NOTA BENE

The Hermeneutics of Bibliographic Data and Cultural Metadata

Edited by
Jens-Morten Hanssen and Sissel Furuseth

19

STUDIES FROM THE NATIONAL LIBRARY OF NORWAY

The Hermeneutics of Bibliographic Data and Cultural Metadata

The Hermeneutics of Bibliographic Data and Cultural Metadata

Edited by
Jens-Morten Hanssen
and Sissel Furuseth

NATIONAL LIBRARY OF NORWAY, OSLO 2025

Contents

Preface 9

1. **Introduction: The Crossroads of Bibliography and Digital Humanities** 11
Jens-Morten Hanssen and Sissel Furuseth

2. **Book Printing in Latin and Vernacular Languages in Northern Europe, 1500–1800** 27
Jani Marjanen, Tuuli Tahko, Leo Lahti and Mikko Tolonen

3. **The Rise of the Novel in Norway: Bibliographical Perspectives** 67
Jens-Morten Hanssen

4. **Corpus and the Bibliography: NB DH-LAB as an Infrastructure for Text and Metadata Mining** 96
Magnus Breder Birkenes and Lars G. Johnsen

5. **Getting Meaning out of Metadata – Analysis of Selected Bibliographies at the National Library of Norway** 118
Oddrun Pauline Ohren

6. **Bibliographic Needs in Literary Critical Reception and Periodical Studies: The Database *Norsk Litteraturkritikk* as a Case in Point** 149
Sissel Furuseth

7. **The Use of Bibliographic and Cultural Metadata – How to Investigate Users' Information Search Behaviour** 175
Nils Pharo and Pia Borlund

8. **Subject Matters: Metadata Standards and Subject Access for Library and Museum Catalogues** 204
Ahmad M. Kamal and Koraljka Golub

Contributors 240

Preface

The initiative that led to the publication of the present volume dates back to 2020. The National Library of Norway had already been experimenting with the use of digital tools and computational techniques within the framework of the Digital Humanities Laboratory for some time, in step with the long ongoing process of digitising library contents of all types. It saw the need for a focused and systematic investigation into one of its core activities, namely bibliographic work. How and to what extent has digitisation altered the field of bibliography and the production of library metadata? What are the current possibilities for conducting research on bibliographic data and cultural metadata? On 28 October 2020, in the middle of the Covid-19 pandemic, an international webinar dedicated to this topic was organised by the National Library of Norway. The webinar kick-started the process that led to this book. The chapters written by Hanssen, Ohren, Furuseth, Pharo and Borlund are revised versions of papers initially presented during that webinar, whereas the chapters written by Jani Marjanen, Tuuli Tahko, Leo Lahti, Mikko Tolonen, Magnus Breder Birkenes, Lars G. Johnsen, Ahmad M. Kamal and Koraljka Golub have been specifically written for this book.

Piecing together a volume such as this one is a laborious collaborative process, involving hard work by many people. We would like to take this opportunity to express our deepest gratitude to a few of those who have contributed in significant ways to the end result: Sofie Arneberg for undertaking the painstaking job of editorial assistant; Trond Aalberg for his expertise and knowledgeable contributions; the

authors for their patience and dedication; and the anonymous peer reviewers for their comments and criticisms that led to improved chapters. The staff of the National Library mourn the loss of the former head of the editorial council of the Notabene series, Jon Arild Olsen, who sadly passed away on 16 March 2024. We would like to cherish his memory by thanking him posthumously for his guidance and strong support in a critical phase of the book project.

Oslo, 12 August 2024
Jens-Morten Hanssen and Sissel Furuseth

1. Introduction: The Crossroads of Bibliography and Digital Humanities

Jens-Morten Hanssen and Sissel Furuseth

Technological shifts and the development of new media have the capability to foster visions of a brighter future for mankind, with universal access to knowledge, scientific progress and human advancement. In his *Traité de Documentation*, summing up four decades of work on bibliography and information processing, Paul Otlet (1934) envisaged the realisation of a new kind of encyclopaedia, the Universal Book. Derived from all existing books, it would be the ultimate desideratum of documentation. In 1895, Otlet and his senior colleague, Henri La Fontaine, had set up the International Institute of Bibliography in Brussels. By 1930, the Institute's monumental Universal Bibliographic Repertory contained nearly 16 million bibliographic entries. Otlet envisioned a distributed system of work stations connected to a central hub by telephone, wireless telegraphy, television and telex. He hoped that the "Universal Book formed from all books would have become very approximately an annexe to the brain, the substratum of memory, an exterior mechanism and instrument of the mind" (Otlet 1934, 428; quoted in Rayward 1997, 298).

At the very end of the Second World War, Otlet's line of thinking found an echo in Vannevar Bush's concept of the "memex", which he developed in his famous *Atlantic Monthly* article "As We May Think" (1945). He imagined an electromechanical device for interacting with microfilm documents, enabling individuals to develop their

own private, self-contained library. This library would store all a person's "books, records, and communications", and would be "mechanized so that it may be consulted with exceeding speed and flexibility". The memex would thus constitute "an enlarged intimate supplement to [the owner's] memory" (Bush 1945, 106).

Each in their own way, Otlet and Bush have received plaudits for having foreshadowed hypertext technology, the invention of the internet and, eventually, the World Wide Web. However, they have more in common than that. They were both optimistic on behalf of the development of modern media and technology. Otlet had a sharp eye for the enormous potential for disseminating information created by new forms of media and communication, such as the telegraph and telephone, radio, television, cinema and phonographic records (Rayward 1997, 296). By the time of writing his article, Bush was the director of the Office of Scientific Research and Development in the United States, and he was a strong advocate of international scientific openness and information sharing, although that was hard during times of war and the nuclear arms race. Furthermore, the two men seem very much in agreement that technology and the growing knowledge industry should work in the service of humanity, with a particular concern for promoting progress in the field of knowledge production through scientific, scholarly and intellectual work.[1]

Less well known is the fact that Otlet and Bush were also united in their dissatisfaction with the current state of affairs in library systems. "Even the modern great library is not generally consulted; it is nibbled at by a few," Bush observed (Bush 1945, 103; see also Aalberg 2003, 15). Otlet's biographer W. Boyd Rayward notes that Otlet was critical of libraries, "because of the way in which they restricted access, resisted technical innovation, including new methods of classification and cataloguing, and were conservative in their approach to information service" (Rayward 1997, 295).

1 We are grateful to Elise Conradi for referring to Paul Otlet and Vannevar Bush in a paper she presented at a webinar at the National Library of Norway on 28 October 2020.

The present volume aims to explore the crossroads between two very diverse fields of practice, bibliography and the use of computation in the humanities and social sciences, the one encompassing a period of two millennia, the other introduced only in the second half of the twentieth century. In retrospect, Otlet and Bush appear as transitional figures, articulating some of the shortcomings of the systems in currency during their lifetimes and vaguely pointing towards better ways to manage knowledge and organise the information society of the future. Neither lived long enough to experience the major achievements of the computer age. Both are, however, rightly considered pioneers of information science, an academic field which is key to understanding what bibliography has become today.

The Origin of Bibliography

The history of bibliography is fuzzy. The word is a compound of the two Ancient Greek words for "book" (βιβλίον) and "writing" (γραφία). So, etymologically, bibliography means "the writing of books", and the word was used by Greek writers in the first three centuries CE to mean the copying of books by hand (Onions 1966, 93; Blum 1969, 1022). Gutenberg made this definition obsolete, and since his invention of the printing press, bibliography has come to mean a variety of things, none of them matching the Ancient Greek sense.

Nonetheless, there is a tradition for tracing the origin of bibliography back to Callimachus (circa 305–240 BCE) who was a poet, scholar and member of the Museion in Alexandria, a royal academy of arts and sciences founded by Ptolemy I. During his tenure at the Library of Alexandria, generally regarded as the Museion's most important institution, Callimachus compiled lists of Greek authors and their works by classes, according to literary forms and by scholarly disciplines. Although not preserved and known only from secondary sources, these lists are known throughout the history of classical scholarship as the *Pinakes*, comprising 120 books, or tables (from Ancient Greek Πίνακες) (Pfeiffer 1968; Blum 1991). The *Pinakes* were based on the books held at the Great Library of Alexandria, which in some respects was a forerunner of the modern national library. For example,

it collected not only works published in the territory of Hellenistic Greece, but all works written in the Greek language wherever they were published, including translations from foreign languages. Callimachus divided the entire body of Greek literature into at least eleven classes: rhetoric, law, epic, lyric poetry, tragedy, comedy, philosophy, history, medicine, mathematics and miscellaneous works. The individual authors of every class were arranged in alphabetical order, and each name was accompanied by a few biographical details (Blum 1991, 124–181).

Callimachus created a model for organising knowledge of a nation's literature that was copied for centuries to come. Rudolf Blum notes that it was not until the seventeenth century that the meaning of bibliography shifted from *writing* books to *describing* them (1969, 1022–1023). With the invention of the moveable-type printing press, book production became industrialised; the service of scribes and scriveners was no longer in demand, and the book turned into a mass medium which functioned as a vehicle for the Enlightenment. Bibliography came to mean primarily two things: first, a list of works of a specific author or publisher, or works on a specific subject, published separately as a book or printed as an appendix to a scholarly work; and second, the history or systematic description of books, their authorship, printing, publication, editions etc.

The Diversification of Bibliography

The modern era saw an increasing diversification and specialisation of bibliography, even leading to a lively discussion about the scientific nature of bibliographic work (Tanselle 1974). The field developed alongside disciplines such as classical studies, literary criticism and textual scholarship, as well as areas of study known today as book history, and library and information science. The American bibliographer Fredson Bowers distinguished between no less than five types of bibliography. The most common of these is enumerative or compilative bibliography, which deals mainly with the compiling of lists or indexes.

Historical bibliography constitutes a second type. This is concerned with enquiries into "the evolution of printing (including

type-founding and paper-making), binding, book ownership, and book-selling" (1952, 190).

Bowers singles out analytical bibliography as a third main category, defining it as "technical investigation of the printing of specific books, or of general printing practice, based exclusively on the physical evidence of the books themselves" (1952, 191). In its pure form, analytical bibliography leads directly to the two last categories: descriptive bibliography and critical or textual bibliography.

Descriptive bibliography is concerned with the examination of a book by all the methods of analytical bibliography in order to arrive at a detailed physical description of it, its external appearance and the evidence bearing on the details of this external appearance.

Lastly, Bowers defines critical or textual bibliography as "the application of the evidence of analytical bibliography, (…) to textual problems where meaning of some sort is involved" (1952, 194–195). Ohren discusses twentieth century conceptions of bibliography, including Bowers' categorisation, at greater detail in Chapter 5.

Bowers' typology suggests a highly specialised and advanced field. But as a matter of fact, he mainly describes bibliography from the viewpoint of textual scholarship and literary studies. Considering bibliography from a library perspective adds even more complexity to the matter. On a visit to England in 1895, just as he and La Fontaine were in the process of establishing the International Institute and Office of Bibliography, Otlet became acquainted with Melvil Dewey's Decimal Classification, a system for organising library materials by subject, based on the division of all knowledge into ten main groups, with each group assigned 100 numbers. The Universal Bibliographic Repertory, that is the great card catalogue which was initiated by Otlet and La Fontaine and assembled within their Office of Bibliography, would be arranged in a classified order according to an elaborated version of the Dewey Decimal Classification (Rayward 1990, 25–70; Rayward 1997, 291).

In the era of big data, it is safe to say that size matters. But long before the on-going flood of bits and bytes, the nineteenth century saw a substantial proliferation of printed matter. This was driven by technological advances such as the introduction of paper-making machines

around 1800 and, later, the transition from the hand-operated press to steam-powered and, ultimately, to automated printing presses. During the nineteenth and twentieth centuries, libraries adapted to expanded needs. "Once temples for a clerisy, once scarce and inaccessible places of refuge," as Jeffrey T. Schnapp and Matthew Battles (2014, 19) observe, libraries underwent a wholesale democratisation through a process driven by the spread of literacy, the industrialisation of printing, and the explosion in printed records of ever more heterogenous kinds, whereby the library became what it remains today: "a vast storage container for printed matter, organized around a standardized classification system, to ensure access, flanked by some ancillary support structures (a reference desk, a reading room, an exhibition area, a cataloging and processing backroom)" (Schnapp and Battles 2014, 19–20). In the further course of events, Schnapp and Battles go on, the information age "transformed stack-centric libraries into data centers filled not only with books but also with workstation clusters, the deadly digital doubles of analog reading rooms", and in came "a flood of access to new information and, with it, doubts about where a library begins and where it ends" (Schnapp and Battles 2014, 20).

Computer Science and Library Automation

The overall approach to bibliography and metadata as discussed in this anthology rests on three interconnected premises. First, that the development of library systems over the last hundred years or so was strongly marked by the advent of information and computer science – and by library systems we are referring to the whole range of procedures and set-ups for classification, cataloguing, indexing, collection management, storing, access, search and retrieval. Second, that bibliographical work and metadata registry as a core activity within the world of libraries is permeated with computation and data technology. And third, that this development consequently opens up avenues for computational approaches and data-driven research.

The rise of digitisation in the 1990s, propelled by the advent of World Wide Web, has left a significant mark on academic research and scholarship, particularly on the humanities. Digital Humanities

became a buzzword around the turn of the millennium, and digital approaches proliferated rapidly, affecting virtually all areas of the humanities. Broadly speaking, the history of computational approaches in the humanities falls in two phases: the Humanities Computing phase in the period until around 2000, and the Digital Humanities phase after that. Commenting on the transition from the former to the latter, N. Katherine Hayles (2014) stated that Humanities Computing was felt to be "too closely associated with computing support services" and that the new term, Digital Humanities, was meant "to signal that the field had emerged from the low-prestige status of a support service into a genuinely intellectual endeavor with its own professional practices, rigorous standards, and exciting theoretical explorations" (Hayles 2012, 24). In hindsight, it becomes clear that there are interesting parallels between Humanities Computing and the initial development of computer-based automation in the library sector, with card technology as a common denominator.

The roots of computational work in the humanities stretch back to the end of the 1940s when the Italian Jesuit priest Roberto Busa created a computer-generated concordance to the writings of Thomas Aquinas, resulting in his vast *Index Thomisticus* (Jones 2016; Burdick et al. 2012, 123). Busa, today heralded as one of the forefathers of the digital humanities, used IBM's punch card technology when developing the concordance. Prior to that, Paul Otlet had adopted the standard 3" x 5" card used in the United States and based the Mundaneum's Universal Bibliographic Repertory on a technology of card and cabinet (Rayward 1997). In the late 1950s, the Library of Congress began to investigate the possibility of using automated techniques for library operations such as cataloguing, searching, indexing and document retrieval. In 1965, the Library of Congress launched a project to convert its library cards into machine-readable form. The outcome was the Machine-Readable Cataloguing Record (MARC) (Avram 1975). For many decades to come, MARC was used as the basis for library automation, standardisation and bibliographic communication across the globe.

Throughout the 1980s and early 1990s, libraries everywhere

made great efforts to convert paper-based library cards into digital form, implementing library automation software such as Dynix or Aleph. The World Wide Web made it possible to share digital content on a previously unprecedented scale over the internet. Around the turn of the millennium, large-scale projects like the Internet Archive, Gallica, Google Books, HathiTrust and Europeana endeavoured to provide online access to the cultural record at large. By then, however, library systems, with their catalogues, authority records and bibliographies, had long since made the digital leap. The analogue-to-digital conversion of library catalogues and the trend of dropping enumerative bibliographies in printed form in favour of online bibliographic databases mostly belong to the pre-2000 phase of Humanities Computing.

The massive, worldwide migration of cultural content to digital formats required the creation of new and updated data standards, to ensure interoperability and access across disparate datasets. This was even more imperative in the field of metadata, as metadata were increasingly perceived as key to the functionality of most information systems. From now on, cultural heritage institutions, such as libraries, archives and museums, either joined forces with or leaned heavily upon the web technology industry in developing metadata standards: the Dublin Core for digital resources, the Metadata Object Description Schema (MODS) for bibliographic material, CDWA for works of art, and the VRA Core for visual resources (Gilliland 2008; Riley 2017; Miller 2022).[2]

Research on Bibliographic Data

Today, research on bibliographic data and cultural metadata is a dynamic and versatile field, spanning a variety of approaches and methodologies. We will here distinguish between four main types of data-driven or data-informed bibliographic research.

The focal point of the first type is traditional library metadata in

2 Kamal and Golub provide an in-depth discussion of metadata schemas and standards in Chapter 8.

digital form. For some time now, every library (national, academic, public or other) has maintained some form of open access public catalogue. Many of them also provide open access to their bibliographic databases or enable harvesting through initiatives such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), creating new avenues for research on library metadata. Researchers affiliated with the Helsinki Computational History Group have published several studies demonstrating a new and innovative take on bibliographic data. Extracting data from four bibliographies, the Finnish National Bibliography, the Swedish National Bibliography, the English Short Title Catalogue, and the Heritage of the Printed Book Database, a total of six million entries, Leo Lahti, Jani Marjanen, Hege Roivainen and Mikko Tolonen have examined the rise of the octavo format in printing in Europe and the breakthrough of vernacular languages in public discourse. In doing so, they address some well-known challenges in using bibliographic metadata collections, such as biases, gaps and inaccuracies, which they seek to overcome by specifically tailoring "open data analytical ecosystems that facilitate robust statistical research use of bibliographic collections" (Lahti et al. 2019, 15). Another study originating from the same multidisciplinary team of scholars treats the English Short Title Catalogue as a record of cultural production and applies a data-driven approach to constructing and examining the English canon during the period 1500–1800 (Tolonen et al. 2021).

National bibliographies and union catalogues such as the English Short Title Catalogue are the result of systematic work conducted by multiple generations of librarians, who have produced catalogue information and bibliographic entries according to established norms, mainly covering printed material. The advent of digital technology led to an explosion in digital publishing and a rapidly expanding information industry, and the second type of data-driven or data-informed bibliographic research deals primarily with born-digital material and reflects the fact that there is a strong (commercial) interest in metadata also from the private sector. A report prepared by the Bibliographic Data Working Group of the DARIAH-ERIC

consortium and submitted in 2022 refers to information services provided by corporations such as Elsevier and Clarivate, citation indexes and academic databases such as Scopus, Altmetric, Microsoft Academic and Google Scholar (Umerle 2022). A study by Martijn Visser, Nees Jan van Eck, and Ludo Waltman on scientific documents covered by Scopus, Web of Science, Dimensions, Crossref and Microsoft Academic, respectively, demonstrates some of the possibilities in terms of research on bibliographic data sources (2021).

The third type of data-driven or data-informed bibliographic research places bibliographic data within the larger context of information science and knowledge organisation. To a librarian, bibliographic work is first and foremost a practical matter. Information science opens the way for more theoretical considerations. What, for example, is the difference between knowledge and information? Ever since the days of Denis Diderot (1713–1784), co-founder of the ground-breaking *Encyclopédie*, the enlightened person has been obsessed with a desire of to map "each and every branch of human knowledge" (Diderot 2001, 283). But how do we classify knowledge, and which model offers the best classification system? The information science perspective also stimulates questions regarding information retrieval systems and indexing systems. An anthology, edited by Koraljka Golub and Ying-Hsang Liu, and published in 2022, provides an abundance of fascinating approaches to information studies and knowledge organisation from the perspective of Digital Humanities (2022).

Last but not least, recent developments in artificial intelligence and machine learning techniques have given rise to promising opportunities for automatic extraction of metadata. The manual production of metadata can be very time-consuming, requiring resources and infrastructure not readily accessible to all. Even the most well-equipped of libraries must make hard decisions regarding the level of bibliographic analysis. Mass digitisation has created the possibility of extracting or enriching metadata through computer algorithms, and the prospects seem particularly promising when it comes to newspapers, magazines, journals and so-called grey literature, material which normally escapes fine-grained analysis by bibliographic staff at

libraries. Ali et al. (2023) and Yang and Park (2018) are examples of studies demonstrating the use of AI-based research methods and machine learning approaches relevant to this book's subject matter.

Structure and Contents of the Volume

The present volume combines cross-disciplinary perspectives from library and information science, data science, book history, literary studies and intellectual history, and demonstrates how data-driven methods and research techniques can produce new insights. It delves into the history of bibliographic standards and examines bibliography from the point of view of its technical development. It analyses bibliographic systems from a user perspective and showcases bibliographic data analysis as a particularly illuminating example of data-rich research.

The book consists of seven chapters exploring the hermeneutics of bibliographic data and cultural metadata in different ways. In the first chapter, "Book Printing in Latin and Vernacular Languages in Northern Europe, 1500–1800", Jani Marjanen, Tuuli Tahko, Leo Lahti and Mikko Tolonen use bibliographic data to map the changing hegemonies of publishing languages in Britain, The Netherlands and Sweden in the age of hand-press printing. Although the process of vernacularisation has been discussed frequently by socio-linguists, historians and literary scholars over the years, changes in the use of vernacular languages have, until recently, been difficult to assess due to a lack of reliable data. In this chapter, however, the Finnish research team demonstrates how the Heritage of the Printed Book database can be used to produce a quantitative assessment of changing language relations in Northern Europe. The database is a compilation of 46 smaller, mostly national, bibliographies and consists of 2.1 million entries from the advent of the printed book up to 1828. But rather than providing national trajectories the team focuses particularly on university towns, capital cities and commercial centres in order to shed light on the different speeds of vernacularisation. By integrating data across multiple catalogues, the writers provide a broader and more precise analysis of the linguistic situation in Northern Europe in the early modern period than most previous studies.

In "The Rise of the Novel in Norway: Bibliographical Perspectives", Jens-Morten Hanssen takes a similar approach to the Norwegian book market in the eighteenth and nineteenth centuries, but focuses particularly on the distribution of novels. Hanssen has created a corpus of two thousand novels, including translated books. By applying a quantitative, bibliographical and transnational approach, he questions the commonly held view that Danish publishers dominated the distribution of books in Norway throughout the nineteenth century.

So what is the relationship between bibliography and corpus in a digital environment? In the third chapter, "Corpus and the Bibliography", Magnus Breder Birkenes and Lars G. Johnsen argue that the distinction between the two types of resources is not as clear-cut as it used to be. While bibliographies traditionally do not qualify as corpora, the authors show that bibliographic metadata may be analysed as corpus data. They also demonstrate the ways in which bibliographic data can assist corpus building. Birkenes and Johnsen support their argument by investigating titles and citations.

In "Getting Meaning out of Metadata", Oddrun Pauline Ohren analyses nine specialised bibliographies at the National Library of Norway and argues that bibliographic data are well suited for uncovering new patterns both within particular datasets and across datasets. Most of the bibliographies in Ohren's corpus are person-bibliographies devoted to celebrated canonical authors such as Camilla Collett, Bjørnstjerne Bjørnson and Knut Hamsun. However, she also includes specialised bibliographies related to historical events, epochs and Sámi culture. Ohren compares the bibliographies in terms of size, types of resources, overlap between bibliographies, gender balance, geographical locations, temporal development, genre representation and topics treated in the author bibliographies. She demonstrates how bibliographic data are well suited for painting the big picture. At the same time, Ohren emphasises the benefits of working with small datasets created within one organisation, because designing reliable analyses requires that researchers have some knowledge about the data themselves and insight into how the bibliographies have been created and managed over the years.

A similarly ambivalent approach to big data characterises Sissel Furuseth's chapter. In "Bibliographic Needs in Literary Critical Reception and Periodical Studies", she identifies some lacunas in conventional bibliographies and how such voids complicate certain types of print culture research. While Ohren, Hanssen, Birkenes and Johnsen demonstrate how to make use of digital resources maintained by the National Library of Norway, Furuseth presents a specialised bibliography owned by the University of Oslo and argues that the database *Norsk litteraturkritikk* (Norwegian Literary Criticism) is a valuable supplement to the digitised newspaper and periodical archives at the National Library. Providing metadata that are often lacking in standard library records, such as publication frequency, *Norsk litteraturkritikk* can be used to shed light on texts and writers that normally escape the attention of literary scholars. Furuseth presents both a historical background for the database and examples of how *Norsk litteraturkritikk* may be utilised in contemporary periodical studies.

The user-perspective is the main focus of the last two chapters in the book. In "The Use of Bibliographic and Cultural Metadata: How to Investigate Users' Information Search Behaviour", Nils Pharo and Pia Borlund demonstrate how interactive information retrieval studies can complement projects such as "*Norsk litteraturkritikks historie 1870–2000*" (The history of literary critical response in Norway, 1870–2000), and serve as inspiration for collaboration on research methods. Pharo and Borlund describe test designs and best practices in managing user studies. The ambition is to improve scholars' access to the collections of the National Library of Norway. Pharo and Borlund describe three hypothetical research designs, inspired by actual research projects identified via the Research Council of Norway's project bank database, that might provide important knowledge about how researchers interact with the systems, including how they query, browse and examine the material. Pharo and Borlund explore different methods for investigating searchers' information needs, analysing search strategies and identifying factors that influence both the success and failure of searches and how searchers interact with the collections.

In the book's final chapter, "Subject Matters: Metadata

Standards, Online Search, and Subject Access for Libraries and Museums", Ahmad M. Kamal and Koraljka Golub take a closer look at the descriptive tools that have been developed to enable libraries and museums to catalogue their collections by subject. They have noted that users today are able to search vaster and more varied collections than ever before, but receive hardly any help from metadata to optimise their search and discovery. Kamal and Golub explore these issues, provide background on metadata standards and their role in enabling subject access, review the historical marginalisation of these standards in online searching, summarise literature that demonstrates the persistent (and growing) need for subject representation for users. They then offer guidelines for improvement based on a literature review as well as several case studies. The chapter also takes into consideration the new role of linked data and its impact on the situation.

References

ALI, DILAWAR, KENZO MILLEVILLE, STEVEN VERSTOCKT, NICO VAN DE WEGHE, SALLY CHAMBERS AND JULIE M. BIRKHOLZ. 2023. "Computer Vision and Machine Learning Approaches for Metadata Enrichment to Improve Searchability of Historical Newspaper Collections." *Journal of Documentation* 27 Feb 2023. DOI: 10.1108/JD-01-2022-0029.

AVRAM, HENRIETTE D. 1975. *MARC; its History and Implications*. Washington, DC: Library of Congress.

BLUM, RUDOLF. 1969. *Bibliographia: Eine wort- und begriffsgeschichtliche Untersuchung*. Frankfurt a. M.: Buchhändler-Vereinigung.

BLUM, RUDOLF. 1991. *Kallimachos: The Alexandrian Library and the Origins of Bibliography*, translated by Hans H. Wellisch. Madison, WI: University of Wisconsin Press.

BOWERS, FREDSON. 1952. "Bibliography, Pure Bibliography, and Literary Studies." *The Papers of the Bibliographical Society of America* 46 (3): 186–208.

BURDICK, ANNE, JOHANNA DRUCKER, PETER LUNENFELD, TODD PRESNER AND JEFFREY SCHNAPP. 2012. *Digital_Humanities*. Cambridge, MA: MIT Press.

BUSH, VANNEVAR. 1945. "As We May Think." *The Atlantic Monthly* 176 (1): 101–108.

DIDEROT, DENIS. 2001. *Rameau's Nephew and Other Works*, translated by Jacques Barzun and Ralph H. Bowen. Indianapolis, IN: Hackett Publishing Company.

GILLILAND, ANNE J. 2008. "Setting the Stage." In *Introduction to Metadata*, edited by Murtha Baca, 1–19. Los Angeles, CA: Getty Research Institute.

GOLUB, KORALJKA, AND YING-HSANG LIU, EDS. 2022. *Information and Knowledge Organisation in Digital Humanities: Global Perspectives*. London: Routledge.

HAYLES, N. KATHERINE. 2012. *How We Think: Digital Media and Contemporary Technogenesis*. Chicago: University of Chicago Press.

JONES, STEVEN E. 2016. *Roberto Busa, S. J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. New York: Routledge.

LAHTI, LEO, JANI MARJANEN, HEGE ROIVAINEN AND MIKKO TOLONEN. 2019. "Bibliographic Data Science and the History of the Book (c. 1500–1800)." *Cataloging & Classification Quarterly*, DOI: 10.1080/01639374.2018.1543747.

MILLER, STEVEN JACK. 2022. *Metadata for Digital Collections: A How-To-Do-It Manual*. Chicago: ALA Neal-Schuman.

ONIONS, C. T., ED. 1966. *The Oxford Dictionary of English Etymology*. Oxford: Oxford University Press.

OTLET, PAUL. 1934. *Traité de Documentation: Le Livre sur le Livre – Théorie et Pratique*. Brussels: Editiones Mundaneum.

PFEIFFER, RUDOLF. 1968. *History of Classical Scholarship: From the Beginnings to the End of the Hellenistic Age*. London: Oxford University Press.

RAYWARD, W. BOYD, ED. 1990. *International Organisation and Dissemination of Knowledge: Selected Essays by Paul Otlet*. Translated and edited by W. Boyd Rayward. Amsterdam: Elsevier.

RAYWARD, W. BOYD. 1997. "The Origins of Information Science and the International Institute of Bibliography/International Federation for Information and Documentation (FID)." *Journal of the American Society for Information Science* 48 (4): 289–300.

RILEY, JENN. 2017. *Understanding Metadata: What Is Metadata, and What Is It For? A Primer*. Baltimore, MD: National Information Standards Organization. http://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf.

SCHNAPP, JEFFREY T., AND MATTHEW BATTLES. 2014. *The Library Beyond the Book*. Cambridge, MA: Harvard University Press.

TANSELLE, G. THOMAS. 1974. "Bibliography and Science." *Studies in Bibliography* 27: 55–89.

TOLONEN, MIKKO, MARK J. HILL, ALI ZEESHAN IJAZ, VILLE VAARA AND LEO LAHTI. 2021. "Examining the Early Modern Canon: *The English Short Title Catalogue* and Large-Scale Patterns of Cultural Production." In *Data Visualization in Enlightenment Literature and Culture*, edited by Ileana Baird, 63–119. London: Palgrave Macmillan.

UMERLE, TOMASZ (LEAD AUTHOR). 2022. *An Analysis of the Current Bibliographical Data Landscape in the Humanities*. Bibliographical Data Working Group/DARIAH-EU. DOI https://doi.org/10.5281/zenodo.6559857.

VISSER, MARTIJN, NEES JAN VAN ECK AND LUDO WALTMAN. 2021. "Large-scale Comparison of Bibliographic Data Sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic." *Quantitative Science Studies* 2 (1): 20–41.

YANG, GI-CHUL, AND JEONG-RAN PARK. 2018. "Automatic Extraction of Metadata Information for Library Collections." *International Journal of Advanced Culture Technology*. 6 (2): 117–122. DOI https://doi.org/10.17703/IJACT.2018.6.2.117.

AALBERG, TROND. 2003. "Supporting Relationships in Digital Libraries." Dr. Scient. diss., Norwegian University of Science and Technology.

2. Book Printing in Latin and Vernacular Languages in Northern Europe, 1500–1800

Jani Marjanen, Tuuli Tahko, Leo Lahti and Mikko Tolonen

Introduction

In this chapter, we use bibliographic data to trace the development of Latin and the vernacular languages of Northern Europe as languages of publication in the age of hand-press printing from the 1500s to 1800. The question of how language dynamics changed in conjunction with the expansion of print media in the early modern period is a complex one. Virtually all societies manifest some form of dynamic multilingualism, through which the functions of different languages and language varieties change over time. While all languages started out as spoken languages, many languages at various times have gained a standardised written form, some of them eventually becoming official or national languages vested with symbolic meaning (Millar 2005, 19; Cobarrubias 1983, 51). Historically, however, most areas and polities have first encountered and adopted writing in languages other than their locally spoken varieties. Sanskrit was the de facto written code in South Asia for centuries, and likewise Latin in Western Europe. Although these cosmopolitan written languages interacted with locally spoken languages, the vernacular languages of these areas only started to gain prominence as written languages alongside Sanskrit and Latin in the second millennium (Pollock 2000; 2006). At the same time, what constitutes a cosmopolitan, literary or vernacular language is context-specific and ever-changing. For

instance, the distinction between Romance languages as local ones vis-à-vis Latin as a cosmopolitan language was historically negotiated in the Middle Ages (R. Wright 1982, 1–4), and a local language may become dominant in the area and develop an asymmetric relationship to less standardised or differently evaluated language forms (for a discussion, see Auerbach 1958, 187).[1]

In the early modern European context, following the establishment of new print technologies, the dominance of Latin as the language of written culture eroded, while local languages became vehicles of politics, religion, culture and science. However, the shift from polities that used Latin as their primary written language to fully vernacular (albeit still multilingual) societies was asynchronous. Local political circumstances, imperial constellations and types of linguistic communities affected vernacularisation in different ways (Burke 2004, 61–159; Millar 2005, 9–30). English and Dutch, for example, gained prominence as written languages much earlier than many smaller European languages, such as Finnish, Norwegian (bokmål and nynorsk) or Czech. Even today, we find complex dynamics between global hegemonic languages, the official languages of states, and minority languages within those states. We therefore argue that the shift from Latin to vernacular languages is best understood as multiple vernacularisations rather than one unified process. Furthermore, we show that the relationship between vernacular languages and Latin was complicated by several incongruous factors in the development of book culture.

We study the changing hegemonies of publishing languages in a Northern European setting (Great Britain, The Netherlands, Sweden) from 1500 to 1800 by quantitatively analysing retrospective bibliographies. As we focus on published works, we are not interested in the emer-

1 We use 'vernacular' in the historical/philological tradition to mean a local or regional language rather than in the sociolinguistic sense of a spoken variety of language that deviates from a standard or accepted norm (see Coupland 2014, 86; 2016, 410–413). In our definition, a 'vernacular' language is not a translocal language such as Latin, Sanskrit or Classical Chinese, but one which may nevertheless be developed into a written or standard variety.

gence of first written texts in a given language but rather in the process through which vernacular languages gained prominence and started to be used for the same written functions and domains as the cosmopolitan languages (Pollock 2006, 4–6, 26). The process of vernacularisation and the internal dynamics between languages in multilingual societies have been discussed in theories of diglossia (Ferguson 1959; Fishman 1967; Hudson 2002), the history and theory of language planning and standardisation (Cobarrubias 1983; Millar 2005; Snow 2013), the history of nation-building (Anderson 2006; Burke 2013), linguistic studies of the creation of new vernacular registers and discourse communities (e.g. Taavitsainen and Pahta 1998), the history of the book (Febvre and Martin 1997, 319–332) and the birth of literary culture in the first and second millennia (Pollock 2006, 1–36). Previously, however, large-scale quantitative assessments of the transformation in the use of vernacular languages have not been possible due to a lack of reliable data. Febvre and Martin (1997, 320) noted in their study, originally published in the 1950s, that on "this subject no overall statistics can be obtained", although they do provide figures from Antwerp, Aragon and Paris (Febvre and Martin 1997, 320–321) which correspond to our findings. Nowadays, we have sufficient data to provide a much broader picture.

We take advantage of our already established bibliographic data science ecosystem in order to utilise vast data collections that cover publication places, page counts, publication languages and other complementary information to date the rise of local written languages in several different printing hubs in Europe (Lahti et al. 2019). The work extends our earlier projects and provides one of the first demonstrations of how large-scale data integration across multiple retrospective bibliographies can be used to analyse shifts in the languages used in print media. In doing so, we hope to contribute to the field of quantitative book history that has been developing rapidly in recent decades (see Suarez 2009; Buringh and Van Zanden 2009).

In this study, we produce comparative trajectories from several different locations and zoom in on different genres and places of publication. By scrutinising the publication records of individual towns, we are able to discern patterns that differentiate university

towns from commercial centres, and capital cities from the rest of the country. This focus on particular publication places has two distinct benefits. First, it allows us to perform in-depth reliability checks for the data concerning each location. Second, while the rise of local languages has been viewed predominantly as a question of national languages, the collecting of national bibliographies is a prime example of the methodological nationalism (Beck and Sznaider 2006) embedded in much of social scientific and humanistic data. In fact, it is crucial to keep in mind that the retrospective character of the nationally delineated bibliographies we use tends to solidify a national outlook, which may not be that relevant to the early modern period.

An analysis conducted from the perspective of different types of towns can highlight conflicting trajectories within national contexts. It also enables comparisons between locations across national borders. When we combine this locational approach with genre information in the bibliographies, we gain an even more granular picture of how publication languages developed, for instance, in various university towns across Northern Europe in the early modern period. Related work exists primarily in the study by Binzel, Link and Ramachandran (2020), who use the Universal Short Title Catalogue to study the effect of the Reformation on religious book production prior to the year 1600. Although similar in approach, our study takes a more nuanced view of the complexities of the language data and suggests that the vernacularisation of book printing in Northern Europe was a complex process, in which factors such as subject matter and intended readership, local legislation and the international book trade all influenced the choice of the language of publication.

Our focus is on the seventeenth and eighteenth centuries, a period when book publishing grew rapidly (Febvre and Martin 1997, 216–222), reading habits changed from reading individual works repeatedly and intensively to a more extensive form of reading as more people became able to enjoy more books (Engelsing 1970; 1974), while printed media came to be seen more and more as a medium for conversation and debate rather than one-sided proclamation (Habermas 1962; Eisenstein 1979; Klein 2004). It is in this period that the relationship between Latin and

many of the vernacular languages of Northern Europe changed. Local languages became standardised to a high degree and were increasingly seen as Latin's equals, or even superiors, as vehicles for producing new knowledge, debating issues and providing enjoyment. As the standardised vernacular became prioritised over Latin, some vernaculars, especially French, started to gain semi-cosmopolitan status as the new lingua franca, while other local varieties of language all but disappeared under the pressure of polity-wide standardisation (see Pollock 2006, 437–467; on French, see Fumaroli 1992, 555–558, 602–606).

Although our bibliographic data clearly demonstrates a gradual rise in the dominance of vernacular languages in Northern Europe, the process varied according to different factors. These include polity, publication place, the dynamics of international book markets, and genre. The rise of the vernacular did not always mean the simultaneous decline of Latin. As late as in the seventeenth century, Latin publishing was increasing along with its vernacular counterparts. Even in the eighteenth century, the overall decline of Latin is better understood as a specialisation of its use in academic genres. Latin still retained a high level of prestige associated with its learnedness, and although vernacular languages gradually took over many of the genres first occupied by Latin, this change took longer for the kinds of literature that were sold internationally. Many of the international 'best-sellers' in the sixteenth and seventeenth centuries were in Latin. Only in the eighteenth century did French books start to take over this market, at least in the Dutch, British and Swedish context. This change is best observed in the import and export records of books but is also visible in book printing at the local level.

Materials and methods

Our analysis draws on five nationally delineated retrospective bibliographies from the hand-press era: the English Short Title Catalogue (ESTC), the Short-Title Catalogue Netherlands (STCN), the Short Title Catalogue Flanders (STCV), the Swedish National Bibliography (SNB), and the Finnish National Bibliography (FNB). Since the STCN is geographically restricted to the current national borders of

The Netherlands, we have supplemented it with the still growing STCV, which covers present-day Flanders as well as Brussels. Similarly, the FNB is used to supplement the SNB. In this text, we use the terms bibliography and catalogue interchangeably, as both terms are used in the names of the datasets we use, although the former is the more apt term in this case.

As the bibliographies are primarily organised according to modern state borders, they inevitably introduce anachronistic elements to the study of book production in the sixteenth, seventeenth and eighteenth centuries. For the sake of simplicity, we mostly talk about Britain when referring to the British Isles with regard to the ESTC, Sweden for both the SNB and the FNB, and the Low Countries or the Dutch Republic for the STCN and the STCV. We will, however, refer to other geographical units when necessary for the sake of clarity in a given time period.

The bibliographies consist of semi-structured data that allows us to use quantitative methods to support the study of the development and languages of book culture before the end of the eighteenth century. They include information about published titles, their authors, publishers, publication places, page counts, publishing languages, book formats and other properties. Table 1 summarises the information fields that we have analysed in this article. Because the ESTC, STCN and STCV only extend to circa 1800, we limited the investigated time period up to and including the year 1799. Data preparation is based on custom workflows that complement critical analysis of historical sources and uncover potentially overlooked information gaps (Lahti et al. 2019; Lahti, Mäkelä and Tolonen 2020). Each bibliography was curated using a semi-automated process and then integrated for joint analysis (Table 1). We enriched language information by indicating the multilingual status of the documents and adding information on whether Latin or the local main vernacular (English/Dutch/Swedish) was catalogued as the primary or secondary language of the document.[2]

2 The implementation source code can be accessed via the URL github.com/comhis/article_2021_vernacular.

Table 1: Summary of the investigated fields. The number of entries refers to the number of unique entries that we have analysed in this article after discarding duplicates and selecting the indicated time period. The data availability per field for each bibliography is indicated by 'x'.

	ESTC	**SNB**	**FNB**	**STCN**	**STCV**
Source of the curated data version	Lahti et al. (2019)	Tolonen et al. (2019)	Tolonen et al. (2019)	Original source	Original source
Entries (n)	468 140	45 922	16 959	193 675	25 828
Time period	1474–1799	1457–1799	1522–1799	1473–1799	1473–1799
Geographical main focus	British Isles	Sweden	Finland	Nether-lands	Flanders
Vernacular language	English	Swedish	Swedish, Finnish	Dutch	Dutch
Primary language	x	x	x	x	x
Other languages	x	x	x	x	x
Publication place (e.g. town)	x	x	x	x	x
Publication country	x	x	x	x	–
Year	x	x	x	x	x
Page count	x	x	x	x	–
Gatherings	x	x	x	x	–
Author	x	x	x	x	–
Title	x	x	x	x	x
Work field	x	–	–	–	–
Publisher	x	x	x	x	x

For our analysis of different publication places within the historical states of Great Britain, the Low Countries and Sweden, we have chosen the capital cities of each state, as well as the university and commercial towns with the largest publishing activity. Thus, from the ESTC, we have included London, Oxford, Cambridge, Dublin, Edinburgh and Glasgow; from the SNB: Stockholm, Lund, Uppsala and Gothenburg; from the FNB: Turku; from the STCN: Amsterdam, Leiden, The Hague, Utrecht and Rotterdam; and from the STCV:

Antwerp, Leuven, Gent and Bruges. Some towns are discussed more thoroughly in the article than others, but all have been examined in our study.

The STCN has been criticised for its incompleteness with regard to the entirety of print production in the seventeenth century Dutch Republic (Pettegree and Der Weduwen 2018). However, because it is based on actual copies of works handled by the compilers, it provides a reliable record of the languages of those works. While the exclusion of newspapers and broadsheets does, of course, limit the uses to which the STCN can be put, it serves our focus on book culture and the shifting languages of book publishing well (see also the Discussion part of this paper). The restriction to books also holds for the other bibliographies we use. The ESTC, FNB and SNB's potential handicaps have been reported in previous research (Tolonen et al. 2021; Tolonen, Mäkelä and Lahti 2022; Tolonen et al. 2019). For our purposes, we have deemed these bibliographies comprehensive enough, although the usability of the data should be evaluated for other studies relying on them.

The book-in-hand method of compiling the STCN and the STCV further means that the information they provide on the subject matter and text types of the publications is more reliable than that of the other bibliographies we have used. A manual inspection showed us that the FNB is also quite thoroughly curated in terms of historically more reliable subject headings. Because the dataset is much smaller, however, it is of less importance for this article. The categorisation by subject heading – and, to a lesser extent, text type – in the STCN is reasonably consistent and comprehensive, enabling fairly reliable analyses by genre. The subject headings form a rich system of assigned indexing in which one document can be assigned to one or many categories.[3] While this sometimes increases the ambiguity of the

3 In the case of a work being assigned many geographical descriptors, we found up to 13 subject headings for one title. However, 89% of the titles in the STCN have from one to three subject headings.

classification,[4] it enables a layered approach to categorising the documents by single or co-occurring subject headings. Although many of the subject headings and text types are strongly connected (e.g. *Public and social administration* nearly always occurs with *State publications*, and *State publications* typically also receive a geographical keyword), some are not (e.g. *Medicine* can appear on its own or in conjunction with *Academic texts*). The STCV is similar to the STCN but smaller and very well documented. It differentiates between target, source and mediating languages for translated works, as well as between subject headings and document types in the keywords. In the STCN, headings referring to text type are not coded separately. Since the STCV is small compared to the other bibliographies and still a work in progress, we have only used it as a complement to the STCN.

Results

As previously pointed out (Febvre and Martin 1997, 331–332; Lahti et al. 2019) and shown below in Figure 1, Latin publishing in Europe was in gradual decline towards the end of the eighteenth century. Yet there are clear differences in the timing and pace of this development across different bibliographies/areas. The top shares of Latin printing vary between around 45 to 100 per cent of all publications in a decade. As late as the mid-seventeenth century, the FNB records consist almost exclusively of Latin items, while Latin appears to have become an almost marginal publication language in the ESTC. Although these differences are real, it must be noted that the bibliographies have very different coverage: the ESTC is dominated by publishing in London, whereas the FNB mostly records the publishing activities of the Swedish university in Turku (Åbo), located quite far from the political centre of Stockholm, and in an area that experienced vernacularisation

4 There is, for example, no fixed order for terms that are commonly used together, and a single work can end up in many categories. Real examples of subject heading combinations include *Theology (church history) + History (Netherlands) + Theology (general)*; *Medicine + Academic texts*; *History (Netherlands) + Period documents + Poetry + Political science*; and *Theology (Bible and Bible interpretation) + Theology (Christian doctrine)*.

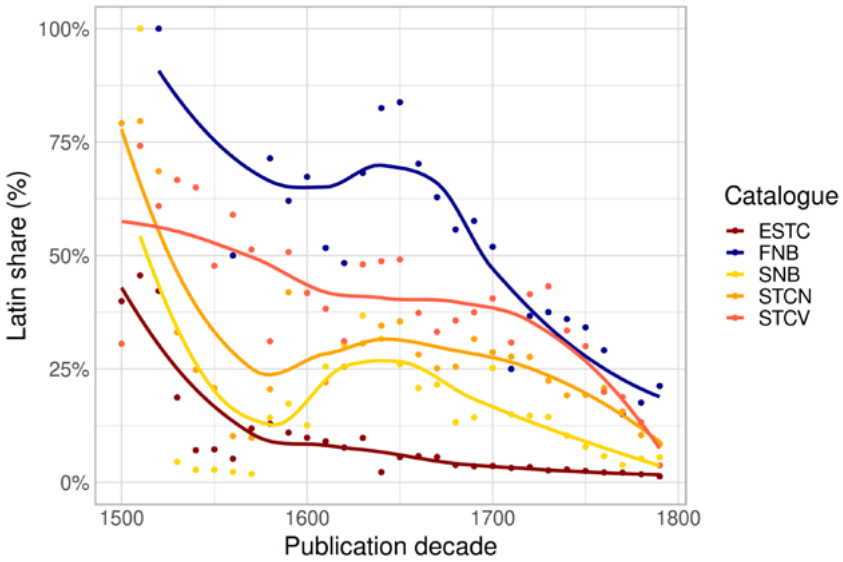

Figure 1: Share of Latin documents per decade in the ESTC, FNB, SNB, STCN and STCV 1500–1800. The loess smoothed curves have been added to highlight broad trends in the data.

twice, first with Swedish and later with Finnish (Kanner, Tahko, and Marjanen 2021).

Figure 1 shows that the share of Latin publishing actually increased in the Low Countries and Sweden in the seventeenth century. Absolute figures also demonstrate that Latin publishing overall increased in that period in all our bibliographies, just not by as much as publishing in the vernacular languages. Although the rise of printing in vernacular languages was definitely connected to the Reformation (Binzel, Link and Ramachandran 2020), we find that the decline of Latin publishing follows a different pattern and is not tied to the Reformation as a turning point. The decline of Latin was really an eighteenth-century phenomenon, and even then, we find some interesting exceptions to the rule (cf. Febvre and Martin 1997, 220–221). It seems that the developments in the seventeenth and eighteenth centuries, when book printing became much more common and diverse, were affected by many other factors besides the sixteenth century political turmoil relating to the Reformation. To understand the decline in the share of Latin books, we need to consider the

interdependence of the different regional traditions through the European book market and the special characteristics of the individual towns. Print location is also connected to genre, as Latin remained very strong in scholarly literature. For this reason, it constantly interacted with local languages even after the latter had become firmly established as written languages (Lindberg 2006). Indeed, Latin's position as the academic lingua franca remained robust even at the end of our period of observation.

Our results cover four themes. We will first discuss the relationship of the international book trade with language dynamics. In the second section, we explore the different publication places recorded in our data. The third section is devoted to connections between the subject, form and language of publications. Finally, we turn to the role of French as a newly emerging cosmopolitan language in the eighteenth century.

The book trade and the culture of the book

At first glance, the ESTC stands out in Figure 1 above as the least Latin of all the bibliographies. However, when estimating the availability of printed materials to audiences in different countries, it is important to remember that unlike trade catalogues, bibliographic catalogue data cannot directly show the effects of the international book trade on local markets. For instance, during the hand-press era, continental printing was cheaper and of better quality compared to Britain's. Continental booksellers took full advantage of this fact, which led to large numbers of scholarly books in Latin being printed elsewhere and imported into Britain (Barnard 2002, 5–7; Hoftijzer 2001). Conversely, any English best-seller published in Latin was likely to be pirated by Dutch printers, and popular works written in English spread on the continent as Latin and French translations published in Amsterdam, Leiden, Rotterdam and The Hague (Hoftijzer 2001).

Initially, the majority of books imported into Britain were written in Latin. In the eighteenth century, French-language books also became important items in overseas trade (Roberts 2002). Thus, at least until the eighteenth century, even as vernacular printing

increased, Latin books remained in wider circulation than the ESTC publishing figures may suggest. We have evidence of this from studies of learned British seventeenth century libraries, such as the one owned by John Locke, that were populated (if not dominated) by foreign-printed, foreign-language books (Harrison and Laslett 1971, 19–20). At the same time, British printers struggled to sell their products to continental readers, most of whom did not read English, which further increased the British focus on the popular vernacular market (Hoftijzer 2001, 91).

The dominance of imports in the foreign-language (predominantly Latin) book market in Britain (Roberts 2002, 141) had a crucial impact on the structural development of the British book trade. Figure 3a highlights the large share of English-language books printed in London. However, because books in Latin were mainly imports, we know that this relatively early vernacular dominance in printing does not necessarily mean that British readers consumed fewer Latin books than their continental counterparts. Conversely, the sixteenth-century growth of Latin printing in the STCN (Figure 1) does not mean that the domestic demand for Latin books in the Low Countries was necessarily increasing more rapidly than in other parts of Europe, since a significant portion of these items were intended for export to France, Britain and elsewhere (Hellinga 2001; Israel 2001, 233).

The decline of Latin printing in the STCN seems to have happened in several waves. In the sixteenth century, Latin book production was still growing but was gradually being overtaken by vernacular printing, as has already been indicated by previous research (Binzel, Link, and Ramachandran 2020). In the seventeenth century, Latin printing in the Low Countries was already dwarfed by vernacular printing, but it still grew and played a significant role in the international circulation of books. By the eighteenth century, most of the Latin books in print were academic theses and other items related to universities. This development can be seen in an analysis of the document counts of academic and non-academic titles in the STCN (Figure 2), which is based on the subject headings and language information included in the data. Latin was destined to dominate increasingly

specialised areas of printing, which meant that it became less prevalent in society at large but still played a prestigious role as the primary language of academic discourse. At this time, French-language publishing for local consumption and for export also started growing (see Figure 13, p. 57) and clearly assumed some of the role of Latin books in the cosmopolitan book trade.

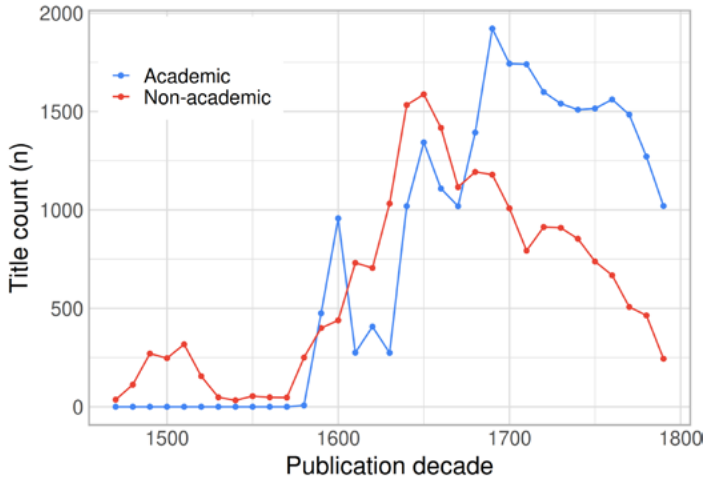

Figure 2: Academic and non-academic titles in Latin per decade in the STCN, 1480–1800.

