EPFL

BIBLIOTHECA HERTZIANA
MAX PLANCK INSTITUTE
FOR ART HISTORY

Mapping Bibliotheca Hertziana

École Polytechnique Fédérale de Lausanne
Master Thesis
by Hannah Laureen Casey

Under the supervision of:
Prof. Frédéric Kaplan
Dr. Dario Rodighiero
Dr. Alessandro Adamou

February 29, 2024

Kennst du das Land, wo die Zitronen blühn
— Johann Wolfgang von Goethe

Acknowledgments

I would like to extend my deepest gratitude to all those who have supported me throughout the journey of completing this thesis. First and foremost, I am profoundly grateful to my thesis supervisors, Prof. Fréderic Kaplan, Dr. Dario Rodighiero, and Dr. Alessandro Adamou, for their invaluable guidance, patience, and encouragement from the initial to the final stages of this research. Their insights and direction helped me navigate through the complexities of my study, and their feedback was vital in refining my work.A special note of thanks to my colleagues and friends at Bibliotheca Hertziana for their encouragement and insightful discussions. Their presence and support have made this journey both enjoyable and memorable. The countless games of foosball on the Hertziana terrace kept me going throughout the year. I would like to acknowledge the contribution of Battsooj Enkhbaatar, who helped with the design, and Martin Hulton Bott, who helped with the words. Last but not least, I'm truly thankful to my family and friends for their unwavering love, understanding, and encouragement.

Rome, February 29, 2024 Hannah Laureen Casey

Abstract

The project introduces an innovative visual method for analysing libraries and archives, with a focus on Bibliotheca Hertziana's library collection. This collection, which dates back over a century, is examined by integrating user loan data with deep mapping techniques to reveal usage patterns and thematic clusters. To achieve this, dimensionality reduction is employed to visualise the catalogue, mapping books based on their loans, and prompt engineering with large language models helps to identify loan clusters with detailed descriptions and titles. This approach not only paves the way for cultural analytics but also provides the basis for dynamic classification and developing a recommendation system. This project offers alternative insights into the art historical research conducted at Bibliotheca Hertziana, capturing the collection's evolution and usage. The method established here provides a flexible framework for visually mapping cultural and academic collections in the digital humanities.

Keywords: Art History; ChatGPT; Cultural Collections; Data Visualisation; Digital Libraries; Knowledge Design; Knowledge Organisation; Natural Language Processing.

Résumé

Le projet introduit une nouvelle méthode d'analyse des bibliothèques et archives en se concentrant sur la collection de la Bibliotheca Hertziana. En retraçant ses origines sur plus d'un siècle, cette collection est examinée en intégrant les données des prêts faits aux usagers de la bibliothèque avec des techniques de cartographie détaillées afin de révéler les habitudes d'utilisation et les regroupements thématiques. Pour ce faire, la réduction dimensionnelle est employée afin de visualiser le catalogue, en cartographiant les livres par rapport à leurs prêts respectifs, et le *prompt engineering* avec des grands modèles de language permet d'identifier les groupes de prêts aux descriptions et titres détaillés. Cette approche ouvre non seulement la voie à l'analyse culturelle mais pose également les bases d'une classification dynamique et du développement d'un système de recommandation. Ce projet propose d'autres perspectives dans la recherche en histoire de l'art menée à la Bibliotheca Hertziana sur l'évolution et l'utilisation de sa collection. La méthode établie ici fournit un cadre flexible pour cartographier visuellement des collections culturelles et académiques dans les humanités numériques.

Contents

Chapter 1

Introduction

As one enters through the *mascherone* from Via Gregoriana, leaving the bustling city behind, one is met by a bright space with shelves brimming with books. The library spans several stories of the buildings constituting Bibliotheca Hertziana. In the middle of it all, there is a courtyard where sunlight falls onto what used to be the garden of Palazzo Zuccari. The knowledge here attracts researchers from across the globe, who sit at desks amidst towering piles of books or are hidden away in the corridors, looking for serendipitous discoveries. The library's serene and almost ethereal atmosphere is enhanced by its exclusivity — it is only accessible to scholars holding a degree in art history. But Bibliotheca Hertziana is not merely a treasure trove of books, it is also home to a renowned institute for art historical research. It hosts scholars with various backgrounds who shape the institution and the library collection through their presence and research.

Libraries are storage houses where books are kept, providing access by organising them with standardised classification systems and offering refuge from flickering screens (Schnapp and Battles 2014). Their role extends to encouraging discoveries by placing books together in intuitive ways that promote findings. Traditional library classification schemes collocate books alphabetically or based on subject (Svenonius 2000), but with the advent of tools that can dynamically represent collections, library classifications are no longer necessarily static (Bowker and Star 2008). Through digital library displays, books can now be arranged and re-arranged indefinitely, reflecting the ever-changing landscape of knowledge and allowing users to discover and re-discover collections repeatedly. The objective of this project is to lay the groundwork for such a dynamic classification focusing on

Bibliotheca Hertziana's holdings.

For over a century, the library collection has grown organically with Bibliotheca Hertziana, reflecting developments in art history and the interests of the researchers passing through. Its evolution is affected by acquisition patterns, trends in art historical research, political changes, and individuals involved with the institute. Bibliotheca Hertziana is a library for experts in the field, so their collective knowledge is represented in the holdings. Their patterns of use can be leveraged to create a library display and to form the basis of a classification schema which provides the foundation of an eventual recommendation system. This project aims to graphically represent Bibliotheca Hertziana's library collection, analyse trends and patterns, and understand its historical development. It will also create a flexible framework for dynamic library classification based on deep mapping techniques and prompting large language models.

Chapter 2 gives insight into Bibliotheca Hertziana's background, establishment, and the changes the collection underwent in the past century. The background to understanding the complexities of the local signature system permits a preliminary analysis of the library's acquisitions over time, showing the collection's thematic focus and factors influencing its growth. The collection is ordered using a classification system developed in the 1960s at Bibliotheca Hertziana, designed to place thematically similar books close together. The system is stable but not adapted to the evolving needs of researchers today. Some categories are outdated, and newly established research fields are not represented.

Chapter 3 introduces projects and principles that are pivotal for the conception of the visualisation. It explores previous work on digital libraries, highlighting the importance of classification schemes and their impact on library usage. The discussion underscores the importance of creating a dynamic display of the collection, as stable systems can be designed to fit existing collections but fall short for evolving ones. Furthermore, Bibliotheca Hertziana's collection is a repository of collective knowledge, and as such, its mapping can represent the field of art history in the broader context of scientific research. Both the methodologies commonly used for science mappings that have inspired this project and previous visualisations of collections that employ dimensionality reduction are explored in chapter 3, which will prove a crucial tool to visualise the book network dynamically. The projects presented provide the basis of the methodology used for visualising and classifying Bibliotheca Hertziana's holdings.

The framework developed in the chapter 4 leverages the collective knowledge of scholars by making use of data collected from user loans over the past ten years to create an embedding of books. It utilises similarity measures and dimensionality reduction to reveal the underlying structure of the data. The embeddings resulting from the parameterisation are evaluated using metadata fields to determine which set of input and parameters results in a mapping representative of Bibliotheca Hertziana's context.

To create a flexible classification from the resulting clusters, chapter 5 explores prompting a large language model using word frequencies and book titles in the clusters in several languages. The generated subjects and descriptions are evaluated by select members of the Bibliotheca Hertziana community to validate the visualisation.

This project was conducted over the course of one year in Rome, Lausanne and Zurich. A four-month internship in the spring of 2023 at Bibliotheca Hertziana in the context of the Digital Humanities Master's program at EPFL was of central importance. The code documentation for the internship and the Master's projects can be found on a publicly available GitHub repository (H. Casey 2023). The data collection process and analysis of the library collection described in 2 were conducted during the internship. The methodology followed in chapter 4 was instigated during the internship and further refined during the subsequent months. The project was completed as a Master's thesis from autumn 2023 to February 2024.

Figure 1.1: The *mascherone* entrance of Bibliotheca Hertziana on Via Gregoriana, which used to be the entrance to the garden of Palazzo Zuccari, now leads directly into the library. (*Photo:* Andreas Muhs)

Chapter 2

Background

> A library is a growing organism.
>
> *S.R. Ranganathan*

2.1 A Brief History of Bibliotheca Hertziana

Bibliotheca Hertziana is a library and research institute focusing on Italian art history. It is located in the centre of Rome at the top of the Spanish steps. Its history dates back to 1913 when it was established as one of the institutes of the newly established *Kaiser-Wilhelm-Gesellschaft*, which after World War II became known as the Max-Planck Society, a German organisation conducting basic research in natural sciences, life sciences and humanities. Bibliotheca Hertziana was the first member institute of the *Kaiser-Wilhelm-Gesellschaft* in the humanities.

The Institute was founded by Henriette Hertz, a wealthy Cologne-born woman who hosted a literary and art historical club after purchasing 16th century Palazzo Zuccari (Rischbieter 2004). The institute's founding was extraordinary, as it wasn't instigated by a German scholar but instead goes back to the friendship between Henriette Hertz and Frida Mond, who were former schoolmates. Their idea of establishing an art historical library in Rome started with their assembling a collection of books and photographs in collaboration with the art historian Ernst Steinmann, their academic advisor and friend (Tesche 2002). This endeavour laid the founda-

tion for the eventual establishment of Bibliotheca Hertziana. Following Steinmann's advice, Hertz left the collection and the Palazzo to the *Kaiser-Wilhem-Gesellschaft* posthumously. In her will, she stated that the focus of the collection and the acquisitions should henceforth lie in art historical literature (Ebert-Schifferer 2013). Later, Steinmann was appointed as the first director of Bibliotheca Hertziana and led the institute until his retirement in 1934.

At the time of its opening, Bibliotheca Hertziana possessed around 5'000 volumes and 12'000 photographs, and it swiftly emerged as a pioneering library and a pivotal centre for art historical research. The original collection mostly originated from the personal libraries of Hertz, Mond, and Steinmann and was systematically extended over the following years. Today, the library contains around 350'000 volumes, of which around 1'000 are periodicals. The collection continues to grow yearly by around 6'000, the acquisition profile focused on highly specialised art historical literature, as the library itself caters mainly for academics at the post-doc level.

Throughout the 1960s, both the library's collection and the number of people using it grew rapidly. In response, a new shelf numbering system was introduced. The institute underwent a physical expansion to house the growing number of books and cater to the increasing influx of researchers. In 1985, the entire library catalogue was made available through an Online Public Access Catalogue (OPAC), having previously been accessible through a card catalogue, which had been common practice. Since then, every volume acquired for the collection has been saved with several metadata fields in the catalogue.

A challenge the library faces today, apart from the physical space of bookshelves running out, is the rapid change the field of art history has undergone in recent decades. As an institute for research of Italian art, Bibliotheca Hertziana represented one of the most important branches of art history until the 1990s. Nowadays, research interest has shifted to the flow between many artistic landscapes, such that the original collection focus only partially reflects current academic issues. The geographic framework of the scholarly literature had to be expanded to the areas with which Italy has been in cultural exchange in centuries gone by. The library has to evolve with the field of art history and simultaneously further develop the collection's main focus so as to build on the work of the past century.

Bibliotheca Hertziana's history provides context for exploring its digital catalogue.

The digitisation of the catalogue since 1985 was a significant advancement in the library's archival methods. This digital repository enables an analysis of the catalogue's content and chronological evolution using the metadata of the library's documents. It offers insight into the interplay between the scholarly community affiliated with the institution, the changes within the field of art historical research and the librarians' effort to maintain the relevance and adaptability of Bibliotheca Hertziana as an important resource for research.

Figure 2.1: View inside the *Neubau* of Bibliotheca Hertziana, where internal and external researchers access books and work during the day. The *Neubau* was completed in 2013 and has since been home to the library collection. (*Photo:* Andreas Muhs)

2.2 Data Collection

This visual exploration of Bibliotheca Hertziana's library collection relied on extensive data gathering throughout one year at the institute by knocking on every door and finding out where data was kept. The project was supported by key figures, particularly Tristan Weddigen, one of the institute's directors, and Dario Rodighiero, who supported the project from the beginning. The data collection started with the invaluable help from Alessandro Adamou, digital humanities scientist at the institute, who provided insights into the institute's structure and library collection and established the necessary connections to gain access to the data sources. Initially, it was considered using the SPARQL (Bayerische Staatsbibliothek 2015) endpoint provided by the *Bibliotheks Verbund Bayern* (BVB) to access the catalogue. The BVB is associated with over 150 libraries, including the library of Bibliotheca Hertziana, and operates a union catalogue and database and provides access to general library data via exports and the SPARQL endpoint. However, accessing local metadata such as acquisition date and shelf number was essential for the analysis and accessing it was only possible through locally saved data. The local catalogue data was accessible only through collaboration with library staff, facilitated by the head of the library, Golo Maurer. After an initial meeting with the library data team, a preliminary export of the catalogue was provided by Klaus Werner, who manages the library digitisation projects. His expertise in digitisation and understanding of library systems was invaluable for understanding and working with the library data. The catalogue data, provided in Machine-Readable Cataloguing (MARC) (Library of Congress 2022) format, was converted into CSV format to streamline analysis through the PANDAS (NumFOCUS, Inc. 2024) (McKinney 2022) library for Python.

The initial export was missing several fields of metadata necessary for the analysis. These fields were made successively available by Sabine Winter, who is responsible for the library data processing at Bibliotheca Hertziana, and who provided a full collection export and patiently explained the complex structure of the cataloguing system. Understanding the catalogue system was only possible with assistance from Pavla Langer, Philine Helas, and Michael Schmitz, specialists in scientific indexing at Bibliotheca Hertziana and responsible for assigning a unique local signature to each book acquired. Deciphering the signatures into categories proved challenging due to the system's many rules and exceptions. The result of this effort will be explained in the following sections on the catalogue structure, signature system, and library collection.

Initially, attempts to visualise the collection using the local shelf numbering system revealed limitations that fell short of our project's expectations. Subsequently, a shift in approach led to a second round of data collection focused on alternative data sources to establish connections between the books. The cooperation of the library staff, particularly Sabine Winter, allowed the extraction of the library's entire loan history of the past ten years, opening new possibilities for this project.

2.3 Catalogue Structure

The catalogue is organised into three sections: the rare collection, periodicals, and magazines, and monographs. The rare collection is being digitised and made available to an online viewer using the International Image Interoperability Framework (IIIF) (Snydman, Sanderson, and Cramer 2015). It contains old and rare documents, such as the most complete collection of old guidebooks to Rome worldwide, and is accessible only upon request. In contrast, the periodicals and monographs are readily accessible to visitors. Distributed over many shelves in the multiple floors and buildings of Bibliotheca Hertziana, they are classified using the unique shelf numbering system developed during the 1960s, which allows for serendipitous discoveries while browsing the library through its thematic grouping.

The library catalogue is accessed through the Kubikat (Kubikat and Ex Libris 2020) information system, the central endpoint to the bibliographical information of the four member libraries: the Bibliotheca Hertziana, the *Kunsthistorisches Institut* in Florence, the *Zentralinstitut für Kunstgeschichte* in Munich, and the *Deutsches Forum für Kunstgeschichte* in Paris. Kubikat is a system for retrieving bibliographically relevant data and plays a central role in providing essential metadata about the documents, including the availability of books across these libraries.

Despite its importance for researchers, the Kubikat system's search functionality is notably restrictive. It is limited to specific queries and does not offer recommendations for documents with similar content or subjects. Users can search through various metadata fields, including GND identifiers (integrated authority file of the German National Library (Deutsche National Bibliothek 2022)), which provide an additional classification layer to each document.

At the beginning of the project, the use of the GND identifiers was considered as

additional information on the documents' contents. However, since these identifiers are assigned a single time by the library that first records that document, they are significantly influenced by that particular institution's cataloguing conventions and thematic focus. Consequently, the initial classification carries a bias shaped by the collection focus of that library, which leads to a noticeable variance in the details of the descriptions. Such inconsistencies underscore the need for a flexible approach to bibliographical classification, which can adapt to each institution's scholarly focus.

Beyond its bibliographical functions, the Kubikat, most importantly, provides users with the documents' local shelf numbers by which they can be located in the libraries. The shelf numbering system developed at the Bibliotheca Hertziana provides a classification tailored to the institution. This system provides valuable insights into the collection, reflecting not just the thematic focus of the collection but also its history.

2.4 Signature System

The signature system used in Bibliotheca Hertziana is an example of a subject language, as described by Svenonius in The Intellectual Foundation of Information Organization. A subject language depicts what a document is about and facilitates users' navigation through the bibliographic universe (Svenonius 2000). As a representation of knowledge, it not only has applications in retrieving information but can also be used to represent knowledge itself in the case of a highly refined language. Svenonius distinguishes between *Alphabetic Subject Languages* and *Classification Subject Languages* (p. 100), the main difference between them being the latter using notations and verbal expressions to designate subjects and ordering them systematically and not alphabetically.

The signature system used at Bibliotheca Hertziana to classify books is an example of such a classificatory subject language. It is meticulously designed, ensuring that documents similar in content will be physically close in the library, a key affordance of classification schemata being to facilitate the retrieval of all and only relevant documents. This principle, also called the objective of collocation — placing similar items together — is one of the main goals of any classification language (Svenonius 2000). Such classifications fundamentally determine the

possible conjunctions among books by the ordering principle that places one book next to another (Schnapp and Battles 2014). Libraries often employ either alphabetical or subject classification languages. However, some, like the one amassed by Aby Warburg, use alternative orderings based on "the law of good neighbourliness" (p. 92).

The signature system at Bibliotheca Hertziana uses a fixed vocabulary of subject-specific terms with several levels, resulting in an in-depth classification of books. Library staff responsible for scientific subject indexing assign a unique shelf number for each document representing this classification. Allocating the signatures requires knowledge of the field of art history as well as of the history of the collection. The librarians working in scientific subject indexing are art historians at post-doc level and have an in-depth understanding of the field.

Since the signature system is *static*, they assign a classification depending on the most prominent and important topic and how similar books have been classified previously. If the document is, for instance, a collection of articles covering several topics, this can introduce significant bias in the classification. Difficulties arise when a document doesn't fit neatly in existing categories, as new categories aren't easily established and existing ones are not easily modified. For instance, the signatures in 'World Topography' reflect the time when the classification system was established, containing names of countries as signatures that no longer exist. Consequently, even today, new acquisitions are still placed in these outdated categories, making it difficult to incorporate acquisitions from new trends in art historical research into the collection.

Despite these constraints, the signature system still offers a crucial advantage by allowing users to find books of similar content close by when browsing the library shelves. Due to the system's structure, books in similar research fields can be found on adjacent shelves, facilitating serendipitous discoveries.

The structure resembles a decision tree, combining letters and numbers which encode the content as well as the location of the documents. Although most users are aware of the system's existence, they can't make use of its full potential as there is no record of all existing categories. To illustrate the signature system's structure, consider the example depicted in Figure 2.3, featuring a book focusing on Byzantine art, designated with the signature MK 1020. The example shows the tree-like hierarchy that characterises the classification system and uses an example

of what will be called the generic signature in the context of this project. There are about as many exceptions as rules to how the signatures are constructed. Still, they all contain several letters followed by digits, and most follow the generic signature pattern. The signature MK 1020 is a multi-tiered code, with each level contributing to an increasingly detailed representation of the book's content. It simultaneously represents how close the books will be found on the library shelves. Books with signatures starting with M can be found on adjacent shelves and are grouped depending on their content. The more letters and digits the signatures of two books have in common, the closer they can be found in the library.

It's worth noting that automatically decoding the system to get the classification of a document is a challenge. For instance, in the example tree, there is a node that is not explicitly represented by the signature itself: the third level of the classification, which represents 'Middle Ages,' lacks a direct numerical or alphabetical counterpart in the signature. To discern this level of classification, a supplementary lookup table is necessary. This lookup table can be found in several copies distributed over several offices of Bibliotheca Hertziana, all with handwritten annotations, additions, and amendments. At some point recently, some newly added signatures, mostly concerning modern artists, were beginning to be recorded in a database. There is currently still no complete version of the translation table available.

Figure 2.2: Books from the rare collection with shelf numbers from the local classification system. The signatures starting with Dg are classified as *Roman Topography* and *Guides*. (*Photo:* Enrico Fontolan)

2.5 Library Collection

The library has kept growing since Bibliotheca Hertziana's collection was established over a century ago. It now provides insight into the development of the field of art history itself through its composition and evolution over time. The acquisition of books and user activity follow trends and patterns that reflect this development and provide a lens through which we can explore the library's history. The catalogue data, central to this research, sheds light on the interplay between researchers' interests, the evolution of the field, and the influence of the scholars associated with the institute.

The number of documents per category provides insight into the library composition, as the signature system was designed to reflect the structure and focus of its collection. The visualisation in Figure 2.4 groups the monographs into their assigned categories and illustrates the clear focus on Italian art and artists through the size of the corresponding circles. The root node in the middle of the graph represents the books in the library. The branches represent the categories of the signature system, which further branch off into their corresponding subcategories. Only the first two levels of the signature system were included in this graph. It doesn't include the journals and catalogues — which comprise a large part of the collection — due to the lack of consistency in recording the corresponding signatures.

Adding time as a dimension to the analysis can give an idea of the recent development of the library collection. Each document's date of acquisition provides insight into when topics became relevant for the research conducted at Bibliotheca Hertziana, as many of the new books are acquired upon request from researchers. Figure 2.5 shows the books acquired since 1985, grouped by category and excluding periodicals and catalogues. The subjects are ranked by the total number of acquisitions, showing the thematic focus of the additions at the time. The purchases show that the library's main focus has not changed drastically over the past decades; it remains an art historical library focusing on Italian and Roman art and artists. However, there are fluctuations worth mentioning.

There was a sudden peak of acquisitions of books in the Italian Artists section around 1998/99, and a later clear trend of acquisition in the Italian Topography section around 2016. These fluctuations showcase some different ways in which the collection is developing. The notable surge in the number of books dedicated

to Italian artists can be attributed to a substantial posthumous donation from a prominent art critic. This remarkable influx of new books demonstrates how the collection is linked to scholars engaged with the library.

Examining the *Italian Topography* section and its subsections unveils a conspicuous peak in acquisitions of books centred on Naples in 2016/17. This sudden uptick is probably influenced by one of the directors of Bibliotheca Hertziana, Tanja Michalsky, who took the post around that time and whose research focus is, among other topics, the city of Naples. This underlines how the presence of researchers significantly shapes the composition of the library's collection and underscores the interplay between scholarship, the library's holdings, and the field of art history research. It is a testament to the living and ever-evolving nature of Bibliotheca Hertziana as a repository of knowledge and a reflection of the researchers actively shaping the institution.

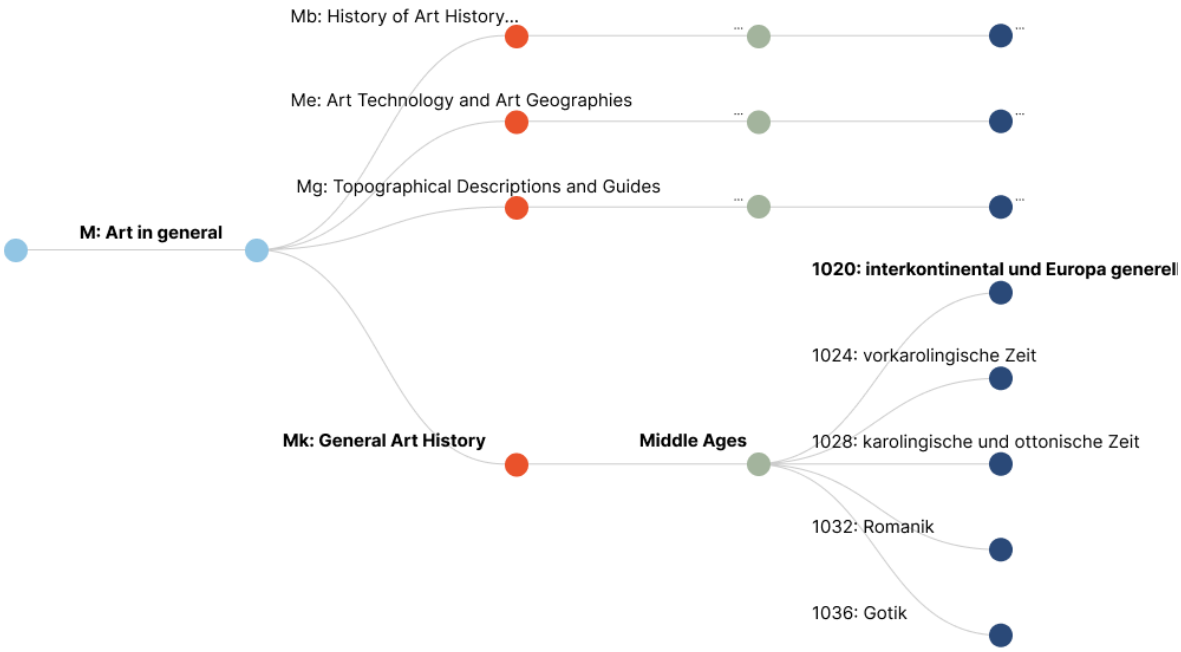


Figure 2.3: The classification tree of example signature MK 1020, designating a book on Byzantine art. The signature is structured similarly to a tree but not all levels of classification have a direct numerical or alphabetical counterpart, making a supplementary lookup table necessary for decoding.

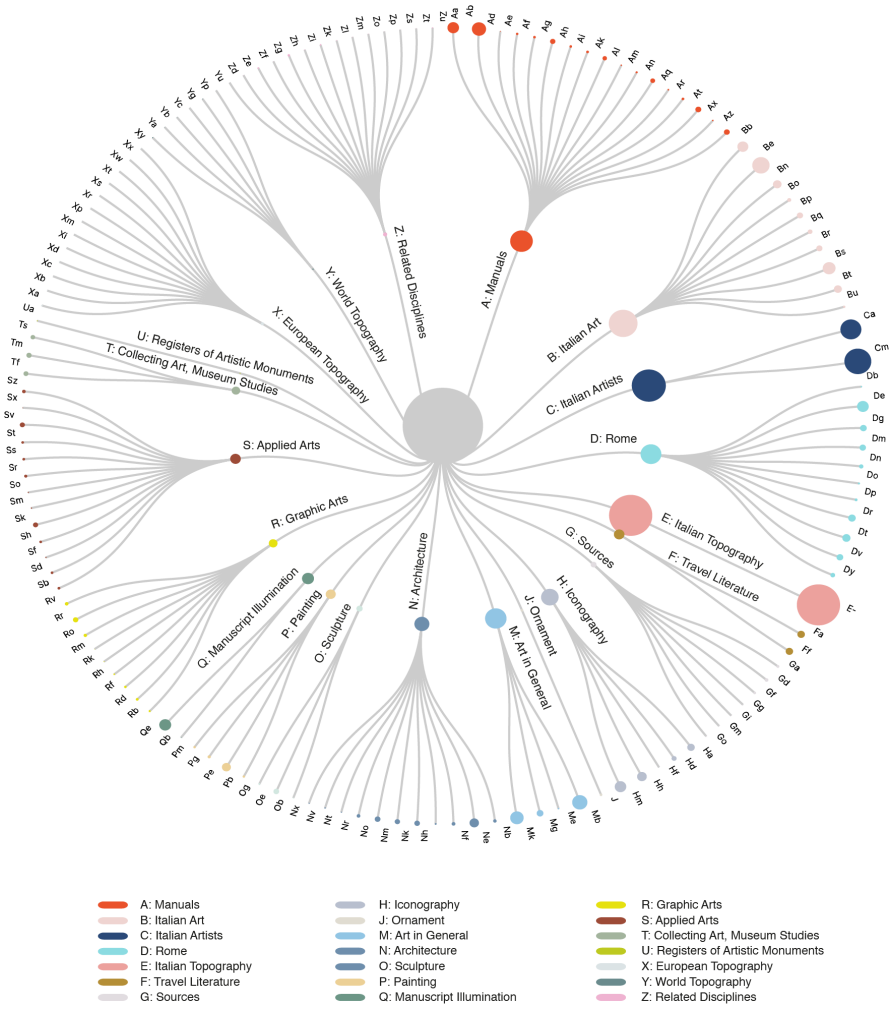


Figure 2.4: The monographs at the Bibliotheca Hertziana by their assigned shelf number. The size of the circles indicate the number of books in the category. The focus of the library collection clearly lies in books about Italian art and artists, although two large categories, Periodicals and Catalogues, are missing from this mapping due to data inconsistencies.

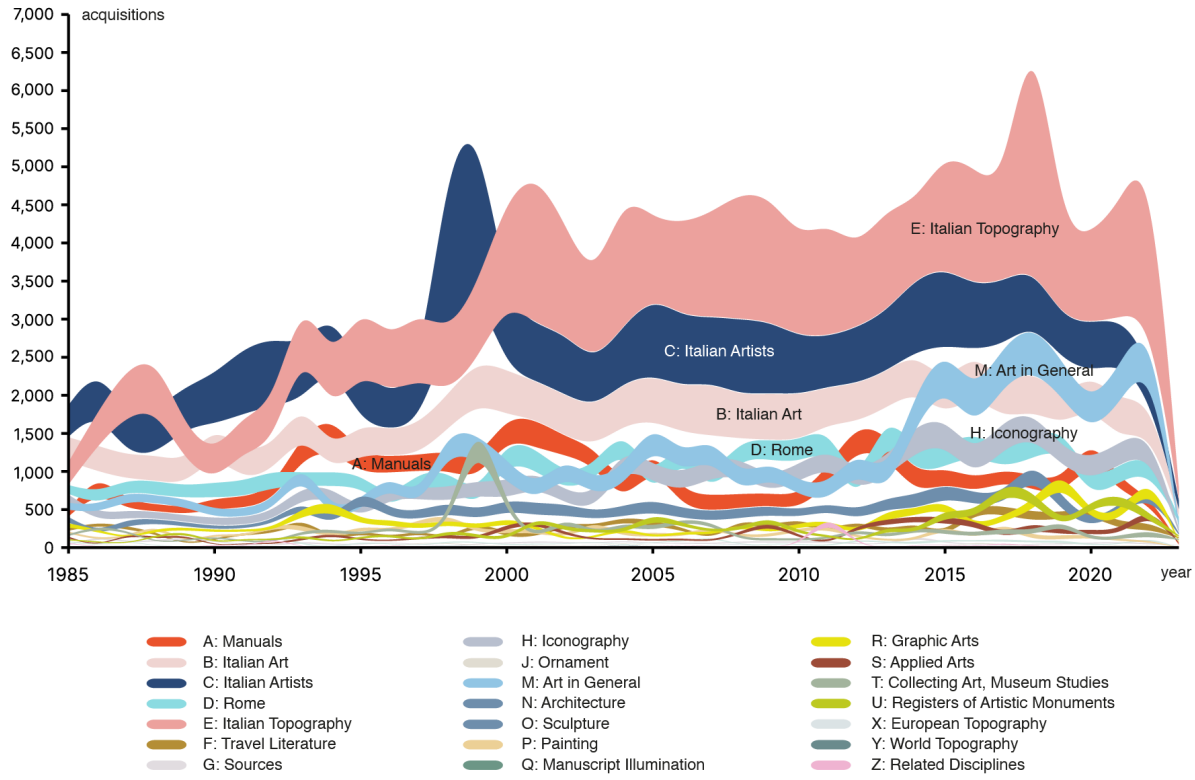


Figure 2.5: Acquisitions since 1985 grouped by signature and mapped over time show the development of the library over the past decades. While the collection's main focus has remained the same, the peaks in the *Italian Artists* and *Italian Topography* section showcase ways in which researchers shape the collection.

Chapter 3

Related Work

> Vision has served knowledge in many ways across the sciences, arts, and humanities in theoretical and applied domains.
>
> *Johanna Drucker*

This section will provide a comprehensive overview of existing research, focusing on studies in library science, the visual principles of networks, visualising science, and a curated selection of example projects. The methodologies and technological frameworks used in the research serve as an inspiration and a basis for developing an alternative visual approach to library collections.

3.1 Digital Libraries

Libraries are undergoing significant transformations driven by the increasing prevalence of digital publications across various disciplines. While the humanities have traditionally clung to paper-based publications, a shift toward digital media is becoming more apparent. Despite this shift, scholars might still be using books and physical shelves in the future, especially for reading lengthier volumes. This evolution in information display demands a re-imagining of library spaces and displays, finding creative ways of making them even more inspiring (Wilders 2017). The

role of the library will be not only to house the information but also to make it accessible. In this context, technology such as dimensionality reduction emerges as a key tool in the digital humanities. As part of a novel information infrastructure, dimensionality reduction can enhance the accessibility and navigability of libraries (Schmidt 2018). It positions similar books together and dissimilar ones far apart. This allows the creation of custom corpora of books, adhering to the collocation principle (Svenonius 2000) and aligning libraries with the evolving needs of research and re-imagination of library spaces.

Part of such re-imagining is understanding how the books are ordered and classified in the library. According to Bowker and Star in *Sorting Things Out: Classification and Its Consequences*, the key to a flexible library is to produce a flexible classification and "only a living classification is a good classification" (Bowker and Star 2008). Because such classifications are expected to evolve dynamically, they should maintain the traces of their construction and usage, and users should be aware of their political and organisational dimensions (Bowker and Star 2008). It should be possible to re-order a library display at any moment, thus reflecting the change in the library and its context.

Libraries aspire to motion, stimulate interest, and allow immersion in an extraordinary environment (Stafford 2012). This should be reflected by a dynamic classification and display of the collection, allowing intuitive connections between books.

3.2 Visual Principles

Graphical representations of knowledge have been crucial to scientific work for communicating results and formulating hypotheses (Drucker 2014). Additionally, they have created standards of visual principles over time, which will be explored in this section.

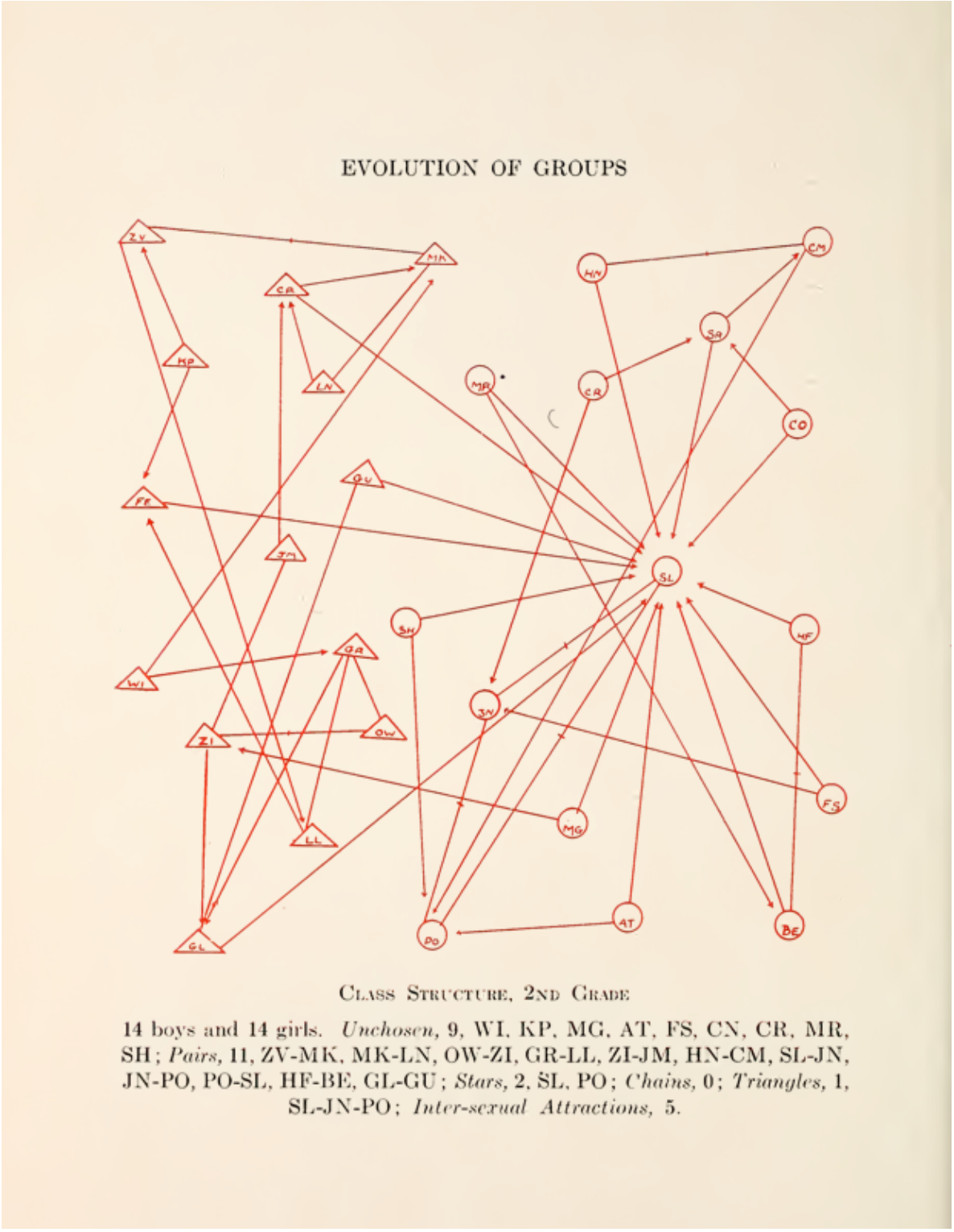
In *Graphesis*, Drucker compares two network diagrams of points and lines, stating that one produces the knowledge it draws and the other only displays information (Drucker 2014). The first diagram she refers to is a historical piece by Athanasius Kircher from 1669, demonstrating a knowledge system. The second is a static image, representing the web traffic on the Internet in 2003, as seen in Figure 3.3. Similarly, Jacob L. Moreno, a pioneer of social network analysis

and creator of the sociograms, considered graphical representations not only a method of presentation but also one of exploration (Lima 2011). One of the first published sociograms in his work *Who shall survive? A New Approach to The Problem of Human Interrelations*(1934) is illustrated in Figure 3.1, mapping the social interactions of second-graders which previously were read in the form of tabular data only.

The most popular graphical pattern visualising knowledge, which evolved into network visualisations, is the tree. Networks have a rich history of visual grammar, and even though they are often limited to a standard graph — made up of links, nodes, and labels (Rodighiero 2021) — they manage to represent a broad range of subject areas (Lima 2011). They are an example of a *regulated representation* (Kaplan 2015), governed by a set of rules of production and usage.

While the size and complexity of networks grow, so has the need to create visualisations effectively communicating the interconnectedness of systems using visual analytics. As network science advances, network visualisations have also increased in complexity, as demonstrated in Figures 3.2 and 3.3, showing the ARPANET and the Internet, two networks of increasing size and complexity.

In the visualisation of scientific practice, citation links and co-authorship have often been used to illustrate the connections between different fields (Lima 2011). Moreover, the dynamic changes in science, which can range from bursts of activity in one field following external events such as funding, are an area of investigation (Börner 2015). Given the ever-changing nature of the library's collection and the institute's research community, this is particularly interesting for mapping Bibliotheca Hertziana. It is crucial to bear in mind the visual principles of network graphics in this process, the goal being to create a visualisation that not only represents knowledge but can also act as a *knowledge generator* itself.



CLASS STRUCTURE, 2ND GRADE

14 boys and 14 girls. *Unchosen*, 9, WI, KP, MG, AT, FS, CN, CR, MR, SH; *Pairs*, 11, ZV-MK, MK-LN, OW-ZI, GR-LL, ZI-JM, HN-CM, SL-JN, JN-PO, PO-SL, HF-BE, GL-GU; *Stars*, 2, SL, PO; *Chains*, 0; *Triangles*, 1, SL-JN-PO; *Inter-sexual Attractions*, 5.

Figure 3.1: One of the first sociograms by Jacob L. Moreno from 1934, showing the interactions of 2nd graders. Before this visualisation, interactions were represented as tabular data only. (*Source*: Moreno 1934, p. 36)

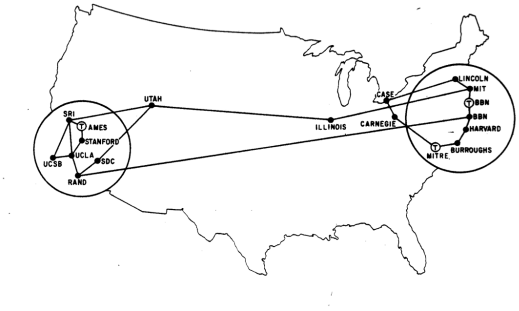


Figure 3.2: A map of the Advanced Research Projects Agency Network (ARPANET), a technical predecessor to the Internet. It was the first large area packet-switched computer network. (*Source*: Heart et al. 1978, p. 80)

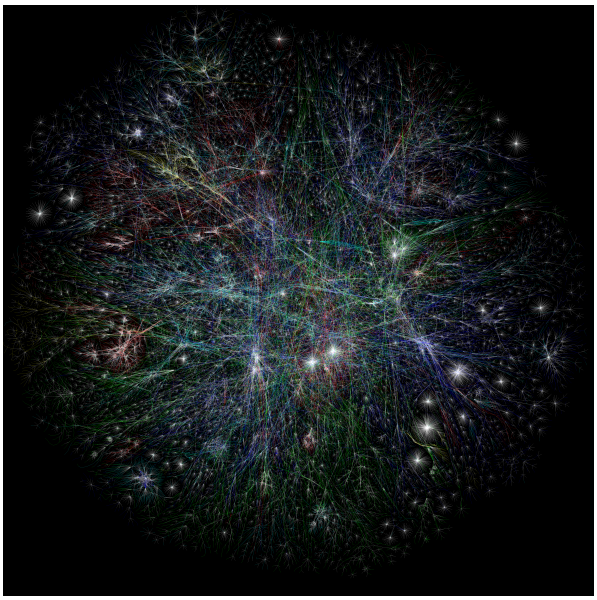


Figure 3.3: The OPTE project's first full map of the Internet in 2003. The colours of the links represent the location of IP addresses. (*Source*: The Opte Project 2003)

3.3 Visualising Science

Science maps are visual representations of the structure of science and aim to show how disciplines relate to each other (Petrovich 2020). Science mapping has a long tradition in the quantitative studies of science. Usually, it consists of a body of scientific literature, a set of analytic tools, and theories that can guide the interpretation of the visualised intellectual structures (Chen 2017).

Science maps are not meant to replace existing classification schemes but rather can provide additional information and have applications not restricted to knowledge organisation (Petrovich 2020). Creating a map of science usually follows a workflow from data collection via network extraction to interpretation (Börner, Chen, and Boyack 2003).

Commonly, scientific papers are mapped via co-citations and inter-citations. Co-citation is counted when two works appear in the references of a third, inter-citation is the count of times that any document cites any other, as well as itself (p. 192). An example of this can be found in *Mapping the Backbone of Science* (Boyack, Klavans, and Börner 2005) by K. W. Boyack et al. They create intuitive mappings of the sciences based on around 7000 scientific journals, using different similarity measures and evaluating the resulting mappings based on accuracy measures and readability of the layout. The result is static mappings of the sciences designed to promote the understanding of the structure of science, seen in Figure 3.4. The structure of this mapping shows larger central clusters of major areas of science and smaller clusters with disciplinary topics (Boyack, Klavans, and Börner 2005).

3.4 Visualising Collections

In the context of visualising large cultural collections, this project has drawn inspiration from the work of the Urban Complexity Lab (UCLAB) at the Potsdam University of Applied Sciences. UCLAB's visualisation of two museum collections from the *Alte Nationalgalerie* and the *Museum Europäischer Kulturen* introduced a more intuitive and user-friendly approach to the collections, employing machine learning techniques to associate objects to one another and facilitating curiosity-driven exploration in their interface (Pietsch 2020a)[1]. The visualisation positions the collection images based on visual similarity and uses handwritten-like annotations to depict clusters, as shown in Figure 3.5. Another visualisation created by UCLAB explored sketches by Prussian king Friedrich Wilhelm IV. (1795 – 1861)[2], employing different visualisation methods on the same data set (Glinka et al. 2016). The sketches can be arranged in a timeline with thematic annotations, in a data cloud using the dimensionality reduction algorithm T-SNE, and as a grid with features similar to the T-SNE visualisation. It allows selecting themes in a bar at the top, which results in sketches being highlighted, showing yet another possible way of clustering and grouping them. The themes are based on metadata and annotations, showcasing the most prominent topics the king covered in his sketches.

Similarly, Davide Picca, senior lecturer at the University of Lausanne, et al. utilised dimensionality reduction techniques to explore Charles S. Peirce's Manuscripts, organising manuscript pages based on semantic proximity via the UMAP algorithm (Picca et al. 2023). The work arranges pages of Peirce's seminal manuscripts on a two-dimensional cartographic plane, creating clusters of similar pages.

In scientometrics, dimensionality reduction techniques are widely used to visualise large datasets and networks, typically using vectorised representations of scientific papers or citation networks. For instance, the work by Maximilian Noichl, a doctoral student at the University Bamberg, on articles from the *arXiv* and the *bioRxiv* archives, visualises scientific preprints in two dimensions based on semantic similarity using UMAP (Noichl 2023). The mapping, shown in Figure 3.6, is annotated using the highest scoring keywords from a TF-IDF representation of the texts, indicating descriptions for each cluster, and suggesting "a gradient from

1. UCLAB's visualisation of two museum collections can be accessed here: https://cpietsch.github.io/smb-vis.

2. The visualisation of sketches by Prussian king Friedrich Wilhelm IV. can be accessed here: https://uclab.fh-potsdam.de/fw4/vis/.

physics, via mathematics and computer science to the life sciences" (Noichl 2023). In the same study, Noichl explores the similarity of the preprints based on mathematical formulas, resulting in a distinctly different mapping that diverges from the structure observed in the semantic mapping.

Similar methods can also be applied to visualising library collections and offer interesting ways for visitors to interact with the collections through interfaces, as demonstrated by RNDR's design for the library of the TU Delft in the Netherlands (RNDR 2022). They map the digital collection of publications as a point cloud, shown in Figure 3.7, using the dimensionality reduction algorithm T-SNE with position determined by text analysis, grouping papers with similar topics together. To overcome the lack of a tangible counterpart or artefact, a unique cover is generated for each publication, its colour depending on the faculty associated with the publication. Visitors can use the interface to explore each publication's closest neighbours, which provides the foundation for a recommendation system.

As these examples show, techniques such as dimensionality reduction using UMAP or T-SNE have previously been employed to gain insight into graphic layouts of large collections. These methods offer a robust framework for exploring underlying structures, particularly when dealing with complex and high-dimensional datasets, such as the library collection at Bibliotheca Hertziana. By extracting a comprehensible visual format from vast amounts of data, dimensionality reduction can enhance the understanding of the data and guide the subsequent analysis.

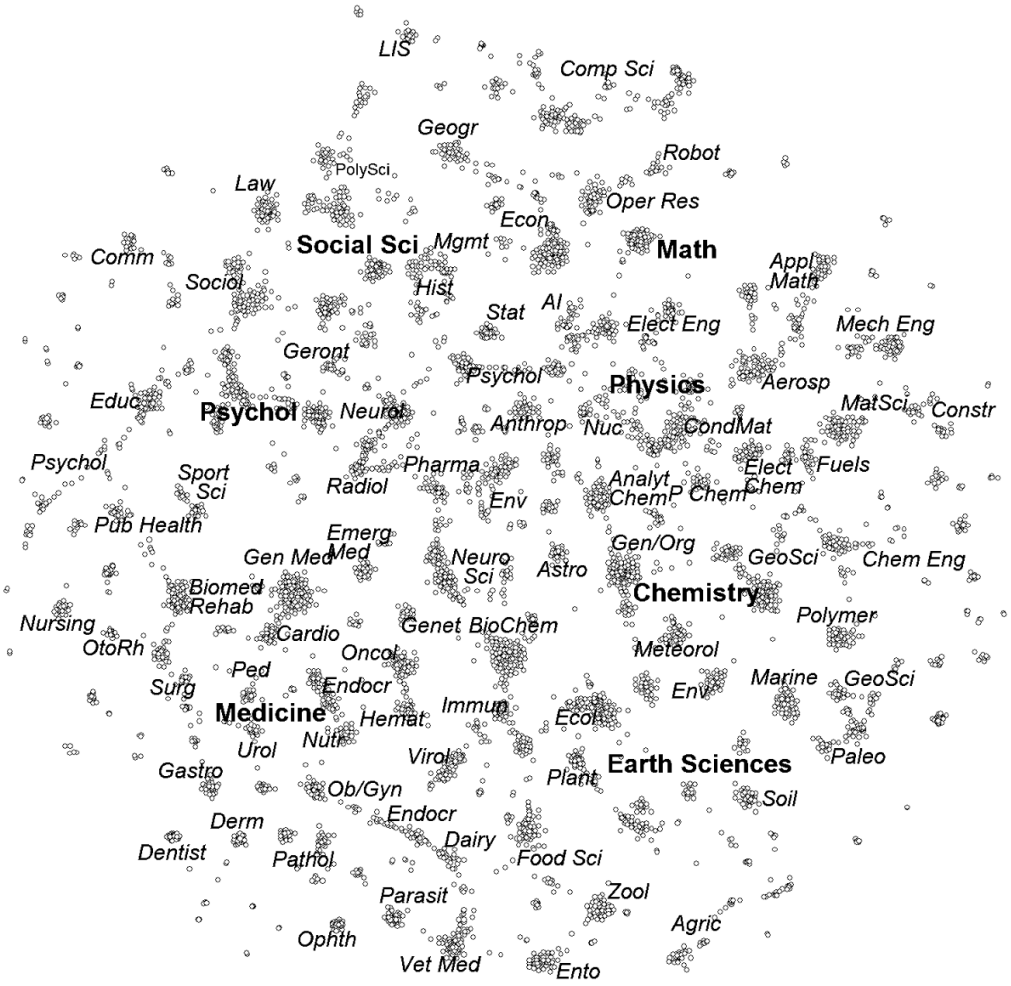


Figure 3.4: This map of science was generated with IC-Jaccard similarity measure from over 7000 journals from 2000 in *Mapping the Backbone of Science*. The cluster centres are the main research fields with smaller clusters around representing different disciplines. (*Source*: Boyack, Klavans, and Börner 2005)

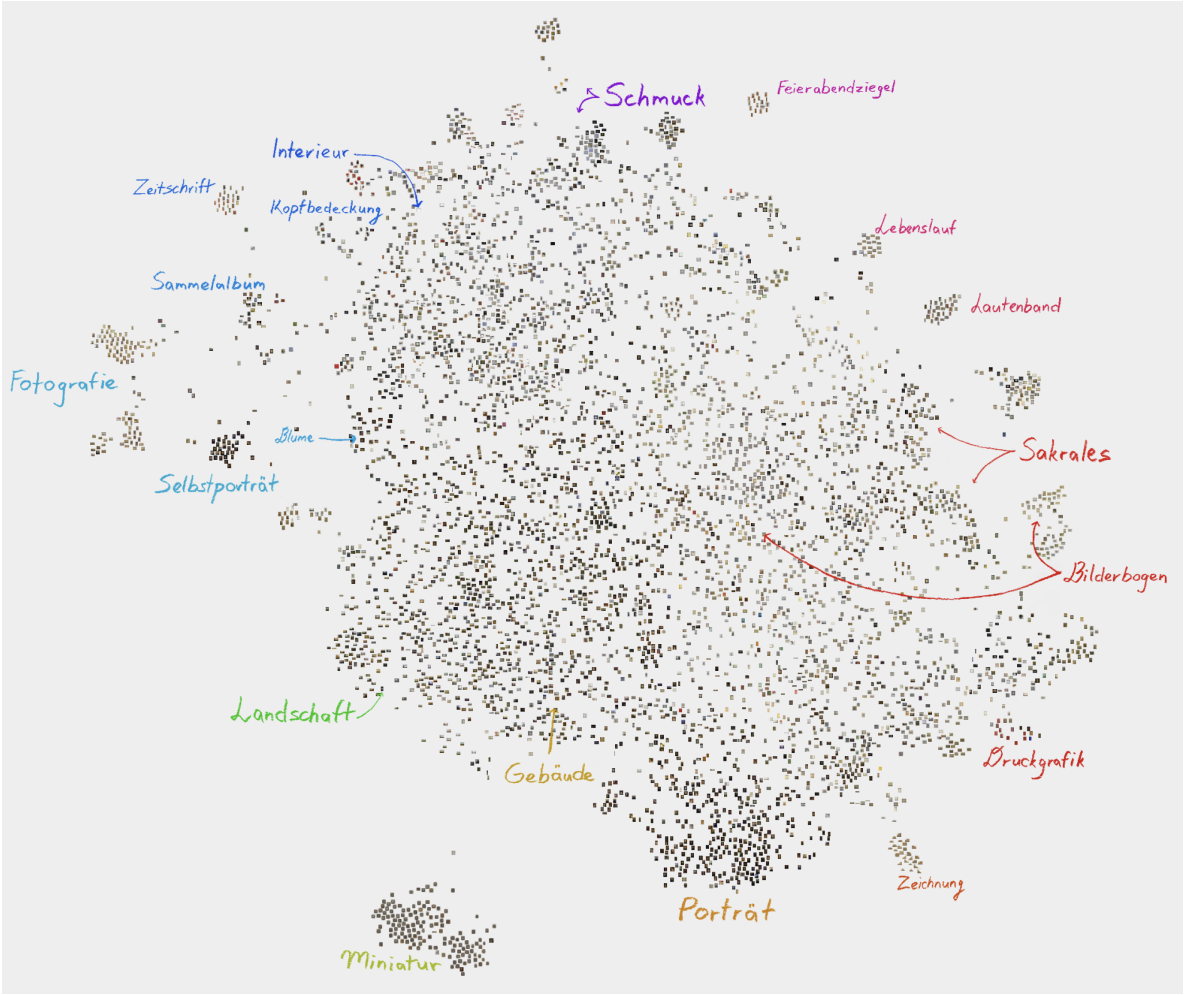


Figure 3.5: The visualisation by UCLAB of two museum collections juxtaposing fine art paintings with everyday artefacts. The annotations give further guidance on where to find specific artefacts. (*Source*: Pietsch 2020b)

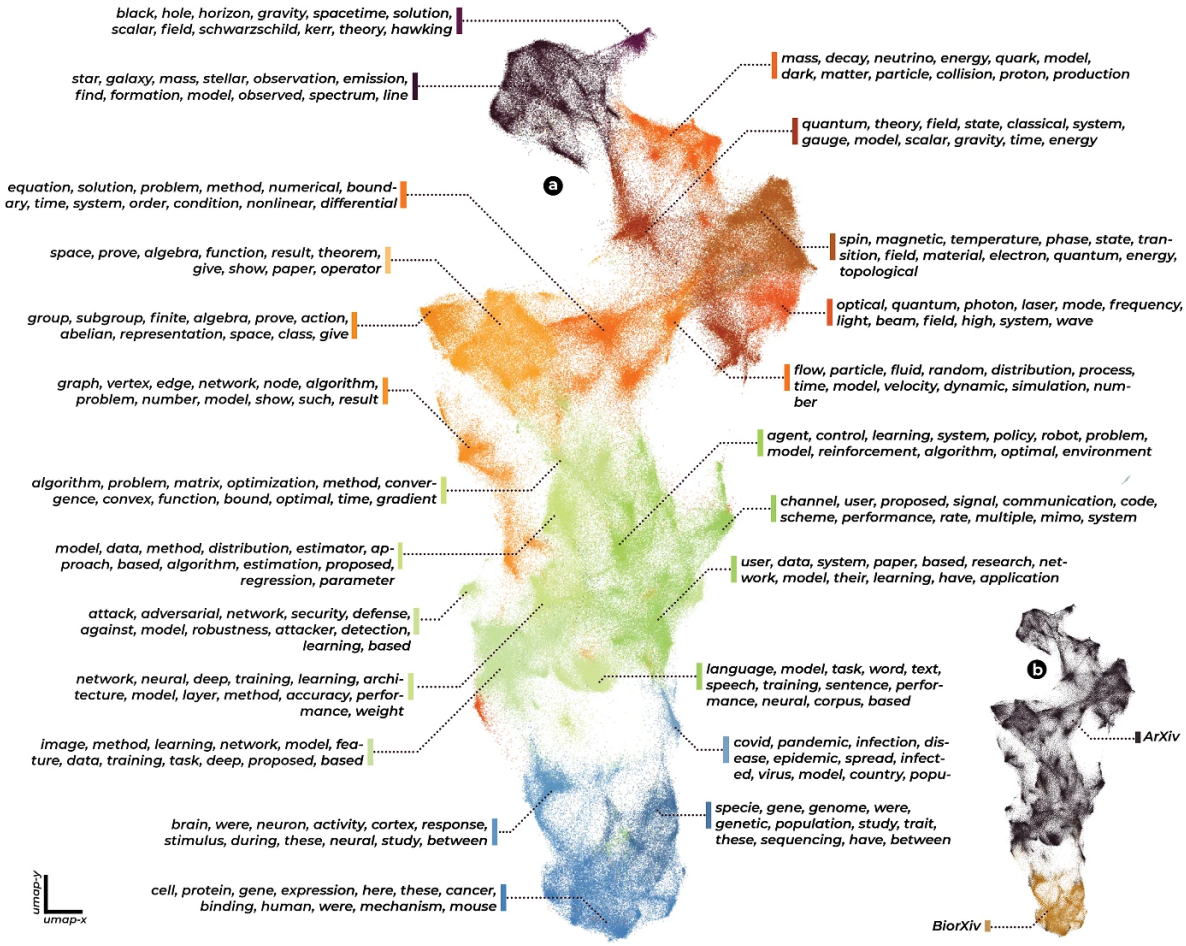


Figure 3.6: Noichl's mapping of sciences based on preprints from *arXiv* and *bioRxiv* archives. The mapping suggest a gradient from physics via mathematics and computer science to the life sciences. Keyword annotations further enhance the understanding of the map of science. (*Source*: Noichl 2023)

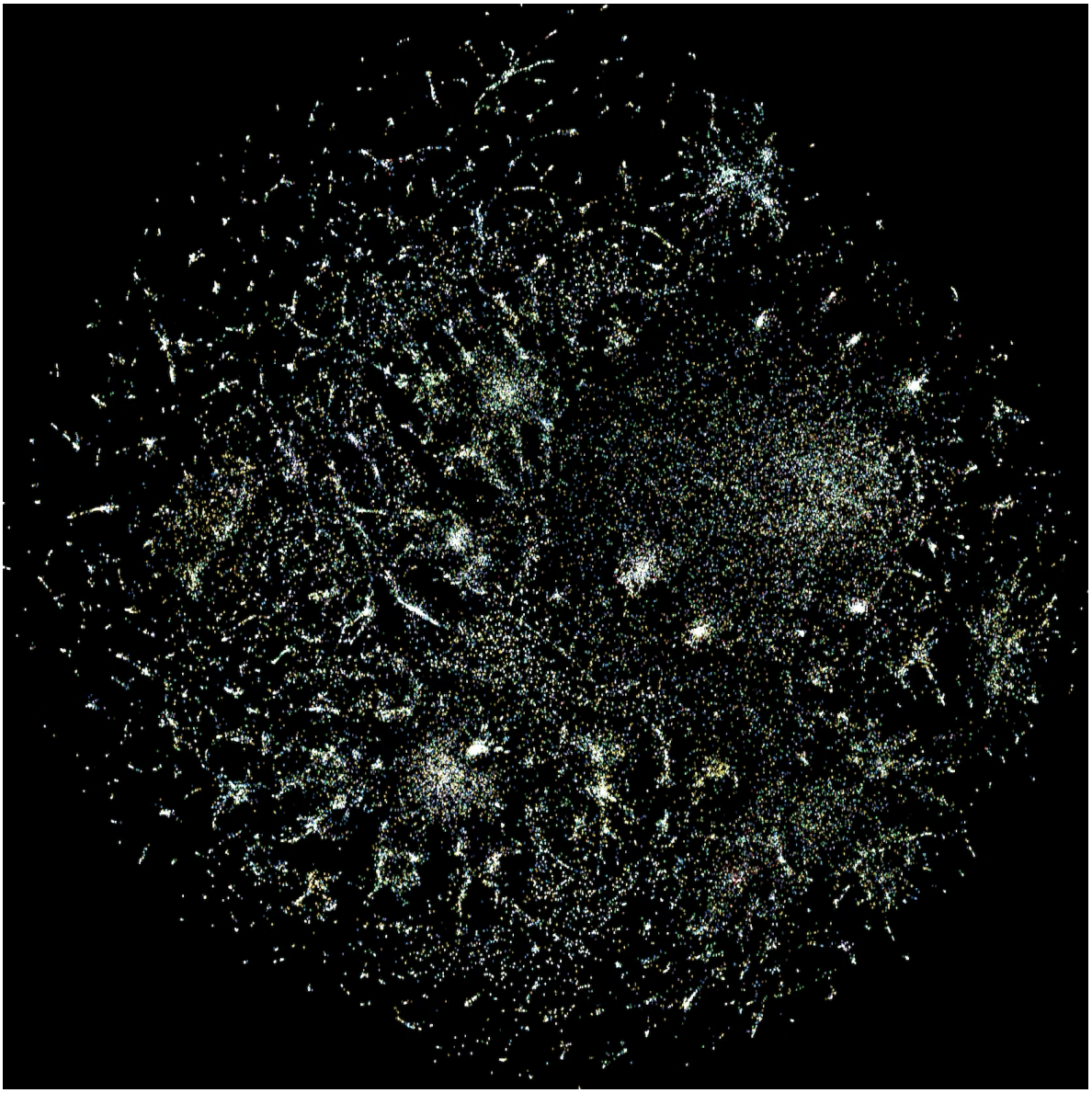


Figure 3.7: T-SNE visualisation by the RNDR design studio, mapping 150k papers in the library of the TU Delft based on content and topics. The visualisation is available as a library display. (*Source*: RNDR 2022)

Chapter 4

Mapping the Collection

> Libraries are beehives of activity, but much of that activity is invisible.
>
> *Jeffrey T. Schnapp*

In *The Library Beyond the Book*, Jeffrey T. Schnapp and Matthew Battles introduce a scenario for a dynamic library classification, where "bookshelves dance and weave at the librarians' behest" (Schnapp and Battles 2014, p. 94). Similarly, they introduce the term *Livebrary*, a touchscreen that allows one to experience the library in real-time (p. 17). This section introduces the methodology employed to create the basis for such an experience, by mapping users' interaction with the library collection.

Mapping the collection of Bibliotheca Hertziana's library required an innovative way of establishing a measure of distance between the books and a way to position them in a two-dimensional space. The collocating objective — creating sets of similar elements — plays an important role here along with the principle of user convenience, stating that bibliographic descriptions should be made with the user in mind (Svenonius 2000). To create a dynamic library classification, it seems to make sense to map each book within the library once and group them based on a measure of similarity.

Traditional methods involving citation networks or text analysis, commonly used in mapping science or visualising cultural collections, proved unsuitable due to

unavailable data—full texts and abstracts were inaccessible — leaving only book titles for textual analysis. This idea was abandoned quickly since titles tend to be short and generally don't provide a good estimate of the book's content.

Another consideration was analysing books within the library's categories using the shelf numbering system. Unfortunately, the lack of a full record of the signature system rendered this approach unfeasible. While the signature system could aid the evaluation of a mapping and the distribution of clusters over the library sections, it failed as a foundation for a mapping. Due to its complexity and inaccessible data sources, the approach based on citation networks was also dismissed.

However, the dataset of library loans offering an alternative was never explored. Like citation networks, it allows measuring pairwise connections between books borrowed by the same users. Identifying researchers with similar reading patterns and exploring their behaviour over time can unveil insights about the recent development of art history as a field. This approach will be explored in detail in this section, focusing on the data sources, methodology, and evaluation of results.

4.1 User Loans

Over the past decade, Bibliotheca Hertziana's book loans in the library have been documented meticulously. Until 2019, library policy stipulated that all books taken from the shelves be checked out and logged into the library system by internal and external researchers. At the end of each day, externals were required to return the books, whereas fellows and library staff had borrowing privileges and could keep books on their desks almost indefinitely.

In mid-2019, the policy changed, and external users were no longer required to borrow books but could keep them at their desks in the library for the day. Internal researchers were still required to check out borrowed books, allowing the books to be tracked within the institute. The loan data provides insight into researchers' interests, and since it can be traced back to either internals or externals, it can also show potential differences between the two groups. Each loan can be traced to a document using its local signature number and to a user using their identification number. The user identification number doesn't allow for distinguishing the externals from the internals directly. The two groups can be told apart using information on how and when the IDs were assigned. A potentially incomplete list of external

user IDs was used to differentiate the loan data of the two groups.

The potential differences between the two user types and their loan behaviour were disregarded in the first attempt at mapping the entire data set. Later on, a mapping based only on the internal users' loan data was created, which made the effect of the loan policy on the visualisation apparent.

Sabine Winter exported all the user data used for this project from the library database. Retrieving a decade of library loan data is unusual, as most libraries delete such records after only a brief period. Several Bibliotheca Hertziana fellows and staff raised concerns about confidentiality, as the loans can be traced to user IDs, which point to researchers and can be used to track their research. To address this, data will not be published, and the user IDs have been anonymised.

The exported data contains the check-out records of 737 users borrowing 95'107 unique documents between 2013 and September 2023. Around 70% of the total of 464'360 documented loans can be traced back to externals and approximately 30% are associated with internal researchers. This is remarkable, considering only 99 external compared to 638 internal researchers are present in the data. Additionally, due to the change in loan policy in 2019, the external loans are only documented until the middle of that year. Until then, the externals were busy borrowing and returning books daily, as illustrated in Figure 4.1.

The loan activity follows clear monthly fluctuations, evident in Figure 4.2. Notably, in July, just before the typical holiday time in Italy, there is an increase in loans followed by a sharp drop in August — when internal and external researchers tend to be absent from the library and on holiday. As the years after the policy change are not directly comparable, they are represented using dashed lines. The change in loan policy substantially impacted users' behaviour, reducing the external checkouts to almost zero. Furthermore, the drop in total loans was reinforced due to the library's closure in 2020 amid the Covid-19 pandemic. This behavioural shift between the two user groups is illustrated in Figure 4.1, showing their separate trends over the years. It's important to note that the reduction of checkouts does not necessarily imply an overall decline in library usage. The library continues to thrive as a dynamic and vibrant hub for research and meaningful connections among scholars. It's probable that internal researchers sometimes omitted to formally borrow the books they keep on their desks.

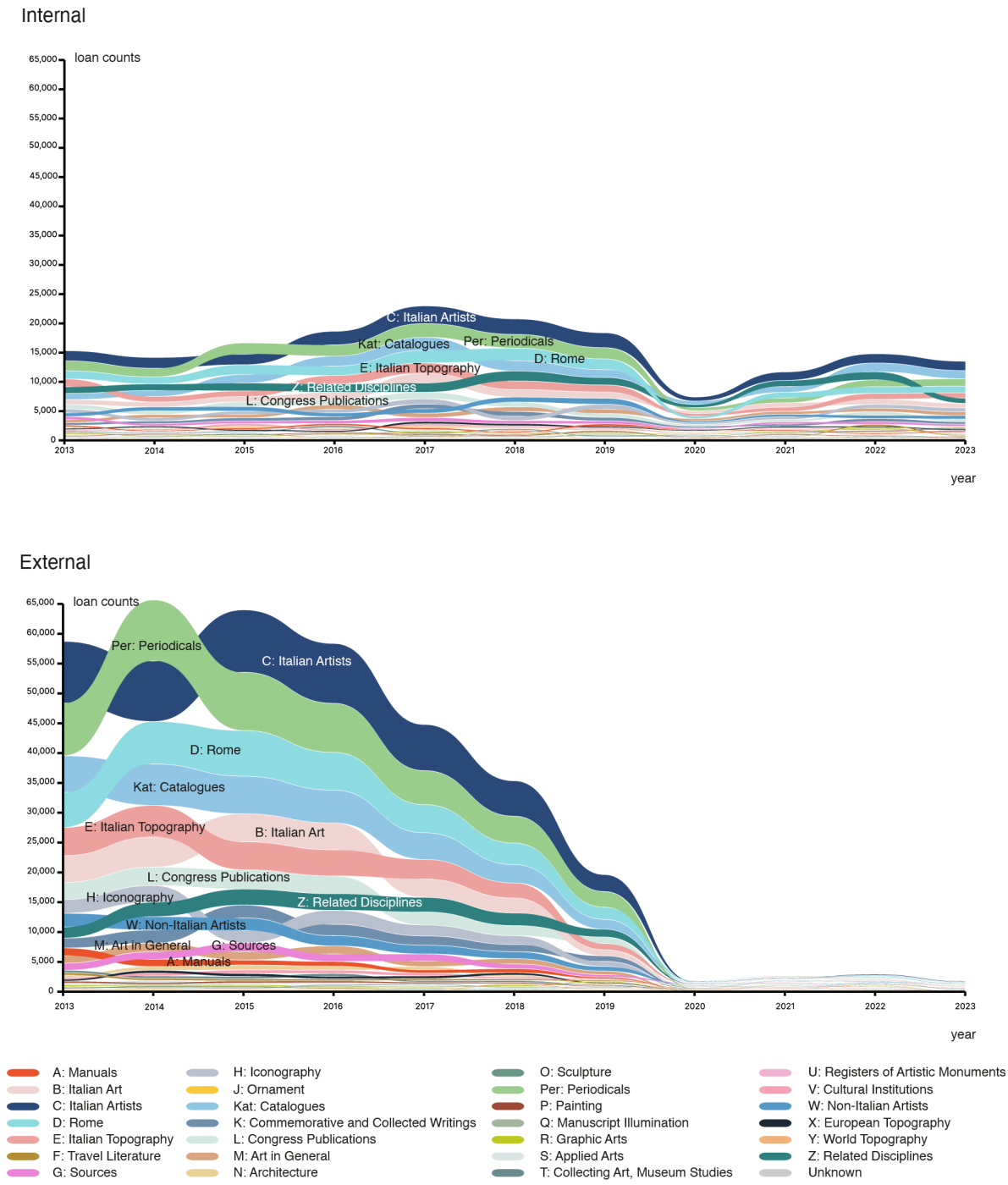


Figure 4.1: The user loans grouped and ranked by subject from 2013 to September 2023. Internal and external users have different loan patterns due to loan policy and research interests. The effect of the loan policy change in 2019 and of the COVID-19 pandemic are clearly visible by drops in loans to externals and internals.

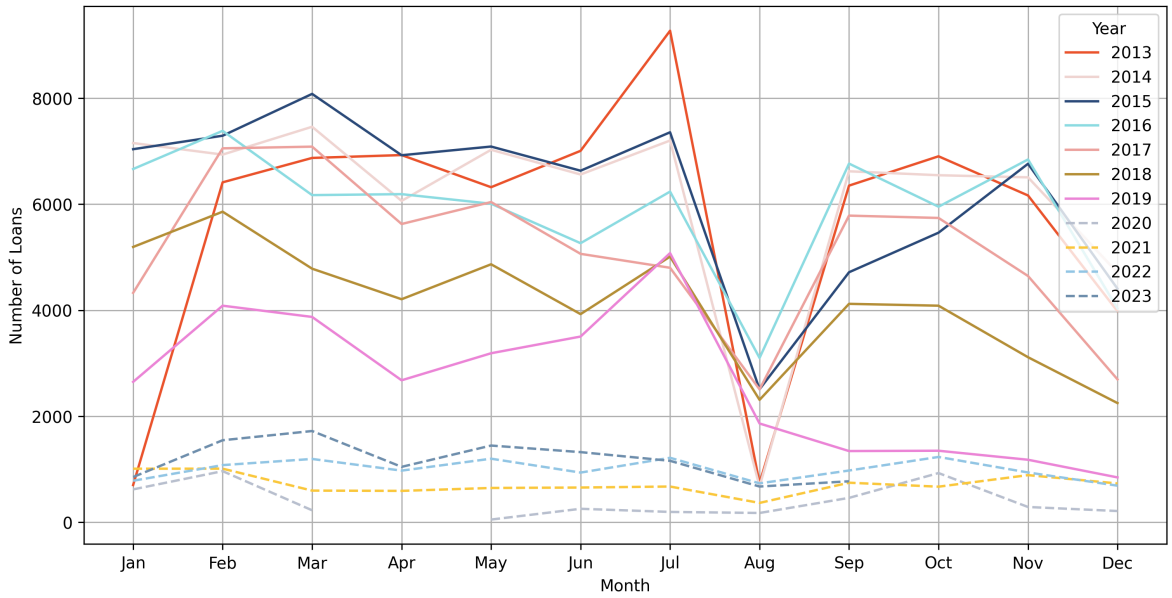


Figure 4.2: Loan activity by month over the past ten years. The dashed lines represent the years after the change in loan policy. Before that period, the years resembled each other in terms of loan pattern.

4.2 Methodology

The user data was enriched with additional information from an export of the library collection, containing language, authors, subtitles and other fields of metadata. There were only a few anomalies in the data that had to be cleaned. For instance, several documents have been loaned over a thousand times over the past decade, all belonging to the *Periodicals* section. Each loan record of magazines and journals only points to a record of the parent document instead of the specific edition, resulting in inflated numbers. As a result, documents with over five hundred loans over the whole time span are considered outliers and were excluded from further analysis.

To map the library collection effectively, an appropriate metric of pairwise distance between the books had to be found. Given the available data and the related work discussed above, it became clear that applying text analysis techniques might not yield meaningful results. The only textual data available for the books in the collection are titles and subtitles; only very rarely are there abstracts of articles available. An alternative approach considered was leveraging the local signature system, as it provides a keyword-like categorisation of the content of the books. However, the lack of the necessary data to create a lookup table to decode each signature made this approach unfeasible. The loan data collected over the past decade opened up a novel approach to the mapping process: relating books to each other according to the users who borrowed them. Central to this approach was the assumption that two books used by the same person are related, and the more common readers they have, the closer they should appear in the mapping.

Constructing a pairwise distance matrix between all the books loaned over the past ten years involved creating a Term Frequency - Inverse Document Frequency (TF-IDF) matrix. TF-IDF was originally coined in 1972 as *term specificity* and proposed that "less frequent, more specific, terms are of greater value than matches on frequent terms" (Sparck Jones 1972). The representation of user loan data used as a basis for the mapping considers the user loans as terms appearing in documents. It measures the importance of each user loan in relation to the entire loan data.

A list of users was generated for each book, containing each user as many times as they borrowed the book. Using a TF-IDF representation of the book loans in this context gives more weight to users who appear only for a few very specific books, possibly creating clearer clustering around niche subjects. Users who borrow many

books and appear together with many other users are given a lower weight. The lower the weight, the less significant the user is compared to other users in the context of that document. This is a direct result of the properties of the TF-IDF scoring.

To calculate the TF-IDF matrix, the *term frequency* is divided by the logarithm of the *document frequency*. *Term frequency* is obtained by counting the number of times a user has borrowed books or, equivalently, how many times a user appears in the loan lists of the books. *Inverse document frequency* is calculated by dividing the total number of books by the number of books a user borrowed and then taking the logarithm of the result. This is done to increase the weight of users who appear infrequently in the loan list, who are more useful for differentiating books than users who appear frequently. The resulting matrix is in a high-dimensional space, where books with common users cluster together. To get from this high-dimensional book-user matrix to a two-dimensional visualisation, we use dimensionality reduction, an unsupervised machine-learning technique.

Dimensionality reduction is a technique commonly applied to high dimensional data sets to extract the most important features and underlying structure of the data. It's unsupervised, as there are no predefined labels the algorithm learns from, but the result is based entirely on the data and the algorithm used. A commonly used method of dimensionality reduction was developed before the Second World War and is called *Principal Component Analysis* (PCA). It's a linear method that uses the eigenvalues and eigenvectors of a given matrix to determine its principal components. The current state-of-the-art methods are no longer linear, but use more sophisticated mathematical principles to retain the most dominant features of the data.

Considering the complexity of the data and the related work, the *Uniform Manifold Approximation and Projection* algorithm (UMAP) (McInnes, Healy, and Melville 2018) was preferred over PCA or *t-distributed Stochastic Neighbour Embedding (T-SNE)*, as it outperformed the other techniques in terms of computational time and visual quality of the result. UMAP is known for having no computational restrictions on embedding dimension and creating visually appealing results. Explaining the precise workings of the algorithm is beyond the scope of this project, but is brilliantly explained by McInnes in a talk he gave at the 2018 SciPy conference[1].

1. McInnes' talk on the workings of the UMAP algorithm can be found here: https://www.youtube.com/watch?v=nq6iPZVUxZU.

Applying UMAP makes it possible to project each book from the high-dimensional user loan space onto a two-dimensional plane. This task can rarely be achieved perfectly, and the results of such mappings and their interpretability for natural data sets have recently led to criticism (Chari and Pachter 2023). This does not necessarily undermine the effort to understand the processes involved in the data generation and gain insight into the collection, however interpreting the distances in the resulting mapping has to be done with care.

There are several hyperparameters of UMAP that affect the outcome of the resulting mapping, of which arguably the most important is the measure of distance applied to compute similarities of points in the high-dimensional space. In the case of the TF-IDF user loan matrix, the similarity between two books was determined by calculating either the *cosine-similarity* or the *hamming distance* between the rows of the matrix. To compare the difference between the distance metrics, several experiments were run on the full data set, while also varying the *minimal distance* and *number of neighbours* parameters. These hyperparameters have an effect on how closely points can group together and how much of global versus local structure is preserved respectively. The result of these experiments can be seen in Figures 4.3 and 4.4.

Mapping the data set using a range of hyperparameters and distance metrics can provide insight into the features of our data, especially if they remain stable regardless of the parameters and layout process. The results of the experiments show how points gravitate to a larger centre cluster and smaller clusters around it. This remains consistent regardless of which distance metric and other parameters are used.

Importantly, the mappings resulting from applying UMAP by no means perfectly represent the original data and not all distances between the points are necessarily interpretable. If a point were completely disconnected from all others, the UMAP algorithm would fluctuate it randomly and place it somewhere in the mapping during the layout process. In the case of this project a disconnected point would be a book borrowed only by one user, who in turn only ever borrowed this one book. The shortcomings of the UMAP algorithm and the caution that has to be used when interpreting the results do not necessarily mean it is unfit for this purpose and context. Since the global structure of the resulting mapping remains robust independently of parameters, it seems reasonable to pick the mapping with the highest visual quality and clearest clustering and conduct further analysis. Because

of its clear separation of smaller clusters and visual appeal, the mapping resulting from using the *hamming distance* as a distance metric, the number of neighbours set to 100, and a minimal distance of 0.1 was selected for a more in-depth evaluation.

To analyse and interpret the chosen mapping, a clustering algorithm was run on the resulting UMAP embedding. HDBSCAN (Campello, Moulavi, and Sander 2013) was preferred over other clustering algorithms, as the data is noisy, and the clusters are not of similar size or shape. HDBSCAN works by clustering high-density areas that are separated by lower-density areas, and thus it manages noisy and non-linear data better than other algorithms (Berba 2020). The resulting clustering after parameterisation distinguishes clusters of varying sizes and shapes. This implementation has the advantage of assigning probabilities to all points indicating the likelihood of their belonging to the cluster (McInnes, Healy, and Astels 2016), which gives additional insight into how well the clustering is working and helps to detect noise and outliers.

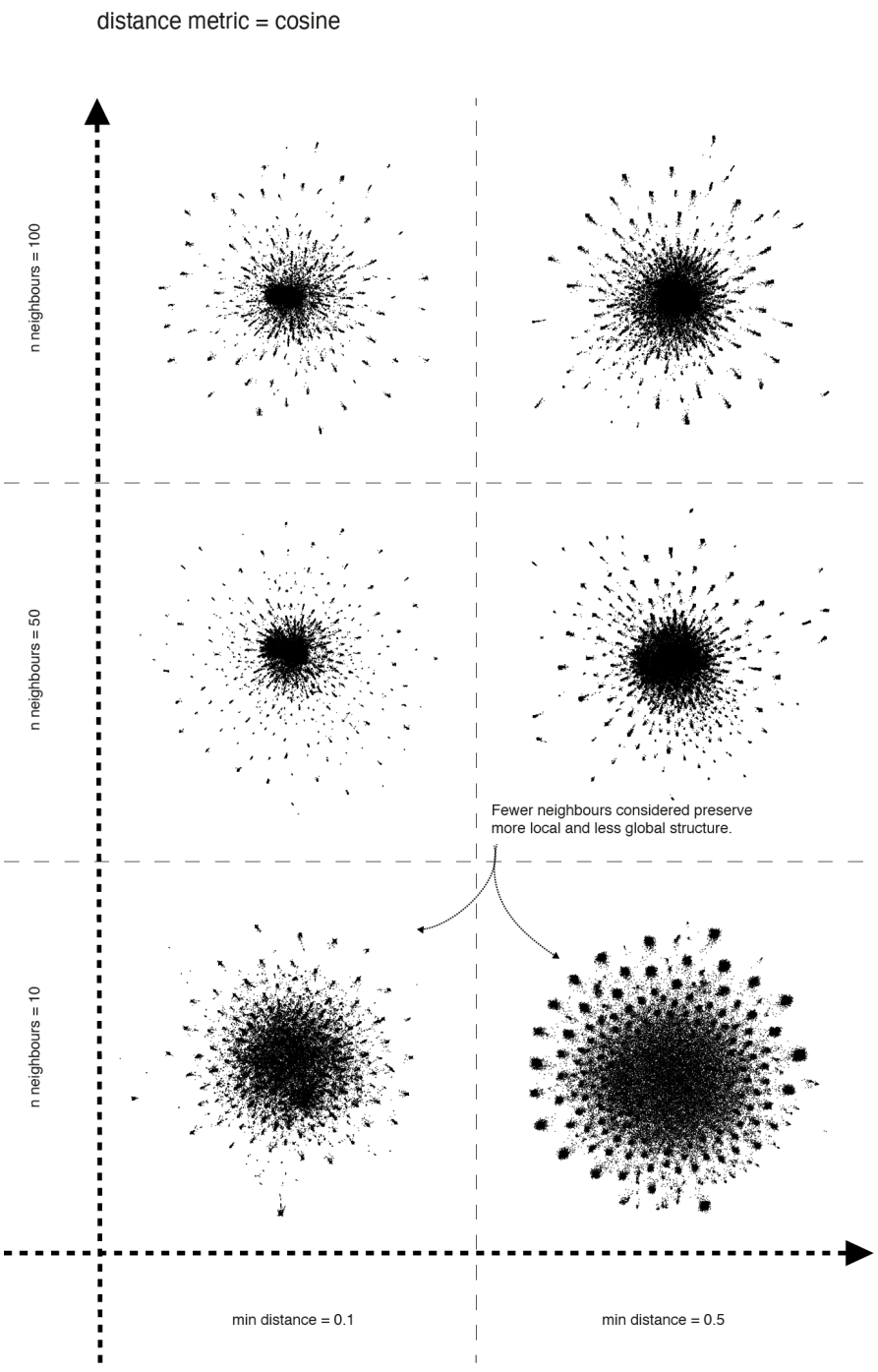


Figure 4.3: The effect of the hyperparameters on the UMAP algorithm illustrated through embeddings using the cosine distance as a similarity measure. The overall structure is relatively robust, with a larger cluster in the centre and smaller ones surrounding it.

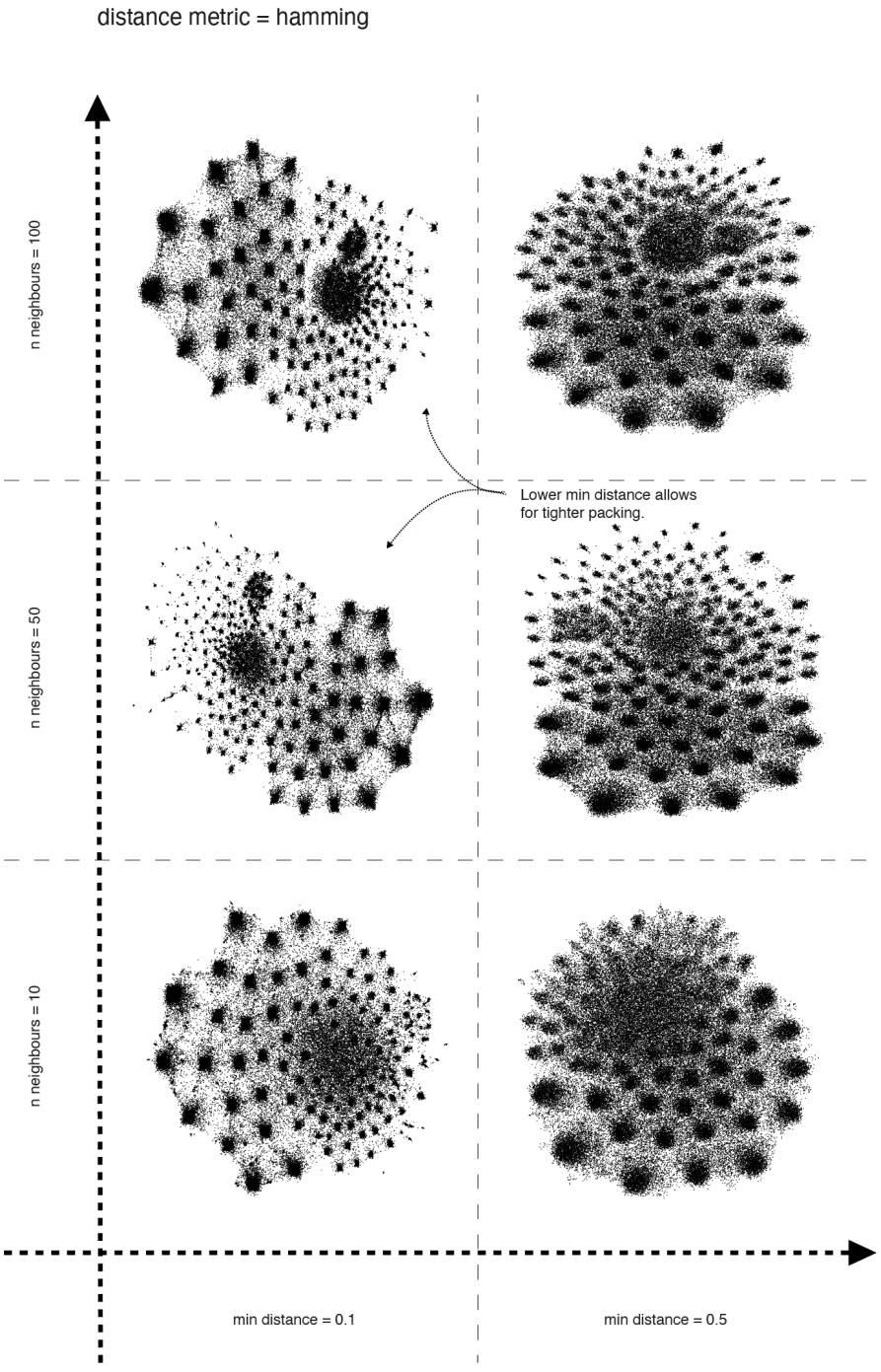


Figure 4.4: The effect of the hyperparameters on the UMAP algorithm illustrated through embeddings using the hamming distance as similarity measure.

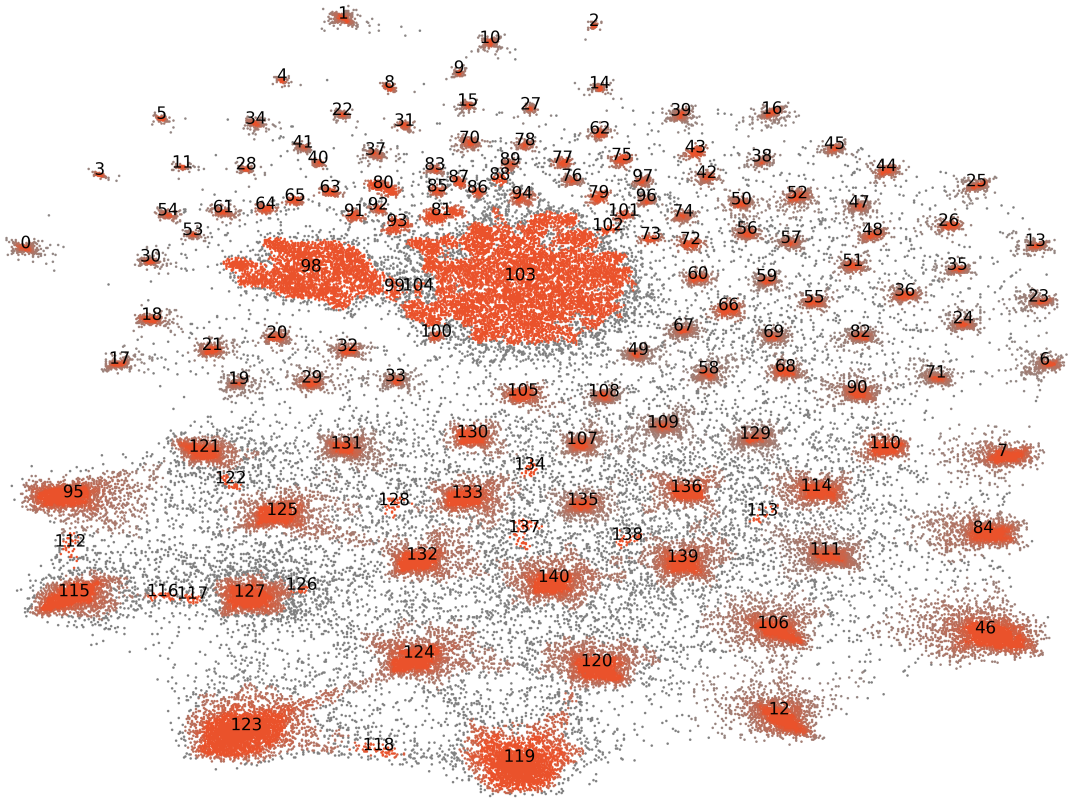


Figure 4.5: The resulting clusters after running HDBSCAN on the two-dimensional embedding. The colour intensity of the points is determined by the probability of their belonging to the cluster they're assigned to. Points in grey are not assigned to any cluster.

4.3 Analysis and Evaluation

The resulting clusters from the previous step were then analysed using the metadata of the books and general information about the clusters to understand how meaningful the embedding is. For instance, the number of users associated with each cluster can indicate a specialised niche research area or larger thematic group. The thematic content of the clusters can be extracted from the classification in the local signature system. The distribution of language across the clusters helps determine if the mapping represents the multilingual community of Bibliotheca Hertziana. By contrast, the temporal distribution can indicate if clusters can be linked with ongoing research projects. This analysis shows how representative of the data and the context of Bibliotheca Hertziana the mapping is.

4.3.1 Subject

By analysing the local shelf numbers of books in the clusters, we gain insight into user interaction with the collection. The aim is to find out whether users predominantly borrow books from the same few shelves or if they explore a diverse range of signature groups for their research. This analysis can reveal potential trends within categories over time and the primary thematic focus of researchers engaging with the collection.

Using colour to distinguish the subjects and visualising them as a small multiple (Tufte 2013) illustrates the subject distribution across the mapping and clusters. For instance, subjects like *Italian Artists* are consistently present across all clusters with similar density. By contrast, the *Periodicals* are primarily concentrated in the lower section of the mapping. Figure 4.8 illustrates these patterns, suggesting users engage with books from several signature groups. This may indicate underlying patterns in user engagement, possibly reflecting a preference for multiple library sections, encouraging further investigation into the user interaction dynamics with the library collection.

Similarly, Figure 4.6 depicts the percentage of each signature group of the classification system across the clusters. Notably, there is an evident prevalence of books attributed to *Italian Artists*, *Italian Topography*, and *Catalogues* sections across most clusters. This could suggest a frequent borrowing pattern from these

sections for all research conducted at the Bibliotheca Hertziana. Another possible interpretation is that these categories complement each other, hence their concurrent use. Some clusters show high percentages exclusively attributed to only one subject category, visible through brighter bars in Figure 4.7. For instance, cluster 137 focuses on books indexed as *Periodicals*, indicating a niche interest among users. It's worth noting that there are only a few clusters with high percentages for any given category, suggesting that the clusters reflect broader subjects rather than very specialised topics.

4.3.2 Language

The collection contains books in several languages, and the visiting researchers, as well as the fellows of the institute, are a multilingual crowd. Exploring the language composition of the clusters can help to understand if they reflect this multilingualism or if there are pockets in the mapping where one language dominates.

Most books have an assigned language tag. However, many books in the rare collection do not. To mitigate this, language detection was run on the titles of the books with no assigned language. Plotting each cluster against the percentage of books in a specific language in 4.7 shows a clear preference of users for books in Italian, followed by German, English, and French. This reflects the context of the Bibliotheca Hertziana, a German institute at the heart of Rome with researchers from all over the world coming to access the collection. The mapping, however, lacks any cluster that focuses on literature in Spanish, which would be expected since a large group of fellows is working on Latin America.

4.3.3 Loan activity

Mapping the loans over time and marking high-activity regions can show which clusters and topics were busy at any given point. This can provide insight into art historical research over the past years. Figure 4.9 shows a clear shift between active clusters before the change in loan policy in 2019 and after. Clusters in the lower part of the mapping appeared very busy before 2019, followed by a significant decline in activity. The upper part of the mapping shows the opposite trend, suggesting that these clusters can be assigned to the research of internal users.

The evaluation results suggest patterns known from the library context that are not represented by embedding the complete data set. Clusters focusing on Spanish books are missing, and the subjects present in the clusters suggest that they represent broader subject areas instead of niche research interests. Additionally, there is a clear difference in loan activity in the upper and lower parts of the mapping, indicating that the higher number of loans to external visitors is masking the research of internals.

A second mapping was created to address these issues and obtain a more insightful embedding, more representative of the research being conducted at Bibliotheca Hertziana, using only the internal user loans.

The internal loan data has been more stable over the years. It was not unduly affected by the change in policy in 2019 and reflects the research conducted within the institute.

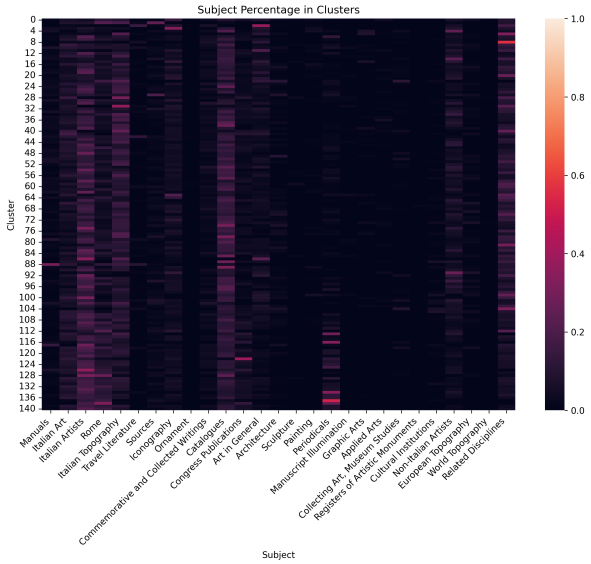


Figure 4.6: Heat map of the percentage of a subject within each cluster. There are only very few clusters with a high percentage in a subject, suggesting that the mapping and resulting clusters are very broad.

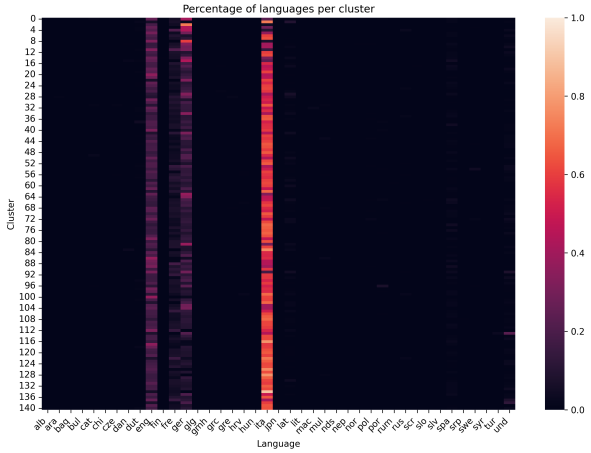


Figure 4.7: The percentage of books of languages within clusters showing a clear preference for Italian books. There is a distinct lack of clusters focusing on Spanish books, indicating that the external loan data might be masking the internal research.

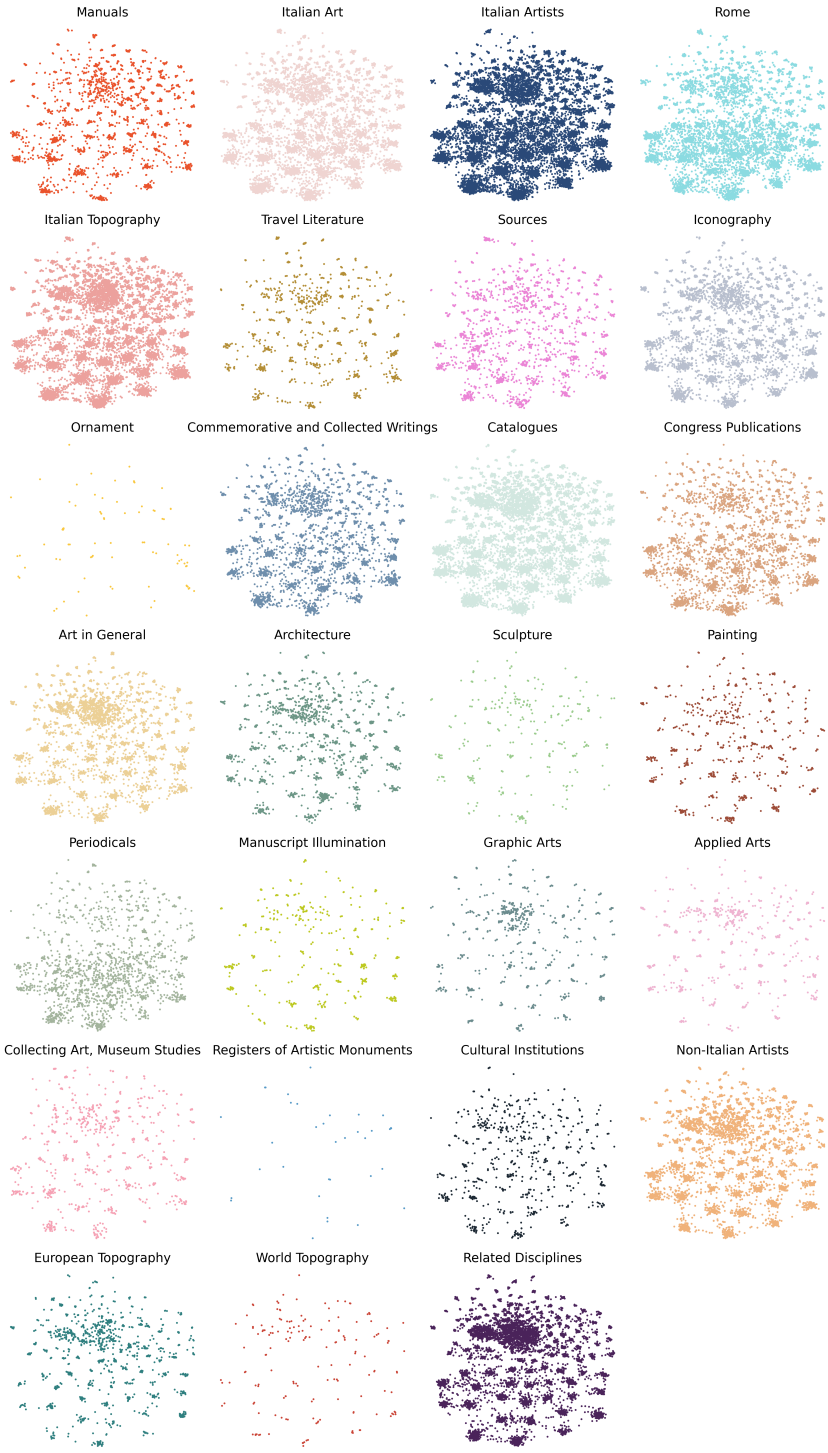


Figure 4.8: Distribution of shelf number classification throughout the embedding. Some subjects are evenly distributed, indicating importance for most research done at the Bibliotheca Hertziana, while other signature groups tend to appear in only one area of the mapping.

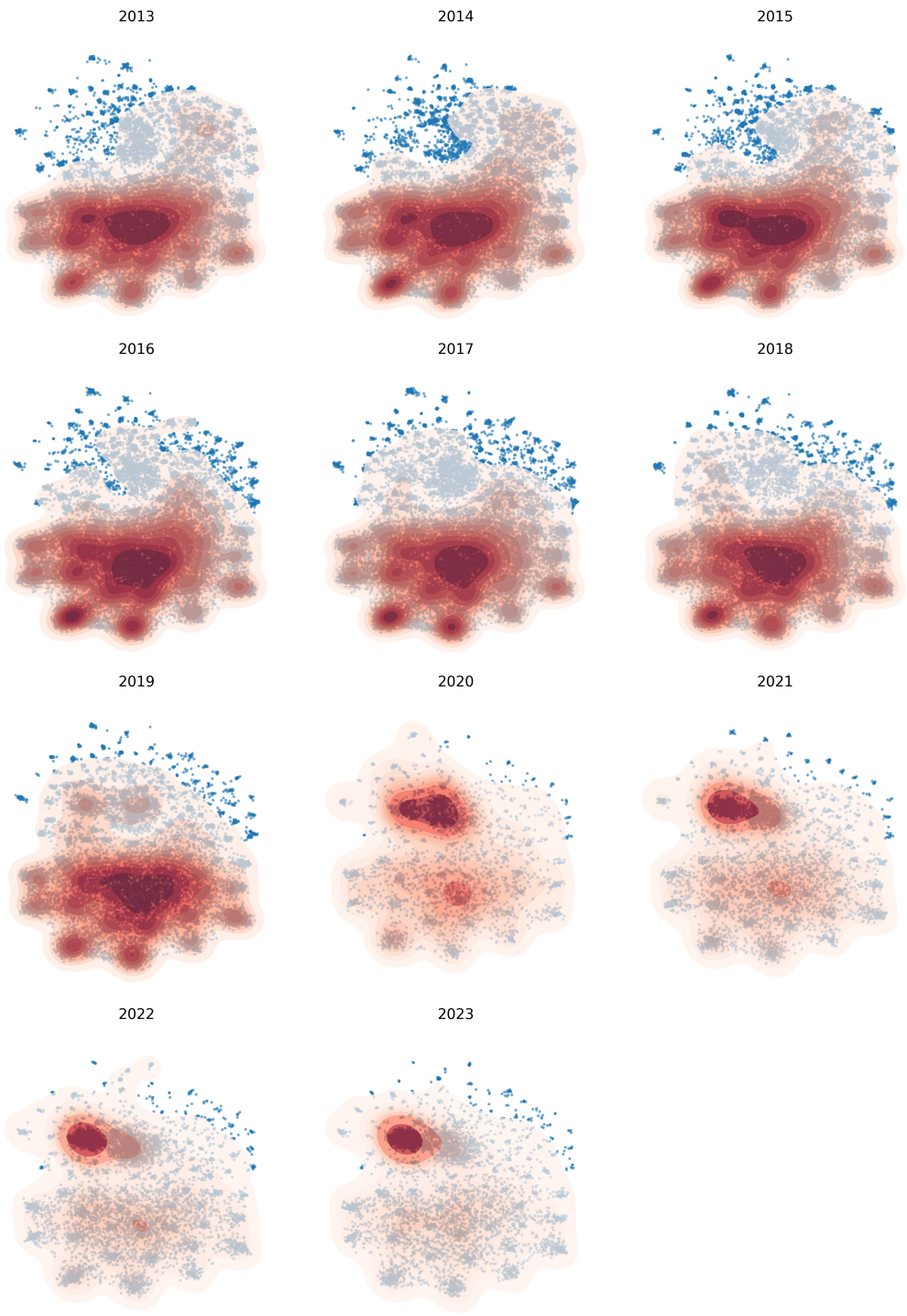


Figure 4.9: Density estimate of loan activity over the years shows a clear difference between the upper and lower parts, suggesting that external users' loan activity is predominantly mapped in the lower part of the mapping.

4.4 The Internal Research Community

In light of the mapping analysis in the previous chapter, showing that the high loan volumes by external users might obscure internal research trends, a mapping based solely on the loan data from internal users at Bibliotheca Hertziana was generated and evaluated. It would, it was hoped, unveil clearer clustering around specific research areas, depicting the interests of researchers and fellows of the institute.

The mapping process of the internal loan data followed the same methodology outlined for the complete data set but excluded the check-out data for external users. The embedding obtained from creating the high dimensional TF-IDF matrix and projecting it into two dimensions using UMAP has a structure similar to the upper part of the previous map. This section was suspected to represent the loan activity of internal users, featuring a prominent central cluster with smaller satellite clusters around it.

The same steps were used on this mapping to evaluate its representation of collection and context at Bibliotheca Hertziana. The heat maps presented in Figure 4.12 and 4.11 illustrate the distribution of languages and subjects across clusters.

Faint yet visible bars in the Spanish language column in Figure 4.12 show a more prominent presence of Spanish books within clusters, potentially indicating an association with user activity linked to the institute's projects investigating the connections of Italian art and art in South America. This suggests that the mapping based on the internal loans offers more apparent distinctions between the research groups and their corresponding loan activity than the mapping based on the complete data.

The thematic focus of clusters was again extracted via the signature system. Figure 4.11 illustrates an improvement compared to the mapping on the entire loan data, as more of the clusters have higher percentages in specific subject areas, indicating a clustering around users focusing on distinct subjects. This suggests a sounder clustering for the books based on only the internal loans, showcasing the niche research interests of the fellows of Bibliotheca Hertziana.

In addition to the visual analysis of subjects and languages, a temporal perspec-

tive on the clusters shows how the mapping on internal loans better represents the context at the Bibliotheca Hertziana. Research fellows, typically with tenures spanning a few months to two years, leave observable imprints in the clusters, often characterised by bursts of concentrated activity over relatively brief periods. The loan frequency for each book across every year within the data set was gathered to validate this observation. This resulted in a dictionary mapping each year to the number of times the book was loaned during that period. The frequencies were normalised using a Min-Max scaler for each year to make the years comparable and minimise the impact of the policy change in 2019 and the COVID-19 pandemic on total loans.

The heat map in Figure 4.13 depicts this normalised activity across clusters, with brighter colours indicating heightened loan activity that year. The activity of several clusters over the years reveals peaks of intense borrowing spanning one or two years, affirming the correlation between fellows' tenure duration and these concentrated bursts of borrowing within the internal loan mapping.

As described in the previous chapter, visualising the density of loans across the years offers insight into the spatial distribution of activity within the collection mapping. This representation shows patterns of loan activity and may be a promising way to establish where the primary canon of art historical texts and emerging subjects are located.

Figure 4.14 illustrates the spatial loan density observed over the years and indicates an area in the lower part of the mapping characterised by high activity across all recorded years. In subsequent analysis, we aim to ascertain whether these clusters predominantly house the primary corpus of art historical literature. Additionally, smaller clusters show shorter bursts of activity, suggesting potential associations with short-term research projects involving fewer researchers. For example, the clusters in the central area of the mapping display a surge in activity from 2020 onward, calling for further investigation to determine their connection to research projects during that period.

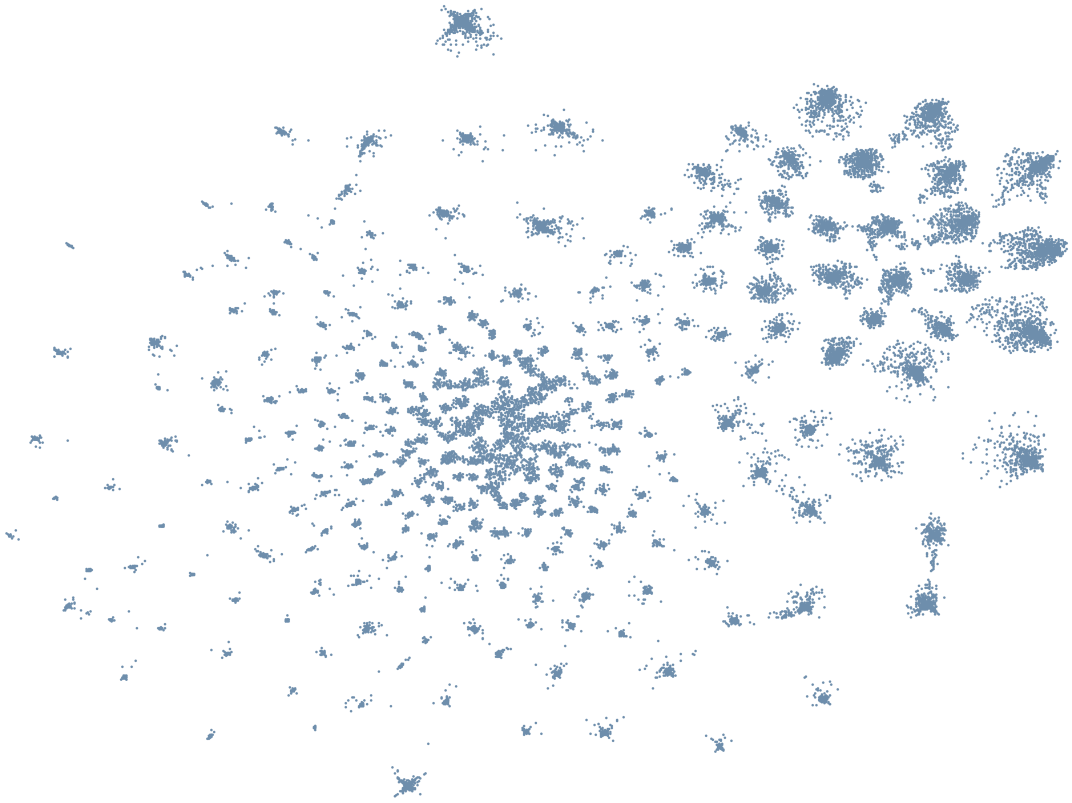


Figure 4.10: The result of the mapping process on the internal user loan data follows a similar structure as the upper part of the mapping of the whole loan data. There is still a larger centre cluster with smaller clusters around it.

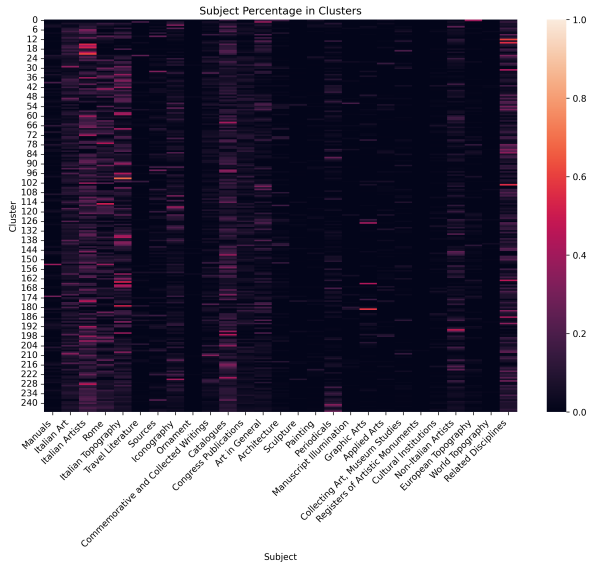


Figure 4.11: Heat map of the percentage of a subject within each cluster. There are clusters with a clearer focus on one subject, and overall, the books are more distributed over the subjects than for the full mapping.

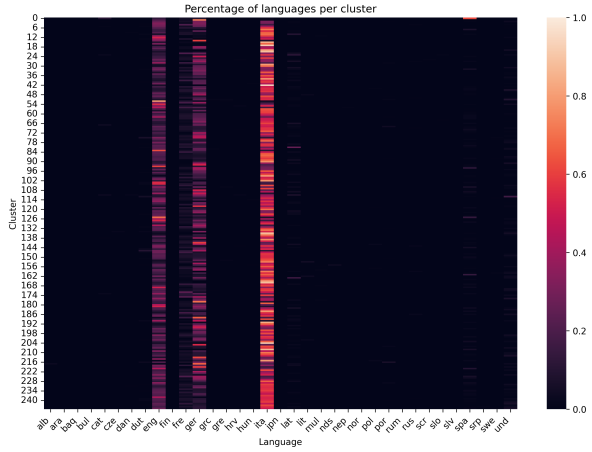


Figure 4.12: The percentage of books of languages within clusters is more distributed than for the full data set. There are some clusters with visible bars for the Spanish language, indicating that some of them represent the fellows' research focusing on South America.

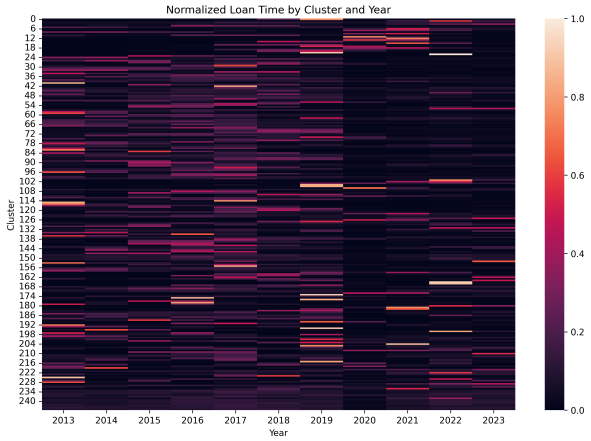


Figure 4.13: Result of normalising the loan activity over each year using a Min-Max scaler to mitigate the impact of the loan policy change and the COVID-19 pandemic. The heat map shows clear bursts of activity in most clusters, indicating a clustering around specific projects and fellows of the institute.

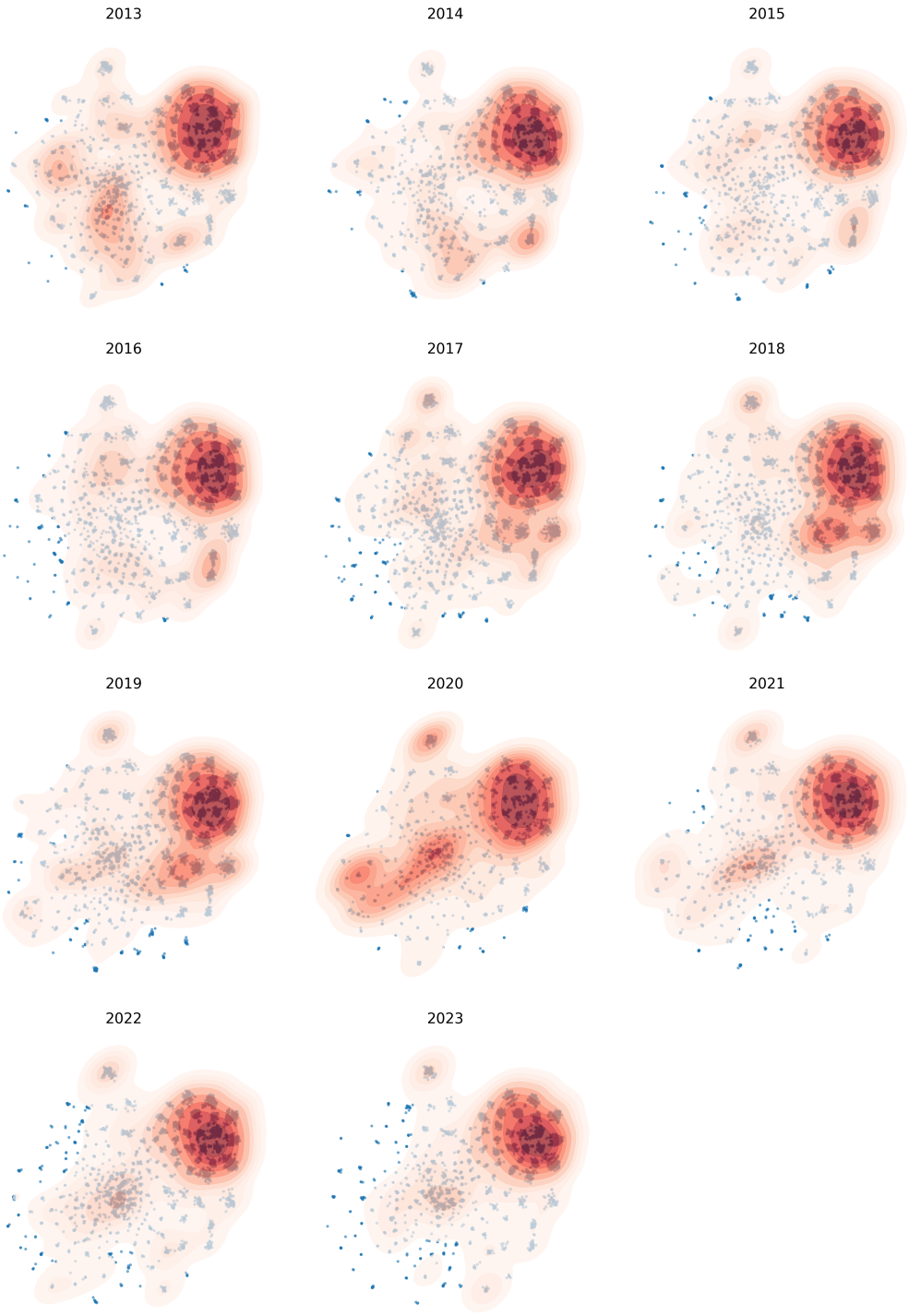


Figure 4.14: Loan activity across the mapping over the recorded years shows a high-density area in the lower part of the mapping. Other areas show shorter bursts in activity, calling for further analysis of the differences between clusters.

Chapter 5

Automated Subject and Description Generation

> The only good classification is a living classification.
>
> *Goeffrey C. Bowker; Susan Leigh Star*

Creating a dynamic library display that reflects the institute's ever-changing research landscape involved establishing a new subject classification system. Central to the approach is harnessing the collective knowledge of researchers at Bibliotheca Hertziana by basing the mapping on their interaction with the collection. This allows the creation of a user-centric classification system and a move beyond traditional bibliographical classification schemata.

5.1 Model Prompting

Given the limitations of the available data, particularly the lack of text to conduct semantic analysis on the documents to establish thematic connections, a different approach was adopted to establish a suitable subject language for the book clusters.

Instead of conducting text analysis on full texts and extracting keywords, the

word frequencies of book titles in clusters were used to generate a subject classification, mapping the natural language expressions in the titles onto a subject terminology. Svenonius emphasises the need for algorithms performing this task to be capable of translating natural language expressions into structured subject-language terminology (Svenonius 2000). Given the recent advancements in large language model technologies, using OpenAI's model GPT-4 for this purpose seemed highly promising. This approach could provide an effective means of interpreting and categorising the clusters, thus offering a scalable solution for dynamic subject classification in the context of limited textual data.

To address the variability of languages comprising each cluster, the analysis was focused exclusively on books written in the four most prevalent languages in the resulting clusters. Subsequently, stopword removal was performed on the remaining titles using, whenever available, the stopword lists from the NLTK library (NLTK Project 2023). By filtering out common words offering little value in the context of subject classification, this approach ensures a more meaningful extraction of keywords important for the mapping from titles to subject terminology. After stopword removal, TF-IDF was applied to derive a list of words and their importance to distinguish clusters, which later acted as input to the GPT-4 model.

The word-score lists generated from titles varied substantially in size, predominantly influenced by the number of titles in each cluster. The number of words and frequencies per cluster was limited to mitigate the varying sizes of clusters and the expenses related to the OpenAI API (OpenAI 2020) usage later on. A maximum of 400 words per cluster was set, allocating a quota of 100 words per each of the top languages in the cluster. Discarding lower-scoring words ensured the retention of the most relevant words for effectively distinguishing the clusters and simultaneously ensuring cost-efficient computation.

To translate the words and scores obtained from the book titles into a coherent subject classification for each cluster, the GPT-4 model was queried using the OpenAI API. The model was prompted to suggest a subject classification for each cluster based on the input of terms and their corresponding TF-IDF scores.

Two messages were sent as a prompt for each cluster. The first specified the request, clearly outlining the desired output and format, while the second contained the list of terms and scores. The full prompts can be found in Appendix A. This approach was preferred to that of sending only a single request and then

transmitting all cluster lists in separate messages, which yielded unsatisfactory responses where the number of subjects did not match the number of clusters. The two-message-per-cluster approach, illustrated in Figure 5.1, proved more effective and reliable, ensuring that the resulting list of subjects was consistent with the number of clusters.

In order to evaluate the impact on its performance, various levels of context were provided to the model during the experimentation process. Interestingly, the model showed reliable capabilities in distinguishing the clusters and generating coherent subject classifications, even without mentioning that the books were from an art historical library. The evaluation and validation of the model's responses with relevant classifications will be discussed in detail in sections 5.2 and 5.3.

Building on the initial prompt response, a second round of queries was made to obtain a more detailed description of the clusters. This followed the same two-message-per-cluster approach shown in Figure 5.1 and involved inputting the previously generated cluster subjects combined with the titles of the 100 most loaned books from each cluster. The rationale behind this limit was to filter out potential noise that might have been introduced during the clustering process and to maintain cost efficiency in light of the OpenAI API's pricing (OpenAI 2023).

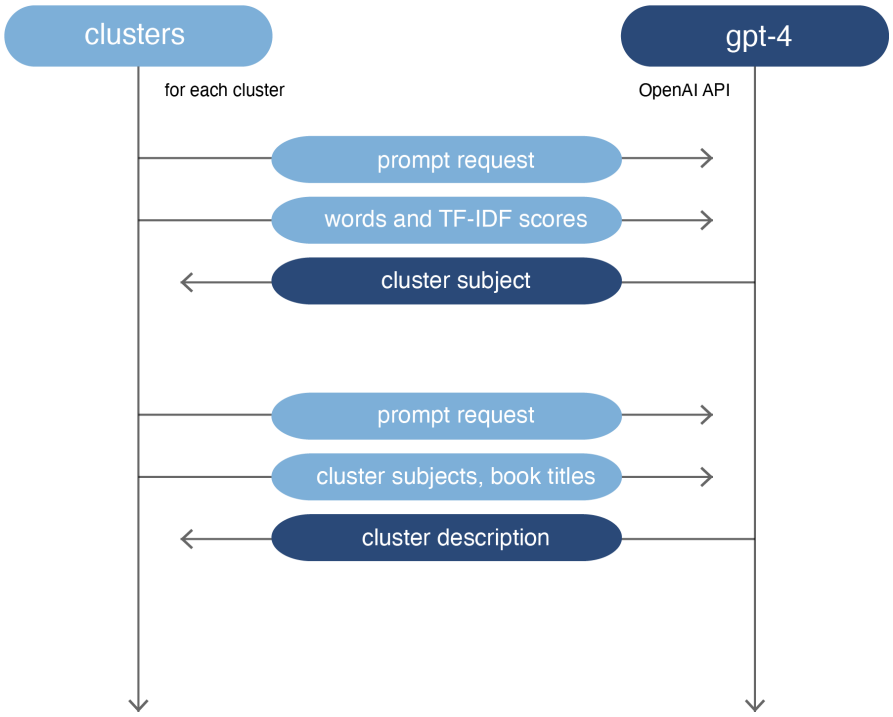


Figure 5.1: The prompting approach used to query OpenAI's GPT-4 model to create a mapping from the natural language expressions in the book titles to a subject terminology for the clusters. The prompting approach illustrated here resulted in stable and meaningful results.

5.2 Cluster Atlas

To evaluate the automatic subject and description generation qualitatively, the results were presented to research staff at Bibliotheca Hertziana visually to ascertain whether they could recognise their own research. This was done by creating a cluster atlas[1], depicting each cluster with its subject classification, description, and most loaned books, as well as plots showing the location of the cluster in the mapping and the time frame when the books in the cluster were loaned.

The cluster atlas was created with a Python script, which translated HTML code into PDF files using the pdfkit library (Stanislav 2021) . This approach streamlined the generation of the atlas but also ensured its coherence. It also allowed for multiple iterations of creation and evaluation of the subject and description, each time assessing the impact of various inputs and prompts for the GPT-4 model.

Once compiled, the atlas was presented to members of the Hertziana research staff, who were asked to identify either their own research or the research of colleagues. Since fellows often stay only for up to two years, the institute's permanent staff were especially helpful in identifying past research projects.

Some of the clusters were immediately identifiable. For instance, cluster 14 was clearly attributable to the research done by Golo Maurer, the head of the library. The cluster showcased a conspicuous combination of topics, including German travellers in Italy, Goethe, and the future of digital libraries. This thematic focus, along with the time frame of the loans, reflects his research interests in recent years. The subject classification, description, and cluster location can be seen in Figure 5.2. Since taking over as head of the library, he has written both a Master's thesis on digital libraries and a book discussing Goethe and German travellers in Italy, titled *Heimreisen: Goethe, Italien und die Suche der Deutschen nach sich selbst* (Eng. Travelling Home: Goethe, Italy and the Search of Germans for the Self) (Maurer 2021). During the discussion, he confirmed that the titles displayed and the generated subject and description stemmed from his research, which also aligned with the period designated by the top loans in cluster 14. The cluster's location in the mapping indicates it is a more niche field of research at Bibliotheca Hertziana, as it is disconnected from other clusters and not part of the larger centre cluster.

1. The full cluster atlas can be found here: https://zenodo.org/records/10700617 (H. L. Casey 2024).

While not all clusters were as distinctly recognisable as Cluster 14, feedback from former and current fellows and one of the directors, Tristan Weddigen, indicated that the atlas was particularly useful for identifying clusters related to more niche subjects. This includes specialised clusters, such as research on Italian cinema or related to the *Italy in a Global Context* projects.

Some examples of directly identifiable clusters, which might be associated with larger ongoing research topics at the institute, are clusters 7 and 11.

Cluster 7 is linked to the *Italy in a Global Context* (Bibliotheca Hertziana 2018) projects in the Weddigen department. Its generated subject and description can be seen in Figure 5.3, which indicates a thematic focus on Italian and Brazilian modern art and the connections between them. As the cluster is not part of the centre cluster, it can be assumed to be a highly specialised field of research. This makes the cluster more disjoint from others and more straightforward to associate with a research project.

Similarly, cluster 11 is located close to cluster 7 but is very different thematically. As seen in Figure 5.4, it focuses on Italian film and the socio-political aspects shaping urban spaces. It can likely be associated with the *Social Reality in Italian Films* project in the Michalsky department (Bibliotheca Hertziana 2019). The loan data mapped in 5.4 reveals a clear peak of activity from 2019 until 2021. This period coincides with the typical post-doctoral fellowship in the department Michalsky, further strengthening the assumption that the cluster results from a specific research project completed in the above-mentioned period.

Browsing the yearly reports, one comes across the research project by Alberto Lo Pinto, a former researcher at the Michalsky department, entitled *Milan as a Site of Social Anxiety: Negotiating Notion of Modernity and Gender in Michelangelo Antonioni's La Notte (1961)*, which is very likely the source of the books loaned in cluster 11. Upon being shown the titles related to cluster 11 and the generated subject and description, Lo Pinto confirmed that he had borrowed the books for this project and asked if the cluster description was developed directly using his publications.

The above examples demonstrate the variety and complexity of research conducted at Bibliotheca Hertziana. They also showcase how the disjoint clusters in the mapping can easily be identified and linked to research projects. Directly identifying research projects relating to clusters in the centre and upper right of

the mapping proved more challenging as the subjects and descriptions are similar. According to the generated subject classification, the clusters mostly contain literature about European art history, the Italian Renaissance, Baroque and medieval periods, reflecting the collection focus and the main research areas at Bibliotheca Hertziana. This aligns with previous maps of science, where the major research areas are represented in larger clusters with disciplinary topics around them. It also corresponds with the activity mapping in Figure 4.14, which showed continuous loan activity in the upper right of the mapping.

14 - Digital Age and Postdigital Art History in Modern German Culture

This collection of books focuses on the broad theme of modern German culture in the context of the digital age and postdigital art history. It explores a variety of sub-themes encompassing the historical evolution of libraries, the significance of paper and book history, and the transformation that occurred with the advent of digital culture. Renowned figures such as Goethe and their relation to Italian culture is a prominent subject, as well as the intersection of German and Italian study. Furthermore, it includes a deep dive into the German imaginary and its perception of the south.

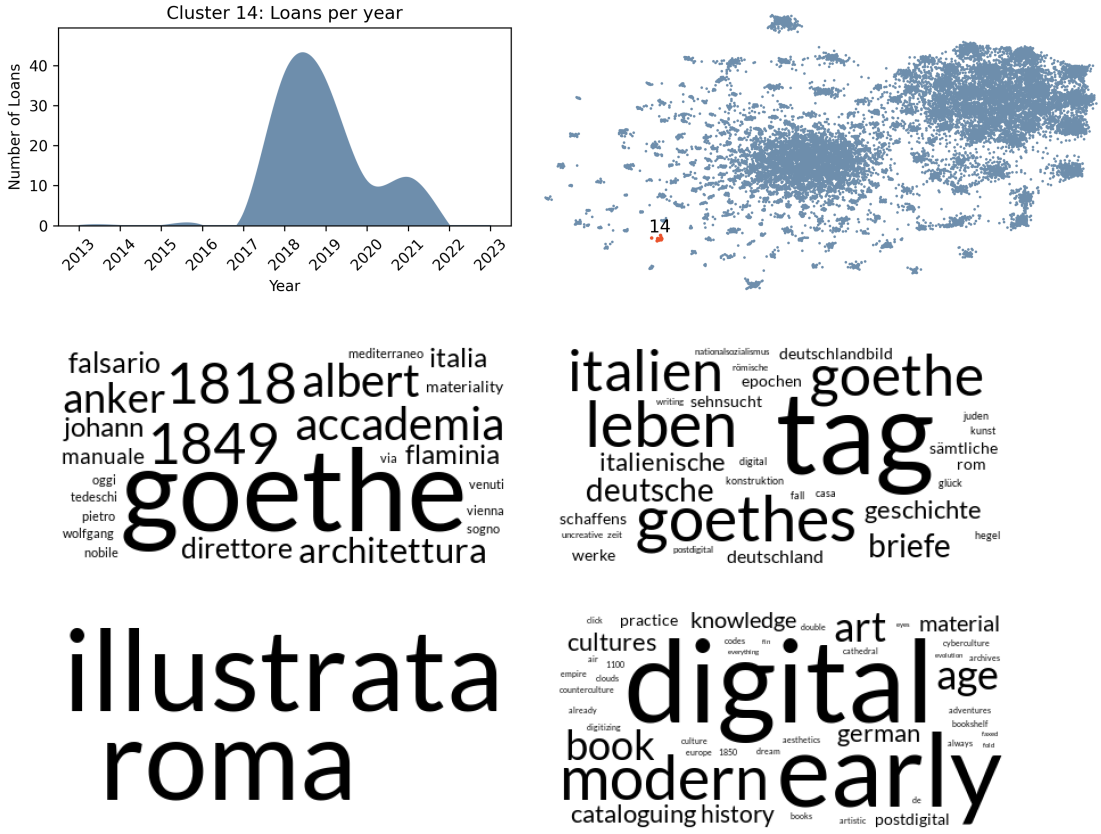


Figure 5.2: The cluster subject classification and description generated via the OpenAI API for cluster 14 based on the word frequencies in book titles. The cluster is disjoint from other clusters and is not located in the larger centre cluster, indicating it is a niche field of research. The timing of the loans and the subject classification are coherent with research undertaken by the head of the library at Bibliotheca Hertziana, Golo Maurer.

7 - Modern Art and Architecture in Italy and Brazil

This cluster of books investigates and illustrates the evolution, variety, and impact of modern art and architecture in Italy and Brazil. The titles cover a wide array of subjects, from the Scuola Romana movement and South American 20th-century art to detailed analysis of architecture in Asmara and the arts scene in Brazil. The collection also delves into specific events, exhibitions, and notable figures within the art and architecture sphere of both countries. A valuable resource for those seeking a comprehensive understanding of modern artistic and architectural influences in Italian and Brazilian culture.

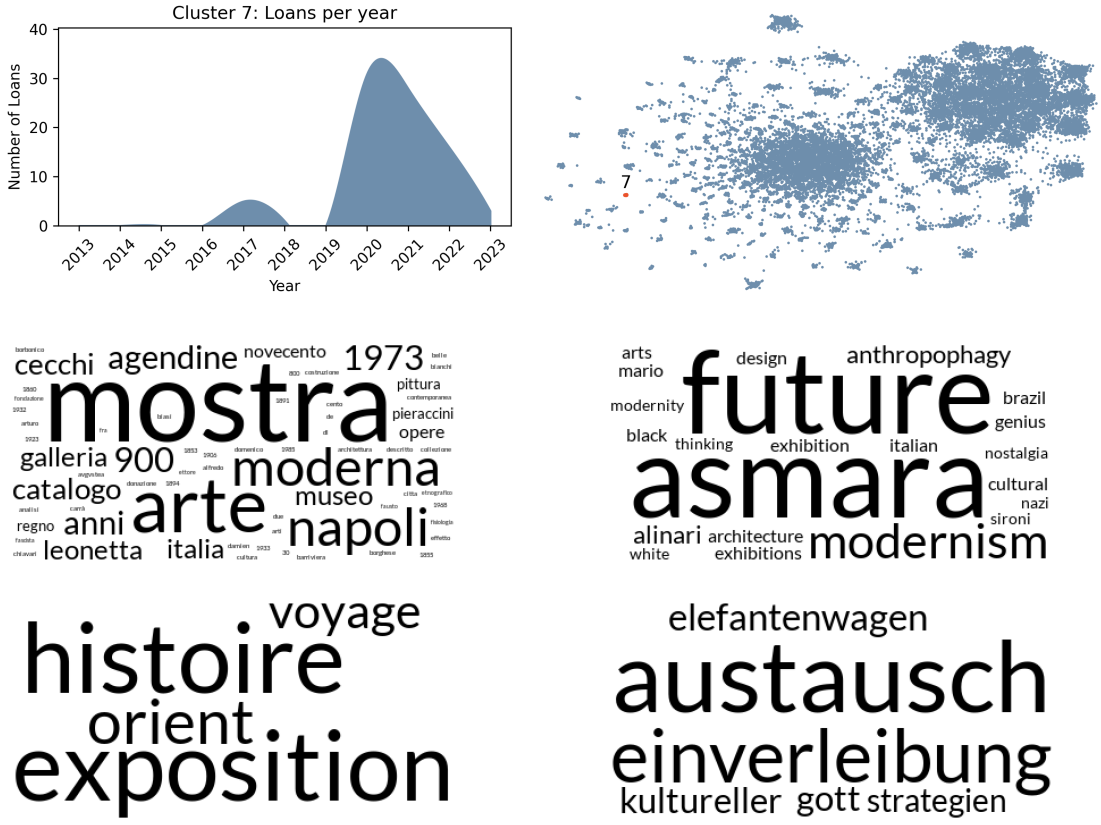


Figure 5.3: Cluster 7 focuses on the connection between Italian and Brazilian modern art, as shown through the word clouds generated from book titles. The automatically generated subject classification and description reflect this focus, albeit in a vague way. It is likely part of the *Italy in a Global Context* project in the Weddigen department at Bibliotheca Hertziana.

11 - Modern Architecture and Cinema in Italian Urban Spaces

This cluster compiles books that examine the intersection of modern architecture, urban spaces, and cinema in Italy - with a particular focus on Milan and Rome. It showcases the evolution of cities over time, reflected in architectural structures and captured via film. These books discuss key architects and filmmakers who've influenced city landscapes and aesthetics, such as Giò Ponti and Michelangelo Antonioni. The cluster also addresses theoretical, historical, and aesthetic discussions around urban planning, architectural design, gender and space, filmic portrayals of cities, and socio-political aspects in shaping modern Italian urban spaces.

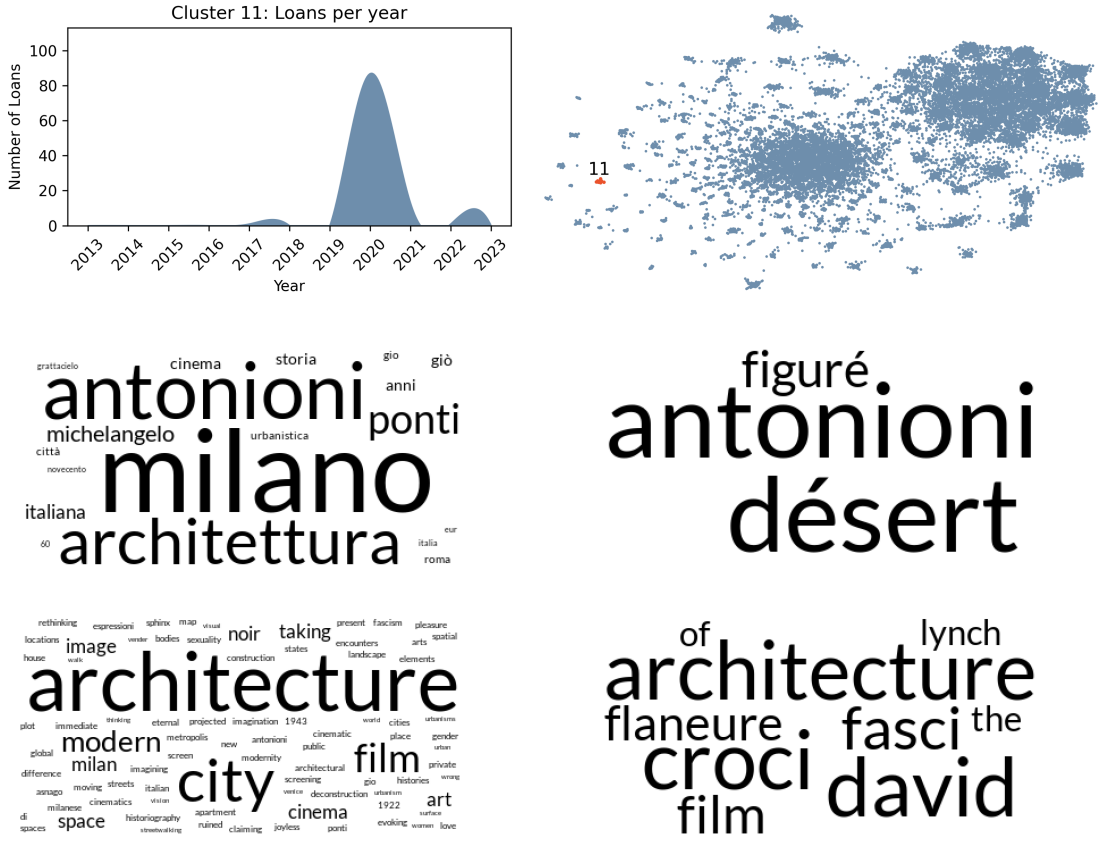


Figure 5.4: Cluster 11 focuses on architecture and urban spaces in Italian film, specifically Milano and Rome. The research project underlying the book loans in the cluster is probably linked with the *Social Reality in Italian Films* project in the Michalsky department.

5.3 Validation

The mapping of the library's collection is for the benefit of the research community at Bibliotheca Hertziana It is intended to reflect the dynamics of the research conducted at the institute. It furthermore provides the basis of a flexible classification system and functions as a map of art history by leveraging scholars' collective knowledge.

This section is dedicated to validating the cluster subjects and descriptions through interviews with selected Bibliotheca Hertziana community members. The scholars at Bibliotheca Hertziana are a valuable source of expertise on the collection and the ongoing research projects. The interviews were conducted in person at Bibliotheca Hertziana and in Zurich with several individuals associated with the different departments at the institute. The participants had varying levels of awareness about the project: a few were directly involved and consulted regularly.

Interviews were conducted with the heads of three departments and their scientific assistants to validate the map. This selection of scholars was due to their direct involvement with the institute, which means they are well informed about past and ongoing projects. The heads of the departments are especially valuable sources of understanding of the institute's context and workings. Similarly, some scientific assistants have been at the institute for almost a decade and can give unique insight into past research projects. As they are also part of the target audience for a potential recommendation system based on the mapping and the automated classification scheme, they were asked questions addressing their reading of the map and its usefulness for their research.

In preparation for the interviews, a set of clusters was pre-selected for each department based on their similarity to descriptions of ongoing projects at the institute. In the case of the head of the library, the selection consisted of clusters from all areas in the mapping, as well as the cluster that could be directly linked to him (see Figure 5.2). Because of his profound knowledge of the library collection and its development in recent years, he was well qualified to respond to the visualisation's representation of the trends in research.

The interviewees were asked questions regarding several possible readings provided by the mapping. First, they were asked to link research projects they know about to the selected clusters using the subject classifications, descriptions,

borrowing time frame, word clouds, and list of book titles from the cluster atlas. This helped to determine if the departments and the institute were represented in the mapping. The participants were not asked to find their projects in particular, as some had not been at the institute long enough to be shown in the clusters, although many did. A printed version of the cluster atlas and the mapping served as navigation tools during the interviews.

Part of the interview was focused on evaluating the representation of recent developments in the field, as one of the mapping's possible readings is as a map of art history. In addition, as the generated subject classification described in section 5 is supposed to represent these developments dynamically, the clusters' subjects are evaluated compared to the current static signature system.

The interviews were structured into several parts. The questions and the interview structure can be seen in Figure 5.5. The answers given by the participants can be seen in Figure 5.6 and the most relevant comments can be found in appendix B.

Most participants could immediately link book clusters to specific research when looking at the word clouds and time of borrowing activity. The longer each participant has been present at the institute, the further back they were able to recognise projects. One participant in particular was able to link almost all projects among the first 30 clusters to one or two scholars and projects. In particular, the clusters in the lower left part of the mapping were easily identifiable, as they seem to result from the borrowing activity of a few scholars. Clusters with broader subject descriptions and the clusters in the upper right part were more difficult to identify, as they appear to contain books borrowed by many researchers working on various projects.

The interviewees agreed that the subject classification and description of the book clusters are too vague and broad to represent the research projects accurately. According to one of the interviewed scholars, for the subject classification to become usable, it should contain the date, geography, and medium. Others heavily criticised the fact that some of the subject classifications indicated different time frames, such as *Renaissance* and *Fascism* together.

The descriptions were considered unclear and imprecise but still somewhat connected to the research projects. One comment suggested that the list of book titles and the corresponding word clouds are insightful, but the researchers themselves

should do the classification and description in order for them to be meaningful.

Most participants agreed that the cluster atlas and the mapping represent the research in the departments and the institute. The different research focuses of the departments and the developments over recent years are represented in the clusters. However, one participant stated that their focus in research, contemporary art, was underrepresented in the clusters. A possible explanation is that few researchers focused on contemporary art at the institute until a few years ago. The same scholar implied that they find most of the literature relevant for their projects in other libraries instead of the library at Bibliotheca Hertziana because the collection contains only a few books on contemporary art. This was underscored by one head of department stating that the acquisition focus of the library excludes certain topics intentionally, because other libraries already have established collections. This indicates the need to integrate other library collections and online resources to fully understand the research conducted at Bibliotheca Hertziana: a requirement that, from the perspective of an ecosystem of interoperable research libraries, is actually desirable.

As a map of art history, the cluster atlas shows certain developments in the field. The interviewees mostly agreed that the mapping and cluster atlas show that the field has changed over recent years and that some topics are more fashionable than others. One participant complained that the subject classification and description lack a scientific methodology, which is a crucial part of art historical research. Unfortunately, methodological descriptions are rarely contained in the book titles. Similarly, one participant stated that the mapping only shows the titles of the books consulted but not how, why, and from which perspective the books were read.

As for the mapping's usefulness, the participants appreciated the potential for a recommendation system and bibliography-sharing tool. They expressed the wish for an interactive version of the mapping to explore the clusters further. They clearly enjoyed going through the cluster atlas and finding past and present research projects of colleagues and friends. They agreed that the descriptions and classifications were too vague to be useful for their research. Still, they valued the prospect of having a dynamic display of the research at Bibliotheca Hertziana. They saw it as a tool for exploring the library collection and current research.

The interviews not only provided a validation of the mapping, but also helped as-

sign clusters to research projects. The result of harnessing the collective knowledge of the scholars is shown in Figure 5.7. Many clusters in the lower left part of the mapping were easily assigned to research projects or scholars. The larger cluster centre and the top right of the mapping were vaguer and broader in content and most likely represent the canon of art historical literature consulted at Bibliotheca Hertziana. This aligns with the borrowing density shown in Figure 4.14, which shows continuous high activity in those areas.

	Questions
Individuals	1. Can you recognise any research projects?
	2. Does the subject classification represent the research?
	3. Does the description represent the research?
Institute	4. Can you assign clusters to department projects?
	5. Is your department represented?
	6. Does the mapping represent Bibliotheca Hertziana?
Field	7. Can you recognise fields of art history?
	8. Does the mapping show developments in the field?
	9. Does the subject classification represent these developments?
Usage	10. Is the mapping useful to you?
	11. Do you see it as a tool to explore the library?
	12. Is it useful for a generic public?

Figure 5.5: The interviews were divided into four parts. The first three evaluated possible readings of the mapping, while the last assessed its usefulness.

	Questions	H1	H2	H3	A1	A2	A3	A4	A5	A6
Individuals	1. Can you recognise any research projects?	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2. Does the subject classification represent the research?	✗*	✓*	✓	✗*	✓	✓	✗	✓	✗
	3. Does the description represent the research?	✓	✓*	✓	✓	✓	✓	✗*	✗	✓
Institute	4. Can you assign clusters to department projects?	✓	✓	✓	✓	✓	✓	✓	✓	✓
	5. Is your department represented?	✓	✓	✓	✓	✓	✓	✓	✓	✓
	6. Does the mapping represent Bibliotheca Hertziana?	✓	✓	✓	✓	✓	✗*	✓	✓	✓*
Field	7. Can you recognise fields of art history?	✓	✓	✓	✓	✗*	✗	✓	✓	✓
	8. Does the mapping show developments in the field?	✗*	✓	✓	✓*	✓	✓	✓	✓	✓
	9. Does the subject classification represent these developments?	✗	✓	✓	✗	✓	✓	✗	✗	✗
Usage	10. Is the mapping useful to you?	✓*	✗*	✓*	✓	✗*	✗	✓	✓	✓
	11. Do you see it as a tool to explore the library?	✓*	✓*	✓	✓	✓	✓	✓*	✓*	✗
	12. Is it useful for a generic public?	✓	✓	✗	✓*	✗	✓	✓*	✓	✗*

Figure 5.6: The results of the interviews conducted at Bibliotheca Hertziana. The asterisks behind the answers indicate that a relevant comment was given for this question, which can be found in the appendix B.

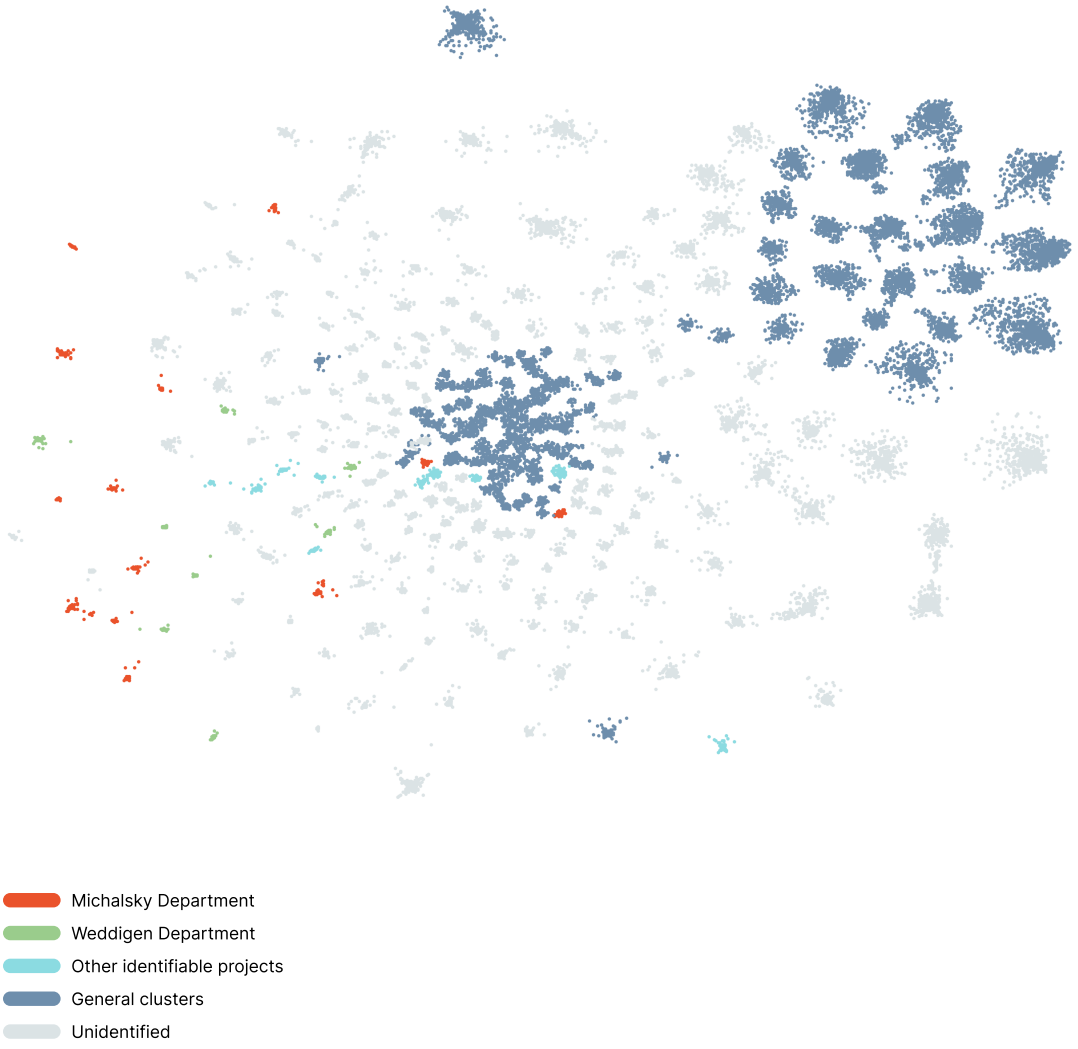


Figure 5.7: The mapping of the internal research community with clusters assigned to the institute's departments by members of the Bibliotheca Hertziana community. The clusters were identified using the cluster atlas in the interviews. The *general clusters* denote clusters with content reflecting Bibliotheca Hertziana's main collection focus.

Chapter 6

Conclusion

Mapping Bibliotheca Hertziana is a graphical representation that lays the groundwork for a dynamic display of library collections. It uses deep mapping techniques and the prompting of large language models to create a dynamic classification system. Analysis of patterns in research, exploration of the history and context of Bibliotheca Hertziana, and leverage of user loan data together provide a framework for mapping academic collections and visualising scientific fields.

Throughout the project, it has become clear that static classification, such as the signature system in use at Bibliotheca Hertziana, is inadequate for evolving collections. Although classification systems can be designed with users of existing collections in mind, the increasing digitisation and accessibility of documents have created the need for more flexible systems. The library holdings only partially reflect developments in art history. A dynamic library display can provide additional ways to interact with and discover the library collection. By leveraging user loan data, this project has established a workflow to harness the collective knowledge of the scholars present at Bibliotheca Hertziana. The user loan data provides the basis for an innovative way to measure the similarity between books and collocate them intuitively. The use of dimensionality reduction and clustering to construct sets of books has been demonstrated as a flexible tool to create the basis of a potential recommendation system. Furthermore, prompting large language models using TF-IDF scores of words in book titles has been shown to have great potential in generating a dynamic subject classification and description where only limited textual data is available.

The flexible framework developed in this project, based on the limited availability of bibliographical data, provides the basis for establishing a recommendation system, offering further paths of discovery for researchers interacting with library collections. While this is a case study of a library with users who are experts in their field, it can be applied to other libraries focusing on other fields. The scholars at the institute conduct highly specialised research, a fact which is undoubtedly reflected in their usage patterns. The applicability of this framework to public or general knowledge libraries will be fundamental to any future work related to this project.

As a map of science, *Mapping Bibliotheca Hertziana* offers only a partial view of contemporary research in the field of art history. Based on the loan activity for users present at the institute, the map reflects only the research conducted there. Integrating additional art historical libraries and data from other catalogues would complete the map of art history and offer a more complete display of the field. Incorporating bibliographical data from libraries in other fields could give rise to displays of entire research organisations, such as the *Max-Planck Society*.

Furthermore, the mapping currently contains only the books that have been loaned at least once in the past ten years and excludes books that were not assigned to any cluster. Integrating the entire library collection would be crucial to creating a representation of the library holdings. In a potential library display and recommendation system, these books can be imagined as points floating around the mapping, shooting across the screen to their designated location as soon as they have been loaned, thus offering an incentive for users to explore the unexplored parts of the collection. Similarly, the arrangement of book clusters could transform and change its shape as the collection does, mirroring the dynamics of the library. As scholars arrive at the institute and new projects are launched, they will be reflected in the display, offering a live view of the research.

The mapping was validated through interviews with selected members of the Bibliotheca Hertziana community, which showed that the subject classification and description are mostly too unspecific to be useful to scholars. The cluster atlas displaying the time frame of borrowing activity and word clouds based on the book titles proved more insightful and precise than the automatically generated cluster descriptions. Overall, the scholars interviewed confirmed that the mapping represents the institute and the field of art history. However, they also indicated the need for more precise subject classification, a dynamic and interactive display, and a recommendation system.

Such a display and recommendation system, designed with users' needs in mind, would represent a valuable new resource for academic researchers. Considering the sheer number of books and articles published each year and the fact that space in libraries such as the Bibliotheca Hertziana is running out, there is an evident need to recreate the affordances of the library shelf in a digital setting. Navigating the digital library should provide researchers with an experience similar to navigating the physical shelves in Bibliotheca Hertziana, which can be further customised to their needs. A dynamic library classification can provide the essential contextualisation that combing online catalogues for literature cannot. With fields in research changing rapidly, such flexible systems are becoming indispensable.

Mapping Bibliotheca Hertziana not only re-imagines the future of users' interaction with libraries and the way they explore academic collections but also provides an innovative approach to the dynamic classification and discovery of scholarly resources. In an age where the digitisation of knowledge and the development of academic disciplines demand innovative approaches to information management, this promises to transform how we visualise and navigate the library of the future.

Bibliography

Bayerische Staatsbibliothek. 2015. *SPARQL-Endpoint*. Accessed January 5, 2024. http://lod.b3kat.de/doc/en/sparql-endpoint.

Berba, Pepe. 2020. *Understanding HDBSCAN and Density-Based Clustering* [in en], January. Accessed November 22, 2023. https://pberba.github.io/stats/2020/01/17/hdbscan/.

Bibliotheca Hertziana. 2018. *Italy in a Global Context* [in en]. Accessed February 6, 2024. https://www.biblhertz.it/en/dept-weddigen/global-context.

———. 2019. *Social Reality in Italian Films* [in en]. Accessed February 12, 2024. https://www.biblhertz.it/en/dept-michalsky/italian-films.

Börner, Katy. 2015. *Atlas of knowledge: anyone can map* [in English]. Cambridge, MA: MIT Press.

Börner, Katy, Chaomei Chen, and Kevin W. Boyack. 2003. "Visualizing knowledge domains" [in en]. *Annual Review of Information Science and Technology* 37 (1): 179–255. Accessed January 26, 2024. https://onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370106.

Bowker, Geoffrey C., and Susan Leigh Star. 2008. *Sorting things out: classification and its consequences* [in eng]. 1. paperback ed., 8. print. Inside technology. Cambridge, Mass.: MIT Press.

Boyack, Kevin W., Richard Klavans, and Katy Börner. 2005. "Mapping the backbone of science" [in en]. *Scientometrics* 64, no. 3 (August): 351–374. Accessed November 18, 2023. http://link.springer.com/10.1007/s11192-005-0255-6.

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander. 2013. "Density-Based Clustering Based on Hierarchical Density Estimates" [in en]. In *Advances in Knowledge Discovery and Data Mining,* edited by Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, 160–172. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer.

Casey, Hannah. 2023. *hanaCasey/BHVizData*. Original-date: 2023-03-07T14:35:55Z. Accessed February 20, 2024. https://github.com/hanaCasey/BHVizData.

Casey, Hannah Laureen. 2024. "Mapping Bibliotheca Hertziana - Cluster Atlas." Publisher: Zenodo (February). Accessed February 24, 2024. https://zenodo.org/records/10700617.

Chari, Tara, and Lior Pachter. 2023. "The specious art of single-cell genomics" [in en], edited by Jason A. Papin. *PLOS Computational Biology* 19, no. 8 (August): e1011288. Accessed December 13, 2023. https://dx.plos.org/10.1371/journal.pcbi.1011288.

Chen, Chaomei. 2017. "Science Mapping: A Systematic Review of the Literature" [in en]. *Journal of Data and Information Science* 2, no. 2 (March): 1–40. Accessed January 26, 2024. https://www.sciendo.com/article/10.1515/jdis-2017-0006.

Christopher Pietsch and UCLAB. 2014. *FW4 Visualisierung*. Accessed November 23, 2023. https://uclab.fh-potsdam.de/fw4/vis/.

Deutsche National Bibliothek. 2022. *The Integrated Authority File (GND).* Accessed January 10, 2024. https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html.

Drucker, Johanna. 2014. *Graphesis: visual forms of knowledge production*. MetaLABprojects. Cambridge, Massachusetts: Harvard University Press.

Ebert-Schifferer, Sybille. 2013. *Die Geschichte des Instituts 1913 - 2013* [in ger]. 100 Jahre Bibliotheca Hertziana : Max-Planck-Institut für Kunstgeschichte 1. München: Hirmer Verlag.

Enthought. 2018. *UMAP Uniform Manifold Approximation and Projection for Dimension Reduction | SciPy 2018 |*, July. Accessed December 13, 2023. https://www.youtube.com/watch?v=nq6iPZVUxZU.

Glinka, Katrin, Christopher Pietsch, Carsten Dilba, and Marian Dörk. 2016. "Linking structure, texture and context in a visualization of historical drawings by Frederick William IV (1795-1861)" [in en]. Number: 2, *International Journal for Digital Art History,* no. 2 (October). Accessed February 15, 2024. https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/33530.

Heart, F., A. McKenzie, J. McQuilian, and D. Walden. 1978. *Arpanet Completion Report*. Accessed January 25, 2024. https://web.archive.org/web/20230527095942/https://walden-family.com/bbn/arpanet-completion-report.pdf.

Kaplan, Frédéric. 2015. "A Map for Big Data Research in Digital Humanities" [in en]. *Frontiers in Digital Humanities* 2 (May). Accessed September 12, 2023. http://www.frontiersin.org/Digital_Humanities/10.3389/fdigh.2015.00001/full.

Kubikat and Ex Libris. 2020. *Kubikat - Einfache Suche*. Accessed February 15, 2024. https://aleph.mpg.de/F?func=file&file_name=find-b&local_base=kub01.

Library of Congress. 2022. *MARC 21 XML Schema*. Accessed January 5, 2024. https://www.loc.gov/standards/marcxml/.

Lima, Manuel. 2011. *Visual complexity: mapping patterns of information* [in eng]. New York: Princeton Architectural Press.

Maurer, Golo. 2021. *Heimreisen: Goethe, Italien und die Suche der Deutschen nach sich selbst*. Originalausgabe. OCLC: on1269197100. Hamburg: Rowohlt.

McInnes, Leland, John Healy, and Steve Astels. 2016. *How HDBSCAN Works — hdbscan 0.8.1 documentation*. Accessed November 22, 2023. https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." Publisher: arXiv Version Number: 3, accessed November 21, 2023. https://arxiv.org/abs/1802.03426.

McKinney, Wes. 2022. *Python for data analysis: data wrangling with Pandas, NumPy, and Jupyter* [in eng]. Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly.

Moreno, Jacob L. 1934. *Who shall survive?* Washington: Nervous / Mental Disease Publishing Co.

NLTK Project. 2023. *NLTK: Natural Language Toolkit*. Accessed January 24, 2024. https://www.nltk.org/.

Noichl, Maximilian. 2023. “How localized are computational templates? A machine learning approach” [in en]. *Synthese* 201, no. 3 (March): 107. Accessed September 12, 2023. https://doi.org/10.1007/s11229-023-04057-x.

NumFOCUS, Inc. 2024. *pandas - Python Data Analysis Library*. Accessed January 5, 2024. https://pandas.pydata.org/.

OpenAI. 2020. *OpenAI API*. Accessed December 21, 2023. https://openai.com/blog/openai-api.

———. 2023. *Pricing* [in en-US]. Accessed January 24, 2024. https://openai.com/pricing.

Petrovich, Eugenio. 2020. “Science mapping.” Edited by Birger Hjørland and Claudio Gnoli. *Encyclopedia of knowledge organization,* https://www.isko.org/cyclo/science_mapping.

Picca, Davide, Antonin Schnyder, Eri Kostina, Alessandro Adamou, Dario Rodighiero, and Jeffrey Schnapp. 2023. “Orchestrating Cultural Heritage: Exploring the Automated Analysis and Organization of Charles S. Peirce’s PAP Manuscript” [in en]. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media,* 1–4. Rome Italy: ACM, September. Accessed January 9, 2024. https://dl.acm.org/doi/10.1145/3603163.3609066.

Pietsch, Christopher. 2020a. *cpietsch/smb-vis*. Original-date: 2020-08-29T17:04:11Z. Accessed February 15, 2024. https://github.com/cpietsch/smb-vis.

———. 2020b. *Visuelle Exploration zweier musealer Sammlungen* [in en]. Accessed December 10, 2023. https://cpietsch.github.io/smb-vis.

Ranganathan, Shiyali Ramamrita, Pazhamaneri Sundaram Sivaswamy Aiyer, and William Charles Berwick Sayers. 2006. *The five laws of library science* [in eng]. Facsimile ed. Ranganathan series in library science 12. New Delhi [India] Bangalore, India: Ess Ess Publications Published for Sarada Ranganathan Endowment for Library Science.

Rischbieter, Julia Laura. 2004. *Henriette Hertz: Mäzenin und Gründerin der Bibliotheca Hertziana in Rom*. Pallas Athene, Bd. 14. OCLC: ocm57541494. Stuttgart: Steiner.

RNDR. 2022. *Oracle* [in en]. Accessed December 10, 2023. https://rndr.studio/projects/oracle.

Rodighiero, Dario. 2021. *Mapping Affinities: Democratizing Data Visualization* [in en_US]. Accepted: 2021-06-28T15:35:34Z. Métis Presses. Accessed September 15, 2023. https://dash.harvard.edu/handle/1/37368046.

Schmidt, Benjamin. 2018. "Stable random projection: lightweight, general-purpose dimensionality reduction for digitized libraries." *Journal of Cultural Analytics*.

Schnapp, Jeffrey T., and Matthew Battles. 2014. *The library beyond the book*. metaLABprojects. Harvard University Press.

Snydman, Stuart, Robert Sanderson, and Tom Cramer. 2015. "The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images." *Archiving Conference* 2015 (May).

Sparck Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." Publisher: MCB UP Ltd, *Journal of Documentation* 28, no. 1 (January): 11–21. Accessed November 18, 2023. https://doi.org/10.1108/eb026526.

Stafford, Barbara Maria. 2012. "Reconceiving the Warburg library as a working museum of the mind." *Common Knowledge* 18 (1): 180–187.

Stanislav, Golovanov. 2021. *pdfkit: Wkhtmltopdf python wrapper to convert html to pdf using the webkit rendering engine and qt*.

Svenonius, Elaine. 2000. *The intellectual foundation of information organization*. Digital libraries and electronic publishing. Cambridge, MA ; London: MIT Press.

Tesche, Doreen. 2002. *Ernst Steinmann und die Gründungsgeschichte der Bibliotheca Hertziana in Rom*. Römische Studien der Bibliotheca Hertziana, Bd. 15. OCLC: ocm51770374. München: Hirmer.

The Opte Project. 2003. *The Internet* [in en-US]. Accessed January 25, 2024. https://www.opte.org/the-internet.

Tufte, Edward R., ed. 2013. *Envisioning information* [in eng]. 14. print. Cheshire, Conn: Graphics Press.

Wilders, Coen. 2017. "Predicting the Role of Library Bookshelves in 2025" [in en]. *The Journal of Academic Librarianship* 43, no. 5 (September): 384–391. Accessed March 12, 2022. https://linkinghub.elsevier.com/retrieve/pii/S0099133317301234.

Appendix A

OpenAI prompts

Listing A.1: Prompts used to query GPT-4

```
def get_title_from_frequencies(frequencies):
    try:
        completion = client.chat.completions.create(
        model="gpt-4",
        messages=[
            {"role": "user", "content": f"Based on the following
               ↪ terms and tf-idf scores, suggest a subject
               ↪ classification (up to 10 words) for the
               ↪ associated cluster of books. Give me the response
               ↪  in format: Subject: subject"},
            {
                "role": "user",
                "content": f"These are the terms and frequencies: {
                   ↪ frequencies}"
            }
        ]
        )

        return completion.choices[0].message.content

    except Exception as e:
```

```
        print(f"An error occurred: {e}")
        return None

def get_description_from_titles(cluster_title, titles):
    try:
        completion = client.chat.completions.create(
        model="gpt-4",
        messages=[
            {"role": "user", "content": f"Based on the following
               ↪ cluster subject and book titles, suggest a
               ↪ description (up to 100 words) for the associated
               ↪ cluster of books. Give me the response in format:
               ↪  Description: description"},
            {
                "role": "user",
                "content": f"This is the cluster subject : {
                   ↪ cluster_title}"
            },
            {
                "role": "user",
                "content": f"These are the book titles : {titles}"
            }
        ]
        )

        return completion.choices[0].message.content

    except Exception as e:
        print(f"An error occurred: {e}")
        return None
```

Appendix B

Interviews

	Comments
H1-2	The subject classification shows themes but not research.
H1-8	The questions asked by scholars evolve and change, but the book titles remain the same, so they don't reflect any developments in research.
H1-10	The mapping is too predictable and doesn't help in getting an overview.
H1-11	Prefers exploring the physical shelves and that the books are arranged according to 'good neighbourliness'.
H2-2	Subjects are very broad, they show the field but not the research questions.
H2-3	Descriptions add together the titles but don't connect the content.
H2-10	At the current state the mapping is too abstract and requires some meta-analysis.
H2-11	The mapping should become interactive.
H3-10	In retrospect the mapping shows the developments at the institute. It also shows what kind of books the fellows want and need.
A1-2	The subjects and descriptions mostly represent the research but are not very precise.
A1-8	The cluster atlas shows that the field changed and when scholars arrived at the institute.
A1-12	The usefulness depends on the mapping's visual material and if it's intuitive to use.
A2-7	The methods used in art history are missing.
A2-10	A recommendation system would make the mapping very useful.
A3-6	The atlas does not show which projects from the institute are not represented in the mapping.
A4-3	The descriptions are evidently AI generated and almost nonsensical.
A4-11	The mapping is a way to explore the research through the library collection. It could be seen as a dynamic bibliography sharing system.
A4-12	The mapping could be very useful if subject and description were assigned by researchers themselves.
A5-11	The classification is too broad and too specific at the same time.
A6-6	You can clearly see the shift in research focus and change between different directors of the institute.
A6-12	To read the mapping currently you need institutional memory of Bibliotheca Hertziana.