AN ANALYSIS OF THE CURRENT BIBLIOGRAPHICAL DATA LANDSCAPE IN THE HUMANITIES

A Case for the Joint Bibliodata Agendas of Public Stakeholders

BIBLIOGRAPHICAL DATA WORKING GROUP

DARIAH-EU

About the report

This report has been prepared by the "Bibliographical Data" Working Group of the DARIAH-ERIC consortium, which develops public digital research infrastructure for the arts and humanities. The Group consists of more than 30 members from 15 countries, most of whom are researchers and curators in the public sector who are engaged in bibliographical data ("bibliodata") research and curation.

This report is aimed at all active stakeholders in the humanities bibliodata landscape, especially public sector entities who may benefit from the Group's insights and engage in cooperation to identify common interests, shape joint agendas, and achieve common goals. Those goals include creating shared infrastructure solutions, harmonising existing standards, and building partnerships to meet major challenges for contemporary bibliodata stakeholders.

The bibliodata landscape is a dynamic ecosystem including the many stakeholders who produce, process, and use diverse bibliographical resources (datasets, tools, services). Following the digital revolution, this landscape has been reconfigured and a critical era is now upon us that demands closer investigation. This report analyses the state of the art by defining current bibliodata (*Chapter 1*), mapping the contemporary landscape (*Chapter 2*), identifying crucial stakeholder challenges and opportunities (*Chapter 3*), and offering recommendations for future cooperation (*Chapter 4*).

This report presents an overview of issues in the bibliodata landscape. It is intended to provide a foundation for more detailed reports and case studies on the issues identified in this document.

BIBLIOGRAPHICAL DATA WORKING GROUP

DARIAH-EU

Report Authors:

Tomasz Umerle (lead author), ORCID: 000-0002-7335-0568
Institute of Literary Research, Polish Academy of Sciences, Poland

Giovanni Colavizza, ORCID: 0000-0002-9806-084X
Media Studies Department, Faculty of Humanities, University of Amsterdam, The Netherlands

Elżbieta Herden, ORCID: 0000-0001-5981-3725
Institute of Library and Information Science, University of Wrocław, Poland

Rindert Jagersma, ORCID: 0000-0002-5515-8273
Radboud University Nijmegen, The Netherlands

Péter Király, ORCID: 0000-0002-8749-4597
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Germany

Beata Koper, ORCID: 0000-0002-6166-7971
Institute of Literatures, Faculty of Philology, University of Opole, Poland

Leo Lahti, ORCID: 0000-0001-5537-637X
Department of Computing, University of Turku, Finland

David Lindemann, ORCID: 0000-0002-8261-6882
UPV/EHU University of the Basque Country

Jakub Maciej Łubocki, ORCID: 0000-0002-1957-0682
Department of Publishing Art, National Museum in Wrocław, Poland

Vojtěch Malínek, ORCID: 0000-0002-9553-5993
Institute of Czech Literature, Czech Academy of Sciences, Czech Republic

Alexandra Milanova, ORCID: 0000-0002-7532-5745
Institute of Balkan Studies & Centre of Thracology, Bulgarian Academy of Sciences, Bulgaria

Róbert Péter, ORCID: 0000-0002-7972-4751
Department of English Studies, University of Szeged, Hungary

Nanette Rißler-Pipka, ORCID: 0000-0002-0719-9003
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Germany

Matteo Romanello, ORCID: 0000-0002-7406-6286
Institute of Archeology and Classical Studies, University of Lausanne, Switzerland

Marcin Roszkowski, ORCID: 0000-0001-7396-4685
Faculty of Journalism, Information and Book Studies, University of Warsaw, Poland

Dorota Siwecka, ORCID: 0000-0003-1649-7579
Institute of Library and Information Science, University of Wrocław, Poland

Mikko Tolonen, ORCID: 0000-0003-2892-8911
Department of Digital Humanities, University of Helsinki, Finland

Ondřej Vimr, ORCID: 0000-0002-9364-0685
Institute of Czech Literature, Czech Academy of Sciences, Czech Republic

doi.org/10.5281/zenodo.6559857
May 26, 2022

Table of contents

Summary 5

1. WHAT IS BIBLIOGRAPHICAL DATA? 6
- 1.1. Producing bibliodata 6
- 1.2. Using bibliodata 7
- 1.3. Processing bibliodata 8
- 1.4. The need for a contemporary bibliodata landscape analysis 9
- 1.5. Bibliography 10

2. MAPPING THE CURRENT BIBLIODATA LANDSCAPE 11
- 2.1. Main categories of bibliodata stakeholders 11
- 2.2. Dimensions of the contemporary bibliodata landscape 15
- 2.2.1. Understanding the public/private dimension 16
- 2.2.1.1. The public-private division and interdependence 16
- 2.2.1.2. Divisions inside stakeholder groups 19
- 2.2.2. Understanding the production-use dimension 21
- 2.2.2.1. Continuous bibliodata reuse across different services 22
- 2.2.2.2. Reconfiguring traditional bibliodata production and use 24
- 2.2.2.3. Inefficiencies in the reuse ecosystem 25
- 2.3. Overarching trends in the bibliodata ecosystem: The data influx, automation, and data-driven research 28
- 2.4. Bibliography 30