Given the trend of importing foreign-language books from the continent, the language profile of the book market in fifteenth and sixteenth century Britain was not necessarily all that different from the Low Countries or Sweden. One might say that the culture of printing in seventeenth century London was peculiarly English, but the book culture was European and still featured plenty of Latin. However, it is reasonable to think that the strong tradition of printing books in the vernacular in England laid the groundwork for a particularly English Enlightenment. English-language printing and book culture were further strengthened in the eighteenth century when exports to America grew strongly. At this time, British book production gained a more cosmopolitan outlook than ever before (Sher 2006, 503–540).

Publication places and their linguistic profiles

Large cities dominate the top publication places for each bibliography. This is particularly the case for the ESTC, in which London, with its roughly 300,000 titles, accounts for 66 per cent of the catalogue's content. Similarly, Stockholm's nearly 29,500 documents (published before 1800) account for 64 per cent of the records in the SNB. Stockholm and London were the absolute political centres of their respective realms, and we see this reflected in how they dominated book production. In the Low Countries, the situation was slightly different. In the STCN, Amsterdam is the most prominent publishing location, with close to 50,000 titles, but they account for only 25.8 per cent of the catalogue's content. The publications recorded in the STCN and STCV are much more evenly distributed, as were the functions and major organisations hosted by towns in the Low Countries. From the point of view of Latin publishing vis-à-vis vernacular publishing, our data shows clear differences depending on whether the town was a political and demographic centre, whether it hosted a university or a science academy, and whether it was a commercial centre that fostered trade with books. These types are regularly mentioned in historical typologies of towns and cities, albeit not as the only ones (Albrecht, Majorossy and Rau 2023). Any typology of this kind is only intended to serve the purposes of a particular analysis, in our case the linguistic profile of early modern book printing. We do not assume that the types are inherently transhistorical, nor do we assume that they are evenly useful. In our case, the academic towns are the most distinct in terms of linguistic profile, whereas the commercial towns are the most diverse group. As a rule, the commercial towns produced fewer books than the academic towns or the political centres. They became active in the printing trade later but switched quicker to more popular genres and book formats.

Looking more closely at London, the absolute epicentre of English book production, we must remember that there are some factors that slightly skew the data. One of these is the elevated number of titles relating to the Civil War resulting from the inclusion of the Thomason Tracts in the ESTC. These and other potential sources of bias in the ESTC have been accounted for in previous scholarship (Tolonen et al.

2021), and they do not challenge the conclusions we draw, as the trend is very clear. Overall, we see that the printing of English-language books, pamphlets and other items recorded in the catalogue grew gradually, especially in the post-Civil War era. In London, the growth in printing happened mostly in English. It is nevertheless worth remembering that Latin publishing also grew, just not as rapidly, until the eighteenth century when it started to decline (Figure 3a).

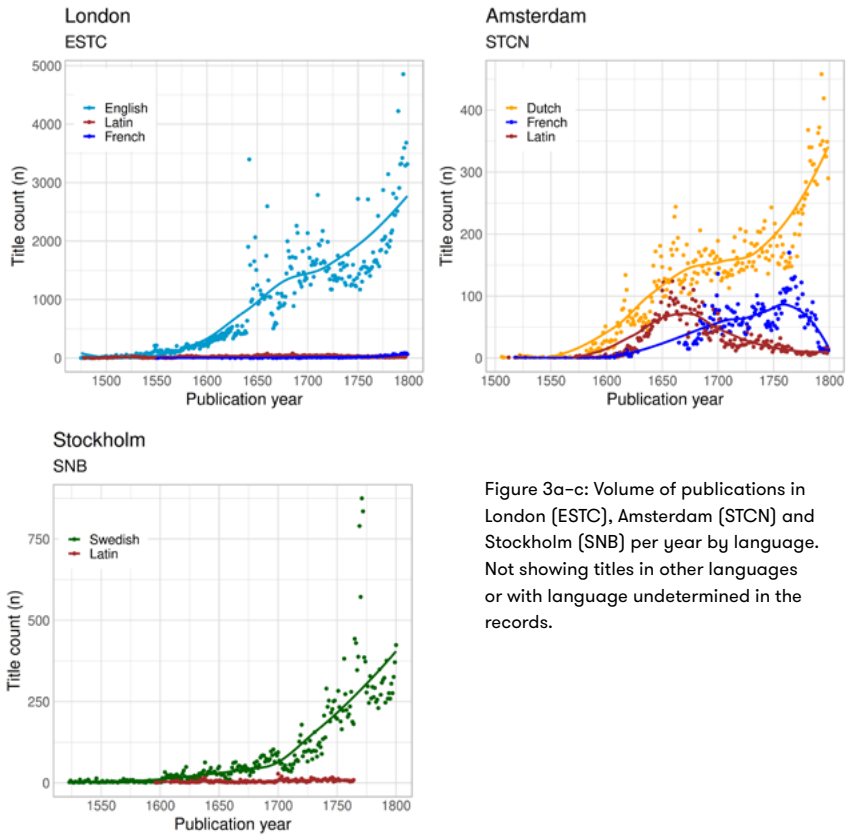

Figure 3a–c: Volume of publications in London (ESTC), Amsterdam (STCN) and Stockholm (SNB) per year by language. Not showing titles in other languages or with language undetermined in the records.

In Amsterdam and Stockholm, Latin publishing did fairly well in the seventeenth century, and only really started declining in the following century (Figure 3b–c). Nevertheless, the vernacular languages dominate in both of these demographic centres. For Amsterdam, we also see a rise of French-language items towards the end of the eighteenth

century. Due to the exceptional liberties of Dutch publishing, many Latin and French books were printed in the Dutch Republic for export. For the same reason, French books were sometimes given a false imprint location in the Netherlands or London, but were in fact printed elsewhere (Mathis 2009; Shaw 2002). In a sense, French titles in Britain and the United Provinces of the Low-Countries in the eighteenth century constitute, at least partly, a cosmopolitan market in the same way as some of the Latin literature before.

Freedom of print is also behind one of the outstanding features of publishing in Stockholm, namely the enormous spike in titles in the period 1766–1774. The surge in debates concerning politics and economy, among many other things, happened almost exclusively in Swedish. It is clear that this increase in vernacular printing happened in a more pronounced way in Stockholm than the rest of the country, not only due to the closeness of major political institutions, but also because of the sheer size of the city compared with other towns in the realm (Tolonen et al. 2019, 62–73).

The most obvious difference between university towns and virtually all other towns is the lingering dominance of Latin. The rise of the vernacular is slower in centres of academic publishing than in other demographic centres. Indeed, Latin *was* the dominant publishing language until the mid-to-late 1700s in Swedish and Dutch university towns. In Oxford, Cambridge and Edinburgh, the number of vernacular titles exceeds Latin ones from the late 1500s onwards, but the percentage of all titles printed in Latin in Oxford or Cambridge does not fall permanently below 25 per cent until the mid-1700s. (As the former capital of Scotland, Edinburgh is not solely a university town and does not fully follow the same publishing trends.) University towns, unsurprisingly, produced large quantities of academic literature, the default language of which remained Latin for most of this period. Oxford and Cambridge also stand out with higher shares of publications in Greek (Figure 4a–b).

Interestingly, we find considerable growth in the number of Latin publications in Edinburgh towards the end of the eighteenth century (Figure 4c). Most of these publications are academic theses of different

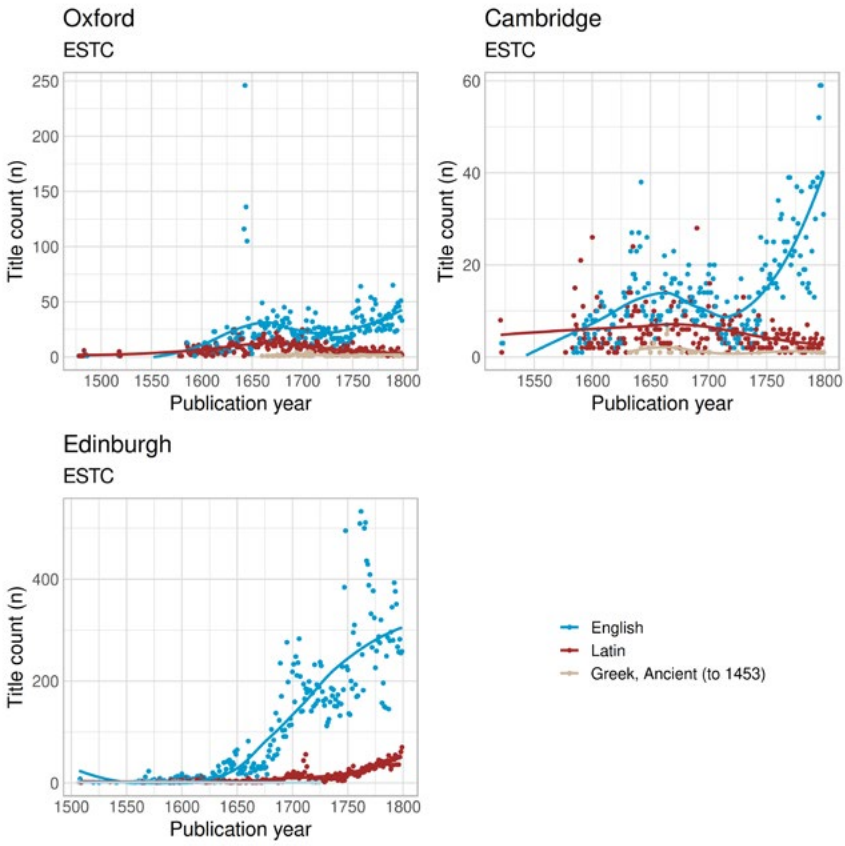

Figure 4a–c: Title count in Oxford, Cambridge and Edinburgh per year by language in the ESTC.

kinds, the majority of which belong to the fields of medicine and law. When we divided all Latin titles published in Edinburgh into two groups, those that have a formulaic academic title and those that do not, it was evident that the growth in titles consists entirely of these formal dissertations.[5] From 1750 onwards, formal theses rose to 100 published titles per year at the end of the century, whereas other Latin books

5 Compared to Latin titles from Oxford and Cambridge, it appears that the those from Edinburgh were subject to a much more formalised naming policy. Most of them followed a similar naming pattern, including the words "*dissertatio*", "*disputatio*", "*theses*" or "*tentamen*", indicating that they belonged to the sphere of formal academic education.

remained at between 10 and 30 titles per year. It is possible that there is a cataloguing bias that overrepresents academic dissertations in Latin. However, it also seems that the expansion of the university with the addition of faculties for Law, Medicine and Arts in the early years of the century may have impacted the publishing culture in Edinburgh. In general, universities started favouring English as the eighteenth century progressed, but due to the new faculties, more dissertations were published and these continued to be written mostly in Latin (Emerson 2012).

Nevertheless, we see varying rates of growth in vernacular publishing even in university towns. In Britain, English-language publishing grew rapidly in the context of the Civil War, although it was not until the latter half of the eighteenth century that vernacular publishing really started to take off. Quite practical issues could affect production levels. For instance, the drop in Cambridge in the late seventeenth century is to a large degree explained by a lack of local practitioners. From 1655 to the 1690s, only two people were active printers in Cambridge, and they were mainly focusing on the London part of their trade (McKitterick 2002, 199). Printing in seventeenth-century Oxford was similarly underdeveloped and vulnerable to contingent factors.

The eighteenth-century expansion of vernacular publishing is also evident in the other university towns in our datasets. Uppsala, Turku, Leiden and Edinburgh (Figure 6a–b, 5a, 4c) all experienced rapid growth in vernacular printing in this period. In Utrecht and Leuven, the trend started earlier or is less pronounced overall (Figure 5b–c). Again, the vernaculars did not take over the domain of Latin publishing. Rather, the overall expansion in genres, printers and markets seems to have produced a more complex trajectory. In Leiden, Utrecht and Leuven, we see a clear decline of Latin publishing already from the late sixteenth century, but this is not the case for Uppsala, Turku and Edinburgh. In these university towns, Latin publishing continued to grow for most of the eighteenth century. In the case of Uppsala, this development must have been even clearer because we know that some of the dissertations that were published are missing from the SNB (Tolonen et al. 2019, 70–71). Similarly, Pettegree and Der Weduwen (2018, 22) estimate that the STCN is missing information

on around 5,000 dissertations from the seventeenth century and that as many as 55,000 dissertations from that period have been lost.

Leuven, and indeed the STCV in general, shows a general decline in publishing activity from the mid-1600s onwards, before the numbers grow again towards 1800. Simons (2001, 36–38) explains that Antwerp's "golden century" as a hub of publishing came to an end around 1650, weakened by a long period of war and economic scarcity coupled with religious unrest in the Southern Netherlands. Publishing in Flanders only began to recover in the mid-1700s and even then it was subdued by the increasing importance of French as the language of religion in the area.

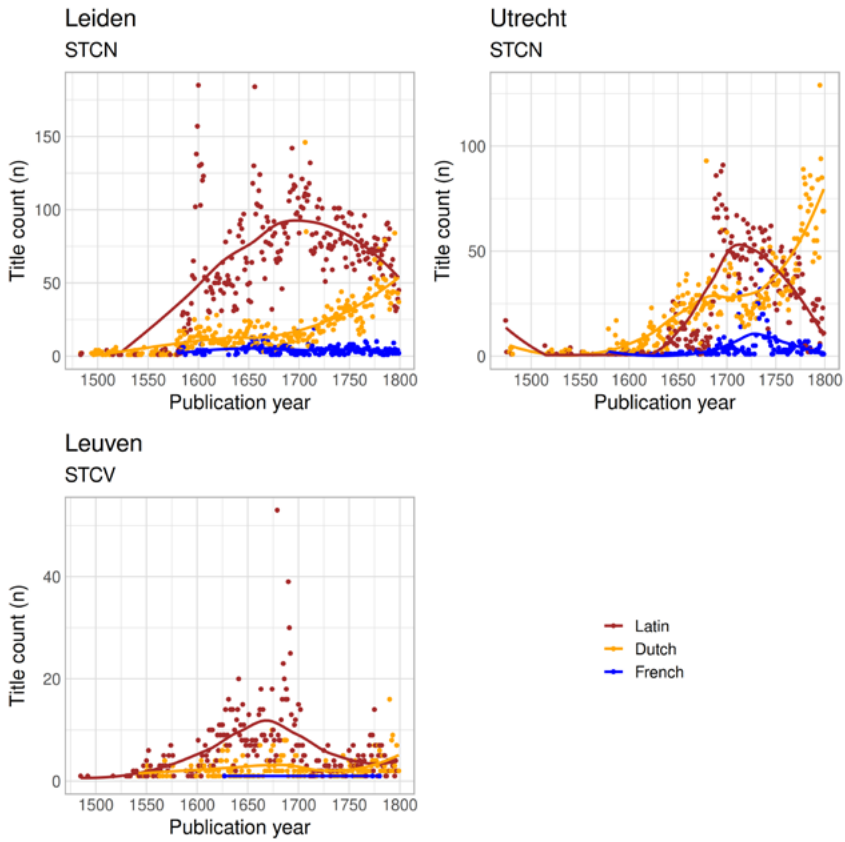

Figure 5a–c: Annual title count in Leiden, Utrecht and Leuven per year by language in the STCN/STCV.

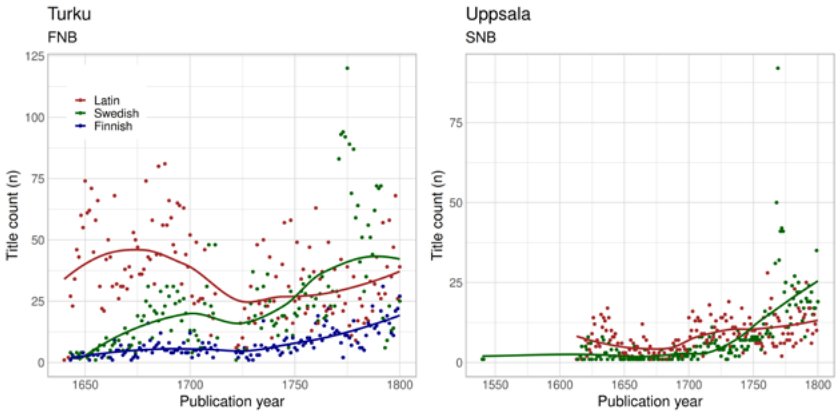

Figure 6a–b: Annual title count in Turku and Uppsala by language in the SNB/FNB.

It should be noted that the classification of towns into separate categories of demographic centres (or capital cities) and university towns is necessarily a simplification. The capital cities of London and Stockholm also hosted scientific institutions, such as the Royal Society and the Royal Swedish Academy of Science, and these will have produced academic literature within the cities, although both institutions promoted the use of the vernacular (Dear 1985; Lindroth 1967). As already mentioned, Edinburgh is a university town but also a demographic centre. In the Low Countries, the situation was even more complex because state institutions and civic organisations were more evenly distributed across the country and the provinces had more power at the local level (Kloek and Mijnhardt 2004).

In towns that did not host state institutions or universities and were not demographic centres, there was less book printing overall, and the number of Latin titles was small. This is evident from the linguistic publishing profiles of Gothenburg in Sweden, Glasgow in Britain, and Rotterdam in the Low Countries, which are quite similar. Latin printing remained at under 30 per cent of all publications for the whole period and vernacular printing dominated the market (Figure 7a–c). For Glasgow and Gothenburg, the rise of the vernacular was more pronounced in the eighteenth century, whereas

Rotterdam, which overall has rather low publication figures for the size of the city, shows gradual growth from early on. In all of these cases, we can assume that commercial towns in general did not produce many books in the genres that required Latin. Obviously, whenever commercial towns also hosted organisations that pushed for such genres, things changed. Turku, for instance, was initially one of the main trading towns in the kingdom of Sweden, but the foundation of a university in 1640 came to dominate its publishing industry. Genre, then, is a crucial aspect in understanding preferences in the language of publication.

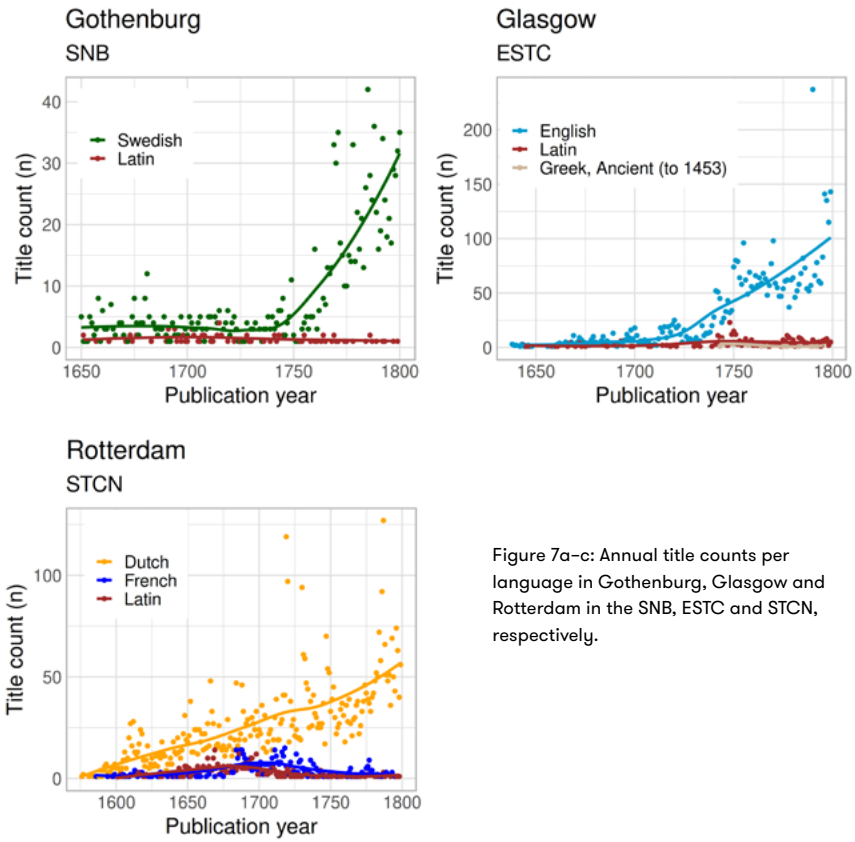

Figure 7a–c: Annual title counts per language in Gothenburg, Glasgow and Rotterdam in the SNB, ESTC and STCN, respectively.

Genre and the choice of language

As the analysis above already suggests, there is a connection between the expected readership of a work and the language in which it is written and published. Different types of towns catered to different audiences and produced different genres and registers. A growing readership and a growing literary output also enabled specialisation within the publishing trade: vernacular books on topics that concerned the (newly) literate local audience, and Latin texts for the highly educated and/or international audience. While genre is notoriously difficult to define and hard to account for in historical records, all of the bibliographies we use provide some information about the subject headings of the titles. The keyword indexing in the STCN is comprehensive enough to allow analyses by subject heading and text type, with the obvious caveat that the keywords signify a cataloguer's interpretation of the content and form of the work within the bounds of a controlled vocabulary, rather than a comprehensive and unambiguous truth about its nature. In addition to keyword indexing, and especially for bibliographies with less systematic indexing, we have previously used different proxies, such as page counts or gatherings (the book format) – or, indeed, language – to estimate the genre of a work, and even categorised some works manually in order to categorise the rest automatically (see Tolonen et al. 2021). In this article, we focus on genre as indicated by STCN subject headings in order to avoid biases based on mismatched systems of categorisation.

The most common subject headings (with considerable overlap between each other as well as other headings) in the STCN are indicated in Table 2. The most common subject headings denoting text type are *Academic texts*, *State publications*, *Period documents*, *Occasional writings*, *Catalogues*, *Almanacs*, *Poetry*, *Drama*, *Periodicals*, and *Songbooks*. The text type categories are typically smaller than other subject heading categories. For our purposes, the most important text type is *Academic texts*, with 25,506 occurrences (works) in the STCN.

Table 2: The most common subject headings in the STCN, all sub-categories combined (works published before 1800).

STCN Subject heading	Works
History (all sub-categories combined)	59,381
Theology (all sub-categories combined)	46,948
Public and social administration	38,215
Law	15,836
Dutch language and literature	15,332
Medicine (including veterinary medicine)	10,684
French language and literature	7,458
Philosophy	4,732
Latin language and literature	4,350

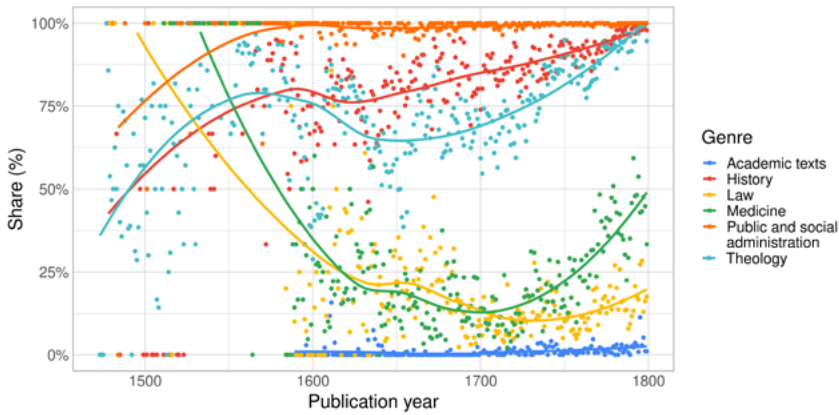

Figure 8: Annual share of vernacular titles according to genre classification in the STCN, 1480–1800. A 100 per cent share indicates that all items are in Dutch, 0 per cent that none of them is in the vernacular (assumed to be in Latin). The categories shown are the same as in Table 2, excluding French language and literature and Philosophy, which would not appear correctly on a scale from Latin to Dutch (Philosophy begins to appear in large numbers in French before Dutch), and Latin and Dutch language and literature (nearly 0 per cent and 100 per cent vernacular, respectively). The subcategories under History and Theology have been combined. The loess smoothed curves have been added to highlight broad trends.

The growing number of works assigned to the category of *Dutch language and literature*, and the prominence of Dutch in *Public and social administration* as well as *History*[6] in the STCN from the mid-1500s onwards suggest that the Low Countries are, at this time, already a largely vernacular society, i.e. one that has a vernacular literary culture and government. Dutch was thus seen as a language fit for all purposes apart from the academic, where international intelligibility is key. Besides the obviously Latin-dominated category of *Latin language and literature*, Latin is the dominant language of *Law*, *Medicine* and *Philosophy* until at least the mid-1700s, with clear differences between the subjects. *Law* becomes vernacular at a slower rate than *Medicine* throughout the 1700s and is still predominantly Latin in 1800, while vernacular works on *Medicine* account for over 50 per cent of publications by 1800 (Figure 8). In *Philosophy*, Latin declines earlier than in *Law* or *Medicine* but is replaced by French as the dominant cosmopolitan publication language in the mid-1700s; Dutch works categorised as *Philosophy* only start to appear in slightly larger numbers towards the end of the eighteenth century. Febvre and Martin (1997, 331) also propose *Mathematics* and *Astronomy* as genres that were particularly slow to transition from Latin, aligning with our findings. However, this is difficult to quantify using the current genre categories.

The STCN subject heading for *Academic texts* is a fairly reliable indication of works that are tied to the formal aspects of university teaching and publishing, and a helpful aid in understanding how subject matter, readership and language correlate in the data. *Academic texts* are most prominent in the university towns of Leiden and Utrecht. In Leiden they comprise more than half of all records. *Academic texts* from Leiden and Utrecht account for approximately 65 per cent of the whole category in the STCN. It also includes publications from Groningen, Amsterdam and, very occasionally, from the other major publication places in the Dutch Republic. The majority of these

6 *History* here refers to all titles that have 'history' among their keywords. The STCN always adds a geographical or other modifier in brackets after the keyword 'history'.

titles are dissertations of different kinds. This is the most Latin-dominated category in the whole dataset. While there are isolated instances of Dutch academic titles, it is really only in the eighteenth century that Dutch is used in academic literature at all, and even then only in a handful of titles per year.

By looking at *Medicine* in combination with *Academic texts*, we get a more nuanced picture of what is happening in terms of the language used in medical writing. We know from medieval England that the audience for medical writing was already very heterogeneous, extending from academically minded physicians to barber-surgeons and other practitioners, as well as lay people. Latin was the dominant language of medical writing in the British Isles until the mid-1600s, but the number of medical texts in English rose steadily from the end of the fourteenth century onwards (Taavitsainen and Pahta 1998). It seems safe to assume that a similar development took place in the Low Countries. Although there are some cases of medical treatises in Dutch, the real growth in vernacular texts in the field of medicine occurs among titles that are not categorised as academic. At the same time, the Latin academic titles are in slight decline (Figure 9).

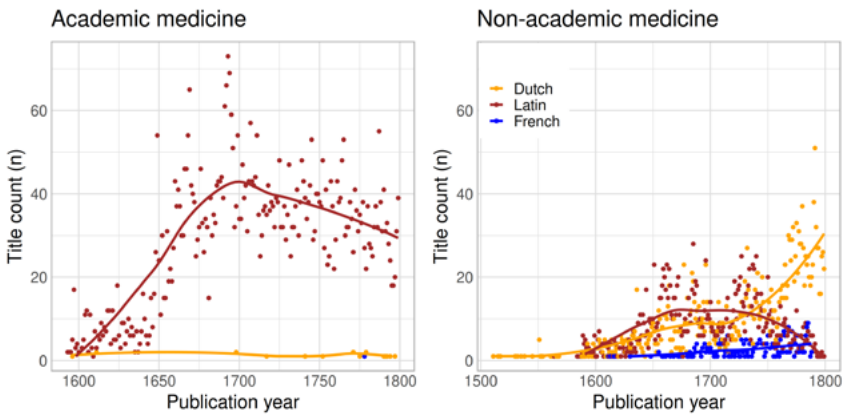

Figure 9: Academic and non-academic publications in the field of medicine by language in the STCN 1600–1800. The loess smoothed curves have been added to highlight broad patterns in the data.

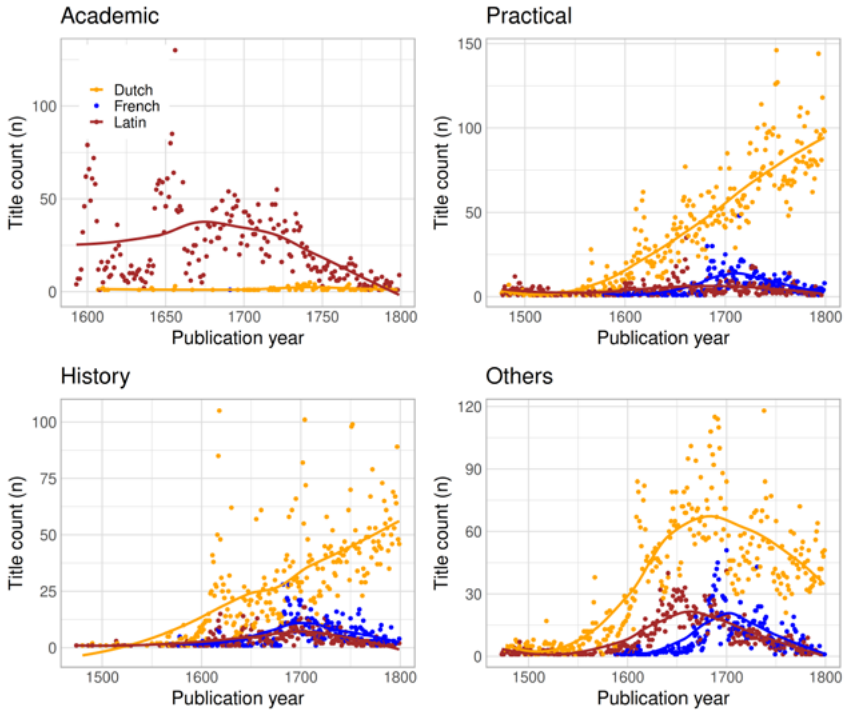

Figure 10: Language by subcategory of *Theology* in the STCN, 1480–1800. The loess smoothed curves have been added to highlight broad patterns in the data.

Of the large subject areas in the STCN, *Theology* in particular is a huge category with many overlapping subcategories and a complicated linguistic profile. Filtering all theology that is also tagged with *Academic texts*, we found that academic theology is, unsurprisingly, written in Latin throughout the period. Presumably, this also applies to other intellectual texts meant for international audiences, which have ended up in the category "Other" in the image above (Figure 10). Interestingly, we found that academic theology as a genre was in decline from

1700 onwards. Practical theology[7] is, as is to be expected, a primarily vernacular genre, as is religious history. "Other" theology has an international dimension first in Latin, then also in French. Although theology overall started leaning towards Dutch as early as the 1600s, the choice of language clearly depended on the text's preferred audience and context.

An inspection of the languages of publication according to subject heading and place of publication deepens our understanding of the language profiles of different towns and confirms that different types of towns focused on different genres in their book production. In Amsterdam, Latin publications are found especially in the fields of language and literature, history and theology. In all of these genres, however, we find a gradual decline in Latin by the end of the seventeenth century. The shift towards vernacularisation takes place earlier here than in the university towns. The university town of Leiden, on the other hand, hosted a vital Latin publishing scene as late as the eighteenth century. This is particularly visible in the works labelled as *Medicine* or *Law*, which are completely dominated by works printed in Latin. Even the category of *History*, which elsewhere is predominantly vernacular, is a primarily Latin genre in Leiden until circa 1700. *Theology* is likewise more of an academic subject than elsewhere (Figure 11). Utrecht, also a university town, shows similar trends, but has a more typical linguistic profile in *History* and *Theology*, with plenty of vernacular publications. The Latin-ness of its works on *Law*, *Medicine* and *Philosophy*, nevertheless reveals the importance of academic activity in the town and confirms the connection between the academic register and Latin. Latin-heavy Leiden and Utrecht differ clearly from

7 Practical theology here consists of works labelled *Theology (practical)* as well as text types we identified as relating to religious practice, such as *Catechisms*, which are sometimes labelled as *Theology (Christian doctrine)* in the STCN. Again, there is considerable overlap between categories, with works sometimes receiving a litany of keywords. To give an example, one edition of *Biblia: dat is*, *De gantsche Heylighe Schriftuere* (record no. 4255) is tagged as *Theology (Bible and Bible interpretation)*, *Poetry*, *Songbooks*, *Theology (practical)*, *Catechisms*, *Prayer books* and *Theology (Christian doctrine)*.

Rotterdam and The Hague, whose Latin output is quite low. Rotterdam produced few works categorised as *Law* and a very moderate number of titles relating to *Medicine*. Similarly in The Hague, *Law* and *Medicine* were marginal subjects compared with *History* and *Theology*, which are notably French categories in the town. At times in the early 1700s, the number of historical or theological titles published in French exceeded that of Dutch titles in these categories.

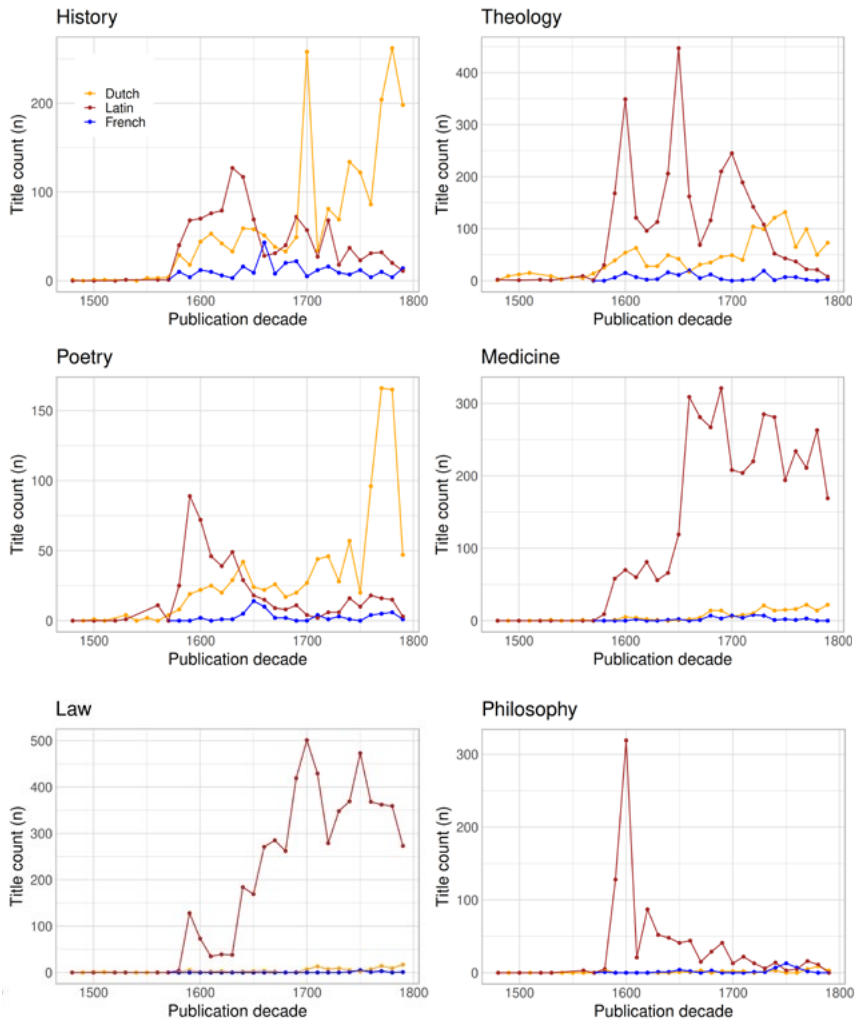

Figure 11: Languages per decade by subject heading in Leiden in the STCN, 1480–1800.

All in all, these genre-related linguistic trends support the thesis of a shift towards the vernacular in stages. The rapid increase in printing, starting in the late sixteenth century, materialised as vernacular works especially in more widely read topics such as history and theology, whereas the more academic genres of medicine and law held onto Latin for longer. By 1800, Latin remained dominant only in academic theses and dissertations. All other forms of literature had effectively been vernacularised (see the section below on the rise of French as a new cosmopolitan language). This did not, however, mean that Latin lost its value completely. It retained much of its earlier prestige, but was limited to very specific genres of publishing. Commercially, however, Latin publishing became practically marginalised. Evidence for this may be found in the fact that while Latin books were still printed, they were rarely reprinted. This is especially clear in the ESTC, which is the only one of our bibliographies that has harmonised and reliable information available about editions (Tolonen et al. 2021, 70–77). The number of editions per book per decade was significantly higher for Latin titles than English titles for the whole of the 1500s. When publishing expanded in the 1600s, the average number of editions remained slightly below two editions per title, with some variance between the languages. In the latter half of the eighteenth century, the average number of editions per English-language title gradually grew, reaching more than 2.5 by the 1790s, whereas Latin titles were reprinted at roughly the same rate as before. In other words, when the book market expanded, the growth in demand was primarily for books in the vernacular.

French as a new cosmopolitan language in the eighteenth century

The growth in book production that can be seen in all of our bibliographies is quite evidently related to many different commercial and intellectual developments. While the expansion is most visible in vernacular publishing, we also see a clear upswing in French publishing in the ESTC, the SNB and the STCN. There are several reasons for this, one of which is that French took on some of the

functions that Latin had had as a cosmopolitan language of communication. Around the Treaty of Westphalia, French became the language of choice in European negotiations and diplomacy, and was often attributed universal qualities. By the eighteenth century, the position of French in court politics and philosophical discourse was indisputable, and some people, like Voltaire, thought that the French "genius of the nation combined with the genius of the language has produced more delightfully written books than one finds in any other country" (Fumaroli 1992, 556, 602).

Chronologically, the rise of French is first visible in the Dutch Republic. We found a large volume of French publications especially in The Hague but also in Amsterdam and Utrecht, notably in theology but also in other genres. Presumably, these books catered to a French-speaking minority of Huguenot immigrants and the market in France for imports of banned French-language religious literature. Perhaps they also reflect the historical position of French as an international language of prestige (Putter and Busby 2010, 2), as well as an interest in fashionable French Enlightenment authors in the late eighteenth century (S. Wright 2004, 118–22). However, the number of

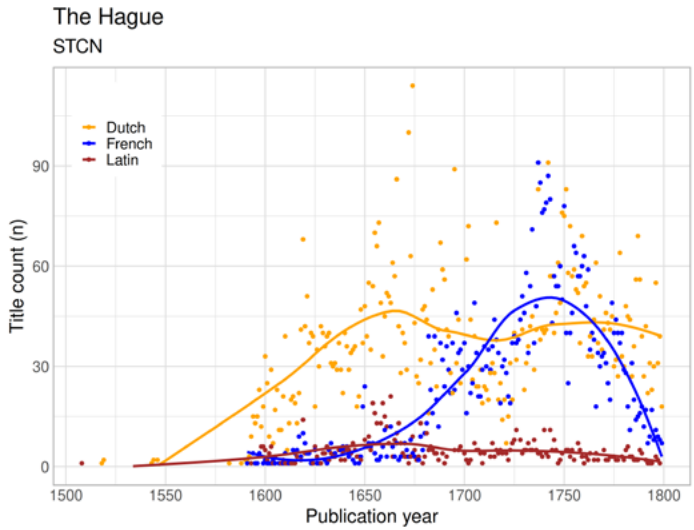

Figure 12: Annual title count per language in The Hague (STCN), 1500–1800. The loess smoothed curves have been added to highlight broad patterns in the data.

French titles in the STCN fell as rapidly as it grew in the latter half of the eighteenth century.

In The Hague, where the number of French publications even exceeded the number of Dutch publications in the early to mid-1700s (Figure 12), publishing in French focused on theology, history and literature. French drama was frequently printed in the Dutch Republic due to its status as modern classics suitable for export. Authors like Molière, Quinault, the Corneille brothers and Racine constituted the foundation of the French Model of theatre, making them highly sought after (Stone Peters, 93–112). In addition, politically sensitive French material was also published, or misidentified as being published, outside of France.

The development of French publishing in the other catalogues is slightly different. The STCV is a clear outlier, as it includes publications from Brussels, where French ought to be seen as a local vernacular. It has therefore been excluded from Figure 13 along with the FNB, in which French holds a marginal position. In the ESTC and the SNB, French publishing is a phenomenon of the late eighteenth century. This suggests that French had by then effectively become a lingua

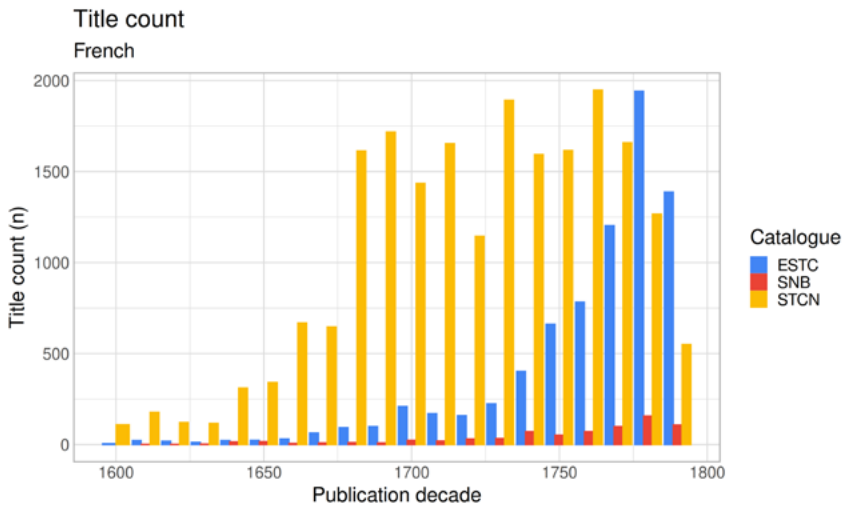

Figure 13: Raw count of French titles published per decade in the ESTC, SNB and STCN, 1600–1800.

franca for intellectual exchange in Northern Europe (S. Wright 2004, 118–22). Indeed, its position in the latter half of the century was so important that books in French were not only imported, but also printed locally in Britain and Sweden (Figure 13). These titles were partly produced for export, but also reflected local demand. In London, French titles overtook Latin imprints by the last quarter of the century. Some printers specialised in serving the French-speaking market, at least in part to cater to "a sizeable French population in the capital: merchants, diplomats, but more numerously the religious refugees of the Huguenot period and the political refugees of post-Revolutionary times" (Shaw 2002, 120, 122). In Stockholm, French publishing had nearly closed the gap to Latin by the end of the century.

Discussion

Using several different retrospective bibliographies prompts questions about the benefits and pitfalls of quantitatively analysing the publication record. While the bibliographies used in this study do not cover the full publication record for their areas, they provide the best and most complete knowledge base that is available concerning publishing activities in the early modern period. Two collections of national union catalogues – the Universal Short Title Catalogue (USTC) for the period up to 1650, and the Heritage of the Printed Book Database (HPB) for the period before 1828 – have been used in similar studies (Binzel, Link and Ramachandran 2020; Buringh and Van Zanden 2009; Baten and van Zanden 2008). However, the retrospective bibliographies that we opted for provide a better data source for studying the tail-end of Latin publishing, in particular for the period after 1650, when publishing became more widespread and complex. Our tests show that the USTC has slightly better coverage of Dutch printed books for the fifteenth and sixteenth century than the combination of the STCN and the STCV, whereas the coverage for Britain and Sweden is similar between the USTC and the retrospective bibliographies. For the seventeenth and eighteenth centuries, the retrospective bibliographies provide better coverage for all countries. Both the national bibliographies and the bibliographic collections are constantly

being developed, but since most of this work is done first at the national level, as evidenced by many projects in the field (Häusner and Sommerland 2019), the nationally delineated retrospective bibliographies are generally a good starting point for analysis. Using retrospective bibliographies also makes it easier to control for particular design principles and biases, which are easily lost in the use of compilation catalogues.

Our analysis of the bibliographies confirms the general story about the rise of vernacular languages as publishing languages and the gradual decline in the share of titles published in Latin. This development naturally relates to policies that promoted the role of local languages (Binzel, Link, and Ramachandran 2020), but also ties in with increasing literacy, the gradual popularisation of scientific knowledge, and a growing need to debate political affairs horizontally outside the confines of feudal society (Burke 2004, 1–14, 43–60; Millar 2005; Habermas 1962). It is virtually impossible to prove how these different processes interacted. One possible chain of events is that ideological positions promoting local languages affected reading skills, reading skills encouraged the spread of scientific knowledge, and increased awareness of the natural world created a need for public debate. In reality, however, these and many other processes were interdependent and produced different trajectories for the vernacularisation of the body of printed works.

Dutch and English were already well-established as written languages by 1500, as attested by their prominence in the data from early on, while Swedish steadily increased its presence from the Reformation onwards and accounted for nearly three quarters of the published record in the SNB in the mid-1600s. Finnish hardly makes an appearance in a dataset that excludes the nineteenth century, although it was occasionally used in texts relating to religion and economy that targeted the peasant population in particular. Even in a protestant, Northern European context, we cannot therefore speak of the "rise of the vernaculars" as a particularly unified phenomenon. Furthermore, as we have shown, the growth of any single vernacular as a publishing language does not automatically mean the simultaneous decline of

Latin, or the decline of Latin in all genres. The use of Latin decreases in stages and is accompanied by an increased specialisation in publications meant for a learned, international readership. The rise of the vernaculars is more linear. It roughly follows the general pattern of growth in book publishing and is connected to the emergence of new genres – most famously, novels – that never became thought of as potentially Latin genres.[8] Latin was likewise never the main language of newspapers and periodicals in the areas we have investigated here. Taking them into account in considering the language of print production as a whole would surely further emphasise the role of vernacular languages (see Pettegree and Der Weduwen 2018, 3–4).

The juxtaposition between a prestige/cosmopolitan language and a vernacular, and the idea of a shift from one to the other in written language use – vernacularisation – are always simplifications. In most, if not all, societies, various codes always co-exist. None of the areas in our analysis became monolingual, even though the balance of power between languages of publication shifted. Other languages or language variants never gained a prominent position in print media at all. Indigenous peoples and their languages are currently undergoing similar processes. In Northern Europe, the development and status of Sámi languages in relation to Norwegian, Swedish and Finnish is especially relevant, as is the role of Scots, Gaelic and Welsh in Britain.

Contemporary struggles relating to language rights and the role of local language variants vis-à-vis standard languages or cosmopolitan codes raise the question of how the choice between using a cosmopolitan or a vernacular language has been linked to the idea of the nation at different times. The term "Herder effect" was coined by Pascale Casanova (2004) to account for the widespread post-French Revolution belief that literature in the national language was crucial for building the nation, and more important than any imaginary international literary community. However, our quantitative analysis clearly shows that

8 See Chapter 3 for an analysis of the rise of the novel in Norway during the eighteenth and nineteenth centuries.

the vernacular had a much earlier breakthrough in a range of genres than the Herder effect would imply. It is also clear that many other factors besides the conscious promotion of national culture influenced language choice. Our analysis suggests that it would be beneficial to pay attention to a wide range of genres when developing an understanding of the nineteenth century preoccupation with nationhood. Casanova assumes that some genres, especially the newly emerging novel, were particularly important for elevating the nation. Indeed, other genres that had previously been considered important to publish in the vernacular and that remained largely vernacular, such as popular religious literature, were only marginally important for national mobilisation in the nineteenth century. We suggest, nevertheless, that the Herder effect could be used to describe not only genres that were new to the modern period, but also genres that most resisted vernacularisation. The nineteenth century push for academic literature to be written in the vernacular could essentially be seen as an example of the Herder effect, although the attempts to vernacularise science have ultimately been unsuccessful. The project of building the nation relied not only on new vernacular genres but also on the vernacularisation of genres that were regarded as prestigious.

References

ALBRECHT, MARA, JUDIT MAJOROSSY AND SUSANNE RAU. 2023. "Typologising Cities: Historical and Methodological Reflections." *Religion and Urbanity Online*, edited by Susanne Rau and Jörg Rüpke. Berlin, Boston: De Gruyter. https://doi.org/10.1515/urbrel.24882548. Accessed 7 May 2024.

ANDERSON, BENEDICT. 2006. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Rev. ed. London: Verso.

AUERBACH, ERICH. 1958. *Literatursprache und Publikum in der lateinischen Spätantike und im Mittelalter*. Bern: Francke.

BARNARD, JOHN. 2002. "Introduction." In *The Cambridge History of the Book in Britain*, edited by John Barnard, D. F. McKenzie, and Maureen Bell, 4:1–28. Cambridge: Cambridge University Press.

BATEN, JOERG, AND JAN LUITEN VAN ZANDEN. 2008. "Book Production and the Onset of Modern Economic Growth." *Journal of Economic Growth* 13 (3): 217–35. https://doi.org/10.1007/s10887-008-9031-9.

BECK, ULRICH, AND NATAN SZNAIDER. 2006. "Unpacking Cosmopolitanism for the Social Sciences: A Research Agenda." *The British Journal of Sociology* 57 (1): 1–23. https://doi.org/10.1111/j.1468-4446.2006.00091.x.

BINZEL, CHRISTINE, ANDREAS LINK AND RAJESH RAMACHANDRAN. 2020. "Vernacularization, Knowledge Creation, and Growth: Evidence from the Protestant Reformation." In *CEPR Discussion Paper*. Vol. DP15454. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3737587.

BURINGH, ELTJO, AND JAN LUITEN VAN ZANDEN. 2009. "Charting the 'Rise of the West': Manuscripts and Printed Books in Europe, A Long-Term Perspective from the Sixth through Eighteenth Centuries." *The Journal of Economic History* 69 (02): 409. https://doi.org/10.1017/S0022050709000837.

BURKE, PETER. 2004. *Languages and Communities in Early Modern Europe*. Cambridge: Cambridge University Press.

———. 2013. "Nationalisms and Vernaculars, 1500–1800." In *The Oxford Handbook of the History of Nationalism*, edited by John Breuilly. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199209194.013.0002.

CASANOVA, PASCALE. 2004. *The World Republic of Letters*. Cambridge, Mass.: Harvard University Press.

COBARRUBIAS, JUAN. 1983. "Ethical Issues in Status Planning." In *Progress in Language Planning: International Perspectives*, edited by Juan Cobarrubias and Joshua A. Fishman, 41–85. Berlin: Mouton Publishers.

COUPLAND, NIKOLAS. 2014. "Sociolinguistic Change, Vernacularization and Broadcast British Media." In *Mediatization and Sociolinguistic Change*, edited by Jannis Androutsopoulos. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110346831.67.

———. 2016. "Labov, Vernacularity and Sociolinguistic Change." *Journal of Sociolinguistics* 20 (4): 409–30. https://doi.org/10.1111/josl.12191.

DEAR, PETER. 1985. "Totius in Verba: Rhetoric and Authority in the Early Royal Society." *Isis* 76 (2): 144–61.

EISENSTEIN, ELIZABETH L. 1979. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early Modern Europe*. Cambridge [Eng.]; New York: Cambridge University Press.

EMERSON, ROGER L. 2012. "Reading in Universities." In *The Edinburgh History of the Book in Scotland. Volume 2*, edited by Stephen W Brown and Warren McDougall, 611–24. Edinburgh: Edinburgh University Press.

ENGELSING, ROLF. 1970. "Die Perioden der Lesergeschichte in der Neuzeit: Das statistische Ausmass und die soziokulturelle Bedeutung der Lektüre." *Archiv für Geschichte des Buchwesens* 10: 944–1002.

———. 1974. *Der Burger als Leser: Lesergeschichte in Deutschland, 1500–1800*. Stuttgart: Metzler.

FEBVRE, LUCIEN, AND HENRI-JEAN MARTIN. 1997. *The Coming of the Book: The Impact of Printing 1450–1800*. London: Verso.

FERGUSON, CHARLES A. 1959. "Diglossia." *Word* 15 (2): 325–40. https://doi.org/10.1080/00437956.1959.11659702.

FISHMAN, JOSHUA A. 1967. "Bilingualism With and Without Diglossia; Diglossia With and Without Bilingualism." *Journal of Social Issues* 23 (2): 29–38.

FUMAROLI, MARC. 1992. "The Genius of the French Language." In *Realms of Memory, vol. III*, edited by Pierre Nora, 555–606. New York: Columbia University Press.

HABERMAS, JÜRGEN. 1962. *Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Herman Luchterhand Verlag.

HARRISON, JOHN R., AND PETER LASLETT. 1971. *The Library of John Locke*. 2nd ed. Oxford: Clarendon Press.

HÄUSNER, EVA-MARIA, AND YLVA SOMMERLAND. 2019. "The Role and Function of National Bibliographies for Research." *Cataloging & Classification Quarterly* 57 (1): 1–4. https://doi.org/10.1080/01639374.2019.1579971.

HELLINGA, LOTTE. 2001. "The Bookshop of the World: Books and Their Makers as Agents of Cultural Exchange." In *The Bookshop of the World: The Role of the Low Countries in the Book-Trade, 1473–1941*, edited by Lotte Hellinga, Alastair Duke, Jacob Harskamp and Theo Hermans, 11–29. Goy-Houten: Hes & De Graaf.

HOFTIJZER, P. G. 2001. "The English Book in the Seventeenth-Century Dutch Republic." In *The Bookshop of the World: The Role of the Low Countries in the Book-Trade, 1473–1941*, edited by Lotte Hellinga, Alastair Duke, Jacob Harskamp and Theo Hermans, 89–107. Goy-Houten: Hes & De Graaf.

HUDSON, ALAN. 2002. "Outline of a Theory of Diglossia." *International Journal of the Sociology of Language* 2002 (157): 1–48. https://doi.org/10.1515/ijsl.2002.039.

ISRAEL, JONATHAN. 2001. "The Publishing of Forbidden Philosophical Works in the Dutch Republic (1666–1710) and Their European Distribution." In *The Bookshop of the World: The Role of the Low Countries in the Book-Trade, 1473–1941*, edited by Lotte Hellinga, Alastair Duke, Jacob Harskamp and Theo Hermans, 233–43. Goy-Houten: Hes & De Graaf.

KANNER, ANTTI, TUULI TAHKO AND JANI MARJANEN. 2021. "Becoming a State Language: Finnish Public Debate and Modal Grammar 1820–1917." In *CEUR Workshop Proceedings*, 2865:236–41. Riga, Latvia: CEUR-WS.org. http://ceur-ws.org/Vol-2865/poster4.pdf.

KLEIN, LAWRENCE E. 2004. *Shaftesbury and the Culture of Politeness: Moral Discourse and Cultural Politics in Early Eighteenth-Century England*. Cambridge: Cambridge University Press.

KLOEK, JOOST, AND WIJNAND W. MIJNHARDT. 2004. *1800: Blueprints for a National Community*. Dutch Culture in a European Perspective 2. Assen: Royal Van Gorcum.

LAHTI, LEO, EETU MÄKELÄ AND MIKKO TOLONEN. 2020. "Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming." In *CEUR Workshop Proceedings*, 2723:280–89. Amsterdam. http://ceur-ws.org/Vol-2723/short46.pdf.

LAHTI, LEO, JANI MARJANEN, HEGE ROIVAINEN AND MIKKO TOLONEN. 2019. "Bibliographic Data Science and the History of the Book (c. 1500–1800)." *Cataloging & Classification Quarterly* 57 (1): 5–23. https://doi.org/10.1080/01639374.2018.1543747.

LINDBERG, BO. 2006. *Den antika skevheten: Politiska ord och begrepp i det tidigmoderna Sverige*. Stockholm: Kungl. Vitterhets historie och antikvitets akademien.

LINDROTH, STEN. 1967. *Kungl. Svenska Vetenskapsakademiens historia 1739–1818. I: Tiden intill Wargentins död (1783)*. 2 vols. Stockholm: Kungl. Vetenskapsakademien.

MATHIS, RÉMI. 2009. "The STCN in a Global Perspective." *Aarboek Voor Nederlandse Boekgeschiedenis* 16: 37–44.

MCKITTERICK, DAVID. 2002. "University Printing at Oxford and Cambridge." In *The Cambridge History of the Book in Britain*, by Maureen Bell, edited by John Barnard and D. F. McKenzie, 1st ed., 189–205. Cambridge University Press. https://doi.org/10.1017/CHOL9780521661829.010.

MILLAR, ROBERT MCCOLL. 2005. *Language, Nation and Power: An Introduction*. 1st ed. 2005. London: Palgrave Macmillan UK.

PETTEGREE, ANDREW, AND ARTHUR DER WEDUWEN. 2018. "What Was Published in the Seventeenth-Century Dutch Republic?" *Livre – Revue Historique*, no. 1: 1–22.

POLLOCK, SHELDON. 2000. "Cosmopolitan and Vernacular in History." *Public Culture* 12 (3): 591–625. https://doi.org/10.1215/08992363-12-3-591.

———. 2006. *The Language of the Gods in the World of Men: Sanskrit, Culture, and Power in Premodern India*. 1st ed. ACLS Humanities E-Book. Berkeley: University of California Press.

PUTTER, AD, AND KEITH BUSBY. 2010. "Introduction: Medieval Francophonia." In *Medieval Multilingualism: The Francophone World and Its Neighbours*, edited by Christopher Kleinhenz and Keith Busby, 1–13. Turnhout: Brepols.

ROBERTS, JULIAN. 2002. "The Latin Trade." In *The Cambridge History of the Book in Britain*, edited by John Barnard and D. F. McKenzie with by Maureen Bell, 1st ed., 141–73. Cambridge University Press. https://doi.org/10.1017/CHOL9780521661829.008.

SHAW, DAVID. 2002. "French-Language Publishing in London to 1900." In *Foreign-Language Printing in London, 1500–1900*, edited by Barry Taylor, 101–22. Boston Spa: British Library.

SHER, RICHARD B. 2006. *The Enlightenment and the Book: Scottish Authors and Their Publishers in Eighteenth-Century Britain, Ireland, & America*. Chicago: University of Chicago Press.

SIMONS, LUDO. 2001. "The Fortunes and Misfortunes of Book Publishing in Flanders." In *The Bookshop of the World: The Role of the Low Countries in the Book-Trade, 1473–1941*, edited by Lotte Hellinga, Alastair Duke, Jacob Harskamp and Theo Hermans, 3. Goy-Houten: Hes & De Graaf.

SNOW, DON. 2013. "Towards a Theory of Vernacularisation: Insights from Written Chinese Vernaculars." *Journal of Multilingual and Multicultural Development* 34 (6): 597–610. https://doi.org/10.1080/01434632.2013.786082.

STONE PETERS, JULIE. 2000. *Theatre of the Book, 1480–1880: Print, Text, and Performance in Europe*. Oxford: Oxford University Press.

SUAREZ, MICHAEL F. 2009. "Towards a Bibliometric Analysis of the Surviving Record, 1701–1800." In *The Cambridge History of the Book in Britain*, edited by Michael F. Suarez and Michael L. Turner, 5:37–65. Cambridge: Cambridge University Press.

TAAVITSAINEN, IRMA, AND PAIVI PAHTA. 1998. "Vernacularisation of Medical Writing in English: A Corpus-Based Study of Scholasticism." *Early Science and Medicine* 3 (2): 157–85.

TOLONEN, MIKKO, MARK J. HILL, ALI ZEESHAN IJAZ, VILLE VAARA AND LEO LAHTI. 2021. "Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production." In *Data Visualization in Enlightenment Literature and Culture*, edited by Ileana Baird, 63–119. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-54913-8_3.

TOLONEN, MIKKO, LEO LAHTI, HEGE ROIVAINEN AND JANI MARJANEN. 2019. "A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52 (1): 57–78. https://doi.org/10.1080/01615440.2018.1526657.

TOLONEN, MIKKO, EETU MÄKELÄ AND LEO LAHTI. 2022. "The Anatomy of Eighteenth Century Collections Online (ECCO)." *Eighteenth-Century Studies* 56 (1): 95–123.

WRIGHT, ROGER. 1982. *Late Latin and Early Romance in Spain and Carolingian France*. Liverpool: Francis Cairns.

WRIGHT, SUE. 2004. *Language Policy and Language Planning: From Nationalism to Globalisation*. New York: Palgrave Macmillan.

3. The Rise of the Novel in Norway: Bibliographical Perspectives

Jens-Morten Hanssen

In Copenhagen, Denmark, the bookseller Johann Nicolai Lossius published a Danish edition of François Fénelon's *Les Aventures de Télémaque* in 1727–28. Fénelon's novel had become a literary sensation in Europe since its first publication in 1699 and would continue to resonate strongly throughout the era of the Enlightenment. According to Patrick Riley (1994), it became "the most read literary work in eighteenth-century France (after the Bible)", being printed in at least two hundred editions and translated into numerous languages in Europe and beyond (Riley 1994, xvi). The man behind the Danish version was Søren Dumetius, who was serving as the parish priest in Vestre Moland, a small rural community on the southern coast of Norway. I begin this chapter on a bold note, in that I claim that the publication of Dumetius' translation of Fénelon's work marks the introduction of the novel as a genre in Norway.

The full story of the rise of the novel in Norway is yet to be told. Previous accounts have left the impression that the emergence of the novel during the seventeenth and eighteenth centuries in nations such as Spain, France, England and Germany went entirely unnoticed in Norway, and that there were no Norwegian novels to speak of before the break-up of the union with Denmark in 1814. In this chapter, I argue that both premises are false and that they have produced a series of misconceptions. I propose to apply a transnational perspective

to the developments that led to the enormous popularity of the novel as a genre in Norway during the latter half of the nineteenth century. I present some preliminary findings of an integrated study of Norwegian novels and translated novels published in Norway during the period up until 1899. My point of departure is the creation of a comprehensive corpus of novels by drawing on data collected by multiple generations of librarians and bibliographers.

My approach is quantitative and based on bibliographical research rather than literary history. Previous accounts written by literary scholars have focused primarily on major works of the literary canon. Numerous studies set out to explore the novels of Maurits Hansen, Camilla Collett, Bjørnstjerne Bjørnson, Jonas Lie, Alexander Kielland, Arne Garborg, Amalie Skram and Knut Hamsun.[1] Existing accounts have also suffered from an overfocus on determining who came first. In this, there are three strands. First, it is commonplace in Norwegian literary historiography to assume that Camilla Collett wrote the first Norwegian novel: *The District Governor's Daughters* in 1854–55. In 1960, literary historian Francis Bull grudgingly admitted that attempts at novel writing were made before Collett, but that these have "sunk into oblivion". *The District Governor's Daughters* thus "stands before our consciousness as the first Norwegian novel".[2] Collett herself is known for having praised Maurits Hansen for being Norway's greatest prose writer, the same Hansen presumably being one of the forgotten novelists that Bull tacitly refers to (Fretheim 2006, 10). Literary scholars who have fought to rescue Collett's predecessors from oblivion have, for their part, referred to Maurits Hansen's *Othar of Brittany* from 1819 as the first Norwegian novel (Beyer and Moi 1990, 46; Hansen 1994, 6).

A third group of scholars claims that the introduction of the

1 See Dahl 1981, 204–15, for references.

2 "Andre norske romanforsøk fra tiden før hennes fremtreden er sunket i glemsel. 'Amtmandens Døttre' lever ennå, og står likefrem for bevisstheten som den første norske roman." (Bull 1960, 90). See Egeland 2020, 35, for references to other literary historians who corroborate Bull's view.

novel in Norway occurred during the period of Norway's union with Denmark. In his 2001 book about Christen Pram, Rolf Nyboe Nettum claims that Pram's four short novels, published in the Danish periodical *Minerva* in the 1880s, made him Norway's first novelist (Nettum 2001).[3] Pram's novels admittedly represent an early stage in the development that introduced the genre in Norway. However, Nettum's line of argument raises some methodological problems and he turns a blind eye to several literary events of crucial relevance which preceded the publication of Pram's short novels in *Minerva*.

Literary genres are a world without clear boundaries. Genres are unstable and non-fixed entities, and novels are no exception to this rule. Henry James is famous for having stated that the novel remains "the most independent, most elastic, most prodigious of literary forms" (James 1909, xxiii). In a similar vein, Mikhail Bakhtin later characterised the novel as an "ever-developing genre" or "a genre-in-the-making" and claimed that "the novel has no canon of its own, as do other genres; only individual examples of the novel are historically active, not a generic canon as such" (Bakhtin 1981, 3–11). The point of the present chapter is to propose and present a bibliographical take on the matter and, on that basis, hopefully shed some new light on the study of the novel from the perspective of eighteenth and nineteenth-century Norway. Organising knowledge about the emergence of literary genres is a task that librarians and bibliographers are regularly expected to perform. They would otherwise fail to do their job of providing the necessary building blocks of a research infrastructure for scholarly activities. Is it possible to combine the two perspectives? On the one hand, theoretical contributions indicating that the novel is a genre impossible to define and determine, and on the other, the daily business of librarians who are routinely expected to tell novels apart from non-novels.

3 Nettum presents three of Pram's short novels, "Jørgen, en Dosmers Levnedsbeskrivelse", "Hans Kruuskop", and "John Thral: Bidrag til Frihedens Historie", and briefly refers to a fourth one, "Den store Skiønhed." They were published anonymously in *Minerva* in 1786, 1787 and 1789, respectively (Nettum 2001, 169–213).

In its initial stage, work on the present chapter began with the creation of a corpus of novels. The corpus includes two thousand novels published in book form over 172 years, covering the period from 1727 to 1899. It consists of 708 Norwegian novels and 1,302 translated novels.[4] The dataset stems mainly from two sources: nineteenth century volumes of *Norsk Bog-Fortegnelse*, and the online bibliography Norwegian Books 1519–1850. The dataset is contained in an Excel spreadsheet available for download through the GitHub account of the Digital Humanities laboratory of the National Library of Norway.[5] In what follows, I will investigate the corpus from five perspectives:

1. examining the books from the point of view of genre label,
2. examining the distribution of the books according to publication year,
3. examining the distribution of the books according to publication place,
4. considering the distribution of works by author,
5. scrutinising the relationship between the Norwegian novels and the translated novels, and the distribution of the latter by source language.

But since the corpus consists of works in a contested genre, it seems only reasonable to begin with an account of my work with the sources.

Extracting Novels from the Norwegian National Bibliography

My main object was to examine a specific subset of the Norwegian National Bibliography, namely pre-1900 book publications that meet the criteria of a novel. I initially tried to extract the novels directly from the Norwegian National Bibliography by using a genre term query, but

4 I owe a debt of thanks to Mari Bentdal for creating the initial version of the dataset to my specifications in 2019–2020. I would also like to thank Ingvild Swane and Anne Kristin Lande for valuable help in complementing the dataset.

5 https://github.com/1420jeha/Norwegian-novels. Accessed 12 June 2024.

the query returned a list of only 211 titles.[6] The fact is that all two thousand novels in my corpus are also contained in the Norwegian National Bibliography, I just had to go elsewhere to find them. This is because the bibliography records lack sufficient metadata when it comes to genres of fictional literature. *Norsk Bog-Fortegnelse* and the bibliography Norwegian Books 1519–1850 came to my rescue in two regards. First, they cover the same areas of publication, in terms of geography, language and author nationality, as the Norwegian National Bibliography. Second, and more importantly, they provide more metadata.

The Norwegian National Bibliography is more than just an online bibliographic database produced and maintained by the National Library of Norway, it also encompasses its own pre-history and is the outcome of a process that began almost two centuries ago. *Norsk Bog-Fortegnelse* and the bibliography Norwegian Books 1519–1850 represent opposite ends of this development, with the year 1814 marking a crucial shift.

Initiated by the booksellers Magnus Wogelius Feilberg and Jens Landmark, and university librarian Martinus Nissen, who served as editor, the very first volume of *Norsk Bog-Fortegnelse* appeared in 1848. It contained a list and catalogue of books issued by Norwegian publishers during the period from 1814 to 1847 (Nissen 1848). The initial generation of bibliographers in Norway, represented first and foremost by Martinus Nissen, was profoundly influenced by Norway's struggle to become an independent nation. To him, 1814 did not mark the beginning of a new chapter in Norway's history, it was rather the beginning of a whole new book. In his line of thinking, Norway had no national literature of its own prior to 1814 (Nissen 1849, 178). Incidentally, Hans Olaf Hansen followed the same line of thought in his account of the literary history of Norway, *Den norske Literatur fra 1814 indtil vore Dage* (Hansen 1862).

6 I clicked on "*avansert søk*" ("advanced search") on this search page, https://bibsys-almaprimo.hosted.exlibrisgroup.com/primo-explore/search?vid=BIBLIOGRAFI&lang=no_NO (accessed 20 August 2024), chose "*sjanger og form*" ("genre and form") and searched for "*romaner*" ("novels") in the period until 31 December 1899.

Norsk Bog-Fortegnelse / *Norsk bokfortegnelse* continued to appear on a regular basis for more than a century and a half, eventually becoming an integral part of the Norwegian National Bibliography. A total of six volumes of *Norsk Bog-Fortegnelse* have been consulted for this chapter (Nissen 1848; Botten-Hansen 1870; Boeck 1877; Feilberg 1885; Feilberg 1892; Haffner 1902). But *Norsk Bog-Fortegnelse* paid no attention to the period before 1814. In 1898, the librarian and bibliographer Hjalmar Pettersen became head of the University of Oslo Library's Norwegian department, and he was instrumental in transcending the mindset of Martinus Nissen and his contemporaries. His four-volume *Bibliotheca Norvegica* (1899–1924) introduced a national bibliography concept that was updated and expanded in two important regards. First, Pettersen traced the history of book printing in Norway back to its very beginning in 1643, when the first printing office was established in Christiania (now Oslo) by Tyge Nielssøn, providing a descriptive catalogue of books printed in Norway between 1643 and 1813 (Pettersen 1899–1908). Second, he introduced the *Norvegica extranea* concept, which includes works published abroad in the following categories, Norwegian works, translated works by Norwegian authors, and works related to Norway, Norwegians or Norwegian subject matters. Norwegian author is defined as a person born in Norway, or born elsewhere but resident in Norway at the time of the publication (Pettersen 1908; Pettersen 1910; Pettersen 1917; Pettersen 1911–1918; Pettersen 1913–1924).

The online bibliography Norwegian Books 1519–1850 represents the most recent stage in this development. The bibliography was produced as part of the research project "Literary Citizens of the World", which was funded by the Research Council of Norway and hosted by the National Library of Norway between 2016 and 2021. In this project, book historians and literary scholars studied the history of the Norwegian book over a span of five centuries and from a transnational point of view (Bjørkøy et al. 2019; Hemstad et al. 2023; National Library of Norway 2024a). Currently, the bibliography Norwegian Books 1519–1850 contains references to 6,700 books printed or published in Norway and books that fall under the *Norvegica extranea*

concept as defined by Hjalmar Pettersen (National Library of Norway 2024b). The bibliography draws on a multitude of printed and digital sources, among them Pettersen's *Bibliotheca Norvegica*, the initial two volumes of *Norsk Bog-Fortegnelse* and Alma, which is the library system currently used by all Norwegian research libraries.

Genre Label

Previous studies of the novel have adequately revealed that it is a genre held low in esteem. The novel was a morally suspicious genre, associated with idleness, cheap sensationalism and female readership. The hierarchy of literary genres put forward by classical thinkers such as Aristotle, Horace, Boileau and Goethe excludes the novel. Even though he wrote one, Ludvig Holberg had a very negative opinion of novels. He considered the novel a "mere pastime for idle people".[7] In an article on the use of "novel" as a genre label during the nineteenth century, Marianne Egeland points to the fact that Camilla Collett deliberately avoided the label "novel" and chose instead the more neutral term "tale" as the subtitle for *The District Governor's Daughters* (Egeland 2020).

Throughout history, fictional works in the prose genre have appeared with a wide variety of genre terms – or with no genre term at all. There is no way to tell "tales" apart from "novels", and there is no reason to place too much emphasis on the lack of a genre label. Besides, there is a well-established tradition in scholarship on the novel not to be too narrow-minded and rigid in this regard. Note, for example, that none of the novels examined by Ian Watt in his classic study *The Rise of the Novel* appeared with the term "novel" on the title page (Watt 2000).[8]

More than a century before Watt conceived his book, Martinus Nissen and his successors recorded fictional works in the prose genre according to a similarly flexible approach. They placed "novels",

7 "*ørkesløse Mennesker for Tids-Fordriv*" (Holberg 1859, 9).

8 At the centre of attention in Watt's book are the novels of Daniel Defoe (*Robinson Crusoe* and *Moll Flanders*), Samuel Richardson (*Pamela* and *Clarissa*), and Henry Fielding (*Joseph Andrews* and *Tom Jones*).

"novellas", "tales" and "adventures" in the same list. In the sections of *Norsk Bog-Fortegnelse* listing titles of fictional works, they appear under subheadings such as "*Romaner og Fortællinger*" ("Novels and Tales") (Nissen 1848, 201–2; Botten-Hansen 1870, xxix–xxxi; Boeck 1877, xxix–xxxi; Feilberg 1885, xxvii–xxx; Feilberg 1892, 501–15), or "*Romaner, Noveller, Fortællinger og andre Prosa-Skrifter*" ("Novels, Novellas, Tales and other Prose Writings") (Haffner 1902, 576–83).

Genre terms are recorded from the bibliographic sources and cross-checked with the title page of all digitised books.[9] A total of 28 different genre labels are represented in my corpus. Counting no genre label as a deliberate choice, the four most represented categories are "*Fortelling*" ("Tale") at 35 per cent, no genre label at 32 per cent, "*Roman*" ("Novel") at 17 per cent, and "*Novelle*" ("Novella") at 5 per cent. Other genre terms include "*Skildring*" ("Depiction"), "*Eventyr*" ("Adventure"), "*Beretning*" ("Account"), "*Erindring*" ("Memoir"), variants of "*Historie*" such as "*Hverdagshistorie*" ("Tale of Everyday Life"), "*Kjærlighedshistorie*" ("Love Story"), "*Livshistorie*" ("Life Story"), and "*Indianerhistorie*" ("Indian Tale"), and variants of "*Billede*" such as "*Livsbillede*" ("Life Portrait"), "*Tidsbillede*" ("Portrait of an Era"), "*Hverdagsbillede*" ("Portrait of Everyday Life"), "*Nutidsbillede*" ("Portrait of Contemporary Times"), and "*Kristiania-Billede*" ("Kristiania Portrait").

Ian Watt ties the advent of the novel to realism, modern individualism and the rise of industrial capitalism in the early eighteenth century. With writers such as Daniel Defoe and Samuel Richardson, the novel appears in Watt's account as a new literary form that needs to be differentiated from the prose fiction of earlier times (Watt 2000). In contrast to Watt, Margaret Anne Doody claims that the novel has a continuous history of about two thousand years. In her book *The True Story of the Novel*, she criticises Watt for parochialism

9 Currently, 592 out of 708 Norwegian novels (84 per cent) and 1,139 out of 1,302 translated novels (87 per cent) have been digitised, cf. https://github.com/1420jeha/Norwegian-novels. Accessed 12 June 2024.

and chauvinism in treating the novel as if it were essentially English, thereby ignoring the Spanish novel of the sixteenth and seventeenth centuries, the French novel of the seventeenth century and the novel in antiquity. The main bulk of Doody's book is a study of four ancient novels (Doody 1996). Mikhail Bakhtin assumes a middle position in that he, like Watt, acknowledges that the novel is a distinctly modern genre. But, like Doody, he traces the roots of the novel back to Greek Antiquity and the Renaissance, more precisely to Menippean satire, serio-comic literature, parody and folklore. With the arrival of the novel, Bakhtin points out, a genre was introduced that was in "living contact" with the "unfinished, still-evolving contemporary reality (the openended present)". He sees the novel as an "ever-developing genre" and a "vanguard of change", whereas, for example, the epic and the tragedy appear as "fixed forms" (Bakhtin 1981, 3–40).

At first sight, genre terms or subtitles such as "Story of Everyday Life", "Life Story", "Portrait of an Era", and "Portrait of Contemporary Times" would seem to resonate well with and support Ian Watt's realism approach. However, the overall distribution of such subtitles is relatively low, and they emerge at a late hour. Pre-1850, there are very few, if any, instances of them. In post-1850 novels, they apply only to some extent.

Distribution by Year of Publication

The bibliography Norwegian Books 1519–1850 documents a history of book production that starts in 1519, when two liturgical books commissioned by the Archbishop of Nidaros, Erik Valkendorf, arrived by boat in Trondheim. According to the bibliography, no book was published during the sixteenth and seventeenth centuries that fits the criteria of a novel. I have filtered through the bibliography using search terms such as "*Roman*" ("Novel"), "*Fortelling*" ("Tale"), "*Novelle*" ("Novella"), and "*Fabel*" ("Fable"). When in doubt, I followed Margaret Doody's definition of a novel. According to her definition, a

work is a novel "if it is fictional, if it is in prose, and if it is of a certain length" (Doody 1996, 16).[10]

In the introduction, I claim that the earliest example in my sample, Søren Dumetius' translation of Fénelon's *Les Aventures de Télémaque*, which was published in Copenhagen in four volumes between 1727 and 1728, marks the introduction of the novel as a genre in Norway, and I do so with reference to the bibliographical record. Dumetius was born in Denmark but spent his career in Norway (Sundtoft 1983, 139–40). He is included in Hjalmar Pettersen's index of pre-1814 Norwegian authors in volume three of *Bibliotheca Norvegica* (Pettersen 1911–18, xx), and the Copenhagen edition of his translation of Fénelon's novel is listed in Pettersen's list of pre-1814 Norwegian books in the same volume (Pettersen 1911–1918, 130–1). Metadata provided by Pettersen and digitised copies of the Copenhagen edition in the collection of the National Library of Norway form the basis for entries in the bibliography Norwegian Books 1519–1850.[11]

So it seems that the novel came to the fore in Norway during the eighteenth century, although the volume of such publications remained modest for a long time. Figure 1 displays the annual distribution of novels by year of publication. There are three different stages in the development: The period until the break-up of the union with Denmark in 1814 saw the publication of only 33 novels. Then followed a period that saw a gradual but still moderate increase in the number of novels; 308 novels were published during the period 1814 to 1869. Finally, the graph shows a significant increase during the last three decades of the nineteenth century; 1,667 novels are recorded during the

10 The novels run to between 31 and 995 pages. The shortest novel is Knut Hamsun's 1877 debut, *Den Gaadefulde: En Kjærlighedshistorie*. The bibliography Norwegian Books 1519–1850 is restricted to books with 48 pages or more.

11 Cf. https://www.nb.no/bibliografi/nor1519/show?id=987cc17ad7201a94f-22f7630523f9f57&bibliography=nor1519 (Vol. 1); https://www.nb.no/bibliografi/nor1519/show?id=94f7d3b7fdab25321a7982b9e1641aaa&bibliography=nor1519 (Vol. 2); https://www.nb.no/bibliografi/nor1519/show?id=766b642c497f1a8abc0331793f17d53a&bibliography=nor1519 (Vol. 3); https://www.nb.no/bibliografi/nor1519/show?id=ad6e58a1bd223c751020d2a9616a2138&bibliography=nor1519 (Vol. 4). Accessed 15 July 2021.

period from 1870 to 1899. Calculated in percentages, the annual distribution across the corpus shows that 83 per cent of the novels were published during the final three decades of the nineteenth century.

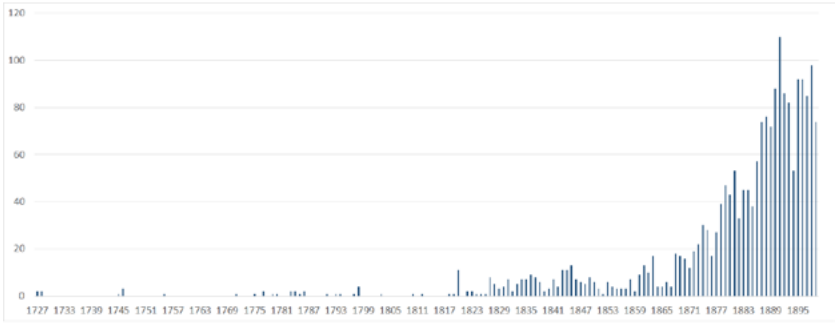

Figure 1: Annual distribution of novels 1727–1899

I will now shed light on the rise and historical development of the novel in Norway, using Figure 1 as my point of departure. The most pressing question is to clarify what explains the steep increase in the number of publications during the latter half of the nineteenth century. There are at least three extraliterary factors that need to be taken into account. First, the rapidly growing popularity of the novel must be considered in the context of the general development of book printing. There was a direct link between the rise of the novel and the nineteenth century industrialisation of book printing. Due to technical developments within the printing industry, the replacement of the hand-press by a high-speed steam-driven press and the invention of the linotype machine, among other things, the late nineteenth century saw an immense increase in the number of books printed compared with earlier times. The Norwegian National Bibliography indicates

that the number of books published in Norway during 1850–1899 was 26 times higher than the equivalent number during 1700–1749.[12]

Second, the increased popularity of the novel in the late nineteenth century must be considered in the context of the general development of print media. Scholars such as Aina Nøding and Jon Haarberg have uncovered that a significant number of novels were published in serialised form in Norwegian newspapers and journals during the eighteenth and nineteenth centuries (Nøding 2010; Nøding 2017; Haarberg 2019). The industrialisation of printing also paved the way for the establishment of many newspapers, not least in the provinces, which in turn led to a diversification of the field of cultural production. Newspapers were an arena for marketing, literary criticism and the publication of serialised novels. The late nineteenth century saw the development of a media-driven mass culture. The novel played a noticeable part in this development and became a mass cultural phenomenon.

A third, related, circumstance ought to be considered as well. The novel became the object of mass reading. The historian Rolf Engelsing has hypothesised that a reading revolution took place at the end of the eighteenth century. Intensive reading of a small collective canon of texts was replaced by an eagerness to consume new and varied reading materials. As a result, extensive reading habits became the dominant cultural norm (Wittmann 1999). Changes in the cultural practice of reading had a profound impact on the book trade. Capitalist principles were introduced, with a strong trend toward sales-oriented book production, new forms of advertising and pursuit of new markets. There was an enormous increase in literacy in the population during the eighteenth and nineteenth centuries, and the book market now tended to address a reading public that was unlimited,

12 According to Alma Library Services reports, the number of published books was 1,243 during 1700–49, whereas the equivalent number was 32,670 during 1850–99, in other words 26.3 times higher. Many thanks to my National Library colleague Lennart Nilsen for help in generating the reports.

heterogeneous and anonymous.[13] *Belles-lettres* or fictional literature stand as the epitome of this development, with the novel appearing as the most prominent literary genre.

Distribution by Place of Publication

The expanded national bibliography concept proposed by Hjalmar Pettersen implied that works by Norwegian authors should be systematically registered, regardless of where they had been published. His four-volume *Bibliotheca Norvegica* (Pettersen 1899–1908) led to the inclusion of works by Norwegians who had emigrated to the United States during the nineteenth century and published books there, such as Lars Andreas Stenholt and Kristofer Janson, or, more prominently, Norwegian works published in Copenhagen.

For many centuries, the history of printing and publishing in Norway was marked by Danish hegemony. The first Norwegian printing office was established in Christiania (now Oslo) in 1643 by Tyge Nielssøn, who was initially based in Copenhagen. A century and a half later, the number of printing enterprises could still be counted on the fingers of one hand. Martinus Nissen notes that by 1814 there were only six printing offices across Norway, three in Christiania, and one each in Kristiansand, Bergen and Trondheim. By 1849, however, the number had risen to 53 printing offices in 30 cities and towns (Nissen 1849, 203). The last quarter of the century saw a doubling of that number again (Jacobsen 1983, 143–236). Norway was late to the party for two interconnected reasons: the market situation and Danish hegemony. Printers who wanted to establish themselves in Norway faced harsh competition from Danish printing companies that also distributed their books in Norway. Norwegian-born authors such as Johan Herman Wessel, Christen Pram, Magdalene Sophie Buchholm and Edvard Storm, who pursued a career within the literary field prior to

13 Rudolf Schenda has estimated that in 1770, the potential reading public was 15 per cent of the total European population, rising to 25 per cent in 1800, 75 per cent in 1870, and reaching a dizzying 90 per cent in 1900 (Schenda 1977, 444–5).

1814, published most of their writings in Copenhagen. The cultural dominance exercised by Denmark lingered on after 1814. Harald L. Tveterås notes that during the period from 1860 to 1890, commonly referred to as the Golden Age of Norwegian literature, around 90 Norwegian writers, most of whom were writers of fictional works, still published their books in Denmark (Tveterås 1964, 409).

But my dataset brings important nuances to the standard narrative of Danish hegemony. The corpus includes novels that appeared in a total of 49 locations across four nations. The great majority of editions, 88 per cent, were published in Norway, 9 per cent were published in Denmark, 2 per cent in the United States, while one novel was published in Paris, France. In addition, 14 novels, 1 per cent, are registered with an unknown place of publication.[14] Unsurprisingly, the major cities in Norway and Denmark, such as Christiania, Bergen, Trondheim and Copenhagen, dominate the picture. However, Christiania outnumbers the other publication places by far. In fact, Norway's capital city saw the publication of seven times as many editions as Copenhagen.[15] The distribution of editions by place of publication inevitably calls into question the notion of a Copenhagen hegemony. The group of Norwegian writers who chose to publish their works in the Danish capital includes distinguished figures such as Camilla Collett, Bjørnstjerne Bjørnson, Alexander Kielland, Jonas Lie, Kristian Elster, Amalie Skram, Knut Hamsun, Thomas P. Krag, not to mention Henrik Ibsen, and they were, without exception, associated with one specific publishing house in Copenhagen, the Gyldendalske Boghandels Forlag under the leadership of Frederik Hegel, who

14 Norway is represented with a good 38 publication places, the United States with seven – Chicago, Minneapolis, Decorah, La Crosse, Grand Forks, Fergus Falls and Orfordville, whereas Denmark features only two, Copenhagen and Odense. Hans Jæger's *Syk kjærlihet*, which was published in Paris in 1893, accounts for the single instance of a Norwegian novel with a French publisher.

15 The top ten list of publication places, with the number of editions in parenthesis, includes Christiania (1,298), Bergen (225), Copenhagen (186), Trondheim (45), Horten (21), Chicago (20), Risør (19), Stavanger (17), Høvik (17) and Kristiansand (14), cf. https://github.com/1420jeha/Norwegian-novels. Accessed 12 June 2024.

occupies a prominent place in history of the Norwegian literature in the late nineteenth century. Norway's leading writers chose Hegel because he was the foremost publisher of his time, not only in Denmark, but across Scandinavia. He offered better publishing terms and secured a far wider distribution than his Norwegian counterparts. Furthermore, publishing with the Gyldendalske Boghandels Forlag was associated with high cultural prestige, securing recognition and awareness from readers and critics alike that publishers in Norway simply could not compete with. Obtaining a publishing contract with Hegel, however, was reserved for the few. The vast majority of Norwegian prose writers turned to publishers based in Christiania to get their novels printed. In view of the evidence contained in my dataset, it is difficult to maintain the often-made claim that most books circulating in Norway during the nineteenth century were books printed in Denmark (Fulsås and Rem 2018, 30; Eide 2013, 110).

Distribution by Author

Engelsing's hypothesis about the transition from intensive to extensive reading habits is not necessarily easy to corroborate. How can one tell whether it applies to eighteenth and nineteenth-century Norway? I will now examine the corpus of novels from the point of view of author attribution. Next to title, publisher and publication date, author is a core element and the foremost guiding principle in bibliography. Historically, a novel's association to its author was not a straightforward one, due to the genre's low esteem. In a bibliographic study by Silas Paul Jones, only 134 out of 946 registered French novels in the period 1700–1750 appeared with the name of the author printed on the title page (Jones 1939, 14). A good century later, Camilla Collett prudently published her spectacular debut novel anonymously. Jorunn Hareide lists 13 Norwegian female prose writers who made their debut in the period between 1850 and 1870, and notes that only two of them published under their own name, Marie Colban and Magdalene Thoresen (Hareide 1988, 50).

Authors who wanted to hide their identity from the public had two options, anonymity or pseudonymity. My dataset shows that the

latter was more usual, and that, contrary to what might be expected, the use of a pseudonym is evenly distributed between the sexes. 390 books contained in the corpus, which amounts to 19 per cent, appeared under a pseudonym. Half of them were written by a female author, while the other half had a male author. In Norway, Hanna Winsnes and Elisabeth Schøyen published books under a male pen name, Hugo Schwartz and Paul Agathon, respectively, thereby following the example of their more famous peers in France and England, George Sand and George Elliot. It was, however, more common to conceal the author's identity by means of initials, as did, for instance, Marie Wexelsen and Elisabeth Lampe with I.L. and A.B., respectively. The gender issue was clearly the main concern for female authors, as writing as a profession was considered an activity unsuitable for women. Male peers who chose to publish under a pseudonym did so primarily for other reasons, the most important of which was to avoid the stigma attached to writing for a mass audience.

In a study of the translated novel in nineteenth-century Denmark, Erland Munch-Petersen applies a bibliographical-statistical method to define a body of foreign authors whose novels were the object of mass reading in Denmark in 1800–1870, a period which he calls "the century of the novel". Munch-Petersen extracts a total of 26 novelists, mainly from Britain, Germany and France, who are represented with 50 or more editions in *Bibliotheca Danica* and *Dansk Bogfortegnelse* (Munch-Petersen 1978). He argues vehemently against the now largely discarded distinction between low and high literature and suggests a methodological approach based on a combination of theoretical perspectives from the sociology of literature and empirical data about the production, distribution and use of literature (Munch-Petersen 1970; Munch-Petersen 1978). I am indeed inspired by Munch-Petersen's work and have applied a similar method here, only have I expanded the view by integrating both Norwegian novels and translated ones. The approach paves the way for an exploration of literature that has largely escaped the attention of literary historians and literary critics.

My corpus contains works that are attributed to 659 unique

authors. The gender distribution shows that 71 per cent were male authors and 23 per cent were female authors, while 6 per cent were unidentified authors of unknown gender. The Norwegian Authority File for Persons and Corporate Bodies has been consulted in the author identification process (Bibliotekutvikling 2023). On average, each author is attributed to three editions. Sorting authors according to the number of books renders a list of 12 authors represented with 15 or more author attributions and who account for a total of 283 editions.

Table 1: Authors with 15 or more author attributions.

Author	**Number of editions**	**Nationality**
Wilhelm Fricke	67	German
Rudolf Muus	26	Norwegian
Jon Flatabø	24	Norwegian
Nikolaus Fries	23	German
Edward Bulwer-Lytton	21	English
Franz Pistorius	20	German
Walter Scott	19	Scottish
Sarah Smith	18	English
Johan Storm Wang	18	Norwegian
Jonas Lie	16	Norwegian
Elisabeth Bürstenbinder	16	German
Victor Hugo	15	French
	283	

The 12 novelists who appear as the most prominent exponents of the mass-market author in pre-1900's Norway may be divided into three groups. The first group includes Wilhelm Fricke, Rudolf Muus, Jon Flatabø and Franz Pistorius. All four of these are strongly associated with scrupulous, commercially oriented publishers, who marketed themselves exclusively toward a mass public. Paradoxically, the most widely published author, the German Wilhelm Fricke, is the person we

know the least about. He consistently published under the pseudonym W. Frey, and it seems that it was the bibliographer Fritz Meyen who had the merit of first identifying him and finding his birth and death dates (Meyen 1942, 29–34). Most of the Norwegian editions of Fricke's books were printed by the publishing houses of Waldemar Kriedt and his younger brother Sophus Kriedt (33 editions) and Peder Omtvedt (29 editions). The same goes for Franz Pistorius, whose editions were also printed by the Kriedt brothers (13 editions) and Omtvedt (6 editions). Furthermore, these editions reveal the dominance of a particular type of publication, namely the pamphlet. The works of Fricke and Pistorius were exclusively published in pamphlet form. In other words, they were printed without book covers and were of restricted length, usually 48 pages (Parelius 1987). The two of them created, or rather churned out, an enormous number of stories following a very recognisable, formulaic pattern. They were strongly plot-driven, set in remote, exotic places such as in the "African Wilderness", on "the Prairie", "the Pacific Islands", or a "cliff island in the Atlantic Ocean". The stories were populated with very stereotypical characters and always ended on a happy note (Frey 1886, 1890a, 1890b, 1891; Pistorius 1895).

In the late nineteenth and early twentieth centuries, Rudolf Muus, Jon Flatabø and Karen Sundt were the most widely read Norwegian authors in terms of popular literature. Sigurd Heiestad assumes that Muus wrote no less than 50 novels, some of which were up to 1,600 pages long, several hundred books in other genres, in addition to numerous translations of German and English novels and Native American stories (Heiestad 1946, 95). He juggled with more than 50 pseudonyms and used them strategically to play competing publishers off against each other. From the 1890s onwards, when he came to the fore with a series of Christiania novels, Muus played a double game by publishing both with the publisher Johan Henrik Kinhult under the pseudonym Julius and, in parallel, with the publisher Sophus Kriedt under the pseudonym Vilhelm or Wilhelm (Knudsen 1995, 33).

Georg Lukács famously defined the novel as "the epic of a world that has been abandoned by God" (Lukacs 1971, 88). The works of the authors belonging to the second group, Nikolaus Fries and Sarah

Smith, may suggest otherwise, as they were known for combining strong religious conviction with fictional writing. The Norwegian editions of Fries made no secret of the author's main profession, as most of them came with the discriptive "pastor" or "pastor in Heiligenstedten", a municipality in Northern Germany, printed beneath his name. The writing of fictional stories was clearly a part-time job, or rather a part of his efforts to spread the Gospel (Fries 1874, 1877). Sarah Smith, who wrote under the pseudonym Hesba Stretton, was an extremely popular English author of religious books for children and young people. Her book *Jessica's First Prayer*, which appeared in a Norwegian translation by Emma and Augusta Hagerup in 1868, allegedly sold a million and a half copies, nearly ten times as many as Lewis Carroll's *Alice in Wonderland* (Stretton 1868; Alderson 1974, 268).

The authors belonging to the first and second groups are largely neglected or appear only at the margin in conventional accounts of literary history. However, the authors in the third group prove that becoming the object of mass reading does not necessarily rule out lasting recognition from literary historians. Victor Hugo is still regarded as one of the greatest French writers of all time; Walter Scott's major impact on both Scottish and English literature is hard to deny, not to mention his significance in pioneering the historical novel; Jonas Lie is still counted among the Four Greats of Norwegian late nineteenth-century literature, though he is constantly in danger of being overshadowed by his more famous contemporaries Ibsen and Bjørnson. The third group, which also includes Edward Bulwer-Lytton, Johan Storm Wang and the German writer Elisabeth Bürstenbinder, who wrote under the pen name Ernst Werner, is a catch-all category, as it is admittedly difficult to pinpoint what they have in common, apart from the fact that they wrote novels that were widely read during the nineteenth century. There is, however, a sense that the authors of all three categories are positioned at a distance from the literary avant-garde in the Bourdieuian sense of the term, as they do not tend to engage in formal and aesthetic experimentation or raise demands in favour of the *l'art pour l'art* concept.

The Novel and World Literature

I now turn to the question of how translation flows and spheres of cultural influence affected the rise of the novel. The novel was introduced in Norway by way of translation, and the majority of the works in my corpus are in fact translated novels. The ratio of translated novels to novels originally written in Norwegian, or Dano-Norwegian, during the entire period under scrutiny is 1.83:1. However, this figure varies over time, across the three stages mentioned above. In the period 1727–1813, the ratio of translated novels to Norwegian novels was 5.0:1, in 1814–69 it was 2.17:1, whereas in 1870–99 it was 1.74:1. This shows that the share of translated novels dropped over time.

In *Atlas of the European Novel 1800–1900*, Franco Moretti draws on a quantitative study of the rise of the novel in Europe, which indicates that the development proceeded in three stages or "take-offs", the first one around 1720–50 with France, Britain and, a little later, Germany constituting the "core". He further claims that "the novel is *the most centralized of all literary genres*" (Moretti 1998, 165–74; italics in the original). At first sight, my corpus seems to support Moretti's argument. The 1,302 non-Norwegian novels were translated from 14 different source languages, with a very uneven distribution. 87 per cent of the novels were translated from either English, German or French.[16] Until the 1830s, these three languages dominate entirely, with only one exception, Latin.

The cultural influence of French, German and English literature marked the introduction of the novel in Norway in consecutive "waves", the initial one being a French wave with Dumetius' translation of Fénelon and Hans Paus' translation of Alain-René Lesage's novel *Le Diable Boiteux*, published in two volumes in 1746, as the most prominent examples. A German wave followed in the initial decades of

16 The 14 languages, with the corresponding number of translated novels in parenthesis, are as follows: English (527), German (423), French (181), Swedish (41), Russian (26), Danish (25), Hungarian (8), Dutch (5), Spanish (5), Low German (4), Italian (4), Finnish (4), Latin (2) and Ukrainian (1). 45 translated novels have an unknown source language, cf. https://github.com/1420jeha/Norwegian-novels. Accessed 12 June 2024.

the nineteenth century, with translations of German novelists such as August Lafontaine, Novalis (*Heinrich von Ofterdingen*), Goethe *(Die Leiden des jungen Werthers*), and Caroline Pichler *(Agathokles*). In the 1820s, Walter Scott inaugurated a strong English wave that would prevail throughout the century, with multiple editions of best-selling authors such as James Fenimore Cooper, Edward Bulwer-Lytton, William Harrison Ainsworth, Charles Dickens and Mary Elizabeth Braddon.

Moretti's core-periphery model seems reasonable and productive, but I would still caution against his line of reasoning. In his 2000 article "Conjectures on World Literature", he reshapes the argument into a law of literary evolution: "in cultures that belong to the periphery of the literary system …, the modern novel first arises not as an autonomous development but as a compromise between a western formal influence (usually French or English) and local materials" (Moretti 2000, 58). Doody accuses Watt of chauvinism, and it is doubtful whether Moretti escapes the thrust of her criticism. It could be argued that Moretti only takes Watt's view one stage further by assuming that the novel is essentially Anglo-French.

Moreover, instances of novels originally written in Latin, of which there are two in the corpus, further complicate the picture. In 1746, a three-volume Danish edition of John Barclay's novel *Argenis*, originally published in Latin in 1621, appeared in Copenhagen. The Danish translation was by the Norwegian lawyer and historian Hans Paus. It is, once again, the work of a Norwegian translator that secures the Copenhagen edition a place in the annals of the Norwegian book (Pettersen 1911–18, lv–120). In 1741, Ludvig Holberg, clearly a citizen of a peripheral corner of literary Europe, published his only novel, *Nicolai Klimii Iter Subterraneum* (*Niels Klim*), in Latin. Holberg may have had several reasons for choosing to write his satirical travelogue in Latin – to avoid Danish censorship, to write a Menippean satire full of quotations from classical texts, or to secure wide international dissemination. This latter may actually have been the most decisive consideration (Jansen 1974, 99). Whatever the reason, the strategy proved successful, although it was not the Latin edition but translations that opened the

door to the public at large, as Thomas Velle has pointed out with reference to Karen Skovgaard-Petersen (Velle 2019). Within a year of its initial publication in Leipzig and Copenhagen, a German, French, Dutch, Danish and English edition of *Niels Klim* appeared, with Swedish, Russian and Hungarian editions following soon after. The Holberg bibliography of Holger Ehrencron-Müller, prepared for his 12-volume encyclopaedia of Danish, Norwegian and Icelandic authors, lists 62 editions of *Niels Klim* in 12 languages (Ehrencron-Müller 1935, 213–326).

Holberg's *Niels Klim* assumes an ambiguous position in literary historiography and in the history of the novel as a genre. In conventional accounts of the literary history of Denmark and Norway, the novel is usually referred to by its Danish title and is mostly quoted in Danish. Although *Niels Klim* was undeniably the first-ever novel written by a Norwegian-born author, it is mostly neglected or only mentioned in passing in historical studies of Norwegian prose fiction because it was first written in Latin. More importantly, Holberg demonstrates the fruitfulness of an approach that combines bibliography, book history and world literature studies. To fully grasp the significance of *Niels Klim*, the work must be considered in light of its wide international distribution and large number of translations – in other words in the context of its status as world literature in Goethe's sense of the term. Holberg's novel was never restricted to classical studies and latinists but belongs in the category of unspecified editions and translations (Velle 2019). The bibliographical record produced by trained librarians provides the best knowledge source for a detailed account of those very same editions and translations.

Conclusion

In this chapter, I have presented some preliminary findings of a quantitative study of the rise of the novel in eighteenth and nineteenth-century Norway. I have applied a quantitative, bibliographical and transnational approach, taking as my point of departure a corpus of two thousand novels which integrates Norwegian novels and translated novels published in Norway or published abroad with a Norwegian

translator. Drawn from bibliographical records produced by multiple generations of trained bibliographers and librarians, the corpus includes metadata such as author name, title and subtitle, genre designation, place and year of publication, and source language for the translated novels.

The creation of a corpus of eighteenth and nineteenth-century novels is not a straightforward task, as the novel is a contested genre with a dubious reputation. Only 17 per cent of the books appeared with the genre label "novel" on the title page. Despite this, there is a long tradition in novelistic studies to consider the novel in the context of works that carried related genre terms such as "tale" and "novella", or works of prose fiction that were published without a genre term. Bibliographers have dealt with the genre problem in a similarly flexible way.

In previous accounts, Maurits Hansen and Camilla Collett have been heralded as the pioneers of the Norwegian novel. I argue instead that the introduction of the novel in Norway occurred during the 1720s, with the publication of Søren Dumetius' translation of François Fénelon's extremely popular novel *Les Aventures de Télémaque*. I demonstrate how an integrated study of Norwegian and translated novels, combined with a transnational perspective, opens fresh lines of inquiry and sheds new light on the rise of the novel as a genre. The great majority of the non-Norwegian novels were translated from English, French or German. But the two instances of novels translated from Latin, *Niels Klim* and John Barclay's *Argenis*, seem to caution against the view of scholars focusing too narrowly on the canon of English and French novels. A more fruitful approach is to acknowledge that literary genres, such as the novel, own no nationality and are not restricted to specific languages. History reveals that Holberg's *Niels Klim* equalled *Les Aventures de Télémaque* in terms of the number of editions and translations. In conventional accounts, however, the tendency has been to disregard Holberg's novel because it was originally written in Latin.

The novel rose to prominence and achieved widespread popularity in the latter half of the nineteenth century. I propose to consider this course of events in the context of technological developments in

and the growing industrialisation of book printing and the general development of print media. More so than the other literary genres, the novel was strongly influenced by and closely linked to the growth of mass culture. A close scrutiny of the distribution of works by author revealed a group of 12 Norwegian and non-Norwegian writers who are represented with 15 or more editions, some of whom were known for producing prose fiction in pamphlet form. Such works – which have attracted next to no attention from literary historians – were manufactured at high speed, distributed in large print runs at a cheap price and targeted at a mass audience. They illustrate that the novel eventually became the object of mass reading and came to epitomise the era of the mass-market author. Most of these pamphlets were printed by commercially oriented publishers in Christiania, and my corpus raises substantial doubts about the commonly held view that Danish publishers dominated the distribution of books in Norway throughout the nineteenth century. The overwhelming majority of editions studied here were printed and published in Christiania, and a publisher based in Norway's capital was the main option for Norwegian prose writers, unless you were among the lucky few who were admitted into the Danish publisher Frederik Hegel's exclusive circle.

References

ALDERSON, BRIAN. 1974. "Tracts, Reward and Fairies: The Victorian Contribution to Children's Literature." In *Essays in the History of Publishing in Celebration of the 250th Anniversary of the House of Longman, 1724–1974*, edited by Asa Briggs, 245–282. London: Longman.

BAKHTIN, M. M. 1981. *The Dialogic Imagination: Four Essays*. Austin: University of Texas Press.

BEYER, EDVARD, AND MORTEN MOI. 1990. *1770–1848*. Vol. 1 of *Norsk litteraturkritikks historie 1770–1940*, edited by Edvard Beyer, Irene Iversen, Arild Linneberg and Morten Moi. Oslo: Universitetsforlaget.

BIBLIOTEKUTVIKLING. 2023. "Felles autoritetsregister for personer og korporasjoner." Accessed 11 July 2023. https://bibliotekutvikling.no/kunnskapsorganisering/vokabularer-utkast/felles-autoritetsregister-for-personer-og-korporasjoner/.

BJØRKØY, AASTA M.B., RUTH HEMSTAD, AINA NØDING AND ANNE BIRGITTE RØNNING, EDS. 2019. *Litterære verdensborgere: Transnasjonale perspektiver på norsk bokhistorie 1519–1850*. Oslo: Nasjonalbiblioteket.

BOECK, THORVALD. 1877. *Norsk Bog-Fortegnelse. 1866–1872*. Christiania: Den norske Boghandlerforenings Forlag.

BOTTEN-HANSEN, PAUL, AND SIEGWART PETERSEN. 1870. *Norsk Bog-Fortegnelse. 1848–1865*. Christiania: Den norske Boghandlerforenings Forlag.

BULL, FRANCIS. 1960. *Norges litteratur fra februarrevolusjonen til første verdenskrig*. Vol. 4, Part 1 of *Norsk litteraturhistorie*, edited by Francis Bull, Fredrik Paasche, A. H. Winsnes and Philip Houm. Oslo: Aschehoug.

DAHL, WILLY. 1981. *Stil og struktur: Utviklingslinjer i norsk prosa gjennom 150 år*. Bergen: Eide.

DOODY, MARGARET ANNE. 1996. *The True Story of the Novel*. New Brunswick, NJ: Rutgers University Press.

EGELAND, MARIANNE. 2020. "Om bruken av romanbetegnelsen på 1800-tallet: Hvorfor kalte ikke Camilla Collett *Amtmandens Døttre* for 'roman'?" *Edda* 107 (1): 34–47.

EHRENCRON-MÜLLER, H. 1935. *Bibliografi over Holbergs Skrifter*, Vol. 12, part 3 of *Forfatterlexikon omfattende Danmark, Norge og Island indtil 1814*. Copenhagen: Aschehoug.

EIDE, ELISABETH S. 2013. *Bøker i Norge: Boksamlinger, leseselskap og bibliotek på 1800-tallet*. Oslo: Pax.

FEILBERG, M. W. 1885. *Norsk Bog-Fortegnelse. 1873–1882*. Kristiania: Den norske Boghandlerforenings Forlag.

———. 1892. *Norsk Bogfortegnelse. 1883–1890*. Kristiania: Den norske Boghandlerforenings Forlag.

FRETHEIM, ARVE. 2006. *Livets kolde prosa: Maurits Hansen og hans samtid*. Oslo: Aschehoug.

FREY, W. [WILHELM FRICKE]. 1886. *Sømandsbruden: En Fortælling fra Sydhavsøerne*. Kristiania: Sophus Kriedt.

———. 1890a. *Kahitschinehs Død: En Fortælling fra Afrikas Vildnis*. Kristiania: P. Omtvedt.

———. 1890b. *Indianerbruden: En Fortælling fra Prærien*. Kristiania: Sophus Kriedt.

———. 1891. *Overfaldet paa Prærien: En Fortælling*. Kristiania: P. Omtvedt.

FRIES, N. 1874. *Byg ikke paa Sand: En Fortælling sigtende til det første Bud*. Translated by P. B. Trondheim: Gram.

———. 1877. *Guds Stad og dens Kilde (Psalm. 46, 5): En Fortælling i to Afsnit*. Translated by J. H. Kristiania: Malling.

FULSÅS, NARVE, AND TORE REM. 2018. *Ibsen, Scandinavia and the Making of a World Drama*. Cambridge: Cambridge University Press.

HAARBERG, JON. 2019. "Da onkel Tom kom til Norge: Oversatte romaner i den nye nasjonalstatens fem første tiår (1814–1857)." *Norsk litteraturvitenskapelig tidsskrift* 22 (2): 182–201.

HAFFNER, H. J. 1902. *Norsk Bogfortegnelse. 1891–1900*. Kristiania: Den norske Boghandlerforenings Forlag.

HANSEN, HANS OLAF. 1862. *Den norske Literatur fra 1814 indtil vore Dage: Et Bidrag til en norsk Literaturhistorie*. Copenhagen: Fr. Wøldike.

HANSEN, MAURITS. 1994. *Othar av Bretagne: Et riddereventyr*. Oslo: Bokvennen.

HAREIDE, JORUNN. 1988. "Den norske romanen vokser fram." In *1600–1900*. Vol. 1 of *Norsk kvinnelitteraturhistorie*, edited by Irene Engelstad, Jorunn Hareide, Irene Iversen, Torill Steinfeld and Janneken Øverland, 49–62. Oslo: Pax.

HEIESTAD, SIGURD. 1946. *Av folkelesningens saga*. Oslo: Halvorsen.

HEMSTAD, RUTH, JANICKE S. KAASA, ELLEN KREFTING AND AINA NØDING, EDS. 2023. *Literary Citizenship in Scandinavia in the Long Eighteenth Century*. Woodbridge: Boydell & Brewer.

HOLBERG, LUDVIG. 1859. *Moralske Tanker*. Copenhagen: Schønberg.

JACOBSEN, GUNNAR. 1983. *Norske boktrykkere og trykkerier gjennom fire århundrer 1640–1940*. Oslo: Den norske boktrykkerforening.

JAMES, HENRY. 1909. *The Ambassadors*. Vol. 21 of *The Novels and Tales of Henry James*. New York: Charles Scribner's Sons.

JANSEN, F. J. BILLESKOV. 1974. *Ludvig Holberg*. New York: Twayne Publishers.

JONES, SILAS PAUL. 1939. *A List of French Prose Fiction from 1700 to 1750*. New York: H.W. Wilson Co.

KNUDSEN, HANS P.S. 1995. *Folkeskribenten Rudolf Muus og hans forleggere*. Oslo: Bladkompaniet.

LUKÁCS, GEORG. 1971. *The Theory of the Novel: A Historico-Philosophical Essay on the Forms of Great Epic Literature*. Translated by Anna Bostock. Cambridge, MA: The MIT Press.

MEYEN, FRITZ. 1942. *Die norwegischen Übersetzungen deutscher Schönliteratur*. Vol. 1 of *Norwegische Bibliographie*, edited by Fritz Meyen. Oslo: Stenersen.

MORETTI, FRANCO. 1998. *Atlas of the European Novel 1800–1900*. London: Verso.

———. 2000. "Conjectures on World Literature." *New Left Review* 1 (January–February 2000): 54–68.

MUNCH-PETERSEN, ERLAND. 1970. *Romantisk underholdning: Triviallitteratur - kvantitetsliteratur*. Copenhagen: Ejler.

———. 1978. *Romanens århundrede: Studier i den masselæste oversatte roman i Danmark 1800–1870*. 2 vols. Copenhagen: Forum.

NATIONAL LIBRARY OF NORWAY. 2024a. "Literary Citizens of the World (LitCit)." Accessed 16 August 2024. https://www.nb.no/en/research/literary-citizens-of-the-world-litcit/.

NATIONAL LIBRARY OF NORWAY. 2024b. "Bibliography Norwegian Books 1519–1850." Accessed 16 August 2024. https://www.nb.no/bibliografi/nor1519/search.

NETTUM, ROLF NYBOE. 2001. *Christen Pram: Norges første romanforfatter*. Oslo: Aschehoug.

NISSEN, MARTINUS. 1848. *Norsk Bog-Fortegnelse. 1814–1847*. Christiania: Feilberg & Landmark.

———. 1849. "Statistisk Udsigt over den norske Litteratur fra 1814 til 1847." *Norsk Tidsskrift for Videnskab og Litteratur* 3:177–207.

NØDING, AINA. 2010. "Fra fabler til føljetong." In *En samfunnsmakt blir til: 1660–1880*, edited by Martin Eide. Vol. 1 of *Norsk presses historie 1–4 (1660–2010)*, edited by Hans Fredrik Dahl, 305–359. Oslo: Universitetsforlaget.

———. 2017. "Periodical Fiction in Denmark and Norway before 1900." *Oxford Research Encyclopedia of Literature*, https://doi.org/10.1093/acrefore/9780190201098.013.293, last visit on 8 July 2021.

PARELIUS, NILS. 1987. *De gamle hefteseriene: Et halvglemt stykke kulturhistorie. En litterær og bibliografisk gjennomgåelse*. Oslo: Bjørn Ringstrøms antikvariat.

PETTERSEN, HJALMAR. 1899–1908. *Norsk boglexikon 1643–1813: Beskrivende katalog over bøger trykte i Norge i tidsrummet fra bogtrykkerkunstens indførelse til adskillelsen fra Danmark*. Vol. 1 of *Bibliotheca Norvegica*. Christiania: Cammermeyer.

———. 1908. *Norge og nordmænd i udlandets literatur: Beskrivende katalog over bøger og tidsskriftartikler om norske forhold. I: Nr. 1–3000*. Vol. 2 of *Bibliotheca Norvegica*. Christiania: Cammermeyer.

———. 1910. *Norge og nordmænd i udlandets literatur: Beskrivende katalog over bøger og tidsskriftartikler om norske forhold. II: Nr. 3001–5000*. Vol. 2 of *Bibliotheca Norvegica*. Christiania: Cammermeyer.

———. 1917. *Norge og nordmænd i udlandets literatur: Beskrivende katalog over bøger og tidsskriftartikler om norske forhold. III: Nr. 5001–7000*. Vol. 2 of *Bibliotheca Norvegica*. Christiania: Cammermeyer.

———. 1911–1918. *Norske forfattere før 1814: Beskrivende katalog over deres værker tilligemed supplement til Bibliotheca Norvegica I: Bøger trykt i Norge før 1814*. Vol. 3 of *Bibliotheca Norvegica*. Christiania: Cammermeyer.

———. 1913–1924. *Norske forfattere efter 1814: Beskrivende katalog over deres i udlandet trykte værker*. Vol. 4 of *Bibliotheca Norvegica*. Christiania: Cammermeyer.

PISTORIUS, FRANZ. 1895. *Klippeøen i det atlantiske Ocean: Fortælling*. Kristiania: P. Omtvedt.

RILEY, PATRICK. 1994. "Introduction." In François de Fénelon, *Telemachus, Son of Ulysses*. Cambridge: Cambridge University Press.

SCHENDA, RUDOLF. 1977. *Volk ohne Buch: Studien zur Sozialgeschichte der populären Lesestoffe. 1770–1910*. Munich: Deutscher Taschenbuch Verlag.

STRETTON, HESBA [SARAH SMITH]. 1868. *Jessicas første Bøn: En Fortælling for Børn*. Translated by E. A. Hagerup [Emma and Augusta Hagerup]. Kristiania: Cammermeyer.

SUNDTOFT, D. S. 1983. *Vestre Moland Kirke: Trekk av dens historie gjennom 800 år*. Arendal: D. S. Sundtoft.

TVETERÅS, HARALD L. 1964. *Norske forfattere på danske forlag 1850–1890*. Vol. 2 of *Den norske bokhandels historie*. Oslo: Den norske bokhandler-medhjelper-forening.

VELLE, THOMAS. 2019. "Ludvig Holberg's *Niels Klim* (1741) and the Irony of Reading and Writing in Latin." In *Neo-Latin and the Vernaculars: Bilingual Interactions in the Early Modern Period*, edited by Florian Schaffenrath and Alexander Winkler, 72–95. Leiden: Brill.

WATT, IAN. 2000. *The Rise of the Novel: Studies in Defoe, Richardson and Fielding*. London: Pimlico.

WITTMANN, REINHARD. 1999. "Was there a Reading Revolution at the End of the Eighteenth Century?" In *A History of Reading in the West*, edited by Guglielmo Cavallo and Roger Chartier, 284–312. Oxford: Polity Press.

4. Corpus and the Bibliography: NB DH-LAB as an Infrastructure for Text and Metadata Mining

Magnus Breder Birkenes and Lars G. Johnsen

Introduction

A corpus is a digital collection of texts together with metadata, normally selected for a specific purpose, e.g. a research question (McEnery and Hardie 2012, 1–2). Unless the documents are specifically characterised, that is, linked to some information concerning topic or content, a collection of Word documents kept in a folder on a computer is normally not considered to be a corpus. Furthermore, to qualify as a corpus, the material needs to be accessible in some way. It may have restrictions on its use, but it is of utmost importance that it has a possible use. A bibliography, on the other hand, is a list of books or documents relevant to a research question or research area, containing only metadata. A bibliography does thus not qualify as a corpus: the corpus consists of the actual texts referred to by the bibliography.

Our main thesis in this chapter is that the distinction between corpus and bibliography is, for the most part, not as clear-cut as might be imagined. Whereas the primary texts and bibliographic metadata of corpora have traditionally been co-located (e.g. as files on a desktop computer), they may be technically separated but still easily available to users by means of a web browser in this era of internet technology. For example, while a hypertext link is a separate object from the webpage it points to, the text is physically just one click or request away. We shall go into more detail about this below and discuss the notion of

a corpus as it is used in the DH-lab at the National Library of Norway. Furthermore, while bibliographies do not qualify as corpora, bibliographic metadata, e.g. titles, may be analysed as corpus data, as exemplified by Marjanen, Tahko, Lahti and Tolonen in Chapter 2. We will show uses of this in the following chapter. We will also show how bibliographic data (e.g. citations) can assist in corpus building.

Below, we will conduct two types of language investigations. One on the bibliography itself, using the titles as language data, and another on the texts pointed to by the bibliography.

NB DH-lab: Bibliographies as Pointers

Through its ambitious digitisation programme, initiated in 2006, the National Library of Norway currently holds one of the world's largest corpora, consisting of nearly all books, newspapers and journals ever published in Norway (for a recent article about the digitisation project, see Gran et al. 2019). The corpus currently amounts to 160 billion tokens, making it three times larger than the German Reference Corpus (DeReKo) and about one-third of the Google Books corpus of English books (Leibniz-Institut für Deutsche Sprache 2024; Lin et al. 2012, 170). This is an impressive size for such a small language as Norwegian, with its five million speakers.

Table 1: The size of the text collections held by the National Library of Norway (figures as of 16 August 2024)

Document type	Documents	Tokens
Books	633,368	33,727,728,639
Newspapers	4,318,439	117,264,611,638
Journals	244,458	16,985,142,717

Due to copyright constraints, however, we cannot publish the corpus as a downloadable resource. Even the corpus of material already in the public domain is so large that it would be hard for most users to utilise it on their machines. For these two reasons, the DH-lab at the National Library of Norway has developed a corpus infrastructure that makes it

possible to search and analyse the digital collections of the National Library without challenging copyright law, while making the texts available as if they were locally stored.

The infrastructure is described more thoroughly in Birkenes et al. (2023). At the heart of the DH-lab corpus infrastructure lies an Application Programming Interface (API) that can be used by developers to build services.[1] The API gives access to metadata, text snippets for all objects and word frequency lists. On top of this API, we have built Python and R packages (e.g. for use in Jupyter notebooks) and easy-to-use web applications for a more non-technical audience. In this chapter, we will illustrate how tasks may be performed on the DH-lab infrastructure using the Python package dhlab.[2]

The DH-lab corpus infrastructure is built around the international FAIR principles.[3] The documents contained therein are

1. findable through a corpus builder, which utilises bibliographic metadata and content data (e.g. text from optical character recognition) to curate material,
2. publicly accessible through the internet,
3. interoperable through the use of an Application Programming Interface (API), by means of which services can talk to each other, and,
4. reusable through the implementation of persistent identifiers.

Each digitised object at the National Library of Norway is given its own so-called URN (Uniform Resource Name), which is a persistent, globally unique identifier used by most national libraries for their digital

1 An API uses one of the protocols available for communication on the internet. It makes it possible to separate code from data, so that a machine (client) can communicate with a remote machine (the server).

2 The package can be installed on a system with a Python distribution on the command line using: pip install dhlab.

3 FAIR stands for Findable, Accessible, Interoperable and Reusable with respect to data and metadata.

resources. For text documents such as books and newspapers, the URN identifies the digitised or born-digital images of these objects. Historically, however, the texts extracted from these objects was not given their own identifiers, since they were regarded more as an index than a relevant object in their own right. This was not a problem as long as the underlying texts were not changed, but since the advent of better OCR technologies, e.g. for historical material in the Fraktur calligraphic style, we now have multiple text versions for the same scanned documents. Therefore, the DH-lab created its own persistent identifiers for derived text objects, which we call the dhlabid.

A corpus in the DH-lab is simply a list of documents with persistent identifiers (i.e. dhlabids) and some basic bibliographic metadata. These corpus definitions are used in all interactions with the infrastructure. Thus, every request to the infrastructure contains the identifiers or pointers that the corpus consists of. This is different from traditional downloadable corpora, where text and metadata are distributed together. Using this setup, a researcher can perform searches in the actual texts via internet communication protocols (the APIs of DH-lab) as if the documents were residing on the researcher's own computer.

Bibliographic Metadata as a Corpus

We can analyse a set of bibliographic metadata as a corpus itself, e.g. by looking at the titles of books in specific categories and extracting typical words for these categories. These typical words can then be used as "proxies" to extend the corpus.

NB DH-lab allows the user to do this quite easily. As an example, we will use Dewey Decimal Classification code 641 (cooking).[4] First, we define a corpus consisting of books in Norwegian Bokmål

4 Dewey Decimal Classification (DDC) are three-digit codes that identify the topicality of documents, with additional decimal places. It exists alongside UDC, Universal Decimal Classification, which serves the same purpose. The books at the National Library are mostly complete with DDC from 1950 and onwards. Books published before 1950 are more sparsely classified.

(= nob) that have been classified as 641 ("Food and drink"). We use the wildcard to include subcategories such as 641.2 ("drinks") and 641.5 ("cookbooks"). We limit the set to 5,000 books. If the query returns more books than this number, the system will create a random sample:

```python
import dhlab as dh
corpus = dh.Corpus(doctype="digibok", from_year=1950, to_year=2024, ddk="641*",
lang="nob", limit=5000)
```

Code block 1: Build a corpus of Dewey 641 books.

In total, we get a set of 4,250 food-related books using the query above which will be used in the following. But to find the most typical words in the titles of food-related books, we need a reference corpus for comparison. We therefore create a sample of 4,250 books in Norwegian from the same period, but without specifying Dewey code:

```python
import dhlab as dh
ref_corpus = dh.Corpus(doctype="digibok", from_year=1950, to_year=2024,
lang="nob", limit=5000)
```

Code block 2: Build a reference corpus from the same period.

The tables returned by the DH-lab package for the queries above contain pointers to the actual text and bibliographic metadata. The first columns may look like this:

Table 2: Corpus object

Dhlabid	Urn	Title	Authors
100052973	URN:NBN:no-nb_digibok_2011081106010	Eplekaker : og annet smått & godt av epler	Hallgren, Ann-Kristin / Mårtensson, Hans / V...
100262769	URN:NBN:no-nb_digibok_2015081948061	Grønne retter : fristende, fargerikt og godt	Olsson, Brita / Dotterud, Aase
100383640	URN:NBN:no-nb_digibok_2018062807117	Fristende smoothies og juice	Doeser, Linda / Bergan, Brit / Doeser, Linda

The **dhlabid** is the primary key to the text document, the **URN** points to the digitised book (images) and **title** and **authors** represent bibliographic metadata from the national bibliography. When looking at the titles in Table 2 above, we observe that certain words are used in multiple places, e.g. the word *godt* 'good' and *fristende* 'tempting' (marked in bold in the table). We also notice the repeated presence of function words like the conjunction *og* 'and'.

We will use the title field as a corpus of its own. If we count the tokens in all titles of the Dewey 641 books, we get a total corpus of 27,615 running words. When aggregating them, the top five are as follows:

Table 3: Top-five word frequencies in the titles of the Dewey 641 corpus (total: 5,335 unique words).

Word	absolute frequency	relative frequency
og 'and'	1296	4.69%
mat 'food'	590	2.14%
fra 'from'	495	1.79%
for 'for'	469	1.70%
i 'in'	409	1.48%

The table shows the top-five words together with their absolute frequency, i.e. the number of times they appear in the titles, and their relative frequency, i.e. how often the word appears in relation to all other words in the titles. Only one of these words, *mat* 'food', points to a topic word, the others being grammatical words such as conjunctions and prepositions. But these grammatical indicators are not uninteresting. It might be imagined that *fra* 'from' often refers to food from different culinary traditions, e.g. *Mat fra Norge* 'Food from Norway', whereas *for* could hint at the intended audience, e.g. *Kokebok for folkeskolen* 'Cookbook for the primary school'. But being function words, we would expect them to be used in many titles, regardless of category, whereas *mat* 'food' would be limited mainly to Dewey 641 books. To

investigate this, we count the words in the titles of the reference corpus in the same way, here resulting in a total of 33,688 running words:

Table 4: Top-five word frequencies in the titles of the reference corpus (total: 11,172 unique words).

Word	absolute frequency	relative frequency
og 'and'	1495	4.44%
i 'in'	1182	3.51%
av 'of'	695	2.06%
for 'for'	621	1.84%
en 'a/an'	433	1.29%

We first notice that although we have created corpora of equal sizes in terms of books (4,250 food-related books and a reference corpus of 4,250 books), they are not similar in the amount of running words: a total of 27,615 words in the food corpus, compared with 33,688 words in the reference corpus. We can thus infer that food-related books have somewhat shorter titles than books in general (an average of 6.5 words, compared with 7.9 words). When comparing the lists, we see that three of the words appear in both lists, *og* 'and', *i* 'in' and *for* 'for', albeit at different ratios. When comparing corpora of different sizes, looking at the relative frequencies is instructive. We notice that *og* 'and' and *for* 'for' have similar relative frequencies, whereas *i* 'in' is underrepresented in the food corpus in comparison with the reference corpus. The preposition *fra* 'from', on the other hand, is used nearly twice as much in the Dewey 641 corpus as in the reference corpus, coming in at 1.79 per cent compared with 1 per cent. If we compare these relative frequencies across all words, by dividing the relative frequencies of the Dewey 641 corpus by the relative frequencies in the reference corpus, and sort by these ratios, we get a list of "typical" words for the Dewey 641 corpus.

We include only words that occur ten times more often in the Dewey 641 corpus than in the reference corpus, and sort this list by the absolute frequencies of the words to reduce the effect of low-frequency words. Table 5 below shows the top-ten list:

Table 5: Typical words in the titles of the Dewey 641 corpus (T = target corpus, R = reference corpus).

Word	freq (T)	rel (T)	freq (R)	rel (R)	Ratio
mat 'food'	590	2.14%	8	0.02%	89.97
oppskrifter 'recipes'	348	1.26%	2	0.01%	212.27
beste 'best'	163	0.59%	8	0.02%	24.86
godt 'good'	140	0.51%	6	0.02%	28.46
vin 'wine'	102	0.37%	2	0.01%	62.22
kaker 'cakes'	87	0.32%	2	0.01%	53.07
verdens 'world'	85	0.31%	8	0.02%	12.96
god 'good'	82	0.30%	10	0.03%	10.00
gode 'good'	80	0.29%	8	0.02%	12.20
fisk 'fish'	77	0.28%	8	0.02%	11.74

We notice several things: *mat* 'food', the most frequent noun in the corpus, is used almost 90 times more often than in the reference corpus, and the adjective *god* 'good' discussed above is also typical for the Dewey 641 corpus (in various inflection forms). Overall, we get the impression that titles in the Dewey 641 carry a positive "sentiment".

To investigate this further, we perform a simple sentiment analysis using the NorSentLex lexicon of positive and negative Norwegian words (Barnes et al. 2019; Language Technology Group Oslo 2024). For each word from the datasets exemplified in Table 3 and Table 4, we look up whether the word is in the set of positive or negative words, and use the relative frequency of the word in the corpus as its weight. By summing the weights of the respective positive and negative words, we get the proportions of positive and negative words.

In the case of the Dewey 641 corpus, 9.9 per cent of the title tokens are classified as positive, whereas 3.2 per cent are negative. In the reference corpus, the numbers are 3.7 per cent and 3.1 per cent respectively. The difference in percentage points between positive and negative is thus noticeably higher for the food-related books than for the sampled reference corpus: 6.7 per cent compared with 0.6 per cent.

This supports our hypothesis that food-related book titles carry a more positive sentiment than books in general. Examples of such titles are *Sunt snop & snacks: 23 fantastiske søte og salte oppskrifter på sunne godsaker som kan nytes med god samvittighet* (Healthy Treats & Snacks: 23 Fantastic Sweet and Savory Recipes for Wholesome Goodies to Enjoy Guilt-Free) and *Maten & Moi: få det beste ut av råvarene: enkelt, godt og bærekraftig* (Food & Moi: Get the Best out of Your Ingredients: Simple, Tasty, and Sustainable) (Hammer 2012; Moi and Lødemel 2023).

But as shown above, the DH-lab not only gives us access to metadata, we also get aggregated full-text statistics. Thus, we can run the same analysis on the full-text of the two corpora to see if food-related books carry a more positive tone than books in general. To do this, we can use the DH-lab's count () function, which gives us word statistics for each book in a corpus.[5] We looked up each word in the Norsentlex lexicon and aggregated the frequencies more or less as we did with the titles. When aggregating the relative frequencies of positive and negative words in the food corpus compared with the reference corpus, we find that 4.9 per cent of the words are positive and 4.8 per cent negative, compared with 4.5 per cent and 4.7 per cent respectively in the reference corpus. There is thus a very slight preponderance of positive words in the full-text of the food corpus, but not comparable to that of the titles.[6]

To investigate this further, we applied the same method to all main Dewey classes in the digital collections of the National Library of Norway. We sampled 4,250 books (similar to the size of the food book corpus) for each class. The results are visualised in Figure 1.

5 Because aggregating frequencies locally is quite memory intensive, and for reasons of space, we will not describe all the steps here.

6 The food corpus has a size of 168,345,077 tokens and the reference corpus comprises 216,359,525 tokens.

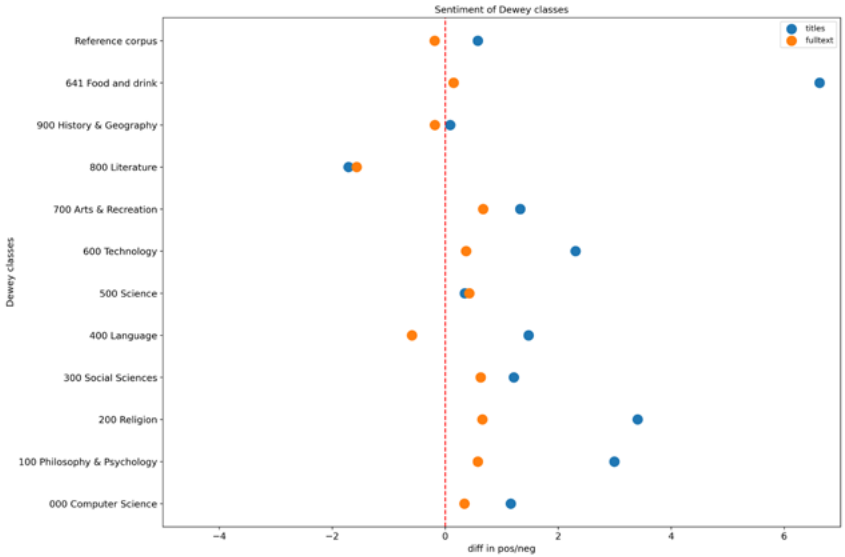

Figure 1: Difference in the percentage of positive and negative words in the titles and full-text Dewey categories.

The overall picture is that we find most data points to the right of the 0 line on the x-axis. This indicates a preponderance of positive words in both titles and full-texts across Dewey categories. Interestingly, this is not the case in 800 Literature, where negative words predominate in both full-text and titles. On the other hand, in 200 Religion, we find a substantial preponderance of positive words in the titles, whereas 500 Science and 900 History, Geography and Auxiliary Disciplines are close to neutral in terms of title sentiment.

Most of the works showing a preponderance of positive words in the title also show a preponderance of positive words in the full-text. A notable exception is 400 Language, where negative words largely predominate. Interestingly, all the categories in which negative words outweigh positive ones belong to the arts and humanities. When comparing this to the food-related corpus, we notice that this corpus has a higher preponderance of positive words than the more general Dewey categories – even more than religious books. Yet this is not followed by an overall positive sentiment in the full-text.

All in all, this first study has shown how we can use the NB

DH-lab to apply corpus linguistics methods to both a bibliography and the texts it refers to using the same API. The sentiment analysis also revealed that works with a preponderance of positive words in the title also typically have a preponderance of positive words in the full-text, despite the fact that titles and full-text are two quite different objects in terms of length and linguistic properties.

Extracting Citation Information through Concordances

In this section, we will use the tools available in the DH-lab for accessing text content via metadata. As described above, the texts are connected to their identifiers (URNs) via protocols, enabling us to view a list of books described in a bibliography as a proper corpus. What we now set out to do is find citations in texts. This is an aspect of bibliometrics, the examination of citations and references within texts.

Quantitative properties of citations and references within texts raise questions about the interconnectedness of various discourses. Specifically, at least two questions emerge:

1. How far back do citations go, or how far back do most of them go?
2. What is the frequency of a text's citation within its citing documents or within a corpus of text?

While we believe the first question, about the timespan metric, represents a novel approach, the frequency of citations was used by Sivesind and Hörmann (2022) to address, among other questions, the centrality of certain documents.[7] Addressing these questions necessitates a move beyond the conventional metadata, where we need to go directly into the textual content itself.

Even though citations and references are part of the text itself, we believe they may be elevated to the level of catalogue metadata. In the future, such information can be entered alongside other metadata

7 A study conducted within the field of education policy.

information provided by libraries. In a sense, we may say that these textual elements belong to what Gérard Genette calls paratext, i.e., they are not part of the text itself, but exist within parentheses and footnotes (Genette 1997).

Bibliographic records extracted from texts represent the relationship between the source text and other texts. As such, they may be incorporated as a building block of metadata. In this way, the information extends the relational aspect of metadata, which already consists of relationships like those between the text and its author or the text and its publisher, to include relationships between documents. The standard for metadata used by the National Library of Norway, MARC, currently lacks the field 504.[8] This field is specified by the US Library of Congress to contain information about the pages in a book where a bibliography exists. It is a first step towards lifting citations and references up into the realm of metadata. Although this field in the MARC records does not specify a convention for listing the cited material itself, such a connection could be made ad hoc, as described in Sivesind et al. (2022) for a subset of public governmental documents in Norway.

Methodological Approach

Historical standards for citation formats have evolved since their inception in the late nineteenth century, undergoing numerous revisions throughout the twentieth century (e.g. Bergmark and Lagoze 2000). Our research adheres to a methodology that enables citation and reference extraction from texts, inspired by Mattia Chenet's thesis on this topic (2017). At the core of this effort lies the construction of a textual network, linking texts through their references. However, our present goal is somewhat more modest, in that we draw a distinction between the citations within the text and the references the citations

8 MARC = Machine Readable Cataloging - a format used for making metadata catalogues. The three digits that identify a field in MARC are unrelated to the Dewey decimal system, and are arbitrary numbers with a conventional interpretation.

point to. While the citations exist within the text and refer to an entry in the reference or bibliography of the text, the references themselves carry full information about the cited work, so that it can be identified amongst the millions of published books. A citation will therefore have a simpler structure, enough to identify the corresponding entry in the bibliography.

The central information for a citation typically consists of a four-digit year and one or more authors, in contrast to the full reference which contains all the information required to identify the cited text. For example, a full reference to a work in a journal may look like the following, where the full names of the authors, the title of the work, its publisher, the journal name and volume, page numbers, as well as online links in the form of a DOI, are specified:[9]

> Bamman, David, Jacob Eisenstein and Tyler Schnoebelen. 2014. "Gender identity and lexical variation in social media." *Journal of Sociolinguistics* 18(2): 135–160. https://doi.org/10.1111/josl.12080

A citation to that work within a text may take the simplified form:

> (Bamman et al. 2014)

Variations in the way citations are written are part of the challenge. In-text citations are the focus of this chapter, leaving the full analysis of reference parsing to a future date.

A key property of a citation is that it is unique within the text. A citation may be ambiguous if an author being referred to has published several texts in a single year. This is resolved by the use of numbering schemes in the citations themselves, for example (Bamman et al. 2014b), if the need arises.

For citation analysis, we use the concordance extraction feature

9 DOI = Digital Object Identifier – typically used for texts, it serves the same purpose as a URN for the library's digital collection. It is intended to be a persistent identifier that will identify the digital object across time.

provided by the DH-lab API. A concordance is a relatively small snippet of text that can be fetched from a text based on a keyword. They are therefore sometimes referred to as keywords in context (abbreviated to KWIC). Concordances have a long history, dating back to thirteenth-century biblical studies, where the concordance contained the index of a keyword in the Bible along with some context (O'Keeffe and McCarthy 2010, 3–4). Concordances have a wide range of applications in language processing, especially in various ways of constructing word models.

A practical problem facing the DH-lab is copyright law, which limits the amount of text that can be transferred from the databases to the end user (researcher). Depending on the law, the size of that snippet can be made copyright-compliant, and thus form the basis for a strategy to perform a large-scale citation analysis within access-restricted texts.[10] By focusing on concise text fragments, or concordances, this methodology facilitates the identification of citations in copyrighted works as well as in freely available texts. The method presented here can therefore be applied to any corpus by any researcher anywhere, as long as they have access to a computer and the internet.

The process has three steps:

1. Select or construct a corpus.
2. Select concordances that contain years as four-digit numbers.
3. Run regular expressions on these concordances to select the citations within them and, if a citation is found, store the concordances with the extracted citation.[11] Pattern matching will determine whether the extracted concordance conforms to a citation or not.

Below, we go through these steps and illustrate them with the code commands in play. The first step is to construct a corpus using subject

10 The snippet size currently used by DH-lab is up to 25 words in sequence that do not cross a page or paragraph boundary.

11 A regular expression is a pattern-finding method used within computer science.

headings as metadata.[12] Here we want to extract books from the 1960s to the present day that have either the subject heading *lingvistikk* ('linguistics') or *språkvitenskap* ('language studies').[13]

```python
import dhlab as dh
corpus = dh.Corpus(doctype='digibok', subject='lingvistikk OR språkvitenskap',
from_year = 1960, to_year = 2020, limit = 1000)
```

Code block 3: Build a corpus for the performance of citation analysis, resulting in a corpus of 531 titles. The subject heading corresponds to the English term "linguistics".

The second step extracts concordances based on years. For this purpose, a string of years is concatenated into one long search string, so we are effectively searching for occurrences of the years 1900, 1901, 1902, …, 2020. The command takes the following form, edited for the purpose of legibility.

```python
import dhlab as dh
number_conc = dh.Concordance(corpus,"1900 OR 1901 OR ... ", limit=4000)
```

Code block 4: Select concordances containing years.

The result is kept in the variable *number_conc* (number concordances) which holds all the concordances. These concordances are then subjected to further analysis. According to the APA style, a simple in-text citation should contain the author's surname and the publication year inside parentheses like this: "(Wertsch, 2002)". A regular expression that captures this format may look like the one in code block 5.[14]

12 Alongside Dewey decimal codes, books are classified using topical words called subject headings. These words come in two varieties, either controlled or uncontrolled. Here we do not distinguish between the source of the subject words, and just look for a match with the subject heading somewhere in the subject metadata.

13 This search is case insensitive, and the search string is matched against any authority source for the term.

14 The regular expressions for this project were crafted by Marie Iversdatter Røsok.

```python
"\([A-Z][a-z])+,\s\d{4}\)"
```

Code block 5

The expression matches parentheses containing one uppercase letter followed by lowercase letters, a comma, a whitespace and four digits. Although similar, this expression does not match expressions indicating lifespans, such as "(Sture 1437–1503)", because the comma is lacking and a dash and an extra year have been introduced. The regular expression in code block 5 will properly distinguish between citations and lifespans, one of the major sources of confusion for a pattern matcher.

Once the concordances have passed through the citation extraction stage, we are left with a dataset that can be analysed, in other words counted and grouped. The data from concordance counting is summarised in a table that can be inspected to check that the citation extraction is accurate, that it contains the names of the cited authors and the year of the cited publication, as well as the number of years that have elapsed (time difference) between the cited and the citing publication. It is this latter information we will use as input for the analysis. An example of what this looks like is given in Table 6 below.

Table 6: Sample from citation analysis. Note that there may be several citations in one line. However, each citation will still be found on its own.

Citations	Year	Diff
following Link 1983 and Landman 1991	1983	33
Askildsen 1957	1957	40
Brown & Gilman 1960, Brown & Ford 1961	1960	38
van Dijk 1972: 153	1972	6
Fra Lindholm og Padilla 1978	1978	25

With this data, we can create statistics over the difference in publication years using the content of the difference columns (named in Diff in

Table 6). Each value in the Diff-column will occur at a certain level of frequency within the corpus. For example, for the figures in Table 6, we get the corresponding values in Table 7 below. The column *Year-difference* contains the difference in years between the cited document's original publication year and its citation, while the column *Frequency* has the number of times this time difference has been observed in the corpus.

Table 7: Example of time differences and how often these differences have been observed. It seems that the smaller the time difference, the higher the frequency.

Year-difference	**Frequency**
33	10
40	8
38	5
6	66
25	11

Figures 2 and 3 show two initial graphs, one for the subject heading linguistics (Norwegian: *lingvistikk*, *språkvitenskap*), and one for the subject heading media science (Norwegian: *medievitenskap*).

In both plots, the horizontal axis shows how many years prior to the citing document the cited document was published (from 0 to 100 years). The vertical axis shows how many instances of that time gap (difference) are found within the corpus. If we look at Figure 2, for example, we see that there are 30 instances of a 20-year difference between the citing document and the cited document's year of publication. The most frequent time difference between citing and cited documents is five years. The graph is weighted heavily towards the lower end of the scale, with the majority of citations referring to documents published one to eight years prior to the citing document. Half of all citations have a time difference of 12 years or less.

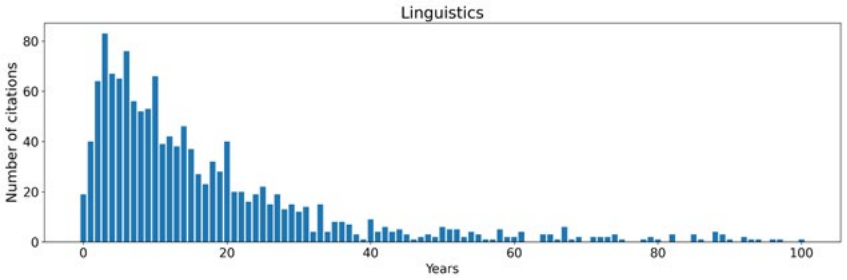

Figure 2: Graph showing the time difference in years between the citing and cited documents.

Although the corpus for media science is slightly larger than the linguistic corpus, the same pattern emerges. In fact, over the 120-year timespan we studied (1900–2020), half of the documents cited were published less than ten years before the citing document.

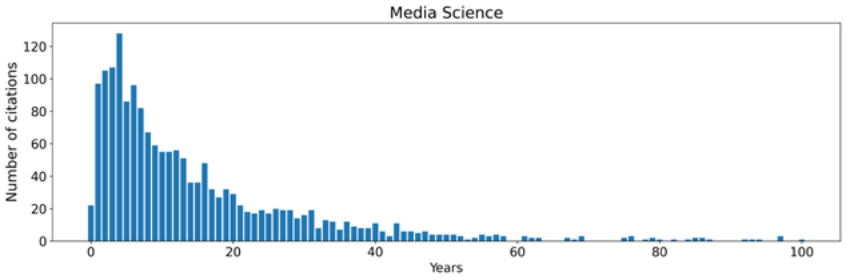

Figure 3: Graph showing citation patterns for *medievitenskap* (media science)

A summary of several categories within the Arts and the Sciences shows the same pattern. In the Figure 4 we have collected graphs for the seven subject headings:

Norwegian search word	**English equivalent**
Naturvitenskap	Natural Science
Litteraturvitenskap	Literary Science
Biologi	Biology
Geologi	Geology
Lingvistikk OR *Språkvitenskap*	Linguistics

Kjemi	Chemistry
Medievitenskap	Media Science

In the visualisation, the bar chart has been replaced with a line graph, and the number of citations has been normalised to the percentage of the total. The total area under each line is the same, so that if a field has a high start it must be lower along the tail.

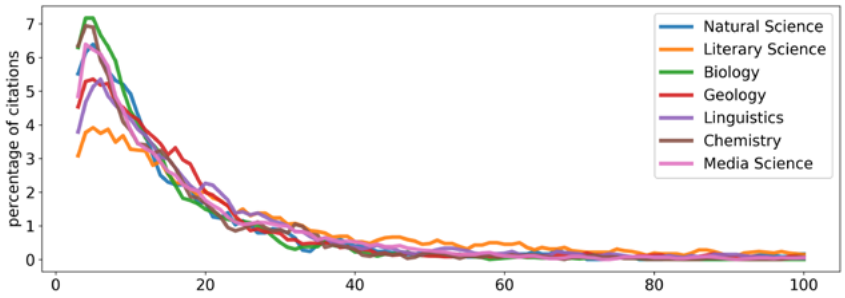

Figure 4: Graph showing the number percentage distribution of citations from seven different subjects. The labels are translations from the Norwegian subject headings. The horisontal axis represents the number of years between a cited document and the document that cites it.

Judging from the shape of the graph, it supports the hypothesis that the sciences seem to be more contemporary in their citations than the humanities or social sciences. Note that both linguistics and literary science have the smallest percentage at the beginning of the curve, while the topmost positions are occupied by chemistry and biology. To make a fuller story of the differences between different fields of study requires a more comprehensive approach to statistics and measurements, but the trend appears to be there.

This citation study offers some insights into the temporal depth of references and the dynamics of citation frequency within academic texts. The findings illustrated by the figures, bar charts and the line graph, reveal a fairly diverse temporal span of citations, and a pattern that holds across various fields. It is noteworthy to observe that authors seem to cite recent texts. Both the humanities and the natural sciences show that half the citations are of works

published within the first ten years or so of the entire 120-year timespan.

Through the use of concordance extraction and classical machine learning techniques (namely regex), this research contributes to the field of bibliometrics, offering a particular understanding of citation patterns. This methodology can be used for future bibliometric studies, opening up for more comprehensive analyses of research papers and documents, in particular for the digitised collections at the National Library of Norway.

Conclusion

In this chapter, we presented two analyses of bibliographies, using different parts of the catalogue data for books, including author, title, publication year and Dewey Decimal Classification. The first analysis focused on the bibliography itself as the object of study, delving into a content analysis of titles and document text using title and Dewey data. In the second analysis, we examined a collection of bibliographies taken from various books, using author and publication year as structuring units. This collection provides a hierarchical view of bibliographies, where an entry in one bibliography (citations) contains or points to another bibliography.

References

BARNES, JEREMY, SAMIA TOUILEB, LILJA ØVRELID AND ERIK VELLDAL. 2019. "Lexicon information in neural sentiment analysis: a multi-task learning approach." *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland. URL: https://www.aclweb.org/anthology/W19-6119/.

BERGMARK, DONNA, AND CARL LAGOZE. 2001. "An Architecture for Automatic Reference Linking." *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, June 2001. DOI: 10.1007/3-540-44796-2_11.

BIRKENES, MAGNUS BREDER, LARS GUNNARSØNN BAGØIEN JOHNSEN AND ANDRE KÅSEN. 2023. "NB DH-LAB: a Corpus Infrastructure for Social Sciences and Humanities Computing." *CLARIN Annual Conference Proceedings 2023*, 30–34. https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf.

CHEN, CHAOMEI, FIDELIA IBEKWE-SANJUAN AND JIANHUA HOU. 2010. "The Structure and Dynamics of Co Citation Clusters: A Multiple Perspective Co-Citation Analysis." *Journal of the American Society for Information Science and Technology* 61, no. 7 (2010): 1386–1409. https://doi.org/10.1002/asi.21309.

CHENET, MATTIA. 2017. "Identify and Extract Entities from Bibliography References in a Free Text." Master's Thesis, University of Twente.

GENETTE, GÉRARD. 1997. *Paratexts: Thresholds of Interpretation*. Cambridge: The University of Cambridge.

GRAN, ANNE-BRITT, EIVIND RØSSAAK AND LINN-BIRGIT KAMPEN KRISTENSEN. 2019. "Digital Infrastructure for Diversity – On Digital Bookshelf and Google Books." *The Journal of Arts Management, Law, and Society* 49 (3): 171–187.

HAMMER, GRY. 2012. *Sunt snop & snacks: 23 fantastiske søte og salte oppskrifter på sunne godsaker som kan nytes med god samvittighet*. Røyse: G. Hammer. https://urn.nb.no/URN:NBN:no-nb_digiebok_3519.

LANGUAGE TECHNOLOGY GROUP OSLO. 2024. "NorSentLex." Accessed 15 August 2024. https://github.com/ltgoslo/norsentlex.

LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE. 2024. "Ausbau und Pflege der Korpora geschriebener Gegenwartssprache." Accessed 15 August 2024. https://www.ids-mannheim.de/digspra/kl/projekte/korpora/.

LIN, YURI, JEAN-BAPTISTE MICHEL, EREZ AIDEN LIEBERMAN, JON ORWANT, WILL BROCKMAN AND SLAV PETROV. 2012. "Syntactic Annotations for the Google Books NGram Corpus." *Proceedings of the ACL 2012 System Demonstrations*, 169–174. https://aclanthology.org/P12-3029.

MCENERY, TONY, AND ANDREW HARDIE. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

MOI, TROND, AND NINA LØDEMEL. 2023. *Maten & Moi: Få det beste ut av råvarene: Enkelt, godt og bærekraftig*. Oslo: Egmont People. https://urn.nb.no/URN:NBN:no-nb_pliktmonografi_000016505.

O'KEEFFE, ANNE, AND MICHAEL MCCARTHY. 2010. "Historical perspective: What are corpora and how have they evolved?" In *The Routledge Handbook of Corpus Linguistics*, edited by A. O'Keeffe and M. McCarthy, 3–13. London: Routledge.

SIVESIND, KIRSTEN, DIJANA TIPLIC AND LARS GUNNARSØNN JOHNSEN. 2022. "Surveying Policy Discourses Across Time and Space: Internationalization of Knowledge Providers and Nordic Narratives." In *The Nordic Education Model in Context: Historical Developments and Current Renegotiations*, edited by Daniel Tröhler, Bernadette Hörmann, Sverre Tveit and Inga Bostad, 308–331. London: Routledge.

5. Getting Meaning out of Metadata – Analysis of Selected Bibliographies at the National Library of Norway

Oddrun Pauline Ohren

The term *bibliography* has had several meanings. Until the twentieth century, the concept of bibliography was 'unprotected' in the sense that almost any finite list of writings or books could be called a bibliography, whatever its target audience and level of detail. Most often, its main purpose was to notify readers of the existence of the listed writings, not to describe them in detail or disclose their whereabouts. More recently, such bibliographies have been called enumerative or reference bibliographies.

During the first half of the twentieth century, a few experienced bibliographers – notably Walter W. Greg, Fredson Bowers and (later) G. Thomas Tanselle – worked towards a scientific foundation for the field of bibliography (Tanselle 1994, vi). In particular, Fredson Bowers established himself as a strong advocate of bibliography as a field of historical scholarship in its own right, and as a critic of the prevailing attitude, which held that describing and listing books was a mere support activity for 'real' scholars and collectors. In his *Principle of Bibliographical Description*, originally published in 1949, Bowers argues that *descriptive* bibliography must be based on – and serve as documentation for – *analytical* bibliography, which, he asserts, concerns itself only with books as tangible objects (Bowers 1994, 31):

> Analytical bibliography deals with books and their relations solely as material objects, and in a strict sense has nothing to do with the historical or literary considerations of their subject matter or content. The findings of analytical bibliography may be used to clarify these considerations, but literary history or criticism is not in itself bibliographical. For example, the determination of stylistic resemblances is not a function of bibliography. Relevant to bibliography, however, is the possibility of explaining misprints in books by peculiarities in the handwriting of manuscripts assigned to a certain author.

Bowers met considerable resistance to his ideas and uncompromising style. The older Geoffrey Keynes criticized – and slightly mocked – his puristic and exclusive approach to bibliography (Keynes 1953, 66): "This tendency to exaggerate the claims of bibliography until it comes to be an end rather than a means is perhaps to be recognised as a psychological frustration." True enough, Bowers' notion of descriptive bibliography, with its emphasis on detailed, codified descriptions of the physicality of books as evidence of their production process, clearly contributes more to the history of printing and publishing, book bindings and other material issues than to our understanding of how thoughts, ideas and ultimately knowledge (as encoded in the books) have developed through the ages.

However, in a later work Bowers presented a more comprehensive structure for the bibliographical field, in the form of five categories: enumerative bibliography, historical bibliography, analytical bibliography, descriptive bibliography and textual bibliography (Bowers 1952, 190–195).

Enumerative bibliography is the simple construction of lists of books and other writings on a given subject. Historical bibliography encompasses research into the social and historical context in which the documents were created. Descriptive bibliography and textual bibliography constitute the two practical applications of analytical bibliography, with the latter application extending the object of study to the meaning of the text, but using only the results of analytical bibliographic endeavours as evidence.

Note that Bowers' categories pertain more to the activities performed by bibliographers, than the resulting bibliographies.

Later generations have to some extent rejected the narrow definition of bibliography of the Bowers' school of thought, wishing to broaden the concept to better reflect actual practices in the field. This implies that not only the physical details, but also the intellectual content, that is the actual works manifested by the books in question should be objects of description. It further implies that not only books, but any text or other expressions in any form should be considered. D.F. McKenzie defines 'texts' to include "verbal, visual, oral and numeric data, in the form of maps, prints, and music, of archives of recorded sound, of films, videos, and any computer-stored information, everything in fact from epigraphy to the latest form of discography" (McKenzie 1999, 13). In doing so, he recognised that bibliographies are relevant tools for literary scholars and critics. All in all, the battle to define the scope of the bibliography domain has led to a more inclusive discipline. Today, we generally acknowledge that there are different types of bibliography rather than tying *bibliography* to one single type. In this way we are disregarding Bowers' claim that a "true bibliography is primarily an analytical bibliography" (Bowers 1994, 34).

The definition of *bibliography* in *Merriam-Webster Dictionary* probably covers our current understanding of what a bibliography is and does, provided we allow "list" to include digital representations like databases. Interpreting *bibliography* as a product of bibliographical activities, rather than the activities themselves, three meanings are listed:

1: the history, identification, or description of writings or publications
2a: a list often with descriptive or critical notes of writings relating to a particular subject, period, or author [...]
2b: a list of works written by an author or printed by a publishing house [...]
3: the works or a list of the works referred to in a text or consulted by the author in its production (Merriam-Webster n.d.)

The first meaning hints at descriptive bibliography, whereas the other meanings – to the extent that they address Bowers' categorisation at all – must be called enumerative, since their distinction as types lies in their scope and degree of annotation.

Another area of discussion has been the difference and similarity between bibliography and cataloguing. Prior to the reign of Greg and Bowers in the bibliography discipline, the difference between cataloguing and bibliography was not clearly conceptualised. However, Bowers did not consider cataloguing as having anything to do with bibliography at all: "Thus I have divorced from descriptive bibliography any consideration of cataloguing: […] This [cataloguing] is a field having little to do with scholarly descriptive bibliography in the analytical sense, and hence I do not concern myself with it" (Bowers 1994, xvii).

Viewed from a more current perspective, a different picture emerges. Yee compares cataloguing to descriptive and enumerative bibliography, respectively (Yee 2007). A set of eleven criteria are used for comparison, including object of the description; functions; scope; collective vs. individual creation; and amount of descriptive detail. The conclusion is that *function* is the essential criterion where cataloguing and descriptive/enumerative bibliography systematically differ. The function of a library catalogue is to be a guide to the library's collection, whereas a descriptive bibliography (in the Bowers sense) is meant to "record the findings of an investigation into the printing and publishing history of a particular work or works". The function of an enumerative bibliography, on the other hand, is "to provide temporary and timely access to citations to works on a subject without guaranteeing that the user will actually be able to obtain those works" (Yee 2007, 314, 319). According to Yee, in all other respects than function, both bibliographies and catalogues vary too much to exhibit any systematic typological distinction.

With the above as a background, the rest of the article will study some bibliography cases at the National Library of Norway. A brief account will be made of their position in the bibliographic context. However, exploring the potential to learn new information through an analysis of the bibliographical records is the main emphasis of the following chapters.

The special bibliographies at the National Library of Norway

Bibliographic work is a core activity at the National Library of Norway, as in most national libraries. While the main focus is to create a national bibliography covering everything published in Norway (or abroad if relevant in a Norwegian context), considerable effort is also put into creating a number of specialised bibliographies. The specialised bibliographies have more or less clearly defined scopes, most delimited thematically, others by publishing time or place.

Several years ago, a decision was made to use the main catalogue as the realisation platform for the special bibliographies, since it was judged to provide sufficient discovery and presentation facilities, as well as a sustainable workflow, in which the combined capabilities of librarians, historians and literary scholars could be utilised with a minimum of extra bureaucracy. This renders the bibliographies completely digital, enabling usage and study not generally available in Bowers' time. This includes the analyses presented in this chapter.

Being integrated into the catalogue, the bibliographies have also become searchable by end-users via the general catalogue discovery service. However, it also has the effect that the bibliographies thus implemented are not 'pure' bibliographies, but partly catalogues in the sense that their entries point to the National Library holdings, although they also include entries not yet held by the library. Another significant effect of the bibliography-integrated-in-catalogue approach is that their descriptions are restricted by the same metadata standards and rules governing any bibliographical record in the catalogue. However, the metadata format used so far has proved rich enough to encompass most of the additional information the bibliographies might need, at the level of detail currently practised.

Currently, not all the special bibliographies have all their records in the main catalogue. For example, the extensive Ibsen bibliography is still in the process of migration. The bibliographies investigated in this study were chosen partly to span the diversity of the National Library of Norway's bibliography work, partly because they exist in the main catalogue in their entirety, making their contents readily available for harvesting. All the included specialised bibliographies are

encoded in the *MARC 21 Format for Bibliographic Data (MARC)*.[1] For each record, its membership in one or more bibliographies is encoded in a dedicated field specific to National Library of Norway – field 913, with a unique code for each bibliography.

The selected bibliographies are described in more detail below. Except for the Sámi bibliography, all are complete, in the sense that they are the result of completed projects. They are nevertheless updated whenever cataloguers become aware of relevant resources.

The 1814-bibliography (*1814*) contains 4,204 records, representing literature from and about the period 1812–1814.[2] It includes Norwegian and foreign books, articles in books and serials, related to the events in Norway and the Nordic countries between 1812 and 1814. Some articles in Norwegian newspapers are also included.

The Bjørnstjerne Bjørnson bibliography (*bjørnson*) contains 5,564 records representing newspaper articles, some journal articles and speeches by Bjørnson, as well as literature about Bjørnson. The records of newspaper articles by Bjørnson are based on an archive of clippings and photocopies, processed and indexed at the National Library. The literature about Bjørnson includes Norwegian and foreign books and articles.

The Collett bibliography (*collett*) contains 1,162 records, representing literature by and about Camilla Collett. Norwegian and foreign books, as well as articles in books, newspapers and other serials are included.

The Hamsun bibliography (*hamsun*) contains 11,224 records, representing literature about Knut Hamsun. Books, articles in books, newspapers and other serials are included, as well as some audio-visual material. It contains material published in Norway, in other European countries and, to some extent, in other parts of the world.

The bibliography Literature on Norwegian-American Subjects (*norsk-amerikansk*) contains 9,760 records, representing literature

1 https://www.loc.gov/marc/bibliographic/.

2 The number of records was counted on 30 March 2022 for all analysed bibliographies.

about Norwegian-American issues, mainly related to emigration from Norway to North America (USA and Canada), and to Norwegian immigrants in North America from 1825. Both books and articles in newspapers and other periodicals are included.

The bibliography Norwegian Books 1519–1850 (*norske bøker*) contains 6,637 records, representing books printed or published in Norway up to and including 1850. Books from the same period that were printed or published abroad, but written, translated or edited by Norwegians, or published on behalf of Norwegians, are also included.

The Alf Prøysen bibliography and discography (*prøysen*) contains 11,419 records, representing literature by and about Alf Prøysen, as well as recordings of his songs. Both Norwegian and foreign books, as well as articles in books, newspapers and other serials are included. The discography contains musical objects by Prøysen and interpretations of his songs by others.

The Sámi bibliography (*samisk*) contains 33,638 records, representing literature and other resources pertaining to the Sámi population and culture in Norway, both fiction and non-fiction (all subjects). Emphasis is placed on local history and personal history. Books, audiobooks, book chapters, articles from selected periodicals, ephemera, governmental publications, sheet music, recorded music and film are included, although the latter to a lesser extent. The Sámi bibliography constitutes a part of the Norwegian National Bibliography.

The Solstad bibliography (*solstad*) contains 2,738 records, representing literature by and about Dag Solstad. Norwegian and foreign books as well as articles in books and serials are included, as are reviews of books by Solstad in Norwegian and some foreign newspapers.

The bibliographies may be subdivided into four, basically thematic categories: personal bibliographies, usually developed in the context of anniversaries of acclaimed Norwegian authors (*solstad, prøysen, bjørnson, collett, hamsun*); bibliographies covering historical events or periods (*1814, norsk-amerikansk*); bibliographies representing the history of publishing during a specific time period (*norske bøker*); and, lastly, the part of the national bibliography focusing on the Sámi population and culture (*samisk*).

The above categorisation relates exclusively to subject scope. While satisfying Bowers' claim that every bibliography should have a "unified subject" and a "definite purpose", this categorisation is as such orthogonal to the enumerative-analytical spectrum (Bowers 1994, 18). Clearly none of the selected bibliographies falls into just one of the categories enumerative, analytical/descriptive, analytical/textual or historical. If we instead consider the Bowers categories as features and evaluate the extent to which each bibliography displays the various features, we see that all the bibliographies are enumerative to a marked degree. Similarly, most of them display aspects of descriptive bibliography, in that material details are described to a varying extent, although probably not to the satisfaction of Bowers' stark requirements. Even so, recognition of publications' material equipment as important information worthy of preservation has grown considerably in the National Library of Norway in recent years (Høsøien 2014, 118). This is especially true of old and rare books, of which only the bibliography *norske bøker* contains a significant number (2,583 books published before 1800). It has also affected author bibliographies that include the works of the bibliographed author, in the sense that the physical properties of their various editions are recorded in some detail. None of the selected bibliographies display evident traces of Bowers' remaining categories (textual and historical bibliography).

In the subsequent parts of this chapter, the individual bibliographies will be referred to by their designator as shown in parentheses after the title in the description above.

Using bibliographical data as research data

In many projects within digital humanities, researchers configure and curate their datasets with their specific project in mind. In this way, they come to know their datasets well and can form a reliable opinion of their strengths and weaknesses. This is essential for judging the datasets' usefulness in enlightening the project's area of inquiry. On the other hand, there is always a danger that purpose-oriented curation creates data that are too adapted to the research question (overfitting) and thereby defeat their real purpose.

Bibliographical data created by libraries over time are a different kind of data. They are created not to facilitate a particular study, but for the general purpose of supporting users in finding, identifying, selecting, obtaining and exploring the resources represented by the bibliographic data (Riva, Le Boeuf, and Zumer 2017, 15). As such, they are useful as finding aids for primary data sources (e.g. texts and other content). However, as structured, and relatively 'clean' data, bibliographical records lend themselves to quantitative data analysis. They can therefore function as research data in their own right. Zeng (2019, 3) points out the "tremendous opportunities for humanities researchers to unearth nuggets of gold" from data held by libraries, archives and museums, including structured and semi-structured data. Semi-structured data are unstructured parts of (an otherwise structured) bibliographical record, typically note fields.

It should, however, be noted that the creation of bibliographic data is governed by an intricate set of rules and regulations, from internationally agreed-upon long-term principles, through international standards, schemas and ontologies for resource description, to local practices guiding the general approach to the description as well as interpretation and use of single information elements. The whole regulatory apparatus that influences the bibliographic descriptions created in libraries should be considered whenever bibliographic data are appraised as potential research data.

In a study by Lahti, Marjanen, Roivainen and Tolonen, enabling the use of metadata as research data is conceptualised as bibliographic data science, inheriting approaches and methods from the more generic field of data science (Lahti et al. 2019). In this field, data harmonisation is put forward as a necessary first step, especially when processing federated bibliographical datasets. While acknowledging that harmonised data increases the possibility of reliable results, no specific harmonisation step has been performed in the data analyses presented in the subsequent sections. Instead, I have aimed for algorithms that are able to handle identified inconsistencies and incompleteness during processing, since the datasets are very small and have all been created by employees at the National Library. Great effort is generally put into

metadata quality and richness of the special bibliographies, making them suitable as datasets for experimental bibliographical data science as they are. That said, cataloguing for bibliographies does have to follow the general cataloguing practice guidelines, so there is no guarantee of complete consistency within one bibliography. Local cataloguing practice is much discussed and subject to relatively frequent changes. Adding to this is the implementation of new cataloguing rules. The National Library started implementing the *Resource Description and Access* (RDA) standard (RDA Steering Committee (RSC) 2022) a few years ago, and this continues to cause changes in cataloguing practice.

Methodology

This study takes an exploratory approach. With the described bibliographies as cases, I wish to demonstrate the possibilities and challenges connected with bibliographical data as information sources. Analysing those relatively small yet rich datasets, my goal has been to discover and compare patterns in the data themselves, illustrate the effect that varying cataloguing practice and metadata quality issues may have on analysis reliability, as well as gain insights into the actual resources represented by the data.

The analyses are available on Github and have been performed by the following process (Ohren 2024):

Firstly, the datasets were collected. The bibliographies were harvested via a standard protocol for metadata harvesting as individual sets of records in MARC. The analyses were performed using the Python programming language in Jupyter Notebook, including dedicated packages for handling MARC records, statistical data and geoinformation.[3]

Secondly, candidate analyses were outlined as a list of potentially relevant information to extract from the bibliographies. Starting with characteristics of the datasets as such, their individual sizes, in terms of number of records as well as pages, are calculated and

3 https://www.python.org/.

compared. Pairwise overlaps between the bibliographies are also identified. The focus is then transferred to the resources (e.g. articles, books) represented by entries in the bibliographies. Important characteristics to identify here are medium (distinguish between text and other media) and bibliographic level (distinguish between monographs and components). Another area of investigation is the agents connected to the resources. As an example, gender distribution among the main authors is analysed. The next analysis involves relating bibliographies to time and place in the 'outside world'. This is performed for selected bibliographies by connecting publishing year to historical events and plotting geographical subjects on a map. Lastly, a closer look is taken at the resource content, by analysing genre and certain types of subjects represented in subsets of the bibliographies.

The third and final methodological step is to program and run the analyses. For each potential analysis, an initial, informal evaluation of its potential outcome is performed by means of spot samples and test runs. The objective is to identify which datasets to include and which metadata fields to investigate, and to outline the overall algorithm for the data processing.

Analysis 1: Sizing up the datasets

One of the advantages of the specialised bibliographies is their relative smallness, making them readily available for experimentation without the need for large computational resources. Indeed, that is a general advantage of handling metadata instead of the digital or digitised resources they represent. This analysis tries to reveal the relationship between the number of records and the size of the accumulated content represented by the records. The diagram below shows the size of the bibliographies, measured as number of records.

In addition, it may be of interest to calculate the bibliographies' volume using content-specific units, in order to come closer to the 'real' volume. How to measure content varies across content types; audio content is measured differently from text content, for example. Though there is reason to believe that the bibliographies mainly reference text content, this must be verified. Consulting a specific

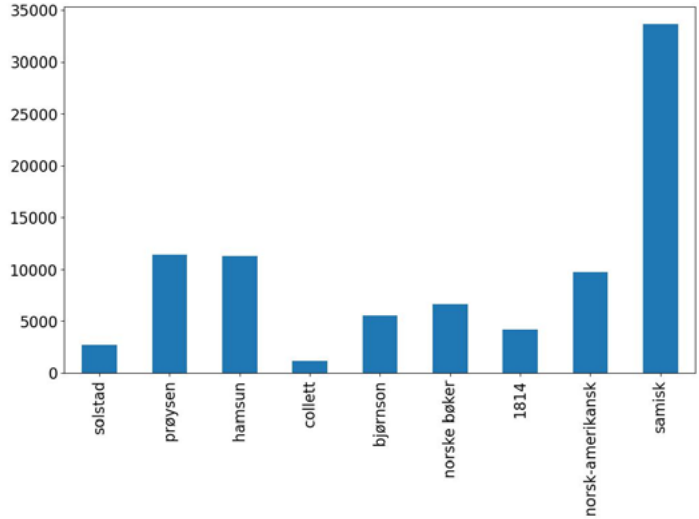

Figure 1: Size of bibliographies – in number of records

element (position 06 in the record leader) gives the following result. As can be seen, all the bibliographies, with one exception, contain almost only text.

Table 1: Material type of the bibliographies, in number of records

	Text	Sheet music	Audio (not music)	Audio (music)	Other
solstad	2725	0	10	0	3
prøysen	5804	934	206	4422	53
hamsun	11166	34	11	1	12
collett	1146	4	6	0	6
bjørnson	5554	3	1	0	6
norske bøker	6634	2	0	0	1
1814	4198	0	3	0	3
norsk-amerikansk	9658	38	8	6	50
samisk	32839	331	156	185	127

Calculating the text volume for each bibliography should therefore give a fair indication of the size of its total content. In a MARC record, the extent of the described resource is encoded in field 300 subfield a. Here, we do not have the benefit of controlled, formal values. Although they adhere to semi-formal syntax rules, the extent is ultimately given as a string, varying from a simple number of pages to a rather complicated expression describing different paginations separately. Examples from *norske bøker*:

15 s.
1 bl., 131, [1] s., 1 fold. pl.
2 b. i 1 (16, 282, 180 s.)
1 bl., S. [279] - 843

To calculate the approximate total number of pages expressed by those strings, the extent expressions in field 300a were decomposed into *extent components* separated by parenthesis and comma.[4] Then the number of pages in each extent component was calculated separately and then summed.

To obtain this result, the most common types of extent components were identified, based on data inspection as well as the MARC specification of field 300. Typical extent components are simple integers with or without a succeeding counting unit (e.g. 15 s.; 4 blad; 16 (as in 16,282,180 s.)), and numeric intervals, with or without a preceding counting unit (e.g. S. [279] - 843). Then regular expressions for each component type were designed and applied to extract the actual extent components in a given bibliography.

As mentioned, the text volume calculated by the above analysis is approximate. For example, Roman numerals are not interpreted, hence are not included in the final sum. Moreover, *blad* (leaves) are counted as pages, even though there is no telling whether the leaves

4 In the examples, 's.' stands for pages, 'bl.' for leaves, 'fold. pl.' for folded charts and 'b.' for volumes.

have content on one side or both. However, even with these shortcuts, it is probable that the calculations performed paint a reliable picture of the text content in the bibliographies.

Below, the size of the bibliographies is shown in terms of text pages.

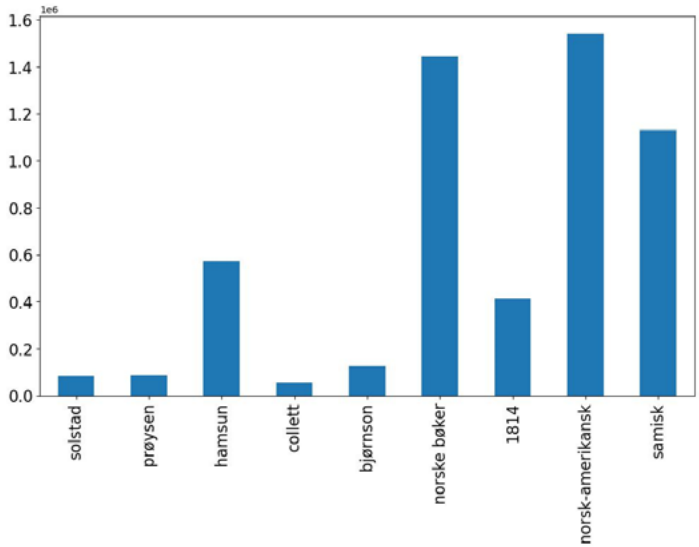

Figure 2: Size of bibliographies – in number of pages

It is evident that the size ranking is different when counting text pages instead of number of records. In particular, the bibliography *samisk*, which has almost three times as many records as the second largest in our sample, contains by no means the largest volume of text. Both *norske bøker* and *norsk-amerikansk* contain much larger volumes of text than *samisk*. Analysing the bibliographic level of the resources represented in the bibliographies may shed some light on this issue.

Analysis 2: Bibliographic level of the resources

The bibliographic level of a record is encoded in the leader element, position 07, and basically distinguishes between whole monographs or serials on one side, and monographic components parts or serial

component parts on the other. In all bibliographies, monographs and monographic components dominate the picture. Serials, articles in serials and integrating resources are combined into Other, of which serials form the majority.

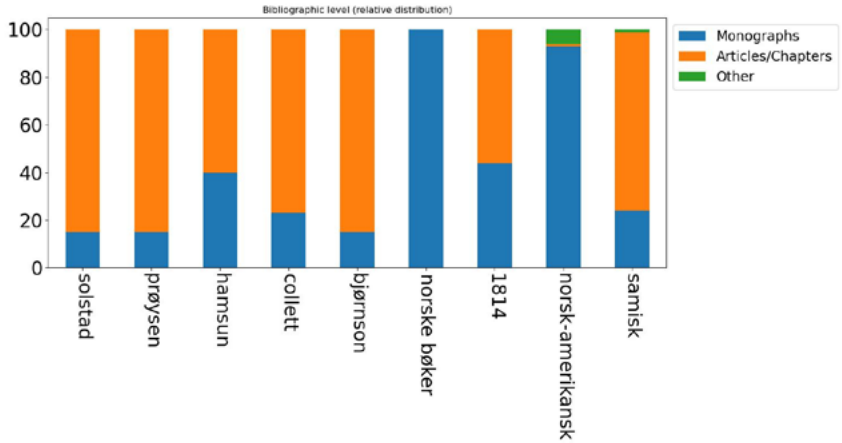

Figure 3: Bibliographical level distribution in the bibliographies

The figure shows that a large part (75 per cent) of the records in the bibliography *samisk* represent chapters or articles. The dominance of articles over monographs is a characteristic shared with most of the other bibliographies, except two. In contrast, *norske bøker* and *norsk-amerikansk* distinguish themselves by containing almost exclusively monographs, partly explaining why *norske bøker* and *norsk-amerikansk* rank highest in terms of the total volume of textual content.

Analysis 3: Overlap between bibliographies

As indicated in the initial descriptions of the selected bibliographies, each one is defined by their specific domain. Looking at the domains in a historical context, there is an obvious potential for overlap between some of the bibliographies. For example:

The authors Bjørnstjerne Bjørnson and Camilla Collett were approximate contemporaries and corresponded by letter.

Knut Hamsun represented the generation of authors following Bjørnson. As a national icon of his time, Bjørnson was someone that

the new generation would naturally have to relate to. Knut Hamsun did so both positively and negatively.

To investigate whether partly shared domains are reflected in any way in the bibliographies, an analysis identifying any pairwise overlap between the bibliographies was carried out. In the metadata records, membership of a bibliography is encoded in a dedicated data field 913, proprietary to the National Library main catalogue.

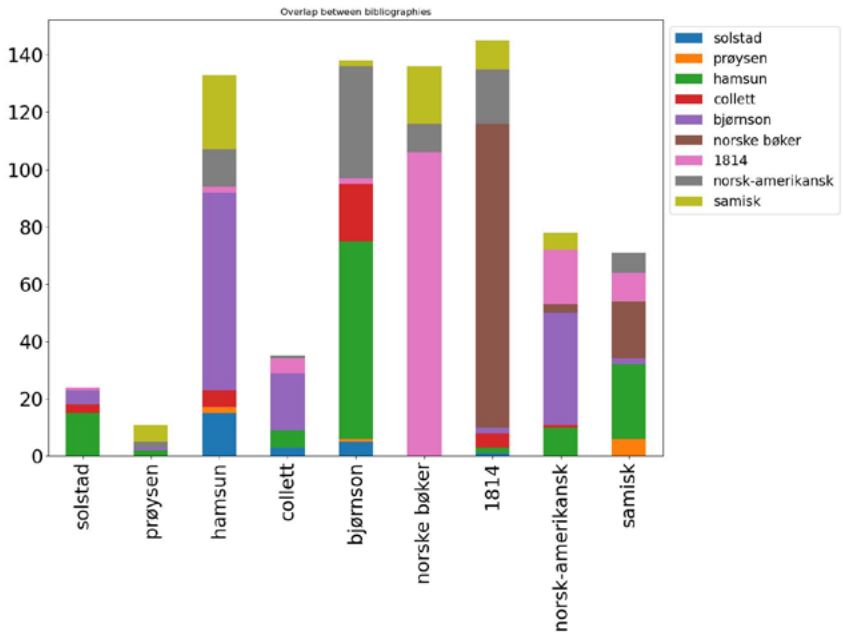

Figure 4: Pairwise overlap between bibliographies

In general, no bibliography shares a large number of records with any other. The maximum overlap between any two bibliographies contains 140 records, constituting a very small portion of the total number. That said, some of the bibliographies share more records than others. For *hamsun* and *collett*, the largest portion of their overlapping records are shared with *bjørnson*. In *bjørnson*, on the other hand, a significant part of its overlapping records is also shared with *norsk-amerikansk*. Bjørnstjerne Bjørnson's active years coincided with the period of mass emigration from Norway to North America. He also visited Norwegian

immigrant communities in the Midwest for several months, during which he made speeches and wrote letters. It is therefore to be expected that he appeared in some of the literature related to emigration, as author as well as subject. Another significant overlap is between *norske bøker* and *1814* (more than 100 documents). Since *norske bøker* is constrained exclusively in terms of material type (books) and publication year (1539–1850), and the thematic focus of the bibliography *1814* is an event taking place in the publication period of *norske bøker,* a certain overlap between the two is inevitable.

Using external sources in the analyses

Bibliographic data partly describe resources or documents in the form of textual descriptions in semi-natural language. However, they also connect the described resources to entities of other types, such as persons, organisations, geographical places or topics – entities which play specific roles in the context of the described resource.

For example, a bibliographic description of a published translation of the novel *Markens grøde,* may connect to *agents* like the original author, a translator and a publisher; to *places* like the place of publication or the place in which the plot unfolds; or *abstract concepts* like genres or topics.

To an increasing degree, such connections are realised by *referring to* the connected entities rather than importing them into the bibliographic data. Typically, the related entities are then managed separately in so-called authority registries, in which a fuller description may be stored. The *Norwegian Authority File: Persons and corporate bodies* (National Library of Norway n.d.) contains information about agents connected to any resource described in the library catalogue shared by the National Library and a large set of other academic libraries. In the bibliographic record, personal agents are referenced by name, year of birth and death and their identifier in the Norwegian Authority File.

It may therefore be necessary to look outside the bibliographical records to retrieve the information required to analyse certain characteristics of agents related to the documents concerned. For information about agents related to the resources included in the bibliographies, the

Norwegian Authority File should be the first port of call. From there, the path leads to the Virtual International Authority File (VIAF) managed by Online Computer Library Center (OCLC) or the International Standard Name Identifier (ISNI), both rich sources of information about agents.[5]

There are also relevant resources of a more general nature, like various statistics prepared by Statistics Norway (SSB), the national statistical institute of Norway.[6]

Analysis 4: Main authors' gender

In this analysis, I wish to focus on the creators of the resources in the selected bibliographies, particularly on the gender of the main authors represented therein. Information about main authors who are individual persons, not organisations, is encoded in field 100 in MARC records. Note that MARC allows only one occurrence of field 100 per record, leaving eventual co-authors and other contributors to be registered in the less prominent field 700. The latter is not included in this analysis, which restricts its scope to main creators (mostly authors) at the expense of co-authors, illustrators, translators, interviewers, editors and other contributors. Including all personal contributors would no doubt create a richer picture of the 'gender situation' and could be an interesting subject for future study.

The main creator field 100 contains sparse information about the person in question. Only name, lifespan and authority identifier are included. Information about the gender of the various creators must be retrieved from the Norwegian Authority File.

The two million authority records are of mixed quality, ranging from records with almost no information to rich records containing multiple name forms, links to synonym authorities, references to other registries, information about gender, country of origin and more. A record's quality level is encoded explicitly in a local data field according

5 VIAF: http://viaf.org/, ISNI: https://isni.org/.

6 https://www.ssb.no/en.

to a simple three-value scheme. At present 9 per cent of the authorities are assigned the indicator for the highest quality, while 72 per cent are of middling quality. An initial examination of gender information in the authority records, shows that 86 per cent of the authorities represent individual persons. Of these, only 8 per cent include information about gender. On the other hand, narrowing down to person records of the highest quality, as many as 81 per cent include gender information. In consequence, a significant proportion of the authority records representing main authors in the selected bibliographies must be expected to lack gender information.

As anticipated, the proportion of main authors with missing gender information is high, ranging from 11 per cent to 67 per cent across bibliographies. To fill in some of the missing information, data from Statistics Norway's statistics on first names in use in Norway during the period 2013–2023, subdivided into female names and male names (Statistics Norway 2023), were consulted. Matching the first names of main authors to the names from Statistics Norway's lists reduced the proportion of main authors with missing gender information from 11–67 per cent to 6–29 per cent across bibliographies.

The resulting gender distribution is shown in Figure 5.

The dominance of documents by male authors is evident. In the author bibliographies, this must be seen in the connection with the percentage of documents written by the celebrated authors themselves. For example, Camilla Collett has written about 30 per cent of the items in 'her' bibliography, accounting for the relatively large proportion of female authors. Since all the other person bibliographies are about male authors, any item written by them personally will of course add to the male count. The bibliographies *norske bøker* and *1814* focus on a time period during which women played only a small role in public life. Female published writers were therefore few and far between. This is reflected in *norske bøker,* in that only 1 per cent of the books are written by female authors.

As the national bibliography for the Sámi population and culture, *samisk* is the most generic and contemporary of the selected bibliographies. As such, there is reason to expect that *samisk* will be the one

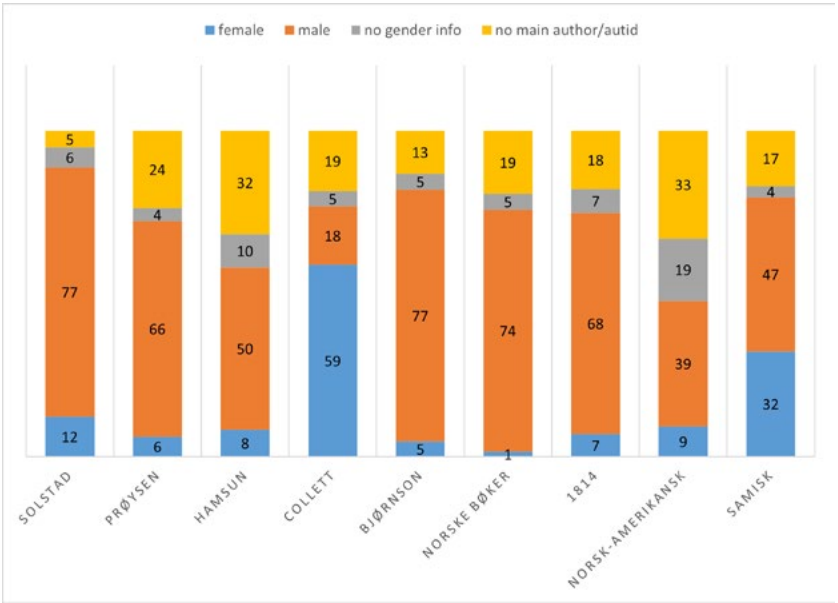

Figure 5: Distribution of main author's gender in each bibliography.

that best reflects the current situation when it comes to the female share of total intellectual output. In this bibliography, 32 per cent of the described resources were created by women as opposed to 47 per cent by men, which is not quite as disproportionate as in the other bibliographies. Data from Statistics Norway about professional writers, show that the average distribution of male and female writers during the time period 2016–2024 is close to fifty-fifty (Statistics Norway 2024). While the percentage of female authors does not necessarily correspond to the percentage of intellectual output by females, it is anticipated that the current even gender distribution of writers will gradually be reflected in the general Norwegian National Bibliography, including the Sámi bibliography.

Analysis 5: Placing the documents on a map of the world

None of the bibliographies are, as a whole, thematically related to specific places. Nevertheless, an initial analysis reveals that some of them include a significant number of resources about geographical places, i.e. resources having geographical places as subjects. Since its main

topic is emigration, the bibliography *norsk-amerikansk* forms a suitable focus for a geographical analysis. It should nevertheless be noted that although 25 per cent of its documents relate to geographical topics, it is not the bibliography with the highest geographical score.

Geographical places as subjects are encoded in a dedicated metadata field (field 651) in the form of place names, sometimes as a single name only ('Egersund'), sometimes with a contextual qualifier, like 'Binford (North Dakota)'. While naming consistency has obviously been aimed for, some place names are registered in a variety of linguistic forms. For example, 'North Dakota', 'North-Dakota' and 'Nord-Dakota' all occur in the metadata, as do 'USA' and its Norwegian counterpart 'Forente stater'.

In contrast to the agents in Analysis 4, geographical places are merely encoded as textual strings with no explicit connection to external registries. However, there are several geographical services against which place names may be matched. For this purpose, an open, multilingual resource with global coverage and easily available programming interfaces is needed. The geographical database GeoNames[7] was chosen. This is a community-driven resource of many years' standing, based on official national resources[8] and with prominent users like the UK's BBC and its Norwegian equivalent, NRK. Figure 6 illustrates the result of matching the geographical topics in the bibliography *norsk-amerikansk* to GeoNames and rendering the places on a map. Each red dot indicates geographical coverage of the corresponding place, and the dot size is scaled according to the number of documents about the place.

As expected, the majority of the documents concerning places are about Norway and North America, but other countries also appear as subjects. Zooming in to Norway, as shown in Figure 7, we see that the topical coverage of places in Norway also to some degree illustrates the extent of emigration from the various parts of Norway.

7 Geonames.org.

8 For Norway: Norwegian Mapping Authority (kartverket.no) and Statistics Norway (ssb.no).

Figure 6: Geographical subjects in the *norsk-amerikansk* bibliography.

According to emigration statistics (Sosialdepartementet 1921), the majority of the Norwegian emigrants to North America during the period 1836–1915 originated from southern Norway, both the interior and along the coast as far north as Trøndelag. As Figure 7 shows, this fact is strongly corroborated by the geographical subjects covered by the bibliography *norsk-amerikansk*.

Figure 7: Documents about places in Norway in the bibliography *norsk-amerikansk*.

Interpreting geographical names in bibliographical records

Automatic, retrospective matching of place names in metadata against a geographical database is challenging and involves numerous ambiguities. Most place names designate many places in the world, typical examples of which are the many cities in Europe with one or more namesakes in the USA. GeoNames supports a number of parameters

for configuring the matching method. For instance, searches may be restricted to places of certain types (administrative units, inhabited places etc.). However, this is not enough to provide an error-free map from the geographical information in the bibliography *norsk-amerikansk*. Two issues in particular cause problems. Firstly, some placenames with qualifiers have the most specific name as a qualifier instead of the name of the surrounding area. Hence, a book about San Pedro in California, should, according to cataloguing practice, be assigned the subject term 'San Pedro (California)'. However, 'California (San Pedro)' has been registered instead. Secondly, many place names occur in metadata without qualifiers at all, for example 'Valle' and 'Ål'. From such place names it is virtually impossible to decide which part of the world is referred to. In the case of Valle, which probably refers to Valle in Setesdal, GeoNames put forward a department in western Colombia, namely Valle del Gauca.

After some experimentation with various search configurations, the best results were obtained by prioritising search results of type administrative unit (including municipality, county, state and country), and using the subject's qualifier as the primary search term in GeoNames. Hence, for the subject 'Bison (South Dakota)', 'South Dakota' was searched for among administrative units, and accepted when found. This creates a less granular map than the subject terms indicate and is faulty for the inverted qualified names. All in all, however, it was deemed to be the lesser of several evils, and gave a fair overall picture of which places on Earth the resources in *norsk-amerikansk* are about.

Nevertheless, it seems clear that the only reliable way to place documents correctly on the world map is to encode places in a unique manner, by using globally unique identifiers from some geographical authority database, or names combined with geographical coordinates. As long as mere place names are used as indicators of geographical topics, the geographical placement of documents necessarily requires a certain element of interpretation, with an unknown number of errors as result.

Analysis 6: A chronological perspective on publishing patterns in the bibliography 1814

This analysis studies how the volume of published works in a bibliography varies through time, and how this may relate to the domain of the bibliography in question. *1814* focuses on a very important historical event, namely the creation of the Constitution of Norway in 1814. The year of publication is uniquely encoded in field 008, and Figure 8 illustrates the number of resources published each year from the earliest publication year represented in the bibliography.

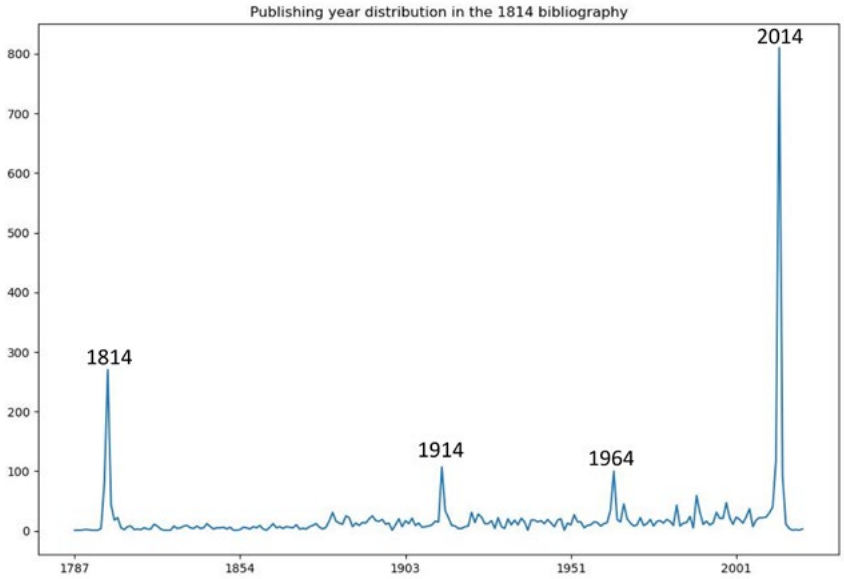

Figure 8: Publication years in the bibliography *1814*

As clearly shown in the figure, documents relating to the events of 1814 have been published steadily but at a very low level during the whole period, from the earliest publication in 1787 to the present time, peaking markedly around the 100th, 150th and – in particular – 200th anniversaries of this occasion.

Analysis 7: Genres represented in the bibliographies

Bibliographic datasets like the ones analysed here are, of course, interesting to study from several perspectives.

In this analysis, my perspective will be the intellectual content of the resources. In traditional bibliographic records, the intellectual content is usually described by assigning descriptors representing subjects or genres. Such descriptors come in many forms, from carefully defined and uniquely identified concepts to natural language terms or phrases. In both cases, the descriptor may be *controlled*, meaning that it belongs to some knowledge organisation system, for example a thesaurus or classification system, within which context it must be interpreted. Such descriptors may have defined relations to other descriptors within the same system. If so, the existence of such a descriptor in a metadata record may give more information than meets the eye. This is because any relations to other descriptors in its originating system constitute part of its semantics, and therefore add to its apparent meaning whenever used in metadata descriptions.

In contrast to controlled descriptors, there are *uncontrolled* descriptors, indicating that the choice of descriptor is entirely up the individual metadata creator.

In the bibliographies analysed here, both subject and genre information originates from controlled vocabularies of natural language terms in the language Norwegian Bokmål. The controlled vocabularies are flat term lists with little information except the terms themselves. Their reliability in use therefore depends entirely on consistent interpretation of natural language – consistent through time as well as across communities.

Because each bibliography has its separate subject vocabulary, the subject term lists are not suitable for analysing subjects across bibliographies. In contrast, for information about genre and form, the same vocabulary is used in all the bibliographies, except *samisk*. This makes it possible to compare genre coverage between bibliographies. In MARC 21, genre and form are encoded in field 655, with term in subfield a and a designator for the vocabulary used in subfield 2. In

this analysis, only terms from the Norvok vocabulary are assessed.[9] For historical reasons, the corresponding vocabulary designator encoded in 655 is sometimes 'norvok', sometimes a bibliography-specific designator. At the time of data collection, the newly developed *Norwegian thesaurus on genre and form* (National Library of Norway et al. 2019) has recently replaced Norvok as source of genre concepts. This change in practice has, as yet, left few traces in the bibliographies at large, and is therefore ignored in the analysis described here, namely a comparison of genres in *1814* and *norske bøker*.

Meaningful visualisation of the frequency of several hundred ungrouped genre terms is a challenge. Below, the genres represented in the bibliographies *norske bøker* and *1814* are visualised as word clouds, which generally display high-frequent textual values in a clear and intuitive manner.

Figure 9: Genres represented in the bibliographies *1814* (left) and *norske bøker* (right)

The word clouds above indicate a somewhat sharper genre profile in the bibliography *1814* than in *norske bøker*. While *1814* includes 56 genres, of which a handful (*Debatt, Grunnlov, Oversiktsverk, Brev*)[10] are used at a significantly higher frequency than the rest, *norske bøker* displays a much more even genre profile. Of the 180 genres used, none dominates to the same extent, with the possible exception of *Oppbyggelig litteratur*[11]. The genre profiles in the two bibliographies seem to

9 https://bibliotekutvikling.no/content/uploads/sites/8/2021/01/norvok_sjangre20210119_cleaned.pdf.

10 Debate, Constitution, Overviews, Letters

11 Spiritual literature

align well with their domains. The domain of *norske bøker*, constrained by time of publication only, naturally implies a broader and more evenly used set of genres than does *1814*, whose domain is defined as literature related to a certain historical event.

Analysis 8: Subjects in the author bibliographies

As mentioned above, no common topical vocabulary has been used to denote general subjects covered in the bibliographies. However, in the five person bibliographies, *bjørnson*, *collett*, *hamsun*, *prøysen* and *solstad*, the persons themselves frequently represent the included documents' topic, as do the persons' works. This analysis investigates each of the person bibliographies according to their distribution of documents about the person, documents about the person's works as well as documents by the person.

Persons as subjects are encoded in a dedicated metadata field (field 600), with the person's name registered in subfield a. A person's work as subject is encoded in the same field type, with the title of the covered work encoded in subfield t. Hence, a 600-field containing both person name and title indicates a work as subject, whereas a 600 field without the title field indicates a person as subject. The result is visualised in Figure 10.

Note that all three distribution groups are counted separately. For some documents, both the person and their works have been registered as separate subjects, which may cause the total to exceed 100 per cent. On the other hand, it is not always the case that the three groups span the whole bibliography. This is particularly the case in *prøysen*, where a large proportion of the included documents are songs and other musical resources catalogued with the composer as the main creator and Prøysen as contributor (lyricist or author). As a partial remedy for this, records in which Prøysen is registered in field 700 with lyr or aut in subfield 4 (i.e. is the lyricist or author of the described resource) are counted as 'by Prøysen' in this analysis. However, it is clear that including only field 100 as creator for this analysis, combined with the cataloguing practice for musical resources as manifested in the data, creates a less than reliable

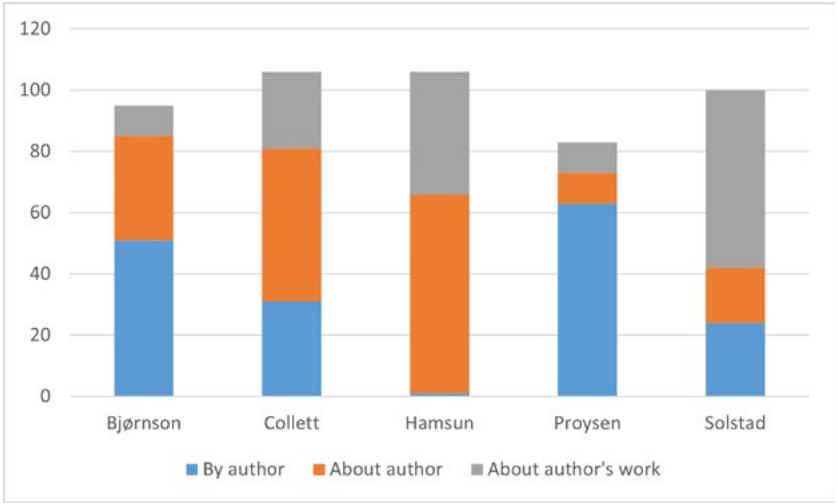

Figure 10: Distribution of documents by a person, about a person and about a person's works

picture of the contributions of writers of lyrics for which others have composed melodies.

For most of the person bibliographies, this grouping gives an informative overall picture of their topical content, which may also be useful input for further development of the collections behind individual bibliographies. However, to judge whether the distribution of these three groups is as desired for a certain bibliography, its declared scope as paraphrased in an earlier chapter must also be considered. For example, works by Hamsun fall outside the declared scope of *hamsun*. The small proportion of works by Hamsun should therefore not be of concern, rather a surprise that they are included at all.

Comparing the bibliographies of the contemporaries Bjørnstjerne Bjørnson and Camilla Collett, there are relatively fewer documents about Bjørnson's work than about Collett's. Since *collett* is a much smaller dataset than *bjørnson*, no firm conclusions can be drawn from this analysis alone. However, it might be worth reflecting on the relatively small percentage of documents about Bjørnson's works in *bjørnson*. It might express lack of interest for the works of Bjørnson among literary scholars, but it might also indicate deficiencies in the bibliography. However, a much deeper knowledge of

Norwegian literary history than this author possesses is necessary to form any informed opinion about this.

Conclusion

This paper has tried to explore how datasets composed of bibliographical records may be used to glean information about both the resources they represent and related resources, such as authors and geographical places. The analyses described indicate that bibliographic data are well suited for painting the big picture, for example by uncovering patterns both within a single dataset and across several datasets. Moreover, many of the analyses involve extraction of the range of values assigned to specific fields. Any erroneous values are thereby discovered, which could greatly support efforts to retrospectively enhance metadata quality.

On the other hand, relatively intimate knowledge of the data themselves, as well as the processes around their creation and management, is necessary to design reliable analyses. This prerequisite is easy to fulfil acceptably well when handling small datasets created in one's own organisation. However, projects intending to analyse large datasets from several, possibly external, sources will have to devote resources to pre-analyses and harmonisation of the datasets.

References

BOWERS, FREDSON. 1952. "Bibliography, Pure Bibliography and Literary studies." *The Papers of the Bibliographical Society of America* 46 (3): 186–208.

———. 1994. *Principles of Bibliographical Description*. Winchester: St. Paul's Bibliographies.

HØSØIEN, HEGE STENSRUD. 2014. "Bibliografi + Nasjonalbiblioteket = sant." *Biblioteca Nova* 2014 (4): 117–121.

KEYNES, GEOFFREY. 1953. "Religio Bibliographici: Presidential Address to the Bibliographical Society, March 1953." *The Library* s5-VIII (2): 63–76. doi: 10.1093/library/s5-VIII.2.63.

LAHTI, LEO, JANI MARJANEN, HEGE ROIVAINEN AND MIKKO TOLONEN. 2019. "Bibliographic Data Science and the History of the Book (c. 1500–1800)." *Cataloging & classification quarterly* 57 (1): 5–23. doi: 10.1080/01639374.2018.1543747.

MCKENZIE, D.F. 1999. "The book as an expressive form." In *Bibliography and the sociology of texts*, 9–29. Cambridge, UK: Cambridge University Press.

MERRIAM-WEBSTER DICTIONARY. n.d. "bibliography". Accessed 9 May 2024. https://www.merriam-webster.com/dictionary/bibliography.

NATIONAL LIBRARY OF NORWAY. 2009. "Hamsun bibliography. Literature on Knut Hamsun." https://www.nb.no/bibliografi/hamsun.

NATIONAL LIBRARY OF NORWAY. 2010. "Bjørnstjerne Bjørnson bibliography. Newspaper articles and speeches by Bjørnson, literature on Bjørnson." https://www.nb.no/bibliografi/bjornson.

NATIONAL LIBRARY OF NORWAY. 2011. "Solstad bibliography. Literature by and on Dag Solstad." https://www.nb.no/bibliografi/solstad.

NATIONAL LIBRARY OF NORWAY. 2012. "Collett bibliography. Literature by and on Camilla Collett." https://www.nb.no/bibliografi/collett.

NATIONAL LIBRARY OF NORWAY. 2013. "1814 bibliography. Literature from and about 1812–1814." https://www.nb.no/bibliografi/1814.

NATIONAL LIBRARY OF NORWAY. 2014. "Alf Prøysen. Bibliography and discography." https://www.nb.no/bibliografi/proysen.

NATIONAL LIBRARY OF NORWAY. 2019. "Norwegian books 1519–1850. Bibliography." https://www.nb.no/bibliografi/nor1519.

NATIONAL LIBRARY OF NORWAY. n.d. "Literature on Norwegian-American subjects." https://www.nb.no/bibliografi/noram.

NATIONAL LIBRARY OF NORWAY. n.d. "Norwegian Authority File: Persons and Corporate Bodies." Accessed 31 January 2024. https://bibsys-almaprimo.hosted.exlibrisgroup.com/primo-explore/search?vid=AUTREG&lang=en_US.

NATIONAL LIBRARY OF NORWAY. n.d. "Sami bibliography." https://bibsys-almaprimo.hosted.exlibrisgroup.com/primo-explore/search?vid=SAMISK&lang=en_US.

NATIONAL LIBRARY OF NORWAY, BIBLIOTEKSENTRALEN SA, BOKBASEN AS AND UNIVERSITETSBIBLIOTEKET I OSLO. 2019. "Norwegian thesaurus on genre and form." https://www.nb.no/nbvok/ntsf/en/.

OHREN, O.P. 2024. "NLN-bibliographies-analysis." Github. https://github.com/olufine/NLN-bibliographies-analysis.

RDA STEERING COMMITTEE (RSC). 2022. "RDA: Resource Description and Access." The American Library Association, The Canadian Federation of Library Associations, CILIP: Chartered Institute of Library and Information Professionals. Accessed 3 April. http://www.rda-rsc.org/content/about-rda.

RIVA, PAT, PATRICK LE BOEUF AND MAJA ZUMER. 2017. *IFLA Library Reference Model. A Conceptual Model for Bibliographic Information*. Den Haag: International Federation of Library Associations and Institutions.

SOSIALDEPARTEMENTET, NORGE. 1921. Utvandringsstatistikk = Statistique de l'émigration. In *Norges offisielle statistikk*. Kristiania.

STATISTICS NORWAY. 2023. 10501: Persons, by first name, contents and year, edited by Statistics Norway. https://www.ssb.no/en/statbank/sq/10097269 and https://www.ssb.no/en/statbank/sq/10097268.

——— . 2024. 13352: Employees and jobs, by sex, occupation, quarter and contents, edited by Statistics Norway. https://www.ssb.no/en/statbank/sq/10097266.

TANSELLE, G. THOMAS. 1994. "Introduction." In Fredson Bowers, *Principles of Bibliographical Description*, v–xiv. Winchester, St. Paul's Bibliographies.

YEE, MARTHA M. 2007. "Cataloging Compared to Descriptive Bibliography, Abstracting and Indexing Services, and Metadata." *Cataloging & Classification Quarterly* 44 (3-4): 307–327. doi: 10.1300/j104v44n03_10.

ZENG, MARCIA LEI. 2019. "Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article." *El Profesional de la Información* 28 (1). doi: 10.3145/epi.2019.ene.03.

6. Bibliographic Needs in Literary Critical Reception and Periodical Studies: The Database *Norsk Litteraturkritikk* as a Case in Point

Sissel Furuseth

To perform literary critical reception studies typically means analysing texts that examine and evaluate books and authors. These evaluative texts are normally published in periodicals: newspapers, journals and magazines. Conceptually and relationally, one could say that book reviews are partly 'outside' and partly 'inside' items that are listed in the library records. While books and periodicals are counted in, most book reviews and essays are not. For example, *Bergens Tidende*'s review (15 November 1888) of Amalie Skram's novel *Lucie* is not registered in Norwegian library records, whereas the newspaper and the novel itself are. Such omissions do not necessarily mean that book reviews and critical essays are of secondary value for bibliographers, but these short, often unsigned pieces in the press appear to be of a secondary bibliographical order and are therefore not systematically included in the records. The bibliographic hierarchy may reflect an attitude held by many literary scholars as well: literary criticism, as an intellectual practice of evaluating books and coming to terms with processes of canonisation and the ways literary works have been interpreted, is secondary to the literature itself. The term "secondary literature" reflects this order of things.

In recent decades, however, book reviews and other forms of periodical criticism have received increasing attention as primary material worth analysing on their own terms. This is made possible by

the growth of digitised newspaper and magazine archives worldwide, such as the Library of Congress's Chronicling America, ProQuest's Periodicals Archive Online, HathiTrust Digital Library, Deutsches Textarchiv, ANNO (Austrian Newspaper Online), Kungliga biblioteket (National Library of Sweden), and NB.no (National Library of Norway), to mention just a few. How well the metadata of these archives satisfy research needs in periodical and literary critical reception studies varies. In many places, therefore, special databases have been established. *Women Writers in History* is a Dutch initiative that focuses especially on mapping and analysing women's intellectual culture in pre-1900 Europe.[1] The *Modernist Journals Project* is a transatlantic initiative that covers English-language literary magazines from the 1890s to the 1920s, offering rich metadata and high-quality scans of single issues.[2] I will come back to some of these examples in my closing reflections.

In this chapter, I will present the online bibliography *Norsk litteraturkritikk* (https://ub-baser.uio.no/norsk-litteraturkritikk) as a case in point. Since the first version was made in the early 1980s, it has been upgraded several times, both technologically and in terms of typology and structure. The database can thus be approached both as a fascinating historic artefact from the 1980s and as a dynamic work-in-progress with a long life ahead. Even though *Norsk litteraturkritikk* in its current state hardly represents the ultimate solution to literary critical reception studies, I will argue that the collective efforts to improve the database's structure and taxonomy over the years bring us to the very core of what motivates this anthology: to understand the hermeneutics of metadata. What kind of knowledge can be extracted from a bibliography? In the following I will discuss some hermeneutical challenges

1 An early version of the bibliography (dating back to 2009) is available at http://www.womenwriters.nl/index.php/Database_WomenWriters. A more recent version of the database is archived at https://womenwriters.rich.ru.nl/. *Women Writers in History* is also a DARIAH-EU working group: https://www.dariah.eu/activities/working-groups/women-writers-in-history/.

2 The database and other resources related to the project are accessible from the website https://modjourn.org/.

that arise from incomplete metadata and explain why I regard *Norsk litteraturkritikk*, hosted by the University of Oslo Library (UB), to be a valuable supplement to the National Library of Norway's digitised newspaper and magazine archives.

In this context I have basically two roles: first, I am a daily user of several online library services, and second, I have been a project supervisor with the mandate to improve one specific database. Moreover, my role as a teacher might be relevant to note, as digital archives and databases are fundamental resources in the education of students who want to perform analyses of periodicals and literary critical reception. As a user, I search for information in a wide range of archives and databases in Norway and beyond, while as a project supervisor, I have taken an active role in coordinating editions in the online bibliography *Norsk litteraturkritikk*. As a teacher, I can demonstrate how to add entries, make corrections and combine different databases in the best way. I have no formal education in library and information science but as a Scandinavian Studies scholar, I have researched literary magazines and literary critical reception for more than 20 years. The terminology applied in this chapter reflects this background.

Before we delve into the database, it may be useful to keep in mind that there is no obvious English equivalent to the Norwegian compound word "*litteraturkritikk*". It translates more often as "literary critical reception" than "literary criticism" as the latter has an aura of academic theory that is not necessarily implied in the Norwegian "*litteraturkritikk*". Referring to the (semi)professional recurring practice of reviewing new books in newspapers and other media, the Norwegian notion of literary criticism is more in line with the German "Literaturkritik" (Hohendahl 1982, 13). It designates an established cultural institution situated in-between the literary field, academia and the press. This in-betweenness may be one explanation for why book-reviewing and literary criticism tend to fall between various stools in bibliographic terms.

Another explanation is that different sets of metadata will be useful in different projects, depending on their angle and research questions. Scholars of literary reception studies and scholars of

periodical studies tend to approach literary magazines with different priorities. While the former typically search periodicals to find texts about literary works and authors in order to scrutinise verbal reflections on aesthetic quality, the latter are looking for editorial profiles and the cultural impact of magazines and journals as such. In *Norsk litteraturkritikks historie 1870–2010* (Furuseth et al. 2016), the chapters alternate between author-focused literary reception studies, critic-oriented profiles, and journal-oriented analyses of generations and aesthetic programmes.

Influenced by book history and Jerome McGann's sentence that "[m]eaning is transmitted through bibliographical as well as linguistic codes" (1991, 57), contemporary periodical studies are often oriented towards the materiality of literary magazines. In the introduction to *The Oxford Critical and Cultural History of Modernist Magazines*, Peter Brooker and Andrew Thacker define *periodical codes* as a particular subset of McGann's notion of *bibliographical codes*. They refer to a range of features at play in any magazine, distinguishing between "periodical codes internal to the design of a magazine (paper, typeface, layout, etc.) and those that constitutes its external relations (distribution in a bookshop, support from patrons)" (Brooker and Thacker 2009, 6). Not every code translates easily into bibliographic metadata but information about *editorial arrangements* (name of editor-in-chief, for example) and *frequency* (weekly, monthly, quarterly or irregular) are of great value for scholars in the field of literary reception and periodical studies. I will return to some problems that may occur when the database lacks this kind of information.

What literary reception studies and periodical studies have in common is an interest in networks between critics, authors and other agents in the cultural field. We acknowledge that the cultural institution of literary criticism reaches way beyond Henrik Ibsen, Sigrid Undset and a few other canonised authors awarded special bibliographies at the National Library. The grey eminences and the hard-working mediators between authors and readers are often more important to us, as are the shifting trends in how new books are read and evaluated within different groups of readers. The question is, then,

to what extent do existing library databases offer metadata that give researchers and students access to the kind of information we are after? The fact that scholars at the University of Oslo embarked on the ambitious – and perhaps rash – undertaking of establishing a digital database for literary criticism in the early 1980s (Moi and Linneberg 1990) indicates that bibliographic needs for more accurate metadata in literary critical reception studies have existed for decades. Technology has improved along the way, but the same questions regarding fact-checking, interoperability and balance between precision and standardisation are still at issue.

In the following I will first provide a brief historical background and describe the contents of the bibliography, including some of the deliberations we have had during the project. I delve into a couple of examples to illustrate the meaning of metadata. In the last sections, I return to more overarching questions of principle and contextualise the Norwegian case in a broader international landscape.

The Beyer Base

Norsk litteraturkritikk, also known as *Beyerbasen* (the Beyer Base), is an early example of digital humanities in Norway. The database is the result of two externally funded research projects carried out at an interval of more than 20 years. The first, *Den norske litteraturkritikkens historie 1830–1940* (The history of Norwegian literary criticism, 1830–1940), was headed by Professor Edvard Beyer at the University of Oslo in the years 1980–1984. The second, *Norsk litteraturkritikks historie 1870–2000: verdiforvaltning og mediering* (The history of literary critical response in Norway, 1870–2000: value-judgements and mediation), was an inter-university collaboration, coordinated by myself between 2009 and 2015, at the Norwegian University of Science and Technology (NTNU) in Trondheim.[3] Irene Iversen and Arild Linneberg,

3 Before I relocated to Oslo in 2016, Thorstein Norheim represented the project at the Department of Linguistics and Scandinavian Studies at the University of Oslo. He was supported by research assistants Nora Campbell, Ylva Frøjd, Marianne Stensland, and Jenny Moi Vindegg.

who were research assistants for Edvard Beyer in the early 1980s and eventually became professors of comparative literature, expressed high hopes for a digital database, claiming that the new EDP register would give detailed pictures of historical literary life in ways otherwise (and hitherto) not possible (Iversen and Linneberg 1983, 90).[4] They believed that the very quantity of material would reveal new and unknown structures in terms of genres, gender, geographical representation etc.

Now hosted by the University of Oslo Library, *Norsk litteraturkritikk* provides information about 22,065 critical texts from the late eighteenth century onwards.[5] An appendix in the first volume of *Norsk litteraturkritikks historie 1770–1940* describes the sources used in the first project: Jens Braage Halvorsen's *Norsk Forfatter-Lexikon* (covering the period 1814–1880), Reidar Øksnevad's newspaper bibliographies (the liberal daily *Dagbladet* is particularly well covered) and a selection of special bibliographies of individual authors such as Arne Garborg, Olav Duun, Knut Hamsun and Bjørnstjerne Bjørnson. Øksnevad's newspaper bibliographies have been very helpful regarding identification of pseudonyms and initials. Some periodicals are systematically examined by the project participants themselves, for example the conservative daily newspaper *Aftenposten* from the period 1890–1905 and the feminist magazine *Nylænde* 1887–1916 (Moi and Linneberg 1990, 307–311). The result is an accumulation of entries around 1900, while the early 1800s and late 1900s are relatively poorly covered in the database.

4 The database was first created by historians Ivar Fonnes and Kåre A. Andersen at the Faculty of Humanities' EDP centre, in cooperation with members of the project, mainly Arild Linneberg. In the 1990s and 2000s, the Department of Linguistics and Scandinavian Studies at the University of Oslo was responsible for operating and maintaining the database. In 2007–2008, Kåre A. Andersen (then at the Department of Archaeology and History) upgraded the database.

5 This number describes the status as of 15 August 2023. The database is a dynamic structure where titles are still included and sometimes removed if regarded irrelevant or redundant.

Table 1: Distribution of entries decade by decade in the database *Norsk litteraturkritikk*.

Decade	Number of posts	Decade	Number of posts
1790–1799	11		
1800–1809	2	1900–1909	3456
1810–1819	32	1910–1919	3580
1820–1829	66	1920–1929	3336
1830–1839	420	1930–1939	1925
1840–1849	321	1940–1949	308
1850–1859	493	1950–1959	253
1860–1869	634	1960–1969	223
1870–1879	981	1970–1979	192
1880–1889	2436	1980–1989	32
1890–1899	3230	1990–1999	23
		2000–2009	69
		2010–2019	41

In the follow-up project (2009–2015), we realised early on that it would be impossible to register every item of literary criticism published in Norway from 1870 onwards. A nodal principle of historiography centred around exemplary cases came to guide both the updating of the bibliography and the main publication *Norsk litteraturkritikks historie 1870–2010* (Furuseth et al. 2016). Compared with the first project, the second project was less focused on the reception of canonised Norwegian authors and more interested in media developments – particularly the shifting ecology of literary magazines – as the driving force behind the practice of literary criticism. The second project prioritised the registration of literary criticism from a few crucial magazines such as *Samtiden* (1890–) and *Vinduet* (1947–). Both magazines have existed long enough to deserve the status of institutional nodes (Cornis-Pope and Neubauer 2004, 17; Furuseth et al. 2016, 19). The Norwegian critical reception of Fyodor Dostoevsky and Isabel Allende was also

included.[6] Project members were encouraged to send bibliographies from their individual chapters and theses to the research assistants for registration.

In the 1980s and 1990s, it was only possible to search the bibliography from personal computers with a specific disc operating system program installed. The DOS program for the Beyer Base was written by Kåre A. Andersen at the Faculty of Humanities' EDP centre at the University of Oslo. The entries resembled good old index cards (Figure 1).

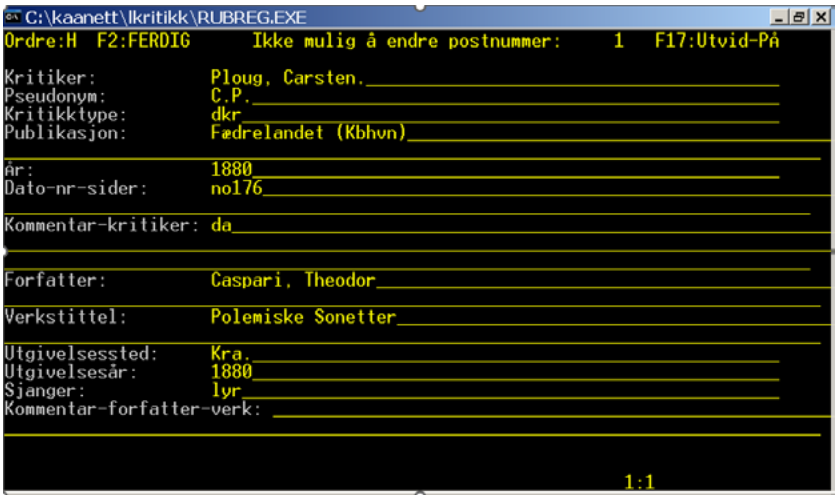

Figure 1: Screenshot from the first entry in *Norsk litteraturkritikk*. Old user interface. Source: Andersen 2007.

In 1999, Herman Ruge Jervell and informatics students at the University of Oslo developed a preliminary version for the internet (Andersen 2007). In 2007, Professor Torill Steinfeld initiated a comprehensive upgrading of the database and Andersen made sure that the bibliography was finally made searchable and available online (Norheim and Stensland 2013). Since then, the database has been hosted by the University of Oslo Library and upgraded several times, both technologically

6 Detailed information about sources is provided online at https://ub-baser.uio.no/norsk-litteraturkritikk/kilder.

and in terms of typology and structure (Norsk litteraturkritikk 2020). In 2019–2020, a working group initiated by Senior Academic Librarian Anne Sæbø and myself, and consisting of representatives from the University of Oslo Library and the Department of Linguistics and Scandinavian Studies, intensified the effort to make the database more accessible and user-friendly.[7] Together, we have functioned as an editorial committee for the database. Today, *Norsk litteraturkritikk* is a Postgres database improved by a self-made application developed by Dan Michael Heggø, built on the Laravel framework. The University of Oslo Library is responsible for the maintenance of the database.

In the years 2012–2018, the database interface reflected the categories established by the first research group. The structure was adopted more or less directly from a card index system by way of an Excel spreadsheet. Although visually not the most appealing, it served its purpose for the second research project as well, at a time when magazines and newspapers were still poorly represented in the National Library's digital archive. Ten years ago, it was difficult to browse the NB digital bookshelf for critics and book reviews, so *Norsk litteraturkritikk* was helpful in directing us to relevant microfilm resources and physical magazine archives. We appreciated the way the old structure placed influential critics in focus, as shown in this old screen shot (Figure 2) displaying the first hits for the impressionist critic Carl Nærup.

This interface illustrates both the strengths and weaknesses of the database at this early stage. *Norsk litteraturkritikk* gives detailed information about the critical texts and their writers: gender, genre, publication,

7 In addition to Anne Sæbø, who is responsible for Scandinavian Studies at the University of Oslo Library (UB), Dan Michael Heggø, Academic Librarian at UB's Science Library, has been the driving force from the library side. From the department's side, I have been responsible for providing funding and coordinating the work carried out by the research assistants Marte Hagen and Even Teistung. The UB's Henrik Keyser Pedersen (German and Philosophy), Signe Marie Brandsæter (Comparative Literature) and Annika Rockenberger (Digital Humanities) have also contributed useful input to the process. My colleagues Aasta Marie Bjorvand Bjørkøy and Thorstein Norheim (Nordic Literature) as well as Nina Marie Evensen and Kristin Kosberg (Centre for Ibsen Studies) have also been consulted.

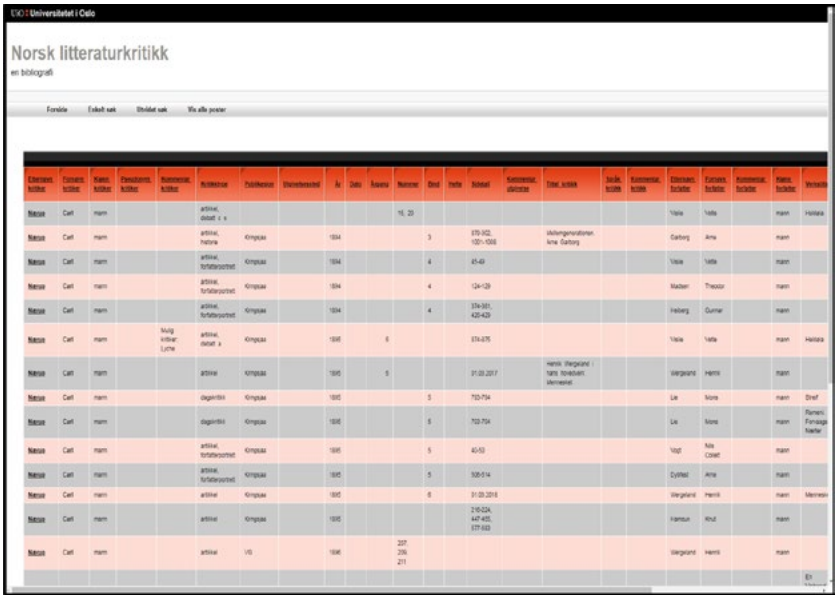

Figure 2: Screenshot from the 2018 version of *Norsk litteraturkritikk*.

page number etc. It even shows empty categories, such as (in this case) pseudonym, place of publication (for the magazine/newspaper), language and additional comments. Titles are reported for only two of the articles. This may reflect the fact that nineteenth century book reviews seldom had striking titles. More likely, though, the scattered schema is the result of unfinished proof-reading. In terms of metadata, the categories are significant. All the red squares at the top, especially the first 19 belonging to the critical text, illustrate what the Beyer group considered important to register in the early 1980s. It also shows how cumbersome it must have been to break new ground for digital humanities and quantitative literary reception studies 40 years ago. They were not prepared for all the technical problems caused by inaccurate proof-reading (Moi and Linneberg 1990). The spreadsheet form endured for 35 years.

Simplification and new interface

In the new interface (operative since 2019), a drop-down menu has been installed so that simplicity and complexity can be combined in a cleaner user-interface. In the present version, there are only three

default categories for the object of criticism: the book's title, year of publication and author. Correspondingly, there are three default categories for the critical text: name of critic, publication in which the text occurs and date of publication. The drop-down menu (*Vis kolonner*) makes it possible to show additional information about the type of publication (newspaper, magazine etc.), genre (book review, essay etc.) and keywords, such as modernism, mysticism, New Criticism, as illustrated in Figure 3 in the case of the post-war critic Erling Christie.

In the working group, there were discussions about structural priorities and which elements should be presented first in the user-interface. While some of us preferred to give the critic priority over the author (as in the old version of the database), representatives from the library argued that most students and researchers approaching them for assistance asked for material on specific authors or books. The present interface reflects this majority view. Since the externally

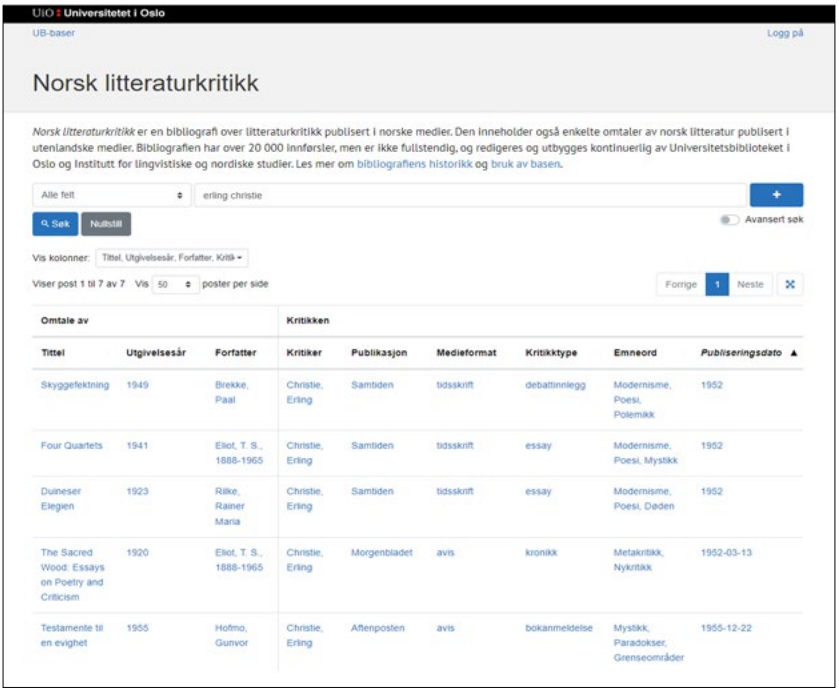

Figure 3: Screenshot of the modernised version of the online bibliography *Norsk litteraturkritikk*, showing entries related to the critic Erling Christie (status 2020, before proof-reading of publication dates).

funded research projects were formally finished, we agreed to take the students' perspective in the modernised version. Trained scholars still have plenty of opportunities when it comes to adding columns to the right and thereby access useful information about a critic's preferred genres or type of medium, for instance.

One of the shortcomings of the early version of the database was the tendency to identify issues of periodicals by numbers and not by dates. This practice probably derives from J.B. Halvorsen's *Norsk Forfatter-Lexikon* and Reidar Øksnevad's newspaper bibliographies from the 1950s, which have been used as sources for the database. As temporality is crucial for understanding literary critical reception, adding dates has been an important task in the later proof-reading process. Precise dating helps us to make sense of the chronology, intensity and development of a critical debate. A case in point is the dispute related to the seizure of Christian Krohg's naturalist novel *Albertine* in December 1886 because of the book's detailed descriptions of how a young woman ends up in prostitution after being raped by a policeman.

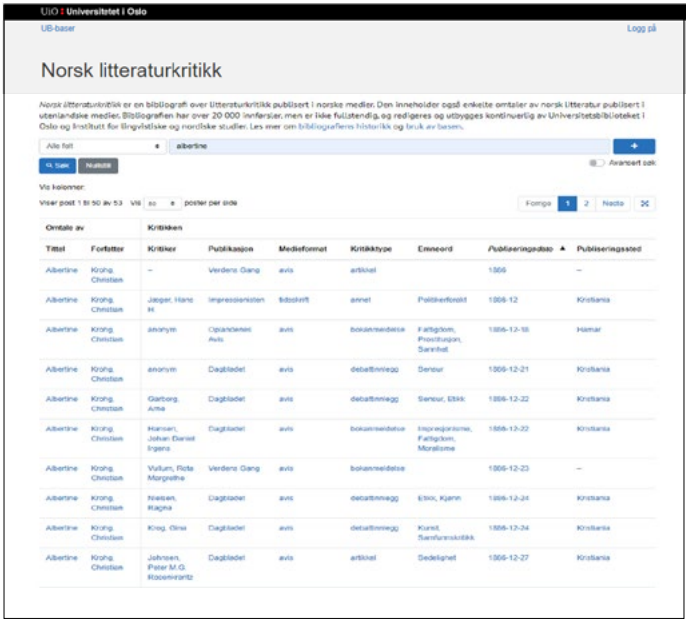

Figure 4: Screenshot showing additional information about the critical response to Christian Krohg's controversial novel *Albertine*, chronologically ordered.

This overview (Figure 4) reveals useful information about the intensity of and engagement in the debate about *Albertine*, especially the involvement of female critics. Regarding genre, it is interesting to note that debate articles replaced book reviews in a situation where the book itself was confiscated by the police. For a long time, the novel was not available to be reviewed.

For today's user of the database, it is particularly helpful that each post has a link to the National Library full-text resources (when available and checked). The combination of two digital resources from two different institutions, UB and NB, makes *Norsk litteraturkritikk* much richer than it was only five years ago. For example, if we are curious about the feminist Gina Krog's opinion regarding the *Albertine* case, we can click on her name and find quite rich metadata on that entry.

Figure 5 shows in detail the metadata for Gina Krog's article on *Albertine*. In terms of genre, the critique was written as a letter to the

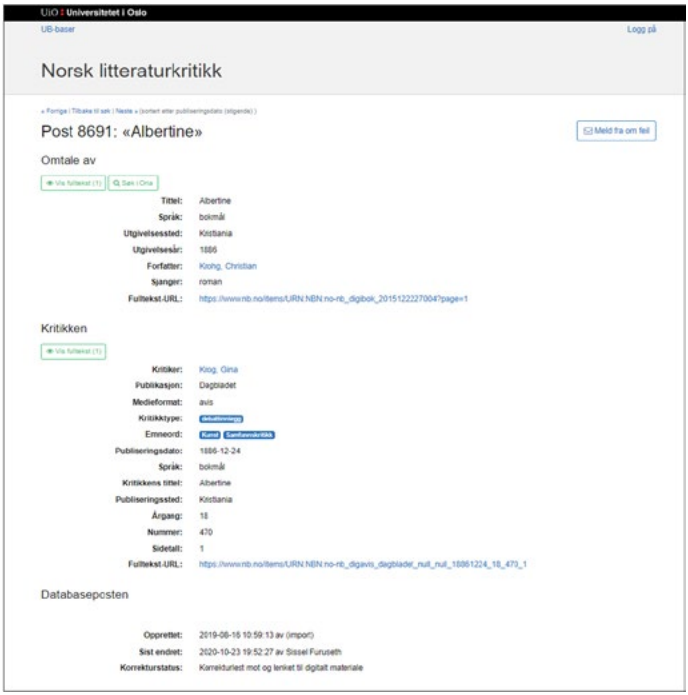

Figure 5: The reverse side of a database entry with links to full text resources.

Nr. 470. Kristiania. Fredag 24. December 1886.

Dagbladet.

Nr. 470. Kristiania. Fredag 24. December 1886.

18. Aarg. 18. Aarg.

„Dagbladet",

redigeret af *J. Halsø.*

Figure 6: Digital reproduction of the front page of *Dagbladet*, 24 December 1886. Gina Krog's piece "Albertine" in the middle of column 4. Source: NB.

editor of the daily newspaper *Dagbladet*, which was published in the Norwegian capital Kristiania (now Oslo). In terms of timing, the piece was published on 24 December 1886 (volume 18). When following the link behind "Fulltekst-URL", the reader is directed immediately to the relevant full-text resource at the National Library: the front-page of *Dagbladet*, 24 December 1886 (Figure 6).

The only metadata that the National Library can provide about the same source is the title of the newspaper (*Dagbladet*), date, language and medium type (newspaper). Researchers or students looking for information about specific articles within that issue may use the search function. Those particularly interested in the critical debate about *Albertine* can search directly in the digital archive at the National Library. However, Albertine is a common women's name, so such searches will retrieve a huge amount of redundant hits. The bibliography *Norsk litteraturkritikk* gives an overview of relevant hits more quickly, which is particularly useful when teaching students about this specific book and the relationship between literary criticism, censorship and confiscation. This is one reason why the database is a valuable supplement to the resources at the National Library.

Volumes or issues: problems of periodicity

While the example above illustrates how the metadata of *Norsk litteraturkritikk* and the digital newspaper archive at the National Library can reciprocally enlighten and support each other, my next example will touch upon how a lack of periodical codes may cause unnecessary frustration for researchers of literary reception published in periodicals. Due to the long tradition of archiving periodicals in bound volumes instead of original issues, many digitisation projects have continued preserving annual volumes instead of monthly or quarterly issues. For most people, who browse magazines for topics or names, this may seem a marginal problem, but for literary reception and periodical studies, this kind of archival practice has caused a lot of headaches over the past couple of decades, both nationally and internationally. In their pivotal article "The Rise of Periodical Studies", published in 2006, Sean Latham and Robert Scholes insist on the autonomy and distinctiveness

of periodicals as cultural objects. Accordingly, "digital archivists and editors should locate original copies of the periodicals they edit and not simply work from bound copies (which regularly exclude advertising) or reprints of those originals" (Latham and Scholes 2006, 525). Unfortunately, this principle is not always followed. As I have shown earlier in an essay about the National Library's digitisation of the literary magazine *Vinduet*, only annual volumes were used for the archive covering the period 1947–1970 (Furuseth 2021, 12). Shortcomings in earlier catalogue systems have been transformed from analogue to digital archives and metadata. This can probably be explained by the semi-automatic way the National Library creates digital metadata – often by simply exporting and reusing analogue objects' metadata (Eugui 2012, 55).

In the case of *Norsk litteraturkritikk*, this problem may be illustrated by the search for "Erling Christie" as shown earlier (Figure 3). The search finds that Christie had three articles published in the magazine *Samtiden* in the year 1952, but the posts do not give any information about exactly when. Even the special Christie-bibliography provided in the anthology *Modernisme og tradisjon*, leaves us with only volumes for journals and magazines (Christie 1983, 263–264). There is little help to be obtained from the publisher, Aschehoug, as they have disposed of original issues of *Samtiden* from these years.[8] So how can we track down the specific issues?

We know from other sources that the magazine published ten issues a year in the early 1950s, but the proof-reading has been complicated by the fact that the National Library has indexed the *Samtiden* volume as a book and not in terms of its original issues. The digital version does the same: *Samtiden* is not treated as a periodical with the temporal rhythm of a virtually monthly journal, but as a yearbook in which the original periodicity is concealed. The table of contents at the beginning of 'the book' does perhaps hint at a chronology of events/

8 I have been informed by Aschehoug (by email from Atussa Taslimi on 5 September 2023) that the publishing house has only a limited number of old issues in their possession. Issues from the 1950s are no longer available.

texts when informing the reader that Christie's essay on poetry and knowledge in Rilke's *Duineser Elegien* is located on page 162, the essay on mysticism and poetry in T.S. Eliot's *Four Quartets* on page 301, and the polemic article on Paal Brekke's poetic protest on page 572. In other words, the articles have likely been published in three different issues in this order. Consequently, they must have been published in the reverse order of the preliminary listing presented in *Norsk litteraturkritikk* (see Figure 3). Unless dates are specified, the database automatically lists entries in alphabetic order (Brekke, Eliot, Rilke).

In the libraries, both the physical and digital versions of *Samtiden* are mute about periodicity. However, by searching for advertisements in the newspaper archive, we have managed to establish that Christie's Rilke essay must have been published in March, the essay on Eliot in May, and the essay on Brekke in November 1952. The dating is the result of some detective work, assisted by our accumulated knowledge about the journal, the critic and 1950s advertising conventions.[9] We were lucky this time, as advertisements for *Samtiden* in which Erling Christie's name is mentioned do actually exist. The finds from the search leave us with a new order of the essays (Figure 7).

The metadata are still not very precise, but at least they provide a timeline that can be matched with other events in 1952. What becomes clear from this updated and extended overview, is that Erling Christie laid the groundwork for his defence of his Norwegian colleague Paal Brekke by first presenting the international pioneers Rilke and Eliot. Furthermore, published as late as November 1952, the Brekke essay came to function as a prelude to the major debate on

9 Searches for "'Erling Christie' AND Samtiden" narrowed down for 1952 result in 15 hits. Four of these are duplicates, which leaves us with 13 actual hits. The first three occur in the end of March (Dagbladet 27th, VG 28th, and Morgenbladet 31st) and show the same advert from Aschehoug informing the public that the March issue of *Samtiden* is out, including an article by Erling Christie on Rilke's *Duineser Elegien*. The next peak in the archive comes at the end of May with three newspapers referring to *Samtiden*'s issue number 5 (month is not specified this time), mentioning an article by Christie on *Four Quartets* among the contributions. Only one hit – a short note in *Fædrelandsvænnen* on 29 November – lets us know that *Samtiden*'s issue number 9 is out, including a contribution by Christie discussing the poetry of Paal Brekke.

Omtale av			Kritikken					
Tittel	Utgivelsesår	Forfatter	Kritiker	Publikasjon	Medieformat	Kritikktype	*Publiseringsdato*	Hefte
Duineser Elegien	1923	Rilke, Rainer Maria	Christie, Erling	Samtiden	tidsskrift	essay	1952-03	3
The Sacred Wood: Essays on Poetry and Criticism	1920	Eliot, T. S., 1888-1965	Christie, Erling	Morgenbladet	avis	kronikk	1952-03-13	
Four Quartets	1941	Eliot, T. S., 1888-1965	Christie, Erling	Samtiden	tidsskrift	essay	1952-05	5
Skyggefektning	1949	Brekke, Paal	Christie, Erling	Samtiden	tidsskrift	debattinnlegg, essay	1952-11	9
Teslamente til en evighet	1955	Hofmo, Gunvor	Christie, Erling	Aftenposten	avis	bokanmeldelse	1955-12-22	

Figure 7: Screenshot showing a selection of entries relating to the critic Erling Christie (status 2024, after proof-reading of publication dates).

poetic modernism – the so-called "*tungetaledebatten*" (The Gibberish Debate) – the following year. These are details, of course, but important details when the very temporal dynamics of literary debates are the object of research. Another interesting dynamic that can be read from this overview is the critic's alternation between genres and media formats. Christie wrote for both newspapers and journals.

Literary criticism across genres

In the early version of the Beyer Base, more than 35 categories of literary criticism – enquête, letter, bibliography, parody, diary, obituary etc. – were registered, in addition to more expected genres, such as the critical essay and the book review. The first project stressed the diversity of critical expressions and the fact that judgements of literary taste and value could take many forms. The second project inherited this policy, but in the modernised version of the database it proved difficult to keep up the high level of precision regarding "*kritikktype*" (types of criticism). A more realistic understanding of what Digital Humanities could or could not accomplish made us realise that a higher level of standardisation was necessary. Over the past five years, we have been less concerned about entering new posts and more focused on structure and proof-reading. The number of categories of criticism has been reduced to 17: book review, theatre review, chronicle, reader's letter,

critical portrait, essay, article, pamphlet, thesis, work of literary history, causerie, panel debate, parody, poetry, novel, play, and other. The categories are still rich and varied, as we believe it is important that a database covering two and a half centuries of literary criticism reflects the shifting dynamics of the genre landscape. We had to consider the historical material already registered in the database (polemics on verse and theatre reviews, for instance) and at the same time allow the database to include modern forms of literary criticism transmitted through electronic and social media.

In the modernised version, the number of "*kritikktyper*" has been reduced, but a high level of precision is preserved by combining them with a new set of media types. As I see it, one of the main achievements of the second phase of the project was the decision to make a clearer distinction between genre and media format (type of criticism versus type of media). It is not always easy to draw that line. For example, we discussed at length whether the blog was a genre or a medium. Eventually, we decided on medium, on the grounds that a blog allows a variety of content (reviews, essays, poems etc.) to be published on its platform. But the decision was not obvious, as it might be argued that the book blog is characterised by a specific affirmative rhetoric not common in other forms of literary evaluation.

There are some striking parallels between *Norsk litteraturkritikk* and the European *WomenWriters* database, both in terms of typologies and use. Wolfgang Kaltenbrunner, who has followed the European COST Action "Women Writers in History", summarises some of the deliberations that took place: "While bibliographical categories are widely agreed upon as viable abstractions, definitions of genre and reception are considered matters of theoretical debate in literary studies" (Kaltenbrunner 2015, 2018). The established division of labour between literary scholars on the one hand and librarians and technology developers on the other sometimes manifested itself as a frustrating clash of cultures, albeit frustrations to learn from. Anne Birgitte Rønning, herself a contributor to the COST Action "Women Writers in History", points out how the digitalisation of databases has challenged literary scholars to rethink genre categories. She experienced

first-hand how the core category "*resepsjon*" (reception) became difficult to use (Rønning 2014, 286). From being perceived as a separate category, it became a relation between data in the other categories.

The ability to articulate precise needs and describe how the empirical material will be used will be important skills for scholars in projects like this (Rønning 2014, 287). Literary scholars are about to learn how to tag texts so that they can be recognised as quantifiable data. Yet another similarity between *Norsk litteraturkritikk* and the *WomenWriters* database is the fragmentation of the datasets. Participants are still optimistic about the long-term prospects of gaining access to a more coherent empirical picture of literary reception across historical periods, but more important, they have realised that the database – with all its flaws and limitations – is a fascinating research object in itself. I would say that precisely because *Norsk litteraturkritikk* has, in many ways, been a miscalculated experiment there is much to learn from this case.

Metadata from the past for the future?

In 2009, the research project *Norsk litteraturkritikks historie 1870–2000* inherited an unfinished and yet complex digitised bibliography of Norwegian literary critical reception. Less than 30 years after its inception, the Beyer Base showed signs of being a cultural heritage project. This was confusing but also inspiring, in the sense that it became clear to us that bibliographical metadata is historically contingent and therefore should be a topic of interest for researchers of the history of literary critical reception. We cannot ignore the fact that the categories we use to organise our material are in flux. This was also acknowledged by participants in the first project, for instance in their reflections on the historical impact of the transition from anonymous reviewers to critics whose names are known (Moi and Linneberg 1990, 305).

Today we can let the computers do more of the work, but even state-of-the-art projects dealing with born-digital contemporary reviews, such as the Norwegian Review Corpus (NoReC) created by researchers at the University of Oslo's Department of Informatics

(IFI), put substantial time and effort into identifying relevant documents and extracting text and associated metadata. The Language Technology group at IFI has created a dataset of more than 35,000 full-text reviews of books, theatre, music, restaurants etc., that have been published online in the period 2003–2017 in order to train and evaluate models for document-level sentiment analysis. One of the challenges when creating such a corpus is that different "publishing conventions require targeting different types of cues in the document structure, like headers, bold-faced content or die-face images" (Velldal et al. 2018, 4187). Today, many reviews are announced by the genre tag "*Anmeldelse*" (Review) before a more content oriented title. Search-engine-optimised headings and clearly marked consumer information labels are helpful for identifying relevant texts. Yet not every newspaper uses the pips-on-a-die scoring convention ("*terningkast*") or other numerical ratings as a genre cue. Some standard book reviews may be easily identified by machines, but literary critical response comes in so many forms that there is probably no way around analogue metadata creation.

The transition from manual close reading of relevant sources to digital quantitative approaches might have been easier if the first large-scale research project on literary critical response in Norway had used a more standardised definition of "*litteraturkritikk*". But the influence of poststructuralist theory set the stage for more creative interpretations of critical genres and an almost unthinkable union of deconstruction and the (at that time) oft derided discipline of literary historiography (Furuseth 2013, 189). Especially Arild Linneberg, with his combined historicist and deconstructive approach to nineteenth century literary criticism, invigorated the field by demonstrating the performative qualities of literary criticism.

While deconstruction encouraged subtlety and complexity, the disadvantage of the diverse typology developed in the 1980s and 1990s was that it made quantitative research difficult. I would say that the Beyer Base was, from the very beginning, a typical example of what Johanna Drucker has described as "taxonomies in the wild", made in the "optimistic spirit of progressive pluralism" (Drucker 2021, 36).

The participants in the first project were quite honest about this, but it took some time before they realised that their disorganised input led to entropy and reduced information value in the database output (Moi and Linneberg 1990, 303). They discovered that proof-reading for computer-based bibliographies was more difficult than for analogue ones.

A poor consolation is that inconsistent metadata is not unique for *Norsk litteraturkritikk*. Anyone who has tried to compose a reading list in Leganto can testify to the chaos of standards in the library records. If we go to ProQuest's Periodicals Archive Online and search for an internationally known Norwegian author, for instance Knut Hamsun, we will find a good number of book reviews on the list. The sources are grouped into different categories, such as magazines, historical periodicals and scholarly journals. Among other obligatory metadata are date, language, city and title of publication – journal title, that is – whereas metadata information about the reviews as such are more random. In some items, the name of the reviewer may be lacking, even if it is clearly visible in the full text version linked to the index, as in Allan Monkhouse's review of *Mothwise* in the New Statesman on 30 April 1921, or in Rebecca West's review of *Wanderers* in the New Statesman on 29 April 1922. Duplicates tend to show different information, for instance either the title of the review or the title of the reviewed book. Older reviews in particular appear to be indexed according to different standards, while recent scholarly articles are tagged in a more systematic way. Again, we see how book reviews are subordinate to other text categories, but Periodicals Archive Online is very particular about dating. Issues of magazines and journals are still tagged with volume and number, but always accompanied by the exact date of publishing. This is also a prioritised task for *Norsk litteraturkritikk*: to situate events of literary criticism in time and space.

Closing remarks

The essential quality of a periodical is its serial form and its "particular relationship to time", Margaret Beetham has emphasised (1989, 97). She draws attention to "the paradox of a form at once so evanescent

and so enduring", referring, among other things, to the fact that "the materials of which periodicals are made have been designed for speed of production and cheapness rather than durability" (Beetham 1989, 96). The momentary value of periodicals has complicated the work of archivists, but the material characteristics of the periodical are nevertheless central to its meaning. The first major output of the research project *Norsk litterturkritikks historie 1870–2000* was the article collection *Kritiske portretter: Litterære tidsskrifter etter 1880* (Furuseth, Thon, and Vassenden 2010) which consists of 16 so-called 'portraits' or detailed descriptions of individual literary magazines published in Norway from the late nineteenth century onwards. In this book, periodical codes were important to us. In various degrees, the contributions reflect on layout, size, editorial arrangements, periodicity and other ways in which literary magazines create meaning. Whereas some small magazines die young and can be analysed almost like a trilogy of novels, others – such as *Samtiden* – endure in shifting forms across centuries and are hard to grasp, both in their ephemerality and in their permanence.

Are researchers like me too demanding? There are of course many projects within both periodical studies and literary critical reception studies that can be pursued without a refined set of periodical codes. A wide range of topical research questions can be interrogated with the help of standard search functions. Yet what I wanted to demonstrate by the case above is that print culture should be studied from many angles and that more possibilities are revealed if we consider different sets of metadata. The diversity of digital archives worldwide gives us an idea of what we can expect in the future. The Austrian archive *ANNO* stands out because it is very particular about periodicity, be it "*monatlich*", "*wöchentlich*", "*täglich*", or sometimes "*unregelmäßig*". The visual appearance of the newspapers and journals also seems to be of high priority in ANNO, and even more so in the *Modernists Journal Project* (MJP), where size is default metadata. In MJP, there is almost a three-dimensional quality to the colourful scannings. Furthermore, MJP acknowledges the importance of editors by having editorial arrangement as obligatory information. MJP's

material can be filtered by publication place, editor and contributor, and is in many ways a dream come true for the periodical studies scholar. The genre typology, however, is less refined. In the otherwise impressive MJP database, "articles" is the umbrella term used for all kinds of literary criticism. It seems that one single database cannot have it all.

From its inception, the idea behind the digital bibliography *Norsk litteraturkritikk*, earlier known as the Beyer Base, was to facilitate literary reception studies and more quantitative approaches to literary criticism in Norway by focusing on overall structures in terms of genres, geography, gender, ideology, networks etc. Forty years later, the database is still a fragmented work-in-progress. Although the database has limited value for quantitative research, it has been useful as an aid to teaching, since it provides practical overviews of reviews of specific contested books, such as Christian Krohg's *Albertine*. More important, however, is the database's function as an incentive to professionalise literary scholarship. As shown in this chapter, *Norsk litteraturkritikk* has served as a valuable learning arena for testing out different metadata approaches and defining what literary critical reception and periodical studies might be in the twenty-first century. While most librarians focus on standardisation and interoperability, there are still researchers out there with more idiosyncratic needs. *Norsk litteraturkritikk* is a compromise, trying to mediate the forces of standardisation and specialisation.

References

ANDERSEN, KÅRE A. 2007. "Norsk litteraturkritikk." Unpublished memorandum, 23 October.

ANNO. n.d. "ANNO Historische Zeitungen und Zeitschriften." Österreichische Nationalbibliothek. Accessed 20 August 2024. https://anno.onb.ac.at/.

BEETHAM, MARGARET. 1989. "Open and Closed: The Periodical as a Publishing Genre." *Victorian Periodicals Review* 22 (3): 96–100. https://www.jstor.org/stable/20082400.

BROOKER, PETER, AND ANDREW THACKER. 2009. "General Introduction." In *The Oxford Critical and Cultural History of Modernist Magazines*. Vol. I, edited by Peter Brooker and Andrew Thacker, 1–25. Oxford: Oxford University Press.

CHRISTIE, ERLING. 1983. *Modernisme og tradisjon. Essays og artikler 1949–1959*, edited by Marit Hammersmark, Andreas Lombnæs and Egil Christie Mathisen. Oslo: Aschehoug.

CORNIS-POPE, MARCEL, AND JOHN NEUBAUER. 2004. "General Introduction." In *History of the Literary Cultures of East-Central Europe: Junctures and Disjunctures in the 19th and 20th Centuries*, Vol. I, edited by Marcel Cornis-Pope and John Neubauer, 1–18. Amsterdam/Philadelphia: John Benjamins Publishing Company.

DRUCKER, JOHANNA. 2021. "Viewpoint: Hetero-ontologies and taxonomies in the wild." *Art Libraries Journal* 46 (2): 36–39. https://doi.org/10.1017/alj.2021.2.

EUGUI, LEIRE ARRULA. 2012. *Case studies on digitization and metadata creation and management*. MA thesis. OsloMet.

FURUSETH, SISSEL. 2013. "Litteraturkritikkens performative øyeblikk: Noter til en kritikkhistorisk disputas." *Edda* 100 (3): 187–200. https://doi.org/10.18261/ISSN1500-1989-2013-03-03.

———. 2021. "Noe tapt og noe funnet. Om *Vinduet* i Nasjonalbibliotekets digitale bokhylle." *Vinduet* 75 (1-2): 12–17.

FURUSETH, SISSEL, JAHN H. THON AND EIRIK VASSENDEN, EDS. 2010. *Kritiske portretter: Litterære tidsskrifter etter 1880*. Trondheim: Tapir Akademisk Forlag.

FURUSETH, SISSEL, JAHN H. THON, EIRIK VASSENDEN, TROND HAUGEN, KRISTOFFER JUL-LARSEN AND THORSTEIN NORHEIM. 2016. *Norsk litteraturkritikks historie 1870–2010*. Oslo: Universitetsforlaget.

HOHENDAHL, PETER UWE. 1982. *The Institution of Criticism*. London: Cornell University Press,

IVERSEN, IRENE, AND ARILD LINNEBERG. 1983. "Edb-register over litteraturkritikkens historie i Norge." *Norskrift* No. 38: 84–98.

KALTENBRUNNER, WOLFGANG. 2015. "Scholarly Labour and Digital Collaboration in Literary Studies." *Social Epistemology* 29 (2): 207–233. https://doi.org/10.1080/02691728.2014.907834.

LATHAM, SEAN, AND ROBERT SCHOLES. 2006. "The Rise of Periodical Studies." *PMLA*, 121 (2), March 2006: 517–531. https://doi.org/10.1632/003081206X129693.

MCGANN, JEROME. 1991. *The Textual Condition*. Princeton: Princeton University Press.

MODERNIST JOURNALS PROJECT. n.d. "Modernist Journals Project." Accessed 20 August 2024. https://modjourn.org/journal/.

MOI, MORTEN, AND ARILD LINNEBERG. 1990. "Appendiks. Bibliografisk database. Beskrivelse og bruk." In *Norsk litteraturkritikks historie 1770–1940. Bind I: 1770–1848*, edited by Edvard Beyer and Morten Moi, 297–313. Oslo: Universitetsforlaget.

NATIONAL INFORMATION STANDARDS ORGANIZATION. 2007. "A Framework of Guidance for Building Good Digital Collections", https://www.niso.org/publications/framework-guidance-building-good-digital-collections.

NORHEIM, THORSTEIN, AND MARIANNE STENSLAND. 2013. "Om databasen *Norsk litteraturkritikk* – en bibliografi." Unpublished conference paper, *Changing Functions of Criticism: Writing the Cultural History of Literary Critical Reception*, University of Oslo, 16 August.

NORSK LITTERATURKRITIKK. 2020. "Om Norsk litteraturkritikk." Last modified 17 March. https://ub-baser.uio.no/norsk-litteraturkritikk/historikk.

PROQUEST. n.d. "Periodicals Archive Online." Accessed 20 August 2024. https://about.proquest.com/en/products-services/periodicals_archive/.

RØNNING, ANNE BIRGITTE. 2014. "I skyggen av kanon. Empiri som utfordring i feministisk litteraturvitenskap." *Edda* 114 (4): 278–291. https://doi.org/10.18261/ISSN1500-1989-2014-04-02.

VELLDAL, ERIK, LILJA ØVRELID, EIVIND ALEXANDER BERGEM, CATHRINE STADSNES, SAMIA TOUILEB AND FREDRIK JØRGENSEN. 2018. "NoReC: The Norwegian Review Corpus." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, edited by Nicoletta Calzolari et al., 4186–4191. Miyazaki: ELRA

7. The Use of Bibliographic and Cultural Metadata – How to Investigate Users' Information Search Behaviour

Nils Pharo and Pia Borlund

Introduction

The ambition of this chapter is to show how the success or failure of users' efforts to retrieve information from the National Library of Norway's collections can be studied in various ways. We will also look at ways to measure users' satisfaction with the results returned by their searches. To what extent does the information retrieved meet their information needs? In this chapter, we will use examples stemming from the humanities to illustrate how the interactive information retrieval process may be evaluated, and recommend suitable test designs. We have devised three sample cases, which are all based on genuine research projects identified via the Research Council of Norway's project bank. We will use these cases to present methods and methodologies for the study of interactive information retrieval. Our approach is well suited for a variety of potential research projects, including research collaborations.

We hope that this perspective will inspire readers from different research communities to engage in joint research projects. An important benefit of such interdisciplinary research collaborations is increased information literacy on the part of scholars. Participants become more aware of their information needs, relevance criteria, search strategy formulation and query building, and learn more about the library's collections. In addition, interactive information

retrieval research helps enhance the library's information retrieval systems.

The National Library of Norway preserves and documents important parts of Norwegian cultural heritage via its collections. The library has a variety of collections that contain different types of materials. In addition to traditional library materials, such as books, journals and newspapers, the library has collections of music, film, radio and TV programmes, photographs, theatre material and a variety of other materials. A large proportion of the collections has been digitised and made available online. At the time of writing, books published before 2006, almost 800,000 photographs, and selected newspapers and journals are all publicly available by users with a Norwegian IP-address.[1] Furthermore, scholars may be granted online access to almost all Norwegian books (650,000 titles), newspapers (4.5 million copies) and journals (370,000 copies).[2] The library also has several language sets of Norwegian speech and texts in *Språkbanken*, the Language Bank.

In addition to managing the collections for present and future generations, the library constitutes a "key part of Norway's research infrastructure".[3] The National Library's research activities are twofold. On the one hand, researchers from the library participate actively in national and international research projects in disciplines like history, computer linguistics and literary studies. On the other hand, the library's collections are used in multiple external projects in media science, linguistics and other areas within the social sciences and humanities.

In brief, the research area "Interactive Information Retrieval" concerns the study and evaluation of users' interaction with information retrieval systems, and the users' interaction and satisfaction with the retrieved information. In other words, interactive information

1 https://www.nb.no/hjelp-og-informasjon/rettigheter/. Accessed 4 March 2024.

2 Numbers as of 19 August 2024.

3 https://www.nb.no/en/strategy/. Accessed 4 March 2024.

retrieval focuses on how people search online, whether they are searching the entire World Wide Web or database collections such as those held by the National Library, and whether they perceive retrieved information to be relevant to their needs. Relevance and information needs are core concepts in interactive information retrieval, and are often considered two sides of the same coin. Information needs typically mature and are refined during the search process, possibly changing what is considered relevant, i.e., the users' relevance criteria (Borlund 2003a).

This chapter will discuss methods for investigating how scholars interact with the library's collections, attempt to understand scholars' information needs, analyse search strategies and identify factors that lead to success and failure in their search of the collections. To capture and exemplify these interactive information retrieval instances, we have devised three illustrative cases, which we discuss one by one below. All three example cases are inspired by actual research projects funded by the Research Council of Norway and make use of the National Library's collections.

In the three proposed cases, we define the scholars undertaking the projects as our objects of study. This enables us to gain insights into this group of users, how they search for information, how they acknowledge and articulate their information needs and preferences, as well as their relevance criteria and the reasons for the success or failure of their searches. This knowledge is also useful for the scholars themselves because it helps them advance their information literacy skills, while we obtain empirical evidence that can be used to recommend search system improvements. Last, but not least, it provides us with an excellent opportunity for collaboration, thereby strengthening and extending our networks for future interdisciplinary research activities. In other words, we see this as an invitation to joint research.

When considering appropriate methods for systematic data collection, we employ a model adapted from Mason (2018, 26), see Table 1. The model served to ensure consistency in thinking in the early stages of the research process. It also helps to support decision making about data collection methods for the empirical study by linking research

questions and potentially suitable methods through the identification of data needs and data sources. For each of our three showcases, we present the research questions, data needs, potential data sources and possible data collection methods, illustrated by a table similar to Table 1. In each of the cases, we discuss the strengths and limitations of the methods.

Table 1: Adapted Mason model for the identification of possible data collection methods.

Research questions	Data required	Data sources	Possible data collection methods

Case 1: Information needs and relevance criteria

Case 1 has been developed with the purpose to discuss how to study information needs and relevance criteria. The case was inspired by the QUEERDOM project "*Ordinary lives and marginal intimacies in rural regions. Contrasting cultural histories of queer domesticities in Norway, ca 1842–1972*", a study headed by Professor Tone Hellesund that runs from 2021 to 2026.[4] The project investigates how women and men with same-sex desires organised their everyday lives in the period 1842–1972, when homosexuality was a criminal offence in Norway. According to the description in the Project Bank (*Prosjektbanken*), "QUEERDOM will be the first collaborative and international project to make use of *Skeivt Arkiv* (SkA) at the University Library of Bergen, and also the first project to use the National Library's vast collections and methods from digital humanities, to map and track traces of queer history."[5]

We envision that documentation of daily life may be found in works of fiction as well as in newspapers from this period. The application of filters to the National Library's database limits the works of

4 https://prosjektbanken.forskningsradet.no/project/FORISS/314253. Accessed 4 March 2024.

5 Ibid.

fiction published in the period 1842–1972 to 16,819 titles.[6] Figure 1 shows the top ranked books by relevance, a relevance ranking which could itself be an interesting case to investigate. The National Library's collection of newspapers covering the period 1842–1972 contains approximately 2.7 million issues.[7] Different filtering options are available for books and newspapers, but date filtering can be applied to both collections. In addition, both collections are indexed for full-text querying.

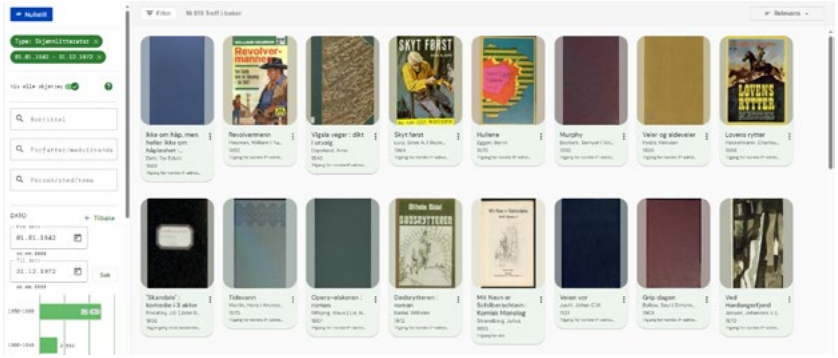

Figure 1: Query performed on the National Library's book collection. The search was for fiction published between 1842 and 1972, with the results ranked by relevance.

The documentation process is our point of departure, as it represents the information needs of the scholar(s), who are searching for evidence of how homosexuals organised their everyday lives in the specified period of 130 years (1842–1972). In this project, the scholars are going to make use of multiple sources to piece together an understanding of the situation in question. Fictional books and newspaper articles are two sources that may be of value to the scholars. In our context, the following research questions seem appropriate:

6 https://www.nb.no/search?mediatype=b%C3%B8ker&fromDate=18420101&toDate=19721231. Access date 1 October 2024. The search system of the National Library of Norway separates fictional books from non-fictional ones by means of a specific MARC field, namely MARC code 08, position 33.

7 https://www.nb.no/search?mediatype=aviser&fromDate=18420101&toDate=19721231. Access date 19 August 2024.

- What information needs do the scholars have and what criteria do they use when assessing the relevance of the retrieved information in response to those information needs?
- How may the relevance criteria change over time, and why?

The concept of an information need is fundamental to interactive information retrieval. A central aim of the research area concerns the retrieval of information relevant to the user's specific information needs (Borlund and Pharo 2019). The studying of information needs provides insight into what people search for and why (Savolainen 2017, 3). By uncovering the criteria against which the information's relevance will be assessed, we learn about the users' preferences and the core elements of their information needs (Borlund 2003b).

Possible methods for data collection – Case 1

Table 2: Adapted Mason model for the identification of possible data collection methods for Case 1.

Research questions	Data required	Data sources	Possible data collection methods
1. What information needs do the scholars have?	Info about information needs	• The scholars • Project results	• Interview • Questionnaire • **Research diary + follow-up interview** • Focus group interview • Document/content analysis
2. What are the scholars' criteria when assessing relevance of the retrieved information in response to those information needs?	Info about relevance criteria	• The scholars • Project results	• Interview • Questionnaire • **Research diary + follow-up interview** • Focus group interview • Document/content analysis

In Table 2 we have applied the adapted Mason model in the following way. It starts by listing the research question, followed by the data needed to answer the research question, then the sources from which

the required data may be obtained, and finally, possible data collection methods. The selected methods are marked in bold.

The succeeding rows are completed in accordance with a given project's research questions. Here we have added the research questions for Case 1.

To answer the two research questions, information about the scholars' information needs and relevance criteria are required. The sources of information are the scholars themselves and, possibly, project results reported by the scholars. The possible methods to obtain that information are: interviews, questionnaires, research diaries and follow-up interviews, focus group interviews as well as content analysis of project reports.

It is worth noting that the following discussion about the efficacy of the different data collection methods relates to the specific research questions in Case 1. It is not a discussion of these methods in general.

Interviews allow for dialogue with the scholars about their information needs (Brinkmann and Kvale 2018). The advantage of interviews is the face-to-face dialogue that provides opportunities for follow-up questions and in-depth answers. Interview-based studies are often considered time-consuming, but that is not really an issue here given the limited number of scholars expected to be working on the project. A significant weakness, however, is that interviews cannot take place at the point in time when information needs are formulated and relevance assessments performed.

Questionnaires can provide open-ended answers to open questions and can give insights into perceptions of information needs and subjective relevance criteria (Hank, Jordan and Wildemuth 2017, 274). As a data collection technique, questionnaires and open questions depend on the willingness and ability of the respondents to supply elaborative replies. Like interviews, questionnaires are completed after the information searching and relevance assessment have been performed.

Research diaries, however, can reveal the scholars' immediate perceptions of information needs and relevance criteria because they are written contemporaneously with need formulation, searching and

relevance assessment. This is considered an advantage in the given context. Furthermore, as research diaries provide information over a period of time, they may shed light on possible changes in information needs and relevance criteria as the research project develops. Often, follow-up interviews with the diarists clarify statements written in the diaries. To be of value, however, research diaries depend on the scholars' routines for notetaking, because important information may be lost if reporting is done after the actual searching and relevance assessment, thereby reducing their value as a data source (Sheble, Thomson, and Wildemuth 2017, 232).

Focus group discussions or interviews are techniques where a researcher assembles a group of individuals to discuss a specific topic, to obtain increased understanding, test assumptions and gain insights. Whereas interviews involve a one-to-one dialogue with the researcher in the role of 'investigator', the focus group discussion is facilitated or moderated by the researcher, and the discussion is a conversation between the group's members (Nyumba et al. 2018, 22). In our case, focus group interviews would serve nicely to uncover information needs and reach consensus about relevance criteria within the group, though the latter is slightly out of the scope of our research agenda. A weakness that this method shares with one-on-one interviews is that it cannot take place at the time of need formulation. The employment of focus group interviews also depends on the dynamics of the group as well as its size. Unfortunately, we lack sufficient information about these crucial factors.

An alternative to obtaining information directly from the scholars would be to utilise documentation they had produced in the form of project reports. A project report will typically include the projects' motivation and research questions, and careful scrutiny of the report may give valuable details about the scholars' information needs and relevance criteria. This careful scrutiny is known as qualitative content analysis (Mayring 2000) and has the advantage of being non-invasive, i.e., the informants (scholars) are not disturbed during their work and can provide reflections that may not appear when questioned during need formulation. This method's biggest disadvantage is the gap in

time between the formulation of information needs and their documentation in project reports, which may lead to important details being lost.

Selected methods for data collection – Case 1

Based on this outline of possible methods, we recommend research diaries with follow-up interviews as the preferred data collection method for obtaining information about the scholars' information needs and relevance criteria. Because research diaries have been used by social scientists since the early twentieth century, their use has become a well-established tradition (Szalai 1972; Wheeler and Reis 1991).

Research diaries resemble ordinary diaries, which are used to capture thoughts and feelings, external events and occurrences, from the perspective of an individual or individuals over time (Sheble, Thomson, and Wildemuth 2017, 228). When diaries are used for research purposes, they share a common set of characteristics. Making use of a largely fixed, mediated format, participants self-report life experiences, observations and reflections repeatedly over time, typically at or near to the occurrence of an event of interest (Bolger, Davis, and Rafaeli 2003). Besides these commonalities, research diaries vary greatly, ranging from highly structured event logs to unstructured personal narratives, and may include a variety of media formats and recording technologies (Sheble, Thomson and Wildemuth 2017, 228). Participants' self-reporting in successive entries over time is a defining feature of research diary studies and proves especially valuable when researchers are interested in the participants' evolving observations or reflections (Bolger, Davis and Rafaeli 2003), as we are in this case.

An important caveat, is that research diaries are usually solicited, i.e. they are kept by request. According to Sheble, Thomson and Wildemuth (2017, 229), this means that research diaries are inherently made for an audience.

As said, research diaries vary greatly in form and structure. The type of research diary we envision for this case is semi-structured diaries and based on an event-contingent study design (Sheble, Thomson,

and Wildemuth 2017). Entries will therefore be recorded in an open-ended fashion at the same time as the actions of formulating information needs and making relevance assessments are taking place. A classic diary is typically written on paper, but for research purposes it may be convenient to have entries recorded digitally, e.g., on the participant's smart phone. The length of diary keeping is another issue to consider. In our case, the phenomenon of interest relates to a specific project, which extends several years into the future and will therefore be unrealistic to follow to its conclusion. However, it might not be necessary to monitor the scholars through the entire project. Benefitting from the momentum of the project's launch, a period of two to three weeks would suffice to capture the types of information needs and identifying relevance criteria in the maturing phase of the project. According to Hernon, Powell, and Young (2004), the norm for keeping work-related research diaries tends to be a period of one to two weeks.

Research diaries are often used in combination with other data collection methods, for example questionnaires, observation, critical incident techniques and/or interviews, in order to elicit a rich description of the phenomenon being studied and to triangulate data. One of the most frequently used multimethod study designs involving research diaries is that of the diary-interview (Zimmerman and Wieder 1977), which pairs diary-keeping with follow-up interviews. This is also the combination we recommend for Case 1. As pointed out by Sheble, Thomson, and Wildemuth (2017, 233), "a diary-interview study is stronger than the sum of its parts" because the interview is grounded in information from the research diaries. They further explain how the follow-up interview can be used as an opportunity to clarify or focus specific diary entries (Sheble, Thomson, and Wildemuth 2017, 233).

Further reading – Case 1

For further reading on studies making use of research diaries, Sheble, Thomson, and Wildemuth (2017) provide two examples. The first study by Koufogiannakis (2012) explores the role of evidence in academic librarians' professional decision making. The second study, by

Pattuelli and Rabina (2010), examines student attitudes toward e-book readers. In the first study, librarians kept online diaries for a one-month period. In the second study, students kept electronic or paper journals for a one-week period. Follow-up interviews were conducted in both studies. In the report of a study that Byström and Järvelin (1995) conducted of work tasks for which information was required in a local government office, the actual diary is depicted.

For further reading on the concept of information needs, the reader is recommended the special issue on this topic edited by Borlund and Ruthven (2020). This covers a variety of empirical and theoretical studies centred on information needs, each with a different research focus and consequently a different research design.

Case 2: Search strategies

The second case concerns information search strategies. It was inspired by the project "*Mediation of Migration: Media impacts on Norwegian immigration policy, public administration and public opinion*", headed by Senior Researcher Kjersti Thorbjørnsrud from 2011 to 2015.[8] The project looked at the media's effect on the public's perception of migration. News stories on migration provided key data for the study. Among other things, the scholars studied the information strategies and production processes behind the news; news form and content; as well as the effect of media coverage on public opinion in general and on the perceptions of minority groups in particular. Norway experienced low levels of both immigration and emigration from the 1950s to the mid-1980s, although the amount of movement increased slightly over the period. From the end of the 1980s, migration, and in particular immigration, numbers rose sharply (SSB 2024). This inspired us to ask whether a potential change in migration-related news stories over time leads to changes in how these news stories are searched for. We are also

8 https://prosjektbanken.forskningsradet.no/project/FORISS/202480. Accessed 4 March 2024.

interested in learning why media scientists apply the given search strategies.

It is possible to search the National Library's newspaper collection using filters that make it possible to query the database in two comparable periods, i.e., 1950 to 1985 and 1985 to 2020. Figure 2 shows the results of the query "*emigrasjon*" [emigration] applied with a filter for the period 1950–1985, ranked chronologically. In all, the query generated 12,752 results. Interestingly, a query on "immigration" in the same period returns only one third of that number, 4,043 hits.[9] This outcome prompts questions about how the media have covered emigration and immigration, respectively, through different periods of time.

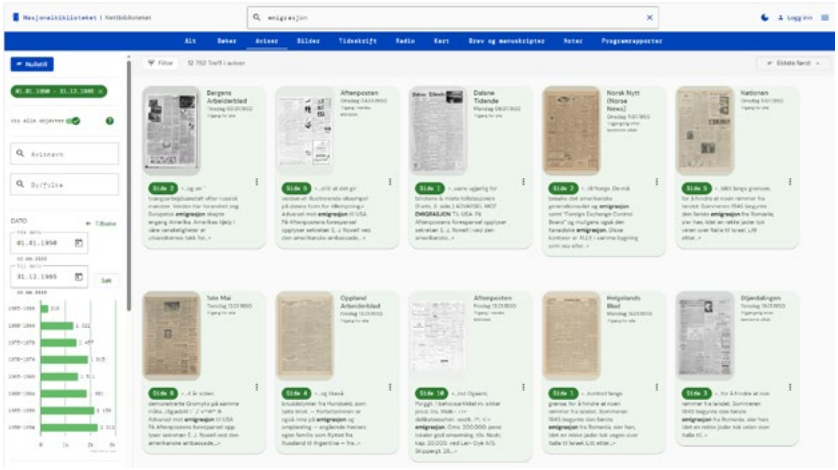

Figure 2: Query performed on the National Library's newspaper collection. The search was for articles on emigration from 1950 to 1985, with results presented chronologically.

Studies of information search strategies constitute another central area of research within interactive information retrieval (Bates 1989; Marchionini 1995; Savolainen 2016). Since search strategies indicate how the searchers intend to interact with information systems,

9 https://www.nb.no/search?q=emigrasjon&mediatype=aviser&fromDate=19500101&toDate=19851231 and https://www.nb.no/search?q=immigrasjon&mediatype=aviser&fromDate=19500101&toDate=19851231. Access date 1 October 2024.

knowledge about information search strategies is essential for system design. Information search strategies are here defined in line with Pharo's (2004) definition of search task strategy, i.e., the searcher's plan for executing a search task. An information search strategy may include a specification of sources and the techniques that should be used in order to explore them. We wish to examine this by observing how searchers interact with the system over time. Search system interactions may reveal changes in search strategies as a result of the user learning about the topic in question, gaining system knowledge and getting a clearer picture of the actual information need.

Possible methods for data collection – Case 2

We have used Mason's (2018) model to present the possible data collection methods for each of the research questions in Case 2 (see Table 3) in a similar way to Case 1.

Table 3: Adapted Mason model for the identification of possible data collection methods for Case 2.

Research questions	**Data required**	**Data sources**	**Possible data collection methods**
1. What search strategies do media scientists apply when searching for newspaper articles on migration in the two periods 1950–1985 and 1985–2020?	• Information about planned search strategies • Background knowledge on scholars' search experience, system knowledge etc. • Data about searcher system interaction	• The scholars • System log files • Screen recordings (video)	• research diary • **interactive information retrieval study**, including **log file analysis**
2. Why do the media scientists apply the given search strategies?	• Explanations about search strategies	• The scholars	• research diary • **Follow-up interviews** as part of the **interactive information retrieval study**

For the two research questions in Case 2 we need information about the scholars' search strategies and their search habits. In addition, we

need background information about their previous search experience, their search system knowledge and their knowledge of the collection(s) in question. Finally, we need their reasons for adopting the specific strategies. Since we are interested in evidence of the scholars' information search interaction, this quickly eliminates the traditional range of data collection methods applied in Case 1. In this case, the point of departure for data collection is a classic interactive information retrieval study. An interactive information retrieval study entails a general methodology that typically includes a pre-search questionnaire, transaction logging and video screen capture of users' search interactions with the information retrieval system, and follow-up interviews, in that order (Kelly 2009).

In theory, a research diary-based approach could have been used. This would have entailed asking scholars to record the queries they formulated, in what order they performed queries, if and how they reformulated queries and what results the queries generated. In addition, they could record their reflections on the reasons for choosing particular strategies. Compared with the use of research diaries in Case 1, it would, however, require more work from the scholars and probably provide a less accurate description of the strategies than could be gained from an interactive information retrieval study.

Selected methods for data collection – Case 2

The interactive information retrieval study, involving use of a pre-search questionnaire, transaction logging and video screen capture of the scholars' search interactions with the information retrieval system, and follow-up interviews, embraces both research questions. Opening with a pre-search questionnaire often functions as an 'ice-breaker', since taking part in an interactive information retrieval study is unusual. The pre-search questionnaire therefore focuses attention on background issues concerning the search experience, the frequency of searching and demographic information about the searchers (such as age and gender). This information makes it easier to understand the scholars' search behaviour. Subsequent information searching and system interactions are logged and all screen activities are captured by

screen capture software as video. The file logging and screen capture process collect data about the applied search strategies, actual search formulations and reformulations (including number of search terms, spellings and use of search operators), documents viewed and time spent on searching. The file transaction logs are recorded on the National Library's server. Since server-side logs only contain information about the interaction with the specific server it will not be possible to use it to capture searching that takes place in systems other than the National Library's database. We therefore recommend the complementary use of screen capture software, which records all system interactions taking place from the scholars' perspective. This means it will be able to document shifts between different systems, e.g., from the National Library's newspaper collection to the web site of a specific newspaper. In addition, screen capture displays the actual search results as presented to the scholars and provides a visual documentation of the scholars' interaction with system and information content. Screen capture software is thus a good alternative to observation, which would require the researcher to be present during the search session. An observer has the advantage of being able to record verbal and non-verbal utterances and reactions, but observation is very time-consuming and may influence the observed scholar. It also requires careful planning with respect to when search sessions are to be conducted (Kelly 2009). We therefore highly recommend the use of screen capture software, with simple start/stop recording procedures. Data collection closes with follow-up interviews with the scholars to elicit explanations about their search strategies. The interviews are carried out as semi-structured interviews in accordance with a developed interview guide. This allows for dialogue with the scholars about their choice of strategies and possible changes in strategies. Furthermore, the opportunity for follow-up questions and in-depth answers makes it possible to obtain a detailed picture of the scholars' information behaviour in a way that other methods, such as questionnaires, cannot. As mentioned above, interview studies are often considered time consuming, which may be considered a weakness in studies with

many informants. However, this does not apply to the present case, which is expected to have a limited number of participating scholars.

Further reading – Case 2

For further reading and practical guidance on evaluating information retrieval systems, the reader is recommended the compendium paper by Kelly (2009) on methods for evaluating information retrieval systems with users. Also the ARIST chapters by Harter and Hert (1997), Wang (1999), and Ruthven (2008) are recommended when considering approaches, issues and methods for the evaluation of information retrieval systems and users' search behaviour.

The study by Vuong and colleagues (2019) is a good example of a recent interactive information retrieval study based on users' genuine information needs. The study combined screen-recordings of ten participants over a period of 14 days. Recordings were made of all activities on the computer and supplemented with logs from the operating system. In addition, the participants filled out diaries in which they described their daily tasks. Other well-designed studies are reported by Koenemann and Belkin (1996), Spink, Greisdorf and Bateman (1998), Nordlie (1999) and Kelly (2006).

Case 3: Search system interaction

Case 3 was inspired by a project on the history of Norwegian literary criticism, "*Norsk litteraturkritikks historie 1870–2000. Verdiforvaltning og mediering*", which was headed by Associate Professor Sissel Furuseth from 2009 to 2015.[10] The project covers the period 1870–2000, it is information-intensive and relies heavily on the National Library's collections for the retrieval of book reviews and articles in newspapers and literary journals. This makes it a clear candidate for

10 https://prosjektbanken.forskningsradet.no/project/FORISS/191134. Accessed 4 March 2024. This is the same project that provides important input for Furuseth's discussion of the database *Norsk litteraturkritikk* in Chapter 6.

the study of scholar's information searching and the reasons for the success or failure of searches, including systems performance.

The National Library's collections of newspapers and journals are examples of sources from which it is natural to retrieve book reviews and other material relevant for the project. The interface of the newspaper collection was presented above, in Figure 2. The National Library's collection of journals facilitates filtering by topic/subject area ("*Tema/emne*" in Norwegian). The subject filter is based partly on controlled vocabularies used in the indexing of the collection and partly on uncontrolled terms from different metadata fields. In Figure 3, we show the results from a query performed on the journal collection on the subject "*litteraturkritikk*" [literary criticism] for the period 1 January 1870 to 31 December 2000. In all, this returns 80 results. A full-text query on the term without the subject filter retrieves 1,863 journals.[11]

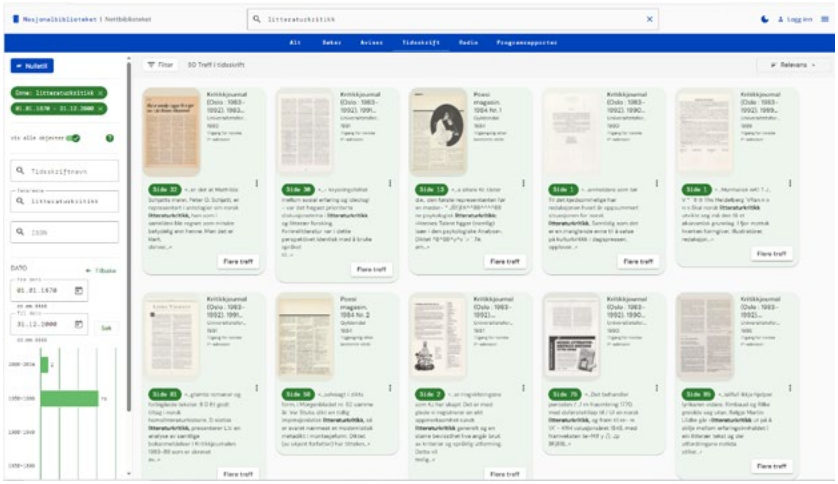

Figure 3: Query performed on the National Library's journal collection. The subject of the search was literary criticism from 1870 to 2000, with the results ranked by relevance.

11 https://www.nb.no/search?q=litteraturkritikk&mediatype=tidsskrift&subject=litteraturkritikk&fromDate=18700101&toDate=20001231 and https://www.nb.no/search?q=litteraturkritikk&mediatype=tidsskrift. Access date 1 October 2024.

Basically, information retrieval is about the retrieval of relevant information. When taking the user perspective, which is what interactive information retrieval does, the primary focus is to understand how people search for information and how satisfied they are with the information retrieved (Borlund 2013). This third case therefore goes to the core of interactive information retrieval by examining the scholars' interactions with the information retrieval system in their search for literary criticism. In addition, it aims to understand the reasons for success and failure in the process of meeting their information needs. The example above illustrates how the use of different retrieval facilities, e.g., the option to filter on subject area, significantly influences the output from the system and therefore the potential for finding relevant documents.

Possible methods for data collection – Case 3

Once again, we rely on the adapted model by Mason (2018) for the identification of possible data collection methods (see Table 4). The data needed to answer the first research question, "What are the reasons for success or failure when searching?", concern searcher system interaction, in other words, captured data from log files and screen videos of the scholars' information search activities.

This calls for a classic interactive information retrieval study, as in Case 2, which reveals how the actual searches take place and the reasons for success or failure. The reasons for success or failure are complemented by the second research question, "How satisfied are the scholars with the search results?", which requires information about the scholars' assessment of the retrieved information's relevance. This, in turn, informs us about their satisfaction or dissatisfaction with the information retrieved, and therefore whether their information needs were met. This information can be obtained in several ways as part of a classic interactive information retrieval study. The scholars can be asked to make notes about relevance judgements while searching, or electronically indicate the relevance assessment integrated in the search interface of the information system. Either way, the assessments are typically clarified and explained in a follow-up interview with the

participants after the search sessions have ended, as outlined in Case 2. We have not included alternative data collection methods for this case.

Table 4: Adapted Mason model for the identification of possible data collection methods for Case 3.

Research questions	Data required	Data sources	Possible data collection methods
1. What are the reasons for success or failure when searching?	• Background knowledge on scholars' search experience, system knowledge etc. • Data about searcher system interaction	• The scholars • System log files • Screen recordings (video)	• **interactive information retrieval study**, including **log file analysis** and searching initiated by **simulated work task situations**
2. How satisfied are the scholars with the search results?	• Relevance assessments	• The scholars • System log files • Screen recordings (video) • Search notes	• **Follow-up interviews** based on **stimulated recall** (as part of **interactive information retrieval study**)

Selected methods for data collection – Case 3

In an interactive information retrieval study, the overall methodology is tailored to the purpose of the research, the research agenda and research questions. In this case, it means the inclusion of simulated work task situations to initiate searching and allow for a comparison of search performance and satisfaction among the participating scholars (Borlund 2003b). Furthermore, the follow-up interviews are conducted with the help of "stimulated recall" (Kelly 2009, 85–86). Stimulated recall is used to collect data about the test participants' thoughts, actions and the reasons for their actions while using the information retrieval system.

To avoid interruption during the search process, data are collected after it has come to an end. Test participants are shown screen recording(s) of their searching activities to stimulate recall. They are then asked to articulate their search-related thinking and decision making as part of the follow-up interview.

To sum up, the interactive information retrieval methodology follows that of Case 2, with a pre-search questionnaire providing background knowledge of the scholars' search experience and system knowledge as well as demographic data. This is followed by information on the scholars' actual search activities in connection with simulated work task situations. All transactions are logged and screen activities captured on video. Finally, the scholars' relevance assessments are explored in follow-up interviews, which are guided by means of stimulated recall. Therefore, the following focuses on the employment of simulated work task situations and stimulated recall.

The concept of a simulated work task situation was introduced by Borlund and Ingwersen (1997) in response to a call put forward by Robertson and Hancock-Beaulieu (1992) for alternative interactive information retrieval evaluation approaches to that of the Cranfield model (Cleverdon, Mills, and Keen 1966; Cleverdon and Keen 1966). The emphasis of the Cranfield model is on controlled laboratory tests. The aim is to keep all test variables controlled and to obtain results which allow conclusions to be drawn about retrieval systems in general. This is clearly evidenced by the Cranfield model's emphasis on the principle of test collections. These include a collection of documents, a collection of queries and a collection of relevance assessments.

The concept, the underlying theoretical assumptions and the application guidelines for the use of a simulated work task situation as the test instrument are discussed in numerous journal articles (e.g., Borlund and Ingwersen 1997; Borlund 2000; 2013; 2016). The objective is to enable interactive information retrieval to be evaluated in circumstances as close as possible to actual information searching and retrieval processes, yet in a relatively controlled evaluation environment. This is achieved through the involvement of potential users as test participants, the application of dynamic and individual information needs (real and simulated information needs), the assessment of multidimensional and potentially dynamic relevance, and by the information searching being based on simulated work task situations.

A simulated work task situation is a short textual description which presents a realistic information-requiring situation that

motivates the test participant to search the information retrieval system (Borlund 2003b). More specifically, it provides the test participants with details of the following:

- The source of the information need.
- The environment and context of the search situation.
- The problem which has to be solved.

It also serves to make the test participants understand the purpose of the search (Borlund and Ingwersen 1997, 227–28). The following textbox shows an example tailored to Case 3:

Example of simulated work task situation tailored to the research project of Case 3

> Your research group has received an NFR grant to analyse the history of Norwegian literary criticism from 1870 to 2000. As part of the project, you are interested in finding evidence of literature critics and literature reviews from this 130-year period in Norwegian daily newspapers. You would therefore like to search the National Library's newspaper collection for this purpose.

In an interactive information retrieval study, the use of a simulated work task situation has two main benefits. First, it opens the way for user interpretations of the simulated work task situation, leading to cognitively individual interpretations of information need, as would be the case in real life. Second, it offers a benchmark against which situational relevance is judged by the test participant.

The simulated work task situation is a stable concept, in that it presents the purpose and goal of the information retrieval system

interaction. The fact that it is the same for all the test participants provides experimental control, making the search interactions comparable across the group of test participants for that task (Borlund 2003b). The use of simulated work task situations therefore affords the interactive information retrieval study both realism and experimental rigor.

It is important that the simulated work task situations describe realistic cases. The issue of realism is essential if the test participants' prompted search behaviour and relevance assessments are to be as genuine as intended. The simulated work task situations create simulated information needs that are meant to replicate genuine information needs. Therefore, realism is emphasised in the requirements for employing simulated work task situations (Borlund 2003b; 2016). In brief, the requirements are as follows:

1. "To tailor the simulated work task situation to the test participants:
   - a situation the test participants can relate to and identify themselves with;
   - a situation the test participants find topically interesting and/or of relevance to them; and
   - a situation that provides enough context for the test participants to be able to apply the situation" (Borlund 2016, 396).
2. "To include test participants' personal information needs as baseline.
3. To rotate the order of simulated work task situation and personal information needs (counterbalancing).
4. To pilot test prior to actual testing (often more than once).
5. To display the used simulated work task situations when reporting the study" (Borlund 2016, 406–7)

Requirement no. 1 addresses the tailoring of the simulated work task situations and the issue of realism. The challenge is to formulate descriptions that are inclusive and have a high abstraction level that people can relate to and identify themselves as being in. The situation being described must also cover a topic of personal interest to the test participants. It should be authentic, relevant and realistic to the

test participants. In our case, the task presented in the textbox above is tailored for scholars searching for literature reviews. Thus, it leads to realistic interpretations and interactions with the simulated information needs. The requirement to tailor the simulated work task situations entails a certain degree of homogeneity in the group of test participants. They need to have something in common, which can form the foundation for the design, tailoring and use of the simulated work task situations.

Requirement no. 2 concerns how to employ a combination of simulated work task situations and the test participants' personal information needs. This applies both when pilot testing and when carrying out the actual evaluation. Borlund (2016, 397) describes how the test participants should prepare a personal information need in advance of the evaluation and bring it with them to search as part of the evaluation. The personal information needs function as a baseline against which the simulated information needs may be evaluated. As such, they serve as a control of the reliability of the search interactions relating to the simulated work task situations. Furthermore, search activities relating to the test participants' personal information needs provide information about the system's effect on genuine information needs (Borlund 2016, 397). The incorporation of personal information needs is also helpful with respect to the pilot test (requirement no. 4), because personal information needs may inspire the creation of realistic simulated work task situations (Borlund 2016, 397).

Requirement no. 3 advises that the order of simulated work task situations and the test participants' personal information needs should be rotated between the test participants. This is to minimise the risk that simulated work task situations and personal information needs are searched for in the same order and thereby neutralise any possible bias-of-order effect on the results based on test participants' search interaction and relevance assessment behaviour (Kelly 2009, 50–54).

Requirement no. 4 is about the good practice of pilot testing prior to actual evaluation. When pilot testing the test setting, the test requisites (e.g., protocol, tutorials, pre-search questionnaire, the simulated work task situations and post-search interview guide), the test

procedure and the collected data are evaluated, and may be adjusted if required. In particular, attention should be paid to the test participants' understanding and use of the simulated work task situations.

Requirement no. 5 reminds researchers to include the actual simulated work task situations utilised when reporting the study. This improves transparency and enables readers to understand and assess the reported results, which depend heavily on the quality and realism of the simulated work task situations.

One finale note on testing by means of simulated work task situations concerns the number of simulated work task situations to use. The rule of thumb is to use no more than five simulated work task situations plus one genuine information need. However, this is something to address and pay attention to in the pilot testing. The reason for keeping to a maximum of five is to avoid exhausting the test participants.

After searching prompted by the simulated work task situations has concluded, the study closes with follow-up interviews of the participating scholars. In this phase, particular emphasis is placed on their satisfaction (or dissatisfaction) with the retrieved information, including their experience with the search system and reasons for the success or failure of their searches. The information is gathered via semi-structured interviews that allow room for a flexible but comprehensive dialogue. The interviews are based on so-called stimulated recall (Kelly 2009, 85–86). Stimulated recall is used to collect experience data from participants, with data being collected during and after the search activity. The researcher records the computer screen as the test participant completes the search tasks. After the task has been completed, the recording, or selected parts of the recording, is shown to the test participant, who is asked to articulate their search-related thinking and decision making.

Further reading – Case 3

For further reading on stimulated recall, Kelly (2009) recommends Rieh (2002) for an example of its application. Rieh studied the problem of judging information quality and cognitive authority by observing

online search behaviour. Fifteen scholars from a variety of disciplines participated. The data collected combined verbal protocols during the searches, search logs and post-search interviews based on stimulated recall.

For additional reflections on how to apply simulated work task situations, we recommend the paper by Wildemuth and Freund (2012) on how to assign search tasks designed to elicit exploratory search behaviours. For a recent empirical study that employed simulated work task situations, we direct the reader to the paper by Sarkar et al. (2020) that studies the users' perceived barriers and desired assistance in different stages of information search episodes. This was done in a controlled lab study, in which each participant performed three simulated work task situations. The participants' implicit behaviours were collected through search logs, while explicit feedback was elicited through pre-task and post-task questionnaires.

Final remarks

The ambition of this chapter has been to introduce interactive information retrieval as an area of research to the community of scholars who, as part of their research, make use of the National Library's collections. We also wished to highlight how interactive information retrieval has a user perspective and concerns the study of information needs, information search strategies, query formulations, relevance assessments, search and retrieval performance, and users' overall search behaviour. Not least, however, our aim was to demonstrate, via the three cases presented, how research projects from different research communities and with different research traditions can be connected and combined for the mutual benefit of the research groups involved. It should be noted that the three cases described here are illustrative and do not limit the approaches that can be taken to address interactive information retrieval research problems. We hope the chapter will serve as inspiration for possible future collaborations between interactive information retrieval researchers and scholars who make use of the National Library's collections as part of their research projects.

References

BATES, MARCIA J. 1989. "The Design of Browsing and Berrypicking Techniques for the Online Search Interface." *Online Information Review* 13 (5): 407–24. https://doi.org/10.1108/eb024320.

BOLGER, NIALL, ANGELINA DAVIS AND ESHKOL RAFAELI. 2003. "Diary Methods: Capturing Life as It Is Lived - ProQuest." *Annual Review of Psychology* 54: 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030.

BORLUND, PIA. 2000. *Evaluation of Interactive Information Retrieval Systems*. Abo Akademis Forlag.

———. 2003a. "The Concept of Relevance in IR." *Journal of the American Society for Information Science and Technology* 54 (10): 913–25. https://doi.org/10.1002/asi.10286.

———. 2003b. "The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems." *Information Research* 8 (3). http://informationr.net/ir/8-3/paper152.html.

———. 2013. "Interactive Information Retrieval: An Introduction." *Journal of Information Science Theory and Practice* 1 (3): 12–32. https://doi.org/10.1633/JISTaP.2013.1.3.2.

———. 2016. "A Study of the Use of Simulated Work Task Situations in Interactive Information Retrieval Evaluations: A Meta-Evaluation." *Journal of Documentation* 72 (3): 394–413. https://doi.org/10.1108/JD-06-2015-0068.

BORLUND, PIA, AND PETER INGWERSEN. 1997. "The Development of a Method for the Evaluation of Interactive Information Retrieval Systems." *Journal of Documentation* 53 (3): 225–50. https://doi.org/10.1108/EUM0000000007198.

BORLUND, PIA, AND NILS PHARO. 2019. "A Need for Information on Information Needs." *Information Research* 24 (4). http://informationr.net/ir/24-4/colis/colis1908.html.

BORLUND, PIA, AND IAN RUTHVEN. 2020. "Information Need: Introduction to the Special Issue." *Information Processing & Management* 57 (2): 102–3. https://doi.org/10.1016/j.ipm.2019.102103.

BRINKMANN, SVEND, AND STEINAR KVALE. 2018. *Doing Interviews*. London: SAGE Publications Ltd. https://doi.org/10.4135/9781529716665.

BYSTRÖM, KATRIINA, AND KALERVO JÄRVELIN. 1995. "Task Complexity Affects Information Seeking and Use." *Information Processing & Management* 31 (2): 191–213. https://doi.org/10.1016/0306-4573(94)00041-Z.

CLEVERDON, CYRIL, AND MICHAEL KEEN. 1966. "Aslib Cranfield Research Project – Factors Determining the Performance of Indexing Systems: Volume 2." Technical Report. Cranfield: Aslib.

CLEVERDON, CYRIL, JACK MILLS AND MICHAEL KEEN. 1966. "Aslib Cranfield Research Project – Factors Determining the Performance of Indexing Systems: Volume 1 Design." Cranfield: Aslib.

HANK, CAROLYN, MARY WILKINS JORDAN AND BARBARA M. WILDEMUTH. 2017. "Survey Research." In *Applications of Social Research Methods to Questions in Information and Library Science*, 2nd ed, 272–83. Libraries Unlimited.

HARTER, STEPHEN P., AND CAROL A. HERT. 1997. "Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods." *Annual Review of Information Science and Technology (ARIST)* 32: 3–94.

HERNON, PETER, RONALD R. POWELL AND ARTHUR P. YOUNG. 2004. "Academic Library Directors: What Do They Do?" *College & Research Libraries* 65 (6). https://doi.org/10.5860/crl.65.6.538.

KELLY, DIANE. 2006. "Measuring Online Information Seeking Context, Part 1: Background and Method." *Journal of the American Society for Information Science and Technology* 57 (13): 1729–39. https://doi.org/10.1002/asi.20483.

———. 2009. "Methods for Evaluating Interactive Information Retrieval Systems with Users." *Information Retrieval* 3 (1–2): 1–224.

KOENEMANN, JURGEN, AND NICHOLAS J. BELKIN. 1996. "Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness." In *Proceedings of the 1996 Conference on Human Factors in Computing Systems, CHI 96*, 205–12. ACM.

KOUFOGIANNAKIS, DENISE. 2012. "Academic Librarians' Conception and Use of Evidence Sources in Practice." *Evidence Based Library and Information Practice* 7 (4): 5–24. https://doi.org/10.18438/B8JC8J.

MARCHIONINI, GARY. 1995. *Information Seeking in Electronic Environments*. Cambridge: Cambridge University Press.

MASON, JENNIFER. 2018. *Qualitative Researching*. Third edition. Los Angeles, California: SAGE.

MAYRING, PHILIPP. 2000. "Qualitative Content Analysis." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 1 (2). https://doi.org/10.17169/fqs-1.2.1089.

NORDLIE, RAGNAR. 1999. "'User Revealment' – A Comparison of Initial Queries and Ensuing Question Development in Online Searching and in Human

Reference Interactions." In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11–18. SIGIR '99. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/312624.312618.

NYUMBA, TOBIAS O., KERRIE WILSON, CHRISTINA J. DERRICK AND NIBEDITA MUKHERJEE. 2018. "The Use of Focus Group Discussion Methodology: Insights from Two Decades of Application in Conservation." *Methods in Ecology and Evolution* 9 (1): 20–32. https://doi.org/10.1111/2041-210X.12860.

PATTUELLI, M. CRISTINA, AND DEBBIE RABINA. 2010. "Forms, Effects, Function: LIS Students' Attitudes towards Portable E-book Readers." *Aslib Proceedings* 62 (3): 228–44. https://doi.org/10.1108/00012531011046880.

PHARO, NILS. 2004. "A New Model of Information Behaviour Based on the Search Situation Transition Schema." *Information Research* 10 (1). http://informationr.net/ir/10-1/paper203.html.

RIEH, SOO YOUNG. 2002. "Judgment of Information Quality and Cognitive Authority in the Web." *Journal of the American Society for Information Science and Technology* 53 (2): 145–61. https://doi.org/10.1002/asi.10017.

ROBERTSON, S.E., AND M.M. HANCOCK-BEAULIEU. 1992. "On the Evaluation of IR Systems." *Information Processing & Management* 28 (4): 457–66. https://doi.org/10.1016/0306-4573(92)90004-J.

RUTHVEN, IAN. 2008. "Interactive Information Retrieval." *Annual Review of Information Science and Technology* 42 (1): 43–91. https://doi.org/10.1002/aris.2008.1440420109.

SARKAR, SHAWON, MATTHEW MITSUI, JIQUN LIU AND CHIRAG SHAH. 2020. "Implicit Information Need as Explicit Problems, Help, and Behavioral Signals." *Information Processing & Management* 57 (2): 102069. https://doi.org/10.1016/j.ipm.2019.102069.

SAVOLAINEN, REIJO. 2016. "Information Seeking and Searching Strategies as Plans and Patterns of Action: A Conceptual Analysis." *Journal of Documentation* 72 (6): 1154–80. https://doi.org/10.1108/JD-03-2016-0033.

———. 2017. "Information Need as Trigger and Driver of Information Seeking: A Conceptual Analysis." *Aslib Journal of Information Management* 69 (1): 2–21. https://doi.org/10.1108/AJIM-08-2016-0139.

SHEBLE, LAURA, LESLIE THOMSON AND BARBARA M. WILDEMUTH. 2017. "Research Diaries." In *Applications of Social Research Methods to Questions in Information and Library Science*, 2nd ed., 228–38. Libraries Unlimited.

SPINK, AMANDA, HOWARD GREISDORF AND JUDY BATEMAN. 1998. "From Highly Relevant to Not Relevant: Examining Different Regions of Relevance." *Information Processing & Management* 34 (5): 599–621. https://doi.org/10.1016/S0306-4573(98)00025-9.

SSB. 2024. "Fakta om innvandring." SSB. Accessed 11 March 2024. https://www.ssb.no/innvandring-og-innvandrere/faktaside/innvandring.

SZALAI, ALEXANDER. 1972. *The Use of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries*. The Hague; Paris: Mouton.

VUONG, TUNG, MIAMARIA SAASTAMOINEN, GIULIO JACUCCI AND TUUKKA RUOTSALO. 2019. "Understanding User Behavior in Naturalistic Information Search Tasks." *Journal of the Association for Information Science and Technology* 70 (11): 1248–61. https://doi.org/10.1002/asi.24201.

WANG, PEILING. 1999. "Methodologies and Methods for User Behavioral Research." *Annual Review of Information Science and Technology (ARIST)* 34: 53–99.

WHEELER, LADD, AND HARRY T. REIS. 1991. "Self-Recording of Everyday Life Events: Origins, Types, and Uses." *Journal of Personality* 59 (3): 339–54. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x.

WILDEMUTH, BARBARA M., AND LUANNE FREUND. 2012. "Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors." In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, 1–10. HCIR '12. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2391224.2391228.

ZIMMERMAN, DON H., AND D. LAWRENCE WIEDER. 1977. "The Diary: Diary-Interview Method." *Urban Life* 5 (4): 479–98. https://doi.org/10.1177/089124167700500406.

8. Subject Matters: Metadata Standards and Subject Access for Library and Museum Catalogues

Ahmad M. Kamal and Koraljka Golub

Introduction

For centuries, catalogues have provided ways for users to find specific items in the collections they are searching (Gnoli 2020). But how is this subject access implemented in practice? There are a myriad possible ways of representing the subject matter of an object, be it a book, a painting, a vase or something other cultural artifact, as well as a myriad means of incorporating such representations into a catalogue. To ensure the systematic, coherent and rational description of objects in and across collections, metadata standards for subject indexing have therefore been established. These metadata standards have evolved over time to accommodate changing needs, growing and diversifying collections, and emerging digital affordances. Yet, for numerous reasons, subject metadata are neither fully nor rigorously implemented in many of today's catalogues – whether in the backend indices or the front-end end-user interfaces – to the detriment of users. Online public library catalogues, discovery services, journal article databases, university repositories and museum catalogues all demonstrate shortcomings in how subject metadata are integrated, consolidated and leveraged (Golub 2016; Golub 2018; Golub et al. 2020; Golub, Ziolkowski and Zlodi 2022). Properly incorporated, subject metadata could greatly improve the discovery of objects within collections and (because collections are increasingly connected online) the reconciliation of

descriptive information across different collections for better cross-database searching. On the other hand, digital technologies could make standardised subject access finally realise the potential always envisioned for it.

To make this argument, we explore subject access and associated metadata standards in library and museum cataloguing. These two heritage sectors are examined due to the contrasting nature of the collections which their respective catalogues must describe. Libraries deal predominately with textual and non-unique objects, where the content of these items takes precedence; museums deal predominately with non-textual and unique items, where context, form and function have great importance. Beyond how these differences manifest in subject description and metadata standards, comparing libraries and museums also highlights different levels of investment in subject access, with the former traditionally investing more than the latter.

In the first section of this chapter, we briefly introduce the nature of contemporary searching, showing the utility of subject access. The section that follows presents the existing categories of metadata standards and their respective role in systemising subject indexing. Notable here, is how these standards have evolved with respect to subject access, which is highlighted at the end of the second section when the Semantic Web is described. The third section then explores the problem of today's library and museum catalogues, tracing the development of online search systems for libraries and museums, and the neglect of standardised subject metadata, despite their persistent importance for users. This leads to the final section of the chapter, which briefly discusses the potential role of contemporary subject standards if they were more deeply incorporated into our catalogues.

Importance of subject access

Appreciating subject metadata requires us to understand the types of searches users undertake when consulting catalogues. Generally speaking, users conduct either a *known-item search* or an *unknown-item search*. A known-item search means that the user is already familiar with the sought-after item. In these cases, users will use their

knowledge of the item, such as its title or its creator's name, to find it in a catalogue. Subject indexing (though still potentially useful) does not play a central role in known-item searches.

An unknown-item search, by contrast, is more about discovery. Here, the user has no clear item in the mind; they may not know what to call the item, or even be confident that such an item exists. Subject access becomes vital, as users search based on the concepts pertaining to the item's content/form/function, such as "medieval Muslim travelogues" or "depictions of rustic life in the English countryside during the late Industrial Revolution". Compared with known-item searching, where proper terms such as author and title suffice, unknown-item searching is more challenging due to the difficulty users face in formulating search queries without sufficient knowledge of the topic concerned, the collection being searched, how the collection is represented in the catalogue, or the search system through which queries are made. Users must rely on natural terms they are familiar with, in the hope that they reflect the topics as indexed in the catalogue.

Many problems can be linked to the inherent ambiguity in natural language. The term "lawyer", "lawyers", "attorney", "barrister", "solicitor" and "advocate", for instance, may each give different results. "Advocate" is both a noun and a verb, with different meanings and associations. Polysemy, homonymy and synonymy are all impediments to successful subject searching. Polysemy ("advocate") can lead to the retrieval of irrelevant results, perhaps too many to review manually, as may homonyms ("bat" referring to a flying mammal or a club swung at a ball). Meanwhile, synonymy requires a searcher to know and apply all possible equivalent terms ("lawyer", "attorney" etc.) to obtain a comprehensive retrieval. These problems are exacerbated as we make our collections available to increasingly diverse audiences, who bring different levels of knowledge, familiarity and motivation when searching a collection. It must therefore be asked whether a collection is best presented through a keyword search when the results can be partial or misleading. In some cases, for example, presenting users with a systematic mapping of the collection's topics may better support unknown-item searches.

These points emphasise the importance of standardising subject access in collections. The next section introduces the standards which are currently used for this purpose by libraries and museums.

Standards for subject access

There are five types of metadata standards that shape subject access in catalogues: conceptual models, schemas, cataloguing rules, knowledge organisation systems, and encodings. While subject access is just one aspect of these standards, it is what we will focus on here. We then explore the latest revolution in metadata, the Semantic Web, which touches on all standards and opens new possibilities for subject access.

Conceptual models

Conceptual models can be challenging to grasp. In essence, they form the theoretical basis for how objects are catalogued for meaningful representation, management, retrieval and integration. Today's main conceptual models for libraries and museums have emerged over the past two decades as traditional cataloguing practices were revised for the new digital landscape. The two models discussed here are the Library Reference Model and CIDOC-CRM.

The Library Reference Model, the latest conceptual model within librarianship, was established in 2017 (Riva, Le Boeuf and Žumer 2017). It defines the role of a catalogue as enabling five user activities – to find, identify, select, explore and obtain – and, except for this last activity, each offer a justification for subject indexing:

- Users can *find* resources by a given subject label.
- Users can *identify* the resources they found and distinguish between related ones by means of subject descriptions, such as items indexed under "drugs—addiction—economics" and from those under "drugs—recreational use—economics".
- Users can *select* the most suitable resource for their needs because the subject metadata describes aspects of the resource (style, period, topic, approach etc.).

- Users can *explore* a collection through the subject relationships, such as browsing related topics, broadening/narrowing the search subject, or following up on associated terms.

Through the Library Reference Model, these user activities underpin bibliographic description, which in turn creates a rationale for subject indexing with respect to end-users.

The other crucial aspect of contemporary conceptual models is a new data model built on *entities* and *relationships*. The entity-relationship architecture of current conceptual models shifts cataloguing from tabular data structures to linked data structures (more below), treating connectivity as the foundation of our information environment. In the Library Reference Model, the entity-relationship structure changes subject indexing by explicitly separating an intellectual work from its associated concept (*res*[1]), and the concept from the names for said concept (*nomens*), with designated relationships connecting each of these entities (see also Figure 1):

WORK [entity]⟶'has as subject' [relationship]⟶RES [entity]
RES [entity]⟶'has appellation' [relationship]⟶NOMEN [entity]
NOMEN [entity] ⟶'is appellation of' [relationship]⟶ RES [entity]
RES [entity]⟶'is subject of' [relationship]⟶WORK [entity]

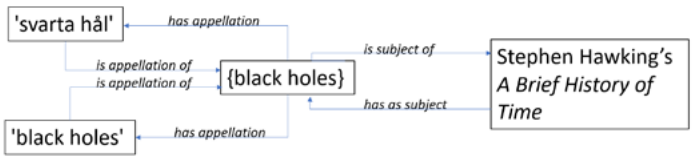

Figure 1: Example of the Library Reference Model entity-relationship subject mapping.

The corresponding conceptual model for museums is the CIDOC-Conceptual Reference Model (CRM). While the Library

1 *Res* (Latin for "thing") is the top-level entity (superclass) for all other bibliographic-related entities in the Library Reference Model, and as such includes both material and conceptual things.

Reference Model focuses on modelling the outcomes of processes (i.e., informational resources and their associated entities/relationships), CIDOC-CRM is an ontology, offering a formal and explicit representation of knowledge relating to cultural heritage. To do this, CIDOC-CRM focuses on events and processes rather than simply outcomes.

As for subject indexing, given the ranges of objects museums must catalogue, CIDOC-CRM offers an assortment of relationships to link objects with concepts (subject matter): "is about", "depicts", "represents", "refers to", "shows visual item" or "has symbolic content". How two different types of entities are connected to their respective concepts is presented below and further illustrated in Figure 2:

> Propositional object [entity]→'is about'[relationship]→CRM entity [entity]
> Physical Man-Made Thing [entity]→'depicts'[relationship]→CRM entity [entity]

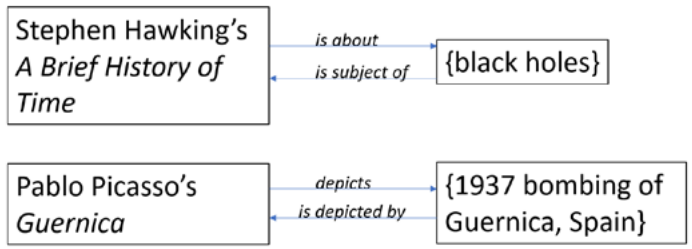

Figure 2: Example of CIDOC-CRM entity-relationship subject mapping for a book (top) and a painting (bottom).

Given the entity-relationship structure, both conceptual models reconfigure catalogues as networks in which any subject matter and its associated labels are made explicit entities.

Schemas

A metadata schema outlines the categories of information to be recorded about resources, with each category called an *element*. The set of elements available in each schema is tailored to the types of resources, collections or setting the schema is developed for. For

example, Metadata Object Description Schema (MODS) is a schema for bibliographic material; Dublin Core is for general digital resources such as webpages; VRA Core is for visual resources; and Categories for the Description of Works of Art (CDWA) is for works of art and other material culture.

Most schemas feature one or more elements for subject description. Metadata schemas for museum resources, for instance, provide a good structure for recording subject-related information. Of the nearly 540 categories and subcategories of information in the CDWA, a group of metadata categories called Subject Matter contains 17 subcategories (Baca and Harpring 2016). According to CDWA, the Hagia Sophia could be described as a "cathedral", "mosque" and "museum" in the Object/Work Type elements and the Specific Subject elements, but should be indexed as "architecture" in the General Subject element. By comparison, MODS has a simple Subject element, though this can also be further distinguished by sub-elements (Topic, Geographic, Temporal, Name or Genre).

Cataloguing rules

While metadata schemas list the relevant elements for describing objects, cataloguing rules guide how the elements are to be used, filled and combined to create a catalogue record. For libraries, the current rules are embodied in Resource Description and Access (RDA). Unlike preceding library rules, which make little mention of subject cataloguing, RDA pays some attention to subject indexing when cataloguing and applying the Subject element – though this attention still lacks a systematic or comprehensive discussion of implementation. This may all seem odd, given that subject access has been discussed in modern library cataloguing since its foundational principles were articulated 150 years ago.[2] Yet it was not until 2010, with the introduction of RDA, that subject representation enters cataloguing rules.

2 In Charles Cutter's *Rules for a Dictionary Catalog*, originally published in 1876.

Even then, without explicit guidance, it has had minimal impact on the practice of subject indexing.

For museums, Cataloguing Cultural Objects is the most influential set of guidelines for cataloguing cultural works and their visual surrogates (Baca et al. 2006). Rules for museum cataloguing accord some attention to subject indexing. SPECTRUM, a UK standard, prescribes the elements Content and Subject as primary steps in creating subject information for an object. However, an important factor differentiating museums from libraries is that most museum objects lack narrative content (as opposed to documents). The subject of a museum object, such as a decorative button or an ivory comb, is therefore designated by the object's form and the function. As such, elements relevant for subject indexing in museum catalogues are not as definitive as those in document-based schemas, where a Subject field would suffice.

Knowledge organisation systems

Knowledge organisation systems (KOS) are especially prominent standards for subject access. Subject headings, classification schemes, thesauruses, glossaries, taxonomies and keywords are all types of KOS. Regardless of its type, a knowledge organisation system represents concepts within a domain (general, sciences, history, medicine, literature, locations and places etc.), often through pre-established sets of terms, symbols and syntax. Many KOS include the relationships between these concepts, such as hierarchies (e.g. denoting "Purism" as a narrower term than "Cubism") or distinctions (e.g. "Marriage – psychological aspects" from "Marriage – religious aspects" or the feline "Jaguar" from "Jaguar automobile").

Some types of KOS represent concepts through a system of notation building (Gnoli 2020). The Dewey Decimal Classification is perhaps the most well-known example of this for libraries. To understand this system, consider the Dewey classification of 173 for Bertrand Russel's 1929 work *Marriage and morals*. The construction of this notation is as follows: 1xx for "Philosophy and psychology", 17x for "Ethics", and finally 173 for "Ethics of family relationships". Besides

providing a systematic representation of subject matter, classifications like Dewey are used for arranging bibliographic material on shelves, resulting in related works being sorted together to enable subject-based shelf browsing by users. This same principle can also be applied online to support online browsing of a collection, such as WebDewey Search.[3] Iconclass is a classification system for art and iconography that is widely accepted for the description of subjects represented in images. Picasso's *Guernica*, for instance, may receive the Iconclass classification of 45M4, with 4 for "Society, Civilization, Culture", 45 for "warfare; military affairs", 45M for "non-combatants – war", and, ultimately, 45M4 for "civilians during bombardment". The Social History and Industrial Classification is used in social and local history museums to link objects through their context, by dividing all aspects of human activity into four top facets (community life, domestic and family life, personal life, working life). This allows an antique carved-wood dumbbell for household exercise to be classified (and thereby contextualised) as 2.74 "Physical fitness in the home" (2 for "Domestic and family life", 2.7 for "Family well-being").[4]

Besides using notation for subject indexing, many KOS use linguistic terms (words or phrases). Whether based on notation or terms, KOS which involve a controlled labelling system are known as *controlled vocabularies*. A controlled vocabulary is made up of a restricted list of terms authorised for indexing (e.g. "airplanes"). This contrasts with natural language, where any number of terms and forms could be used for describing a similar concept ("airplane", "airplanes", "jets", "planes", "aeroplane", "aircraft"). The Library of Congress Subject Headings (LCSH) has a broad scope for subject indexing of bibliographical material and is the world's largest controlled vocabulary, with over 400,000 subject headings. For instance, LCSH articulates two separate concepts of enlightenment: "Enlightenment" for the eighteenth century European cultural movement and "Enlightenment

3 https://deweysearchsv.pansoft.de/webdeweysearch/.

4 https://www.shcg.org.uk/About-SHIC.

(Buddhism)" for Buddhistic spiritual awakening. In contrast to LCSH's wide breadth of topics, the Art and Architecture Thesaurus (AAT) is a domain-specific controlled vocabulary for describing visual material and architecture. It is especially important for indexing cultural heritage objects.[5] AAT contains authorised terms for objects, materials, techniques, styles, periods and other such concepts. As mentioned above, subject indexing for non-textual cultural objects requires an elaborate KOS to represent the subject matter in terms of content, form and function. AAT's Objects Facet, for instance, can distinguish visual works by form (e.g. "medals", "tapestries", "diptychs") and function (e.g. Hagia Sofia as a "cathedral", a "museum" and a "mosque"). AAT is just one of the Getty Vocabularies. A more recent KOS produced by the Getty Research Institute is the Iconography Authority (IA) which focuses on subjects and topics (e.g. "Bouddha couché", "Adoration of the Magi") (Harpring 2018, 2).

The examples illustrate the utility of KOS for users to identify, select and explore works within a collection. Yet there is no one-size-fits-all solution, since the vocabulary ultimately needs to do justice to the collection and users searching it. For instance, to address the minimal indexing of LGBTQI literary fiction in the Swedish union catalogue, the Queer Literature Indexing Thesaurus was developed by adapting the Digital Transgender Archive's Homosaurus Vocabulary for works of fiction and adding terms to reflect the Scandinavian context (e.g. indigenous groups, Swedish judicial terms) (Golub, Bergenmar, and Humlesjö 2022).[6] In another example from the museum sector, Australia's Powerhouse Museum has its own thesaurus of around 9,000 terms which overlap with AAT but is bespoke to their collection (social history and decorative art) and language variant (Australian English).

The primary role of a knowledge organisation system is to ensure consistent and unambiguous descriptions of concepts associated

5 AAT was in fact created in response to LCSH's inadequacy for detailed description in this domain.

6 https://homosaurus.org/v3.

with objects. They can also make explicit the relationships between terms (e.g. "Purism" as a narrower term than "Cubism") and, by extension, objects within a collection (e.g. books on Purism and Cubism). As such, these systems can be useful for refining the scope of users' searches, allowing them to reformulate their queries as well as gain an overview of what a collection houses. This can be especially useful in museum collections, since, as stated previously, subject indexing is relatively underused and under-developed in museum cataloguing compared with libraries. This may be attributed to the complexities faced by museum cataloguers: more diverse and distinct object types, and the complexity of identifying subject matter for non-textual cultural objects, which may need to consider form, function and context.

As might be ascertained from the above, manual subject indexing is a costly and demanding process, limiting its scalability. Unsurprisingly, automated methods for subject indexing have been explored since the earliest electronic documents. Experiments in the 1960s showed that simple terms naturally appearing in documents were highly effective for indexing and led to successful retrieval outcomes by providing ordinary users with easy and flexible subject access (Smucker 2011). Today, the proliferation and volatility of digital documents, such as webpages and electronic articles, necessitate automated subject indexing approaches. Here, instead of drawing from a list of authorised terms (i.e., a controlled vocabulary) full-text indexing applies natural language processing techniques to extract representative terms from documents themselves. This is technically a type of knowledge organisation system, based on simple, uncontrolled keywords for subject indexing.

In its simplest form, this process begins by transforming the text of a document into a list of terms or words (tokenisation). Non-meaningful terms are then removed (stopword removal). The leftover terms are subsequently transformed into their base root (stemming). Lastly, terms are assessed for their importance based on frequency and uniqueness (term weighting). A simplified example illustrates this procedure:

Full-text:	"He decided to bake something so you better ask him what he's baking."
Tokenization:	he, decided, to, bake, something, so, you, better, ask, him, what, he's, baking
Stopword removal:	decided, bake, better, ask, baking
Stemming:	decid, bak, better, ask, bak
Term weights[7]:	decid=0.25, bak=0.5, better=0.25, ask=0.25

Such techniques are the basis of search engines such as Google and textual databases. But despite the efficiency and power of full-text indexing, its reliance on textual information is a limiting factor. While computer vision and image classification systems continue to develop, automatic indexing for non-textual material remains unreliable. Relying entirely on natural terms appearing in documents raises other issues for subject access, however. Different authors may express the same concept in different ways, or express different concepts using the same terms. Two webpages may discuss "stars", for example, but one may refer to celestial objects while the other refers to Hollywood celebrities. Here, too, automated techniques for disambiguation are improving, for instance, by leveraging the co-occurrence of other indicative terms such as "gravity" or "famous". Nevertheless, full-text indexing remains confined to the terms found within a document. It cannot emulate what controlled vocabularies are built to offer: consistent and unambiguous subject description *across* objects in a collection. And while full-text indexing permits users to search using familiar terms, without being exact, and still find useful results, it provides little support for users engaged in unknown-item searches, who are uncertain what appropriate terms

7 Term weighting is usually more than each term's frequency *within* a document; one standard practice is to divide this by each term's frequency *across all documents in the collection* (known as TF*IDF).

they might apply or how to refine them. The subject relations explicated in controlled vocabularies would be effective in such scenarios.

Encodings

Encoding standards play an important role in the era of digital information exchanges by determining how transferable and interoperable data will be. MARC represents an early and still widespread standard for computer cataloguing record encoding (MARC stands for *machine-readable cataloguing*). In MARC, elements like Title, Creator and Subject, for example, are tagged with the three-digit codes (245, 100, 650); the MARC codes 600, 648, 650, and 651 are all subject elements which differentiate person (e.g. "Gustav II Adolf"), time ("Early modern, 1500–1700"), topic ("Kalmar War") and place ("Scandinavia") values, respectively.

More universal formats for data structuring and exchange were developed with the advent of the internet, one example being Extensible Markup Language (XML). XML is thus integrated into many cataloguing systems, leading to MARCXML and the more recent MODS (eliminating the three-digit system). Examples of MODS's encoded Subject element (with terms from LCSH and AAT) is presented below:

```
<subject>
   <topic authority="lcsh"> Spain—History—Civil War, 1936–
   1939</topic>
</subject>
   …
<subject>
   <topic authority="aat">Surrealist</topic>
</subject>
```

More recent advances in web technology, namely the Semantic Web, have ushered in newer encoding standards for library and museum cataloguing. However, these are part of the wider revolution in metadata standards that must be explored in its own right.

Semantic Web

The Semantic Web has led to a re-evaluation of heritage cataloguing, impacting all the standards mentioned above. The goal of the Semantic Web is the formal, explicit and interconnected representation of information across the web, allowing for richer aggregation and inferencing of knowledge online. A foundation of the Semantic Web is linked data, a means of structuring and connecting published data on the web. For heritage institutions already in possession of the structured cultural data inherent in their catalogues, the Semantic Web holds the promise of greater transparency, impact and engagement. Adapting metadata standards to linked data technologies would realise more sophisticated capacities for search, exploration and analysis. This has led to linked data becoming part of numerous national and international heritage initiatives, such as the Swedish Open Cultural Heritage (K-samsök) and Europeana.

The data model unpinning linked data is similar to the entity-relationship basis of conceptual models discussed earlier. More specifically, the Semantic Web's data model is called the Resource Description Framework (RDF). It is composed of three parts – subject, predicate, object – arranged into a statement, as displayed in Table 1.

Table 1: Each row represents an RDF statement for the subject matter of a work.

Subject	**Predicate**	**Object**
Stephen Hawking's *A Brief History of Time*	has subject (MODS)	Black holes (Astronomy)
Pablo Picasso's Guernica	depicts (CIDOC-CRM)	1937 bombing of Guernica, Spain

The next important step is how *things* (entities or relationships) are represented. In linked data, all things are represented by unique and unambiguous strings of characters called a uniform resource identifier (URI). These URIs should follow HTTP formatting (i.e., an internet

address) so that these names can be looked up, offer structurally standardised information and provide links to discover other things.

Table 2: Examples of RDF statements with uniform resource identifiers (URIs).

Subject	Predicate	Object
Stephen Hawking's *A Brief History of Time*	has subject (MODS)	Black holes (Astronomy)
https://id.loc.gov/resources/works/3448552.htm	http://id.loc.gov/ontologies/bibframe/subject	http://id.loc.gov/authorities/subjects/sh85014574
Pablo Picasso's *Guernica*	depicts (CIDOC-CRM)	1937 Bombing of Guernica, Spain
https://www.museoreinasofia.es/coleccion/obra/guernica	https://www.cidoc-crm.org/html/cidoc_crm_v7.1.3.html#P62	https://dbpedia.org/resource/Bombing_of_Guernica

Table 2 adds the requisite URIs missing from Table 1. But understanding how the meaning (semantics) of such data is made explicit requires an understanding of the foundational vocabulary and ontology of the Semantic Web: RDF-Schema (RDFS) and Web Ontology Language (OWL). RDFS and OWL form the building blocks of all other linked data. RDFS defines basic classes (types of entities), properties (types of relationships) and datatypes. OWL complements and extends this basic vocabulary with other essential concepts and rules, such as the cardinality, symmetry or equality of properties linking entities, enabling greater inferential power.

Most of the metadata standards introduced earlier are available as linked data. Returning to Table 2, we see the URIs for the CIDOC-CRM relationship "depicts"[8], the LCSH concept "black holes (Astronomy)"[9], the LC authority file for "Stephen Hawking's *A Brief History of Time*",[10] as well as a URI for the bombing of Guernica[11] from

8 https://www.cidoc-crm.org/html/cidoc_crm_v7.1.3.html#P62.

9 http://id.loc.gov/authorities/subjects/sh85014574.

10 https://id.loc.gov/resources/works/3448552.html.

11 https://dbpedia.org/resource/Bombing_of_Guernica.

DBpedia (one of the largest open repositories of linked data). The Library of Congress Linked Data Service has made their controlled vocabularies available as linked open data, as has the Getty Vocabularies scheme. BIBFRAME is a linked data ontology developed by the Library of Congress, which adapts and simplifies the Library Reference Model and RDA, detailing classes and properties for bibliographic material with the aim of transitioning libraries away from the pre-web standard of MARC. Meanwhile, schemas like MODS, developed using XML, are currently being mapped to BIBFRAME to support the transition as well. This allows the MODS element Subject in Table 2 to have a URI for its role as an RDF predicate.[12] Each of these URIs can be retrieved through a browser for more information. Even in this simple example, we see the interweaving of different metadata standards permitted through linked data.

CIDOC-CRM is formatted as linked data, offering a functional ontology for the domain of cultural heritage. Given CIDOC-CRM's expressivity, it is a much more elaborate standard than BIBFRAME. It is able to describe things, events, actors, places, times, and the numerous relationships between them, to a high degree of granularity (not to mention representing both certainty and uncertainty in knowledge).

The interest in adapting existing knowledge organisation systems (KOS) to linked data led to the development of Simple Knowledge Organization System (SKOS), a data format specifically for reformatting controlled vocabularies as linked data. With SKOS, a controlled vocabulary can be formally expressed and linked to other KOS. This has led to vigorous adoption of SKOS within the heritage sector. Figure 3 shows an example SKOS applied to a hypothetical thesaurus, the "Artistic Styles Vocabulary". Here the entity "Artistic Styles Vocabulary" is defined as "a type" (RDFS) of ConceptScheme (SKOS), within which "Style", "European Style", "Cubism" and "Crystal Cubism" would all be designated Concepts (SKOS) (though only "Style" is shown as such in Figure 3). Note that both preferred and alternate

12 http://id.loc.gov/ontologies/bibframe/subject.

labels for the concept of "Cubism" are indicated (reminiscent of the *res* and *nomens* distinction in the Library Reference Model). The figure also shows how SKOS properties articulate relationships between terms, such as narrower and broader. Finally, using SKOS, we can link this thesaurus's concept of "Cubism" to the very similar (but not exactly similar) concept of "Cubist" in another KOS, the Art and Architecture Thesaurus, making it truly linked data.

Several other linked data vocabularies should also be mentioned. The Dublin Core and Schema.org were both developed to describe web resources, though the former originated from the cultural heritage sector and the latter from the search engine sector. Dublin Core was created as a simple schema for describing resources on the web. It has 15 core elements with broad applicability, including Title, Creator and – of interest to us here – Subject. Schema.org was intended for website labelling for search engines, and as such offers a wider and heterogeneous array of classes and properties for describing people, culture, activities, interests and the like.

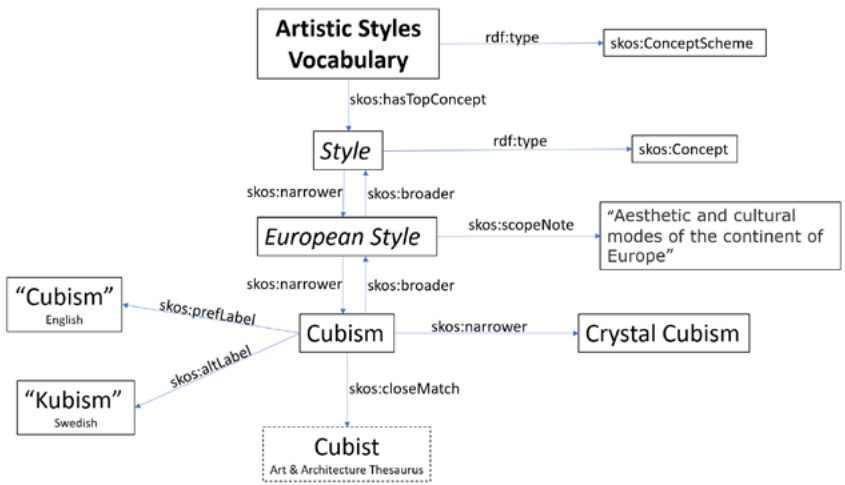

Figure 3: Illustration of SKOS applied to a hypothetical KOS. Note that skos: and rdf: are namespaces standing for <http://www.w3.org/2004/02/skos/core#> and <https://www.w3.org/TR/rdf12-schema/#ch_>, so that "skos:narrower" stands for <http://www.w3.org/2004/02/skos/core#narrower>. All entities (boxes) and relationships (arrows) would also require URIs, except for the boxes containing a literal – a string of text – shown with the use of quotation marks (e.g. "Kubism").

Together, the Semantic Web has revolutionised the possibilities for subject access and revitalised interest in the potential of metadata standards. Nevertheless, in practice, the impact and utility of these innovations for end-users remains negligible, and requires further development.

Catalogues today

Now that we have established the various metadata standards for subject indexing, we must ask: How have these been integrated into catalogue interfaces to support end-users? The answer is "poorly", despite studies showing that subject searches remain a fundamental means by which users explore and navigate collections (Hunter 1991; Meadow and Meadow 2012; Villén-Rueda, Senso and De Moya-Anegón 2007). In this section, these issues will be discussed, alongside the development of online search catalogues and the trend towards subject metadata neglect. In the last section, we will reflect on the types of users now engaged in searches and how they highlight the need to re-visit the role of subject metadata in online searching.

Catalogues go online

Libraries have traditionally supported subject searching and it continues to be a common method for searching in library catalogues. However, subject access has not always been well-implemented. There are several reasons for this.

The first generation of online public library catalogues (OPACs), dating as far back as the 1970s, were developed with a focus on automated efficiency for ongoing institutional practices rather than servicing end-users (Hildreth 1984; Barton and Mak 2012). As such, these early online interfaces often ignored subject searching. The functionalities of early OPACs were restricted to known-item searching, offering exact matching by Author, Title or Control number. As tends to be the case with the early adoption of novel technologies, these new systems reproduced old systems (essentially serving as online card catalogues) rather than fully exploiting the new affordances. This would eventually change. Subject searching using Boolean operators was gradually

introduced. This allowed for queries like "suffragette AND (parade OR march)". But many end-users found such functionalities counterintuitive and difficult to use. These developments would be rendered moot in the 2000s by the popularity of the World Wide Web and services from companies like Google and Amazon which redefined online searching. Users now encountered simple search boxes, attractive website designs, massive recall, relevance ranking, recommendations and directly accessible resources, not to mention full-text indexing as the norm for subject access in documents. As these practices became the industry standard, online catalogues for libraries and other institutions followed suit.

Besides the pressure from without, there were challenges from within that limited the usability of subject metadata in catalogues for the new form of searching. Despite the existence of metadata standards for subject indexing, these were not always applied thoroughly or consistently. Cataloguing rules have generally lacked comprehensive instructions to guide subject indexing or omitted them entirely. The results of this are illustrated in the case of the British Library (Ashton and Kent 2017), where cataloguing was frequently done with outdated or minimal metadata standards that left inadequate subject descriptions, or none at all. In other cases, older subject terms were found inappropriate for reuse in online catalogues, or terms were derived from several different KOS, resulting in inconsistent subject representation across collections. Meanwhile, perennial struggles with institutional budgets impeded investment in cataloguing practices, despite the growing scale and variety of new resources entering collections and requiring cataloguing. Lastly, users raised on simple search and full-text indexing were unaccustomed to using standardised metadata, making it difficult to justify the expense of subject cataloguing. Such conditions led to a steady decline in subject indexing and to less consistent subject indexing, creating a vicious cycle of subject metadata neglect. As a result, library catalogues emulated external search systems, while the potential for subject metadata for search functionalities was left unrealised.

The latest generation of catalogues continued to introduce new

resources (e-books, online journal articles, licensed databases, preprints, digital images, datasets, repositories etc.), which often brought their own metadata. Discovery tools became the norm in this generation of catalogues, allowing a series of catalogues, databases and repositories to be integrated in a single index, wherein the descriptive information of all resources was combined. These discovery tools, which enabled searching across databases, were undoubtedly convenient, saving users from having to search multiple repositories separately. Yet combining metadata from multiple sources presents new challenges, due to a lack of controls within individual databases, let alone mapping the various schemas and knowledge organisation systems across different databases. Lee and Chung (2016) studied the search effectiveness of discovery services, comparing web-scale discovery services with four domain-specific (education and librarianship) databases, and concluded that discovery services were less effective than individual databases. Meanwhile, commercial bibliographic systems claim to provide comprehensive coverage, although their indexing policies are not standardised. Subject, Keyword and Category elements are all used but the differences between them is rarely stated. Nor is any indication usually given of which controlled vocabulary is applied (if any). When searching by means of a discovery system, users are not informed about the lack of mappings across databases. This prevents truly integrated cross-searching, as some resources on a subject will be indexed with terms from one controlled vocabulary while other resources, on the same subject, will be indexed using different terms from another.

In addition to the problem of inconsistent and incomplete metadata, it was found that the blending of controlled vocabularies, free keywords and full-text automatic indexing created the biggest problems for subject searching (Dempsey 2012; Fagan 2011; Golub 2016; Golub 2018). In other cases, the situation is exacerbated when no controlled vocabularies are used whatsoever. Repositories, for instance, often lack international standards or common guidelines for implementing quality-controlled subject access. Instead, a range of different vocabularies are applied across repositories and repository platforms (e.g. Bundza 2014). The National Library of Sweden provides

guidelines for repositories that are made available through the Swedish national repositories service, SwePub.[13] These require works to be assigned a designation from the National Subject Category (a rather modestly sized controlled vocabulary of over 250 terms), while the Keywords element has no pre-defined values.

As a means of redressing some of the problems of modern search interfaces (simple queries, high recall, low precision) faceted navigation has become a standard feature of catalogue interfaces. Alongside a list of results, users are automatically presented with a choice of facets to use as filters. By choosing facets, users can drill down into the query results. Facet categories are built from metadata elements (e.g. Format, Date, Author) and the values from the records within these fields create checkboxes that users can select or remove ("image", "article", "book", "film"). Tellingly, "subject" is often one of these facets (Chickering and Yang 2014). Nevertheless, studies show that users tend to be confused by these features and lack an understanding of how facets work, or the type of terms included in them (Emanuel 2011; Osborne and Cox 2015). This is an expected result when no mappings have been conducted between different subject terms used by incorporated collections.

Searching museum catalogues highlights a similar neglect of subject indexing. Despite the many standards in place, quality-controlled subject indexing and the information retrieval interfaces which leverage them are largely missing from museum catalogues. This could be explained by that fact that while some museums support subject access (Liew 2005), Trant et al. (2006) found that others (especially art museums) do not do so because providing subject access is not considered a priority compared with operations like object registration, inventory or location control. Indeed, it appears that many museum catalogues offer users only the most basic schema: title of work, dimensions, creator's name, and perhaps a picture of the object (Fortier and Ménard 2018).

13 http://www.swepub.se.

In many ways, museums' catalogues may resemble early OPACs. But unlike libraries, which have a stronger tradition of creating catalogues for public use and therefore made more regular use of subject indexing, museums continued creating catalogues primarily for inventory and administrative purposes. According to one study, which assessed 91 Swedish museum websites, subject indexing was largely missing from search services (Golub, Ziolkowski and Zlodi 2022). The same study analysed the search interfaces against a set of 21 criteria. The results showed that effective subject access was largely unavailable in existing services, which matched the findings for journal databases, journal articles, repositories and discovery services (Golub 2018; Golub et al. 2020).

Users

The transition to the online environment has created an unprecedented opportunity for cultural heritage institutions (libraries, museums, archives and galleries) to make their collections accessible and available to a wider range of patrons than ever before. The Getty Foundation noted that while "printed permanent collection catalogues are typically aimed at an exclusively scholarly audience, the Internet allows museums to engage multiple audiences simultaneously" (The Getty Foundation 2012, 19). The pivot towards heterogenous publics begs the question: How do these groups engage with collections, and could more be done to make these collections more accessible to the newer audiences? This is especially pertinent because studies show that, despite the lack of facilitation for subject access, subject searching remains a common type of query in library catalogues (Hunter 1991; Villén-Rueda, Senso and De Moya-Anegón 2007), bibliographic databases (Siegfried, Bates and Wilde 1993), repositories (Heery et al. 2006), discovery services (Meadow and Meadow 2012) and related digital search services (Patel et al. 2005).

In this subsection, we will consider the contrast between general users and experts/scholars, with the latter group focusing on humanities scholars (since humanities material is especially challenging to index by subject and is more representative of heritage collections in

general). Despite the different knowledge, needs and motivations that each of the groups bring with them when searching, they both now make use of same online search systems. And both, as we will see, exhibit a strong tendency to prefer subject searches and would therefore benefit from a more rigorous integration of subject indexing in search interfaces.

With respect to search preferences, a comprehensive study of seven museums showed that the majority (67 per cent of the public and 59 per cent of non-professional researchers) prefer browsing a collection to using a search box (David Walsh et al. 2020). In contrast, results from Denmark's National Museum of Military History observed that users chose browsing only for exploratory searching but preferred free-text searches in all other types of information seeking (Skov and Ingwersen 2014). Subject-related metadata standards implemented in these collections, while appreciated, went largely unmentioned by participants. This suggests either that these elements, and the controlled vocabulary they include, are not prominently integrated into search functionalities, or that users are unfamiliar with how best to exploit them.

Another alternative method to enable subject access can be seen in the use of social tagging in museum interfaces and projects such as steve.museum (Trant 2006). Social tags have been criticised for their subjective nature, as users themselves choose and create labels for items. This leads to a lack of term consistency as well as other features from controlled vocabularies, such as hierarchical relationships between terms (e.g. Srinivasan et al. 2009; Trant 2009). All the same, social tags hold the potential to fill the existing subject access gap, thus improving access to museum collections for the public (Trant 2006). Social tagging has also been found to augment the professional descriptions of art museums' collections: 77–85 per cent of tags found in the research (the ratio depends on the kinds of objects and their specific context) are new terms not existing in the standard vocabularies in use. Most of these are subject terms (39%) and genre terms (27%) (Trant 2009b).

Enabling subject access would be useful for the public, as it

encourages the exploratory searching these users are interested in. All this points to the need for subject indices to better support access to collections. This would provide a better overview of the collections' contents and counter problems of high recall (with hundreds or thousands of hits) but low precision (too many irrelevant hits), which can lead to poor user experiences when searching.

In contrast to general users, most professional researchers require information with a high level of granularity when searching for information in bibliographic databases (de Andrade and Baptista 2014). A comparative analysis of humanities scholars' queries, based on search logs from the Getty catalogue, and sciences scholars' queries reveals dramatic differences in the type of subjects the scholars searched for (Bates 1996). Scientific queries typically comprise common terms and only rarely proper terms (titles, names, places, periods, discipline etc.). In contrast, only 57 per cent of the humanities queries contained any common terms at all, relying more on proper terms. Yet a close examination of the natural language queries shows that most searches were not known-item but unknown-item or subject searches. 91 per cent of statements indicated a subject of some kind, strongly substantiating the need to support subject searching for humanities scholars. Despite this, the evidence showed that the Getty library's catalogue was better suited to known-item searches.

Similar findings have been reported elsewhere. Wiberley (1983; 1988) observed that humanities subject terms are often highly precise proper terms, while Tibbo (1994) noted that single proper terms like authors' names and the titles of works are common in certain disciplines, such as literary studies, while common terms are more characteristic of fields such as religion and philosophy. Likewise, Yi et al. (2006) found that most search terms used in the history databases they studied referred to specific instances of historical events, people and regions, in contrast to the search terms used in a psychology database, which were mostly common terms matching those of a classification system.

While full-text indexing works for some tasks, for others it creates information overload and prevents the searcher from gaining a

comprehensive overview of a topic. This is especially evident in areas such as the humanities, which is heterogenous, parochial and complicated, compared to scholarship in STEM fields that tends to be complex but disciplined. Even relatively well-defined subdisciplines within the humanities exhibit a high volume of distinct, overlapping or associated concepts across literature, culture and the arts. The terminological complexity, exemplified by synonyms, complicates queries by placing the burden on the scholar, who would ideally need to include all possible synonyms in a query if a comprehensive set of results is desired. Homonyms are likewise a challenge, leading to false positives. Furthermore, since humanities scholars frequently work with materials and sources produced over long periods of time, changing terminology presents yet another obstacle.

Consequently, automatically indexed terms may conflate different meanings, such as the same term for different periods (e.g. the "Renaissance", which refers to different timespan in Italy and England) or different terms which refer to the same concept (the "October War" and the "Yom Kippur War"). When it comes to searching by titles, terms may be metaphorical, containing allusions or intertextual references, but lack descriptive power, resulting in low recall as well as false positives. Similarly complex problems are also commonly encountered when it comes to primary sources. For example, literary studies of sexuality do not traditionally employ computational methods on large corpora because the phenomenon studied is not openly manifested in texts, calling for human subject indexing (Bergenmar and Golub 2020). Simply put, the terms extracted from texts commonly lack the conceptual granularity needed by humanists. Tibbo (1994, 609) further notes that humanities scholars tend to use a "dense" rather than "readable" writing style, making it particularly challenging to create representative metadata, the quality of which depends on the cataloguer being a subject expert.

Integrating subject access into search interfaces could afford users a higher level of specificity and context when searching for a given topic. This would be especially useful for large databases and cross-search services with tens of millions of records. For instance,

Knapp (1998) established that the most effective way of searching databases online is to combine free-text searching with the use of controlled vocabulary indexing. Controlled vocabularies are particularly needed in large databases covering many subjects (Markey 2007; Tibbo 1994) as well as in databases of primary sources (Bair and Carlson 2008) such as museum objects, which cannot be simply queried against a full-text index. Tibbo (1994) makes the point that the exponentially increasing volume of information objects available online leads to information overload and information entropy, rather than increasing information access. Applying the tools we already have at our disposal for representing subject matter could resolve these issues while ushering in new affordances.

Reintroducing subject standards

The National Gallery of Art lists a total of 75 functional requirements for museum digital collections, of which 31 relate to search and retrieval (The Getty Foundation, 2012). Of the many suggestions included, such as keyword search, the recommendation includes a specific reference to controlled vocabularies (like the Art and Architecture Thesaurus) as well as several other functions which would directly benefit from subject metadata integration. These requirements are:

- Auto-suggestion of available terms.
- Refinement of search results by modifying the search criteria.
- Support for search and browse functionality with a synonym ring, authority files and provision of alternatives to those entered by the user.
- Expansion of results with broader terms.
- Faceted browse searching.
- Linked terms from search results to other results linked to the same terms.
- Contextual aids such as pop-up descriptions.
- Visualisations (e.g. timelines, maps and networks) of events, places, agents and objects.

Repositioning subject metadata in contemporary search interfaces would be pivotal for realising a new generation of catalogues, catalogues that still leverage contemporary search technologies from outside the cultural heritage sector but also exploit the subject access standards from within the sector. The implementation of search functionalities based on subject metadata, for instance, was attempted with the previously mentioned Queer Literature Indexing Thesaurus (Golub, Bergenmar and Humlesjö 2023). Such applications of subject metadata, substantively integrating them into the search interface, could enable four factors essential for optimising the accessibility of our collections:

- Specificity
- Browsability
- User assistance
- Harmonization

Few search systems today support browsing through the conceptual relationships set out in thesauruses or classification systems; or offer users options for narrower, broader or related concepts; or find which term is best to name a particular concept while disambiguating homonyms.

Faceted vocabularies are highly suitable for improving the search process, since they support high specificity and the stacking of geographical, chronological and disciplinary terms, such as "Protest songs—USA—1970–1979" versus "Protest songs—France—1960–1969" (see Bates 1996; Tibbo 1994). Facet selection and query expansion based on such controlled vocabularies need to be implemented effectively in search interfaces. At present, these seem to be limited to experimental interfaces (e.g. Alani, Jones and Tudhope 2000; Tudhope et al. 2006) rather than applied in practice across online search systems. For a notable example of this, contrast a simple search

of "Great Depression" in the Cornell University Library Catalog[14] and the University of California's Berkeley Library discovery system.[15] Both return results with subject facets, but Cornel's catalogue breaks these subject facets into Subject: General ("Depressions", "Economic Policy", "Social History" etc.), Subject: Region ("United States", "Germany", "New York" etc.), and Subject: Era ("1900–1999", "since 1918" etc.). In contrast, Berkeley presents a largely undisciplined series of facets ("United States", "Political Science", "Business, Economics and Finance – Communications – Newspapers"), due to the unreconciled metadata across databases.

Furthermore, the Cornell Catalog offers contextual information. For instance, a subject browse for "Eliot, George" will lead to an information page populated with data automatically pulled from the Library of Congress Name Authority File and DBpedia.[16] The same page offers a list of works in the catalogue *by* Eliot (203 at present) and works *about* Eliot (497), distinguishing the entity of George Eliot as creator from subject matter within the catalogue. These interface choices demonstrate deeper integration of subject metadata for a better sense of what is being searched for and what the collection has. This could be further built upon by exploiting the technologies of linked data described above to extensively enrich the representation and discovery of resources, as they can be intertwined with ever larger networks of formal subject representation across the Semantic Web.

Hierarchical browsing of classification schemes and other controlled vocabularies could help users gain an overview of the collection, as well as improve their understanding of their information needs. This could aid users to formulate their queries more accurately. In another example from the Cornell Catalog, the subject page for

14 https://catalog.library.cornell.edu/.

15 https://search.library.berkeley.edu/.

16 https://catalog.library.cornell.edu/browse/info?authq=Eliot,%20George,%20 1819-1880.&browse_type=Subject&headingtype=Personal%20Name. Accessed on 26 August 2024.

"Russo-Japanese War, 1904–1905",[17] is also enriched with information from DBpedia and the Library of Congress files. Moreover, it links to a broader subject term in the catalogue ("Eastern question (Far East)"), enumerates the number of works catalogued with this term (521), and shows variant subject headings ("Japan > History > War with Russia, 1904–1905", "Japanese-Russian War, 1904–1905", "Russia > History > War with Japan, 1904–1905"). Another illustration of functionalities enabled through controlled vocabularies can be seen in Sweden's national union catalogue, Libris, which may be navigated through the Dewey Decimal Classification.[18]

Another aspect for potential inclusion in future catalogues would be automated user assistance, offering users suggestions which leverage subject metadata to help them with search strategies, search techniques and query formulation. Given how disoriented many users find themselves, and how little experience most have beyond conducting simple searches, such features could be pivotal in guiding users through a search interface and help familiarise them with the layout, functionalities and catalogue. This is nowhere more important than when conducting an unknown-item search. Subject metadata would therefore be crucial for tailoring assistance and suggesting remedies for the usual problems users encounter during subject searches.

Lastly, there is the struggle to ameliorate cross-search services, where subject searching is undermined by the inconsistency and incompleteness of metadata and the blending of controlled vocabularies, keywords and full-text indexing (Dempsey 2012; Fagan 2011; Golub 2016). Numerous national and international infrastructure projects are working to make cultural heritage collections interoperable with each other. Semantic Web standards and interoperability, as discussed above, provide opportunities for cross-institutional searching and linking data across collections. Many institutions today provide

17 https://catalog.library.cornell.edu/browse/info?authq=Russo-Japanese%20War,%20 1904-1905&browse_type=Subject&headingtype=Topical%20Term. Accessed on 26 August 2024.

18 https://deweysearchsv.pansoft.de/webdeweysearch/.

metadata to portals such as the World Digital Library, which allows cross-searching of dispersed collections. Several bridges are already available to connect metadata elements across schemas (e.g. linking Categories for the Description of Works of Art element to near matches in VRA Core or Dublin Core), as well as knowledge organisation systems (e.g. linking the concept of "Cubism" in LCSH with "Cubist" in AAT). FAST (Faceted Application of Subject Terminology) is a schema which simplifies LCSH and connects it to other sources (e.g. Wikipedia); FAST Linked Data goes further to make available linked data authorities formatted in schema.org and SKOS for the Semantic Web.[19] Finally, the development of a new version of the Lightweight Information Describing Objects (LIDO) data exchange standard should facilitate the display of such integrated metadata and interconnected collections in search interfaces.

Closing caveats

Future research should focus specifically on user interfaces for subject access – be it supporting query expansion, word sense disambiguation or other possibilities – based on specific user needs. Moreover, further testing with users for optimal interface and functionality design is needed. All these refinements should be rooted in user studies, analyses of real search sessions, and the inclusion of all potential user groups (domain-specific and interdisciplinary scholars, students, cultural heritage professionals, and the segments of the public). More needs to be done to understand the reasons for the sorry state of subject access in cultural heritage catalogues. While it is highly likely that much of this underdevelopment can be attributed to historical legacies and strained resources, there remains a persistent need for better subject access. This can only be realised by simultaneously leveraging the powerful tools at our disposal: digital technologies and metadata standards. Properly integrated, these are fundamental for making our collections more knowable, discoverable, shareable and meaningful.

19 For example, http://experimental.worldcat.org/fast/833708/.

Acknowledgments

Our sincerest gratitude to our superb reviewer, wonderful editor and incredible copyeditors for their suggestions (and patience). Their feedback vastly improved this chapter.

References

ALANI, HARITH, CHRISTOPHER JONES AND DOUGLAS TUDHOPE. 2000. "Associative and Spatial Relationships in Thesaurus-Based Retrieval." In *Research and Advanced Technology for Digital Libraries*, edited by José Borbinha and Thomas Baker, 45–58. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45268-0_5.

ANDRADE, MORGANA CARNEIRO DE AND ANA ALICE BAPTISTA. 2014. "Researchers' Information Needs in the Bibliographic Database: A Literature Review." *Information Services & Use* 34 (3–4): 241–48. https://doi.org/10.3233/ISU-140744.

ASHTON, JANET, AND CAROLINE KENT. 2017. "New Approaches to Subject Indexing at the British Library." *Cataloging & Classification Quarterly*, 55(7–8), 549–559. https://doi-org.proxy.lnu.se/10.1080/01639374.2017.1354345

BACA, MURTHA, AND PATRICIA HARPRING, EDS. 2016. *Categories for the Description of Works of Art*. Los Angeles, CA: Getty Research Institute. https://www.getty.edu/research/publications/electronic_publications/cdwa/.

BACA, MURTHA, PATRICIA HARPRING, ELISA LANZI, LINDA MCRAE AND ANN WHITESIDE, EDS. 2006. *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images*. Chicago, IL: American Library Association. http://vraweb.org/wp-content/uploads/2018/08/CatalogingCulturalObjectsFull.pdf.

BAIR, SHEILA, AND SHARON CARLSON. 2008. "Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship." *Journal of Library Metadata* 8 (3): 249–62. https://doi.org/10.1080/19386380802398503.

BARTON, JOSHUA, AND LUCAS MAK. 2012. "Old Hopes, New Possibilities: Next-Generation Catalogues and the Centralization of Access." *Library Trends* 61: 83–106.

BATES, MARCIA J. 1996. "The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions." *College & Research Libraries* 57 (6): 514–23. https://doi.org/10.5860/crl_57_06_514.

BERGENMAR, JENNY, AND KORALJKA GOLUB. 2020. "Subject Indexing: The Challenge of LGBTQI Literature." In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference: Riga, Latvia, October 21–23, 2020*, edited by Sanita Reinsone, Inguna Skadiņa, Anda Baklāne and Jānis Daugavietis, 203–210. Aachen: CEUR Workshop Proceedings.

BUNDZA, MAIRA. 2014. "The Choice Is Yours! Researchers Assign Subject Metadata to Their Own Materials in Institutional Repositories." *Cataloging & Classification Quarterly* 52 (1): 110–18. https://doi.org/10.1080/01639374.2013.852439.

CHICKERING, F. WILLIAM, AND SHARON Q. YANG. 2014. "Evaluation and Comparison of Discovery Tools: An Update." *Information Technology and Libraries* 33 (2): 5. https://doi.org/10.6017/ital.v33i2.3471.

DEMPSEY, LORCAN. 2012. "Thirteen Ways of Looking at Libraries, Discovery, and the Catalog: Scale, Workflow, Attention." *EDUCAUSE Review*. 2012. https://er.educause.edu/articles/2012/12/thirteen-ways-of-looking-at-libraries-discovery-and-the-catalog-scale-workflow-attention.

EMANUEL, JENNIFER. 2011. "Usability of the VuFind Next-Generation Online Catalog." *Information Technology and Libraries* 30 (1): 44. https://doi.org/10.6017/ital.v30i1.3044.

FAGAN, JODY CONDIT. 2011. "Discovery Tools and Information Literacy." *Journal of Web Librarianship* 5 (3): 171–78. https://doi.org/10.1080/19322909.2011.598332.

FORTIER, ALEXANDRE, AND ELAINE MÉNARD. 2018. "What Do Museum Website Users Expect from Linked Open Data?" *Advances in Knowledge Organization* 16: 900–907.

GNOLI, CLAUDIO. 2020. *Introduction to Knowledge Organization*. London: Facet Publishing.

GOLUB, KORALJKA. 2016. "Potential and Challenges of Subject Access in Libraries Today on the Example of Swedish Libraries." *International Information & Library Review* 48 (3): 204–10. https://doi.org/10.1080/10572317.2016.1205406.

———. 2018. "Subject Access in Swedish Discovery Services." *Knowledge Organization* 45 (4): 297–309. https://doi.org/10.5771/0943-7444-2018-4-297.

GOLUB, KORALJKA, JENNY BERGENMAR AND SISKA HUMLESJÖ. 2022. "Searching for Swedish LGBTQI Fiction: Challenges and Solutions." *Journal of Documentation* 78 (7): 464–84. https://doi.org/10.1108/JD-06-2022-0138.

———. 2023. "Searching for Swedish LGBTQI fiction: The Librarians' Perspective." *Journal of Documentation*, 79 (7), 261–279. https://doi.org/10.1108/JD-05-2023-0080

GOLUB, KORALJKA, JUKKA TYRKKÖ, JOACIM HANSSON AND IDA AHLSTRÖM. 2020. "Subject Indexing in Humanities: A Comparison between a Local University Repository and an International Bibliographic Service." *Journal of Documentation* 76 (6): 1193–1214. https://doi.org/10.1108/JD-12-2019-0231.

GOLUB, KORALJKA, PAWEL MICHAL ZIOLKOWSKI AND GORAN ZLODI. 2022. "Organizing Subject Access to Cultural Heritage in Swedish Online Museums." *Journal of Documentation* 78 (7): 211–47. https://doi.org/10.1108/JD-05-2021-0094.

HARPRING, PATRICIA. 2018. "Linking the Getty Vocabularies: The Content Perspective, Including an Update on CONA." In *2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, 1–8. San Francisco, CA: IEEE. https://doi.org/10.23919/PNC.2018.8579460.

HEERY, RACHEL, LIZ LYON, CHRISA TSINARAKI, TIM BRODY, TRAUGOTT KOCH AND MARTIN DOERR. 2006. "Report on Digital Repositories: An Evaluation Study on the Development and Implementation of Community Repositories to Support Research (And Learning and Teaching)." *DELOS2 Network of Excellence on Digital Libraries*.

HILDRETH, CHARLES R. 1984. "Pursuing the Ideal: Generations of Online Catalogs." In *Online Catalogs, Online Reference: Converging Trends (Proceedings of a Library and Information Technology Association Preconference Institute, June 23–24, 1983, Los Angeles)*, edited by Brian Aveney and Brett Butler, 31–56. Chicago, IL: American Library Association.

HUNTER, RHONDA N. 1991. "Successes and Failures of Patrons Searching the Online Catalog at a Large Academic Library: A Transaction Log Analysis." *RQ* 30 (3): 395–402.

KNAPP, SARA D., LAURA B. COHEN AND D.R. JUEDES. 1998. "A Natural Language Thesaurus for the Humanities: The Need for a Database Search Aid." *The Library Quarterly* 68 (4): 406–30. https://doi.org/10.1086/603001.

LEE, BORAM, AND EUNKYUNG CHUNG. 2016. "An Analysis of Web-Scale Discovery Services From the Perspective of User's Relevance Judgment." *The Journal of Academic Librarianship* 42 (5): 529–34. https://doi.org/10.1016/j.acalib.2016.06.016.

LIEW, CHERN LI. 2005. "Online Cultural Heritage Exhibitions: A Survey of Information Retrieval Features." *Program* 39 (1): 4–24. https://doi.org/10.1108/00330330510578778.

MARKEY, KAREN. 2007. "The Online Library Catalog: Paradise Lost and Paradise Regained?" *D-Lib Magazine* 13 (1/2). https://doi.org/10.1045/january2007-markey.

MEADOW, KELLY, AND JAMES MEADOW. 2012. "Search Query Quality and Web-Scale Discovery: A Qualitative and Quantitative Analysis." *College & Undergraduate Libraries* 19 (2–4): 163–75. https://doi.org/10.1080/10691316.2012.693434.

OSBORNE, HOLLIE M., AND ANDREW COX. 2015. "An Investigation into the Perceptions of Academic Librarians and Students towards Next-Generation OPACs and Their Features." *Program* 49 (1): 23–45. https://doi.org/10.1108/PROG-10-2013-0055.

PATEL, MANJULA, TRAUGOTT KOCH, MARTIN DOERR AND CHRISA TSINARAKI. 2005. "Semantic Interoperability in Digital Library Systems." *DELOS2 Network of Excellence on Digital Libraries*. http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/SI-in-DLs.pdf

RIVA, PAT, PATRICK LE BOEUF AND MAJA ŽUMER. 2017. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. Den Haag: International Federation of Library Associations and Institutions.

SIEGFRIED, SUSAN, MARCIA J. BATES AND DEBORAH N. WILDE. 1993. "A Profile of End-User Searching Behavior by Humanities Scholars: The Getty Online Searching Project Report No. 2." *Journal of the American Society for Information Science* 44 (5): 273–91. https://doi.org/10.1002/(SICI)1097-4571(199306)44:5<273::AID-ASI3>3.0.CO;2-X.

SKOV, METTE, AND PETER INGWERSEN. 2014. "Museum Web Search Behavior of Special Interest Visitors." *Library & Information Science Research* 36 (2): 91–98. https://doi.org/10.1016/j.lisr.2013.11.004.

SMUCKER, MARK D. 2011. "Information Representation." In *Interactive Information Seeking, Behaviour, and Retrieval*, edited by Ian Ruthven and Diane Kelly, 77–94. London: Facet Publishing.

SRINIVASAN, RAMESH, ROBIN BOAST, KATHERINE M. BECVAR AND JONATHAN FURNER. 2009. "Blobgects: Digital Museum Catalogs and Diverse User Communities." *Journal of the American Society for Information Science and Technology* 60 (4): 666–78. https://doi.org/10.1002/asi.21027.

THE GETTY FOUNDATION. 2012. *Moving Museum Catalogues Online: An Interim Report from the Getty Foundation*. Los Angeles, CA: Getty Foundation. https://www.getty.edu/foundation/pdfs/osci_interimreport_2012.pdf.

TIBBO, HELEN R. 1994. "Indexing for the Humanities." *Journal of the American Society for Information Science* 45 (8): 607–19. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<607::AID-ASI16>3.0.CO;2-X.

TRANT, JENNIFER. 2009a. "Studying Social Tagging and Folksonomy: A Review and Framework." *Journal of Digital Information* 10 (1). http://hdl.handle.net/10150/105375.

———. 2009b. "Tagging, Folksonomy and Art Museums: Early Experiments and Ongoing Research." *Journal of Digital Information* 10 (1). http://hdl.handle.net/10150/106510.

TRANT, JENNIFER, WITH THE PARTICIPANTS IN THE STEVE.MUSEUM PROJECT. 2006. "Exploring the Potential for Social Tagging and Folksonomy in Art Museums: Proof of Concept." *New Review of Hypermedia and Multimedia* 12 (1): 83–105. https://doi.org/10.1080/13614560600802940.

TUDHOPE, DOUGLAS, CERI BINDING, DOROTHEE BLOCKS AND DANIEL CUNLIFFE. 2006. "Query Expansion via Conceptual Distance in Thesaurus Indexed Collections." *Journal of Documentation* 62 (4): 509–33. https://doi.org/10.1108/00220410610673873.

VILLÉN-RUEDA, LUIS, JOSE A. SENSO AND FELIX DE MOYA-ANEGÓN. 2007. "The Use of OPAC in a Large Academic Library: A Transactional Log Analysis Study of Subject Searching." *The Journal of Academic Librarianship* 33 (3): 327–37.

WALSH, DAVID, MARK M. HALL, PAUL CLOUGH AND JONATHAN FOSTER. 2020. "Characterising Online Museum Users: A Study of the National Museums Liverpool Museum Website." *International Journal on Digital Libraries* 21 (1): 75–87. https://doi.org/10.1007/s00799-018-0248-8.

WIBERLEY, STEPHEN E. 1983. "Subject Access in the Humanities and the Precision of the Humanist's Vocabulary." *The Library Quarterly* 53 (4): 420–33. https://doi.org/10.1086/601405.

———. 1988. "Names in Space and Time: The Indexing Vocabulary of the Humanities." *The Library Quarterly* 58 (1): 1–28. https://doi.org/10.1086/601949.

YI, KWAN, JAMSHID BEHESHTI, CHARLES COLE, JOHN E. LEIDE AND ANDREW LARGE. 2006. "User Search Behavior of Domain-Specific Information Retrieval Systems: An Analysis of the Query Logs from PsycINFO and ABC-Clio's Historical Abstracts/America: History and Life." *Journal of the American Society for Information Science and Technology* 57 (9): 1208–20. https://doi.org/10.1002/asi.20401.

Contributors

MAGNUS BREDER BIRKENES is a Senior Researcher at the National Library of Norway. He earned his PhD at the University of Marburg, Germany, in 2013, with a thesis on the morphology of German dialects. He has published on the field of corpus linguistics and dialectometry, and is involved in developing language technology and building research infrastructure at the National Library of Norway.

PIA BORLUND is Professor of Library and Information Science at Oslo Metropolitan University, Norway. She holds a degree in librarianship and a master's degree in library and information Science (MLISc) from the Royal School of Library and Information Science, Denmark. She also holds a doctorate (PhD) on the evaluation of interactive information retrieval (IIR) systems from Åbo Akademi University, Finland. Her research interests include interactive information retrieval, human-computer interaction and information seeking (behaviour). She is concerned with methodological issues, test design and recommendations for the evaluation of user-based performance and search interaction.

SISSEL FURUSETH is Professor of Scandinavian Literature at the University of Oslo. She earned her doctorate from the Norwegian University of Science and Technology (NTNU) in 2003 and has published widely on poetic modernism, ecocriticism and literary critical reception. In 2009–2015, she coordinated the national research project *Norsk litteraturkritikks historie 1870–2000*. She is currently heading the cross-disciplinary project *Translatability of Oil*, combining insights from periodical studies and the growing field of energy humanities. Among her latest publications, are *Økokritisk håndbok* (Universitetsforlaget, 2023), which she co-authored with Reinhard Hennig, and the article "Petrocultures in the making: Oil in 1920s Scandinavian newspapers" (*JEHRHE*, 2023).

KORALJKA GOLUB is a full Professor and Head of iInstitute at Linnaeus University, Sweden. Her research focuses primarily on topics related to information retrieval and knowledge organisation. She is particularly interested in the integration of traditional knowledge organisation systems with social tagging and/or automated subject indexing, and evaluating results in the context of end-user information retrieval.

JENS-MORTEN HANSSEN is a Senior Researcher at the National Library of Norway. He earned a PhD degree from the University of Oslo in 2018 with a thesis on the early reception of Henrik Ibsen on the German stage. He is the author of the book *Ibsen on the German Stage 1876–1918: A Quantitative Study* (Narr Francke Attempto Verlag, 2018). He has published extensively in the field of literary and theatre studies, particularly on authors such as Ibsen, Bjørnstjerne Bjørnson and Jon Fosse. At the National Library of Norway, he is involved in creating digital research services as well as developing bibliographic resources.

LARS G. JOHNSEN is a Senior Researcher at the National Library of Norway, with a PhD from the University of Trondheim. At the library, he plays a central role in developing the Digital Humanities infrastructure, focusing on methods for analysing data from digitised texts and images. His most recent work includes developing language modelling and analysis techniques, particularly in relation to structured metadata for library objects.

AHMAD M. KAMAL is a Senior Lecturer at the Department of Cultural Sciences, Linnaeus University, Sweden, where he teaches topics such as knowledge organisation, information retrieval and digital collections. He also coordinates a Digital Humanities master's programme and serves as a member of the metadata

working group for Huminfra, a Swedish national infrastructure for digital research in the humanities. He received his PhD in Library and Information Science from the Western University, Canada.

LEO LAHTI is Professor of Data Science and Computational Humanities at the University of Turku, Finland, focusing on computational analysis of complex natural and social systems. Lahti obtained a doctorate in applied probabilistic machine learning from Aalto University, Finland, in 2010. He has led the preparation of the Finnish national policy on open research methods, coordinated international data science networks and training, and currently serves on the executive committee of the International Science Council's Committee on Data.

JANI MARJANEN is a Lecturer in Political History at the University of Helsinki, Finland. His research focuses on discourses of patriotism and nationalism, the history of ideology and isms (together with Jussi Kurunmäki), Digital Humanities approaches to analysing historical data, and the theory and method of conceptual history.

ODDRUN PAULINE OHREN is a Senior Adviser at the National Library of Norway, specialising in knowledge organisation, metadata analysis and metadata models. She holds a Master of Science in Informatics from University of Oslo as well as a Master of Library and Information Science from Oslo Metropolitan University. Currently, she is working to prepare the library sector's transition to the linked metadata framework Resource Description and Access (RDA). Ohren is also involved in international projects, notably the library-driven initiative Share-VDE: linked data for libraries.

NILS PHARO is Professor of Knowledge Organisation and Information Retrieval at Oslo Metropolitan University. He has a master's degree in library and information science from Oslo University College and a doctorate in information studies from the University of Tampere, Finland. Pharo's research interests are in the field of information behaviour, interactive information retrieval and scholarly communication.

TUULI TAHKO, PhD, has worked as a Postdoctoral Researcher at the Department of Digital Humanities, University of Helsinki, Finland. Besides vernacularisation, her research interests revolve around embodied interaction and communities of practice.

MIKKO TOLONEN is Professor of Digital Humanities at the University of Helsinki, Finland. His background is in intellectual history, and he is the Principal Investigator (PI) of the Helsinki Computational History Group (COMHIS). His main research focus is an integrated study of public discourse and knowledge production that combines metadata from library catalogues as well as full-text libraries of books, newspapers and periodicals in early modern Europe. Tolonen also works in other areas of Enlightenment studies, such as the intellectual development of Bernard Mandeville and David Hume.

Nota bene is the National Library of Norway's channel for disseminating research findings built on its collections, and research of relevance to these collections. All manuscripts are peer reviewed. Nota bene has a wide thematic profile. In order to mirror the full breadth of our collection, the publications, which include monographs, critical editions and edited collections, may be based on manuscripts, printed material, film, photography, music, broadcasting, and digital media.

NOTA BENE 1
Det nasjonale i Nasjonalbiblioteket | Marianne Takle | 2009

NOTA BENE 2
The Archive in Motion. New Conceptions of the Archive in Contemporary Thought and New Media Practices | Eivind Røssaak (ed.) | 2009

NOTA BENE 3
Axel Charlot Drolsum. Brev 1875–1926 | Bjørg Dale Spørck | 2011

NOTA BENE 4
Opplysning, vitenskap og nasjon. Bidrag til norsk bibliotekhistorie
Ruth Hemstad (ed.) | 2011

NOTA BENE 5
Latin Manuscripts of Medieval Norway. Studies in Memory of Lilli Gjerløw
Espen Karlsen (ed.) | 2013

NOTA BENE 6
Den engasjerte kosmopolitt. Nye Bjørnson-studier
Liv Bliksrud, Giuliano D'Amico, Marius Wulfsberg and Arnfinn Åslund (eds.) | 2013

NOTA BENE 7
Naturen og eventyret. Dokumentarfilmskaperen Per Høst
Gunnar Iversen | 2014

NOTA BENE 8
Å bli en stemme. Nye studier i Camilla Colletts forfatterskap
Trond Haugen (ed.) | 2014

NOTA BENE 9
Propagandakrig. Kampen om Norge i Norden og Europa 1812–1814
Ruth Hemstad | 2014

NOTA BENE 10
Small Country, Long Journeys. Norwegian Expedition Films
Eirik Frisvold Hanssen and Maria Fosheim Lund (eds.) | 2017

NOTA BENE 11
Reformasjonstidens religiøse bokkultur cirka 1400–1700:
tekst, visualitet og materialitet | Bente Lavold and John Ødemark (eds.) | 2017

NOTA BENE 12
I dørtrekken fra Europa. Festskrift til Knut Sprauten. I anledning 70-årsdagen
22. juni 2018 | Ola Alsvik, Hans P. Hosar and Marianne Wiig (eds.) | 2018

NOTA BENE 13
Litterære verdensborgere. Transnasjonale perspektiver på norsk
bokhistorie 1519–1850 | Aasta M.B. Bjørkøy, Ruth Hemstad, Aina Nøding and
Anne Birgitte Rønning (eds.) | 2019

NOTA BENE 14
Lov og lovgivning i middelalderen. Nye studier av Magnus Lagabøtes landslov
Anna Catharina Horn and Karen Arup Seip (eds.) | 2020

NOTA BENE 15
Notated Music in the Digital Sphere: Possibilities and Limitations
Margrethe Støkken Bue and Annika Rockenberger (eds.) | 2021

NOTA BENE 16
Språk i arkivet. Historier om hvordan språk reflekterer samfunnet
Johanne Ostad and Elise Kleivane (eds.) | 2021

NOTA BENE 17
Silent Ibsen. Transnational Film Adaptation in the 1910s and 1920s
Eirik Frisvold Hanssen and Maria Fosheim Lund (eds.) | 2022

NOTA BENE 18
Old Norse Law Books from a Material Perspective
Jóhanna Katrín Friðriksdóttir and Lukas Rösli (eds.) | 2024

NOTA BENE 19
The Hermeneutics of Bibliographic Data and Cultural Metadata
Jens-Morten Hanssen and Sissel Furuseth (eds.) | 2025

ISBN 978-82-7965-587-9 (printed)
ISBN 978-82-7965-588-6 (e-book)

ISSN 1891-4829 (printed)
ISSN 2535-4337 (e-book)

Design: Superultraplus Designstudio AS
www.superultraplus.com

Print: Erik Tanche Nilssen AS

Computational tools and the digital revolution propelled by the invention of the World Wide Web pervade every aspect of human existence, including the work life of researchers across the humanities, social and natural sciences, and staff at research libraries. The development of library systems over the last hundred years has been heavily influenced by the advent of information and computer science. Bibliographical work and metadata registry as a core activity within the world of libraries are permeated with computation, and this development opens up new avenues for digital approaches and data-driven research.

This book presents a wide variety of approaches that explore the intersection between two very diverse fields of practice, bibliography and the production of library metadata on the one hand, and the use of computation in the humanities and social sciences on the other. How and to what extent has digitisation altered the field of bibliography and metadata production? What opportunities for conducting research on bibliographic data and cultural metadata are currently available?

Nota bene is the National Library of Norway's channel for disseminating the findings of research built on its collections and research of relevance to these collections. All manuscripts are peer reviewed. *Nota bene* has a wide thematic profile. In order to mirror the full breadth of our collections, publications, which include monographs, critical editions and edited collections, may be based on manuscripts, printed material, film, photography, music, broadcasting and digital media.

ISBN: 978-82-7965-587-9

National Library of Norway