and Mikko Kivelä. "pymnet: A Python Library for Multilayer Networks." *Journal of Open Source Software*, vol. 9, no. 99, 2024, p. 6930, DOI: 10.21105/joss.06930.

Padilla, Thomas. "Foreword." *Library Catalogues as Data: Research, Practice and Usage*. Edited by Paul Gooding, Melissa Terras, and Sarah Ames. Facet Publishing, 2025, pp. 20–21.

Petras, Vivien, Ray R. Larson, and Michael Buckland. "Time Period Directories: A Metadata Infrastructure for Placing Events in Temporal and Geographic Context." *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06)*. ACM, 2006, pp. 151–60, DOI: 10.1145/1141753.1141782.

Reed, Frank. "The Centennia Historical Atlas: Academic Research Edition." *Clockwork Mapping*, 2016.

Roller, Ramona. "Theory-Driven Statistics for the Digital Humanities: Presenting Pitfalls and a Practical Guide by the Example of the Reformation." *Journal of Cultural Analytics*, vol. 7, no. 4, 2023, DOI: 10.22148/001c.57764.

Roller, Ramona. "Tracing the Footsteps of Ideas: Time-Respecting Paths Reveal Key Reformers and Communication Pathways in Protestant Letter Networks." *SocArXiv*, 2023, DOI: 10.31235/osf.io/cfqry.

Ryan, Yann Ciarán, and Mikko Tolonen. "Networks of Influence in Scottish Enlightenment Publishing." *Connections*, vol. 44, no. 1, 2024, DOI: 10.21307/connections-2019.034.

Ryan, Yann Ciarán, and Mikko Tolonen. "The Evolution of Scottish Enlightenment Publishing." *The Historical Journal*, vol. 67, no. 2, 2024, pp. 223–55, DOI: 10.1017/S0018246X23000614.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT: A Distilled Version of BERT—Smaller, Faster, Cheaper and Lighter." 2019, DOI: 10.48550/arXiv.1910.01108.

Scebba, Rossana, and Margherita Fantoli. "Integrating Library and Prosopographical Data in the Publication Network of the Old University of Louvain." *Students, Scholars and Their Books at the University of Louvain (1425–1797)*, edited by Violet Soen, Wouter Druwé, Wim François, Ralph Dekoninck, vol. 17, Lectio Series Studium Lovaniense, 1, Brepols, forthcoming 2026.

Schich, Maximilian, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. "A Network Framework of Cultural History." *Science*, vol. 345, no. 6196, 2014, pp. 558–62, DOI: 10.1126/science.1240064.

Tennant, Roy. "MARC Must Die." *Library Journal*, vol. 127, no. 17, 2002, pp. 26–27, https://www.libraryjournal.com/story/marc-must-die.

Tiihonen, Iiro, Leo Lahti, and Mikko Tolonen. "Print Culture and Economic Constraints: A Quantitative Analysis of Book Prices in Eighteenth-Century Britain." *Explorations in Economic History*, vol. 94, 2024, DOI: 10.1016/j.eeh.2024.101614.

Tolonen, Mikko, Leo Lahti, Hege Roivainen, and Jani Marjanen. "A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828." *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 52, no. 1, 2019, pp. 57–78, DOI: 10.1080/01615440.2018.1526657.

Tolonen, Mikko, Mark J. Hill, Ali Zeeshan Ijaz, Ville Vaara, and Leo Lahti. "Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production." *Data Visualization in Enlightenment Literature and Culture*, 2021, pp. 63–119, DOI: 10.1007/978-3-030-54913-8_3.

Valleriani, Matteo, Malte Vogl, Hassan el-Hajj, and Kim Pham. "The Network of Early Modern Printers and Its Impact on the Evolution of Scientific Knowledge: Automatic Detection of Awareness Relationships." *Histories*, vol. 2, no. 4, 2022, pp. 466–503, DOI: 10.3390/histories2040033.