3. KEY CHALLENGES AND OPPORTUNITIES FOR PUBLIC BIBLIODATA STAKEHOLDERS 34
- 3.1. Infrastructure 34
- 3.2. Open science 37
- 3.3. Data management lifecycle: Creation and documentation 39
- 3.4. Bibliography 42

4. CONCLUSIONS: TOWARDS JOINT AGENDAS FOR PUBLIC BIBLIODATA STAKEHOLDERS 43

Acronyms 46

Summary

The bibliodata landscape in the humanities has entered a critical phase of digital transformation characterised by a surge in the amount of data available, greater automation and data exchange capacities, and the increased ability to produce data-based knowledge. To highlight the current state of this landscape, this report presents an analysis organised around two dimensions: public vs. private sector involvement and production vs. use of bibliodata.

Different business and data-ownership models create a natural division between public and private sector. Despite this division, these two types of stakeholders are deeply interdependent. At the same time, they are subject to internal divisions: this includes the split between libraries, archives, and museums (LAM) and research institutions in the public sector and the competition between smaller companies and the consolidating forces of corporations in the private one.

A defining feature of the landscape is the continuous reuse of bibliodata. This has been coupled with the reconfiguring of traditional stakeholders' roles related to bibliodata production and use. However, these processes remain compromised by structural limitations including the replication of bibliodata production efforts and the lack of coordination between metadata aggregators and data providers.

This report identifies bibliodata infrastructure, open science, and data management as the most critical areas of concern for the future. In terms of infrastructure, standardisation, the efficient distribution of innovations, and the identification of new areas for investment are the most pressing requirements. Concerning open science, there is a need for widespread advocacy, outreach to smaller, under-resourced stakeholders, and a deeper understanding of what "openness" means in the bibliodata context. Finally, to support bibliodata management lifecycle stakeholders must develop new methods and sources for creating bibliodata while ensuring rigorous dataset documentation.

Those challenges and opportunities call for closer cooperation, especially among public entities. This report contends that this cooperation would benefit from increased focus on the uniqueness of bibliodata combined with advocacy and education. Stakeholders should take full advantage of the open science movement, which has already proven advantageous for public actors. Finally, cooperation among diverse stakeholders is especially important with the emphasis on innovation sharing in both bibliodata curation and research.

1. WHAT IS BIBLIOGRAPHICAL DATA?

Bibliographical data are structured information about the form, content, and context of documents in any form (textual, graphic, musical notation, etc.) or medium (printed, electronic, etc.).

Bibliographical data are key tools for describing and discovering information resources and cultural objects. The main goal of these tools is to connect users with resources that fulfill their information needs.

The first bibliographical descriptions were limited to basic information about individual works including the author's name and publication title. The scope of these entries gradually increased, however, particularly following the invention of the printing press, which brought with it the need to record the work's physical description, place of publication, printer's name, and format. Later the digital transformation introduced computer-assisted data processing and new types of resources (digital publishing). Today bibliodata take diverse forms that range from card catalogues and printed bibliographies to the bibliographical records in Online Public Access Catalogues (OPACs)[1], digital libraries, citations, references, and Resource Description Framework (RDF) statements in Linked Data databases.

This report applies a **broad definition of bibliodata**, which are understood as "all the data elements necessary for a full description, presented in a specific bibliographic format."[2] Therefore, items such as controlled vocabularies, authority files, subject headings, ontologies, thesauri, (persistent) identifiers are also viewed as bibliodata if they are used to present information in the form of a bibliographical record.

1.1. Producing bibliodata

Historically, the creation of bibliodata was in the hands of those who owned physical collections or could access them to further disseminate information. This included institutions like libraries, archives, and museums ("LAM institutions"), individuals such as researchers and curators (e.g. bibliographers, documentalists) as well as publishers and booksellers. The bibliodata produced by these entities established library catalogues, bibliographies, and booksellers' catalogues as important records of cultural and scientific production.

[1] Where possible, the acronyms and initialisms used in this report are defined in the text. For a full list, please see the Acronyms section on page 46.

[2] Reitz, J. M. (n.d). Bibliographic record. In ODLIS Online Dictionary of Library and Information Science. Retrieved December 7, 2021, from www.products.abc-clio.com/ODLIS/odlis_b.aspx

Today societies produce bibliodata in a systematic way to organise, monitor, understand, and provide access to cultural and scientific outputs. The digital transformation[3] has also introduced **computer-assisted data production and automated bibliodata processing,** which have caused an unprecedented **surge in the volume of bibliographical data**. These developments have led, in turn, to modern information systems, new data production services (e.g. union catalogues, metadata aggregators), new stakeholders (e.g. private tech companies, online users), and new cooperation models between traditional bibliodata producers (e.g. cooperative cataloguing, Cataloguing-in-Publication programmes). Lastly, users of these modern information systems have been empowered to create their own bibliodata thanks to the growing accessibility of tools for creating, processing, and storing metadata. Those tools include reference managers and various kinds of social indexing such as social tagging, transcribing, and annotating.

1.2. Using bibliodata

Users typically rely on bibliodata to find out about the existence of the publications that they represent and access content to meet their personal, professional, educational, or research needs.

Historically, bibliographical data primarily took the form of card catalogues that recorded the physical existence and location of publications. Printed book-length bibliographies and other listings were created mainly to organise information about publications on a specific topic, person, or region.

Digital transformation has made it **easier to encounter and use bibliodata**. Today these data may be found in digital library catalogues, Wikipedia, university repositories and in popular documents like journalistic and web articles. A student or researcher may, for example, browse a library catalogue or digital repository, and use the bibliodata retrieved there to access and/or cite a resource. If they wish, they may also process that bibliodata further to create their own materials including reference lists and bibliographical collections in reference software. This can all be done with services and tools including catalogues, repositories, and databases as well as indexes, reference management software, social bookmarking services, and text editors.

[3] This report understands digital transformation as a process that began with the computerisation of bibliographical resources. At the core of this process was the digitisation of library catalogues, which started in the 1970s and continues to this day.

Bibliodata have always been subject to processes of reuse, and for many traditional data producers, the curation of their resources has relied heavily on the work of others. Since digital transformation, **bibliodata reuse has accelerated**. Bibliodata are, thus, in constant motion, and new datasets are continuously being created from older collections, which they enrich in turn. In this regard, all institutional bibliodata producers – including LAM institutions, publishers, and information services – are in fact also bibliodata users.

Finally, bibliodata are the focus of research. Bibliographical research has a long history that is intertwined with fields such as book history, documentation studies, and information science. Digital technologies have enabled a wider use of bibliodata in data-driven research including quantitative studies (e.g. bibliometrics, cultural analytics, statistical analysis). Bibliodata-based research has also been driven by **growing capacities to produce data-based knowledge** and the rising demand for such information.

1.3. Processing bibliodata

Historically bibliodata were created manually and to some extent this remains the case today. Some of this work was carried out, for example, by cataloguers with documents in hand who analysed the items and created their descriptions. Similarly, researchers performed descriptive, analytical, and enumerative work to create and compile bibliographical information.

Currently this work is significantly assisted by computers including creation, exchange, and data storage software. Bibliographical descriptions are digital units of information, represented in digital formats and circulating in digital data ecosystems.

Given this computer-assisted data production and usage, the standardisation of digital bibliodata processing has been a central concern in recent decades. To this end, bibliographical description standards have been adapted and/or developed, including international cataloguing standards (e.g. ISBD, RDA) and metadata models and standards (e.g. MARC and its variants, MODS, Dublin Core). As the scope of bibliodata has expanded, new areas have required unification. These include standards for citation representation (e.g. ISO 690:2021) and exchange (e.g. RIS, BibTex) and formats utilised by particular stakeholders such as publishers (e.g. ONIX).

While these standardisation efforts have improved data quality, they have not culminated in any single data format, model, or standard. The current landscape is, however, characterised by **diverse initiatives to increase the interoperability of services,**

tools, formats, and standards. In this context, international authority files, systems of persistent identifiers, and semantic web and Linked Data technologies (i.e. RDF models like BIBO and BIBFRAME) are becoming increasingly important.

1.4. The need for a contemporary bibliodata landscape analysis

In recent years, the bibliodata landscape has entered a critical phase of digital transformation. This era is characterised by the **growing amount of data available, greater automation and data exchange capacities, and the increased ability to produce data-based knowledge**.

These overarching trends have also produced a number of **tensions in the bibliodata landscape** that now call for closer analysis. First of all, there is a tension driven by the **need to adapt and change traditional bibliodata curation approaches.** Historical models have been tested by the transfer of older bibliodata descriptions to newer standards and platforms as well as by the pressures of curation for a growing but sometimes elusive user base. Second, there is a **tension between open and closed data exchange ecosystems**, which has also exposed conflicts between the solutions of stakeholders from different (public/private) sectors and different business models. Third, there is a **tension based on the gap between the quality of data currently available** and the **growing expectations of stakeholders** engaged in data-based knowledge and research. For many of these entities, older means of information retrieval are simply no longer enough.

Since this report aims to analyse the bibliodata landscape from the perspective of the humanities community, we need to address the **specificities of the humanities environment.** In particular, we must examine two large groups of stakeholders of critical importance to the humanities: LAM and the research sector. A range of varied stakeholders representing these two groups, including research infrastructures, publishers, information services, libraries, and archives are all equally important to humanities research, and the resources they maintain are viewed as complementary.

To better understand these tensions in the current bibliodata landscape, in the next chapter we provide a landscape analysis that focuses on bibliodata stakeholders in the humanities and their relationships and prevailing trends in the field. This leads us to define the most important challenges and opportunities for bibliodata stakeholders in the following chapter. Finally, we provide recommendations for the future joint agendas of these key public actors in order to meet the challenges of the current bibliographical data landscape.

1.5. Bibliography

1. Bowman, J. H. (2007). Retrospective conversion: The early years. *Library History, 23*(:4), 331–340. doi.org/10.1179/174581607x254811
2. Fallgren, N. J. (2007, February 25). *Users and uses of bibliographic data. Background paper for the Working Group on the Future of Bibliographic Control*. www.loc.gov/bibliographic--future/meetings/docs/UsersandUsesBackgroundPaper.pdf
3. IFLA. (2021). *IFLA Professional Statement on Universal Bibliographic Control*. www.ifla.org/wp-content/uploads/2019/05/assets/bibliography/Documents/ifla-professional-statement-on-ubc-en.pdf
4. Reitz, J. M. (n.d). Bibliographic record. In *ODLIS Online Dictionary of Library and Information Science*. Retrieved December 7, 2021, from www.products.abc-clio.com/ODLIS/odlis_b.aspx
5. Society of American Archivists. (n.d.). Bibliographic description. In *Dictionary of Archives Terminology*. Retrieved December 7, 2021, from www.dictionary.archivists.org/entry/bibliographic-description.html

2. MAPPING THE CURRENT BIBLIODATA LANDSCAPE

In this chapter, we map the current bibliodata landscape in order to gain a better understanding of its active stakeholders, their roles, and significant trends and dynamics in the field. For this purpose, we also provide our own *Map of the Bibliodata Landscape* (*Figure 1*), which is described and analysed in subsequent sections of this report.

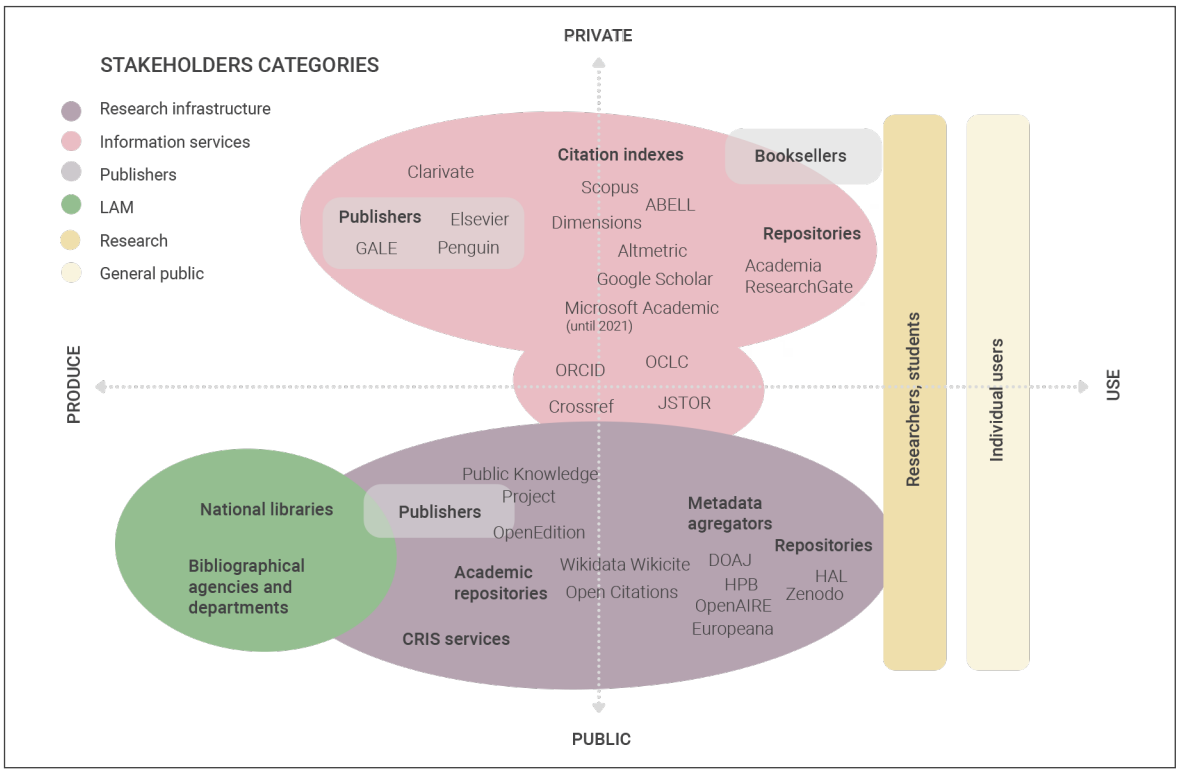
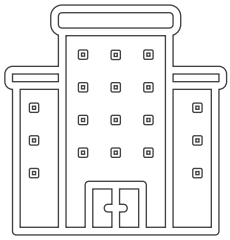


Figure 1: Map of the bibliodata landscape
Stakeholders are positioned along two axes: Production <–> Use (horizontal axis) and Private <–> Public (vertical axis). Neither the relative size of a stakeholder bubble nor its position (centre vs. periphery) reflects its importance or centrality. The map shows the position of specific stakeholders vis-à-vis the two defined axes and the vast range of other participants in the field.

2.1. Main categories of bibliodata stakeholders[4]

In line with *Figure 1*, below we identify the main categories of bibliodata stakeholders in this landscape. Each stakeholder group brings together entities that have similar interests and patterns of activity and/or are similarly impacted by developments in this environment:

[4] We define a stakeholder as any entity that plays a role in the bibliodata ecosystem whether this is as a legally recognised organisation or an initiative, service, or some other kind of a non-formally incorporated entity. We also highlight six basic stakeholder categories (in colour) and different types of stakeholder institutions, initiatives, and service providers (in bold; e.g. publishers, national libraries, citation indexes). Finally we provide examples of these undertakings (e.g. GALE, Scopus).



Libraries, archives, and museums (LAM) are typically public institutions which both produce and use bibliodata. Among this group, libraries are the key institutions in this ecosystem. Their activities have both production and use elements and encompass cataloguing, digitisation, data processing, and preservation as well as information provision, access, enrichment, research and development, and education. Many libraries supplement the active (manual) processing of new acquisitions with the reuse of machine-readable metadata, which may be provided, for example, by different libraries through shared cataloguing or else created by publishers or retrieved from authority files. Other LAM institutions have active plans to migrate to automated text mining-driven solutions for bibliographical description. LAM institutions face technological challenges as they transition from traditional cataloguing to (semi-)automated methods. Many untapped opportunities remain for the management and use of their digitised collections.

Research and academic libraries combine some features of LAM with those of Research Infrastructures (RI). They may, for example, manage both a library catalogue and a scientific publishing service such as an Open Journal System (OJS) or an academic repository.

In *Figure 1*, LAM are represented by national libraries (e.g. the National Library of Finland) and bibliographical agencies and departments (e.g. the cataloguing and bibliographic departments of different LAM institutions).

Research infrastructures (RIs) are institutions, resources, and related services that are used by members of a scientific community to conduct research in their field. In this report, we rely on the definitions of RIs proposed by the European Commission and European Strategy Forum on Research Infrastructure (ESFRI), which closely tie this category to the public sector. There may, however, be some overlap between RIs and private information services (see below). RIs include research consortia, research networks, and organisations that produce and organise shared-access infrastructure for the public. Such infrastructure may relate, for example, to scientific equipment, tools, and knowledge-based resources[5].

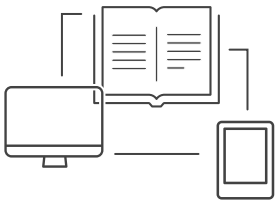
[5] For more information about this infrastructure, see ESFRI. (n.d.). Research Infrastructure (RI). In *Glossary*. Retrieved December 7, 2021 from www.esfri.eu/glossary.

RIs produce and use bibliodata to support researchers in their activities, but they also have responsibilities to the public. Current RIs have reported a surge of interest in digital research and related community-driven initiatives. However, these same institutions face systemic problems, especially in the humanities. Among the challenges the most prominent are persistent LAM – RI divisions, which some consider artificial and outdated, and the tensions between RIs and private (commercial) information services.

In *Figure 1*, RIs are represented by institutional and public repositories such as HAL and Zenodo for research output, Current Research Information Systems (CRIS), and metadata aggregators such as OpenAIRE and Europeana. We situate the publishing platforms used for scientific research (e.g. OpenEdition) and Public Knowledge Project services on the boundary between RIs and publishers.

Information services are typically commercial service providers, but in some cases mixed business models apply. While all these services produce and use bibliodata, their activities centre on the reuse of data produced by researchers, RIs, or LAM stakeholders. Information services offer targeted support to research stakeholders and LAMs including software to manage, publish, and disseminate data and services to enrich, analyse, and reuse datasets. The rapid rise of these commercial information services has provided public entities with new opportunities for R&D collaborations. However in critical fields such as scientific publishing and scientific data reuse, it is hard to reconcile the different approaches to open science in the private and public sectors.

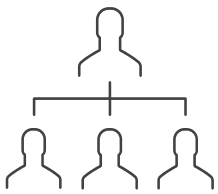
In *Figure 1*, this category is represented by corporations such as Clarivate and Elsevier that target both researchers and LAMs with various services. Other examples include citation indexes and academic databases, which may be general-purpose (e.g. Scopus, Dimensions, Altmetric) or discipline-based as in subject-specific bibliographies (e.g. the Annual Bibliography of English Language and Literature). The research services of big tech companies (e.g. Google Scholar and Microsoft Academic) also offer different (non-subscription-based) access modes. Finally commercial repositories appear on the far right-hand side of the entry for this group; Academia.edu and ResearchGate are key examples.



Publishers, which may be either commercial or public entities, qualify as both producers and users of bibliodata. These stakeholders include publishing houses, publishing platforms (e.g. journal publishing services), and others involved in publishing activities. While publishers play an important role in producing the raw bibliodata attached to their publications, they often rely on authors to supply them with information **and/or create** recommended bibliographical records with national cataloguing agencies. Publishers have a wide range of data exchange systems, which they use to cooperate with LAM institutions (e.g. ISSN, ISBN systems, cataloguing-in-publication systems), booksellers (e.g. the ONIX format), RIs, and information services.

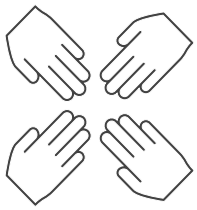
In *Figure 1*, this category is represented by public entities like institutional publishing houses on the one hand, and commercial publishers such as GALE and Penguin on the other.

Booksellers are a distinct group that is closely related to publishers. These stakeholders generally reuse and enrich bibliodata for trade, marketing, and e-commerce purposes. Most bibliodata that originate directly from booksellers are created by antiquarian booksellers or the organisers of book auctions who need to describe rare and unique books of financial or other value to collectors.



Researchers and students are another unique category of public sector stakeholders in bibliodata. This category includes various researchers, scholars, and university students as well as research groups and institutions. These entities use bibliodata mainly to discover information, however there has also been an upswing in bibliodata research[6] in the humanities. Current bibliodata researchers also engage in diverse forms of data production and reuse. This includes creating and enriching data for future research based on the high-level (pre-)processing of datasets. As such, research stakeholders may serve at once as data producers and data users – two identities that are often hard to separate.

[6] We understand bibliodata research to be any research which primarily concerns bibliographical data. This includes scholarly work within the fields of book history, bibliographic data science, bibliometrics, cultural analysis, library and information studies (LIS), and documentation studies.



Individual users/members of the general public may be found in either the public or the private domains where they chiefly interact with bibliodata as end users. Typically, these individuals encounter bibliodata when browsing online or in library catalogues. Users of online bibliodata services form part of this group, but it also extends to any user who interacts with bibliographical data and is not affiliated formally with any scientific community. All public bibliodata stakeholders should provide information retrieval services to the general public. This includes not only ensuring traditional services such as OPACs, but also allowing these users to take advantage of the knowledge obtained through bibliodata analysis.

2.2. Dimensions of the contemporary bibliodata landscape

In this survey, we focus on **two key dimensions of the current bibliodata landscape:**

• **Public vs. private:** This dimension relates to the legal status of stakeholders and initiatives, i.e. whether they are private or public entities. At the same time, it identifies the entity's source of funding, which may be full or partial, and the type of business model it represents, i.e. commercial, non-profit, foundation, and so on. Lastly, it highlights the data policy applicable to the stakeholder including whether access to the bibliodata it produces is limited or open.

• **Production vs. use:** This dimension concerns the actual activities that stakeholders and initiatives perform with bibliodata. At one end of the spectrum, we find the production of new bibliodata while further along there are processing and enrichment activities and finally direct use or reuse, for example, through the public interfaces of online services.

Our focus on these dimensions represents only one approach to organising this ecosystem. Nevertheless it allows us to highlight two important aspects of the landscape: the relationship between different types of entities (i.e. private vs. public actors) and the lifecycle of data (i.e. from production through to use). While this framework and visualisation cannot provide an exhaustive analysis, we believe that they offer a representative sketch of the entire bibliodata ecosystem.

2.2.1. Understanding the public/private dimension

On the public side of *Figure 1*, we find publicly funded institutions that provide public services, often in open-access mode. These include national libraries, academic repositories, and publicly funded metadata aggregators. On the private side, in contrast, there are private for-profit organisations that collect, process, and/or (re)sell data. Key examples of these entities include citation index service providers and publishing houses. Between these two poles, we locate institutions with mixed business models or semi-open systems (e.g. non-profit organisations such as ORCID and CrossRef).

On the far right-hand of the map, we also find bibliodata users who cover the entire spectrum of public- and private-sector activities. They include researchers, students, and individual users/members of the general public who rely on bibliodata services to meet various needs. Such individuals may, for example, be part of private sector institutions or research or LAM organisations.

2.2.1.1. The public-private division and interdependence

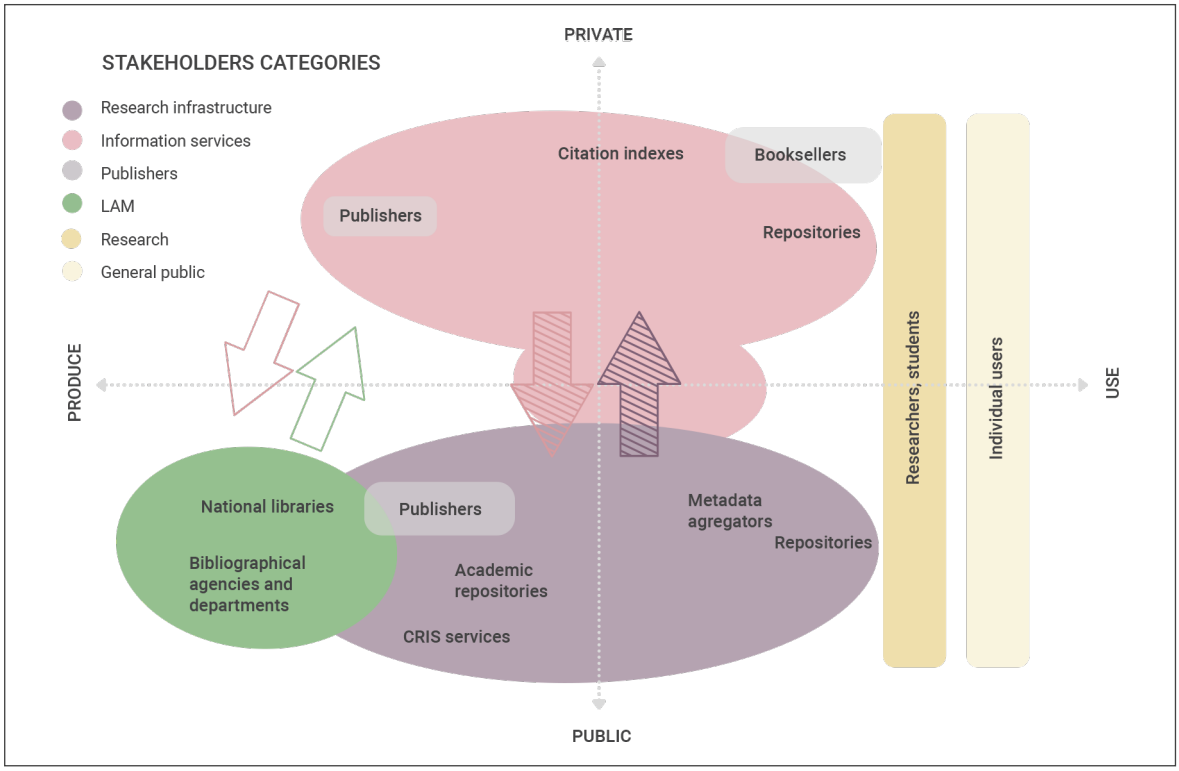


Figure 2: Public/private division and interdependence

Different funding, business, and data-ownership models create a natural division between public and private sector stakeholders. This divide shapes the bibliodata landscape.

Public stakeholders respond to the needs of their main end users such as researchers, students, and members of the public while also observing various duties to other public entities. Those duties are often set out in public policies or statutory or other legal instruments such as Library Acts, Findability, Accessibility, Interoperability, and Reusability (FAIR) principles, and open-access policies. The data evaluation ecosystem is an illuminating example. This framework governs data exchanges among researchers, publishers, CRIS systems, and national research evaluation databases in order to provide data-based knowledge and other metrics that allow for large-scale financial decision-making. Similarly, national data cataloguing ecosystems reflect the obligations of some institutions (e.g. national libraries) to maintain bibliographical control over publications in their country.

Private stakeholders, on the other hand, respond to market needs and opportunities by offering their services to various stakeholders that may be public or private. At present, these entities face competition from many other stakeholders, including other commercial operators since targeting public sector needs is the business model of many companies whose products address similar demands. These commercial players include repositories and networking services such as Researchgate and Academia.edu and metrics services like Altmetric.com and Dimensions. Large corporations clearly also have a growing influence on information services in the bibliodata landscape. Clarivate (which in 2021 acquired ProQuest) and Elsevier are both notable examples.

Despite this division, **private and public stakeholders are deeply interdependent**. For their part, public stakeholders rely heavily on the information services of commercial and privately-owned stakeholders. Over time, universities and libraries have also come to depend on private sector citation indexes and service providers for CRIS and Integrated Library System (ILS) solutions. This reliance on private entities may be attributed partly to the lack of systemic support for the public management of large continuously updated data sources with vast coverage. At the same time, the shift clearly relates to the innovative products and services of private operators, which tend to be more flexible in their decision-making and investments and less constrained by long-term statutory and other legal obligations.

Conversely, private stakeholders often depend on their public counterparts. Many privately-owned services are based on the commercial reuse of – or mediation of access to – data produced by public stakeholders. This is the case, for example, for publishing and discovery services (e.g. EBSCO and ProQuest), citation indexes (e.g. Scopus), and repositories and social networking services (e.g. Academia.edu) as well as analytical solutions like Dimensions and Clarivate's analytics portfolio. Similarly, privately-owned academic publishers depend entirely on the involvement of public institutions, while commercial citation indexes rely on research and data from the research community.

One consequence of this public-private interdependence is significant public funds are paid to private stakeholders in publication costs and subscription fees. The reliance on commercial solutions is not only financially taxing but it has effectively restricted access to knowledge produced with public resources. This issue has been at the centre of debates about open science principles, and it has been one cause for the introduction of open access policies. The same debates have also led to initiatives to highlight gaps in the accessibility of bibliodata and propose solutions. Some examples include the Open Knowledge Foundation's Open Bibliographic Data Working Group and more recently OpenCitations, Open Research Data, and The Initiative for OpenAbstracts.

In addition, these tensions around public–private relations have had at least two other major effects on the bibliodata landscape. First, public entities tend to provide, develop and maintain their own targeted services (e.g. Open Journal System, OpenCitations Index, HAL, Zenodo, GOTRIPLE) along with more general tools such as public metadata aggregators to facilitate open knowledge exchange (e.g. OpenAIRE, Europeana). As a result we can see a number of similar services of public and private nature existing in the current ecosystem. It has also culminated in the building of complex platforms where public stakeholders can integrate their own independently developed solutions. The incorporating of the public OpenCitations index into OpenAIRE Nexus, for example, embeds an open bibliodata metrics service in a public aggregator platform.

Second, the **landscape now hosts "intermediary" stakeholders which occupy a space between the private and public domains**. One function of these entities is to improve the interoperability of bibliodata between private and public sources, for example, by generating persistent identifiers (PIDs: ORCID, DOI) and authority controls (OCLC's

Virtual International Authority File). The Online Computer Library Center is a multi--faceted initiative that supports libraries with metadata aggregation (e.g. Worldcat) and production (e.g. OCLC's cataloguing services). Crossref also plays a crucial role by aggregating and enriching research output and facilitating open access through its discovery services (search engine, Application Programming Interfaces (APIs)).

2.2.1.2. Divisions inside stakeholder groups

While **public stakeholders** in the bibliodata space may share some traits and obligations, they are by no means unified. One crucial distinction between these entities is that some are research-oriented entities (RIs, universities, institutes, etc.) while others are LAM institutions.

This opposition is reflected not only in the most obvious differences between the groups' objectives but also in the main focus of their work (i.e. the types of documents they deal with) and their legal status. It can also be seen from the types of services which are essential to each group's operations.

For research stakeholders, the main purpose of bibliodata services is to enable the publishing, dissemination, and evaluation of research output. Key activities may, thus, include managing CRIS systems and academic repositories or providing data to citation indexes and major metadata aggregators. LAM services, on the other hand, focus on managing and preserving cultural heritage and making institutional physical holdings accessible. To this end, these institutions produce public interfaces for the discovery of these holdings such as library catalogues, ILS, and digital libraries.[7] What is noteworthy, however, is that manual cataloguing (bibliodata production) remains an active practice in LAM institutions while it is more decentralised in research institutions. Indeed, researchers often produce metadata that are then processed and enriched by other parties. The data exchange ecosystems of the two groups, thus, also differ: while Europeana is the lead aggregator for LAM-related information, OpenAIRE focuses on research outputs.

Technology is not, however, the crux of this division, and private stakeholders with enough resources may build business models that target both research entities and LAMs. This has been true, for example, of ProQuest, whose portfolios include both

[7] Significantly, many public, research, and academic libraries have begun to provide their users with access not only to physical holdings but also to electronic items. However the metadata of those electronic publications are mostly accessible via commercial platforms and seldom found in standard library catalogues.

research-oriented pro ducts (e.g. discovery systems for research output) and library--focused solutions (e.g. Alma, and Ex Libris products now owned by ProQuest). Most crucially, the research-LAM divide remains a pressing concern in the humanities where there have been efforts to bridge the gaps between communities and infrastructures. All this stems from the need for humanities researchers to have access to unified and harmonised resources that draw on both LAMs and research institutions.

Private stakeholders clearly compete with one another, but there are important divisions between the "larger" (e.g. Clarivate, Springer, Elsevier, Wiley) and "smaller" (e.g. Academic Analytics, Semantic Scholar) players in the information services and scholarly publishing spaces. In recent years, this has led to a well-documented process of private sector consolidation, driven by a series of acquisitions. Clarivate's takeover of Proquest is just one recent example of this trend and it follows Proquest's own acquisition of Ex Libris. These changes have placed an unprecedented amount of power in the hands of a limited number of private stakeholders that control critical infrastructure for research and cultural heritage institutions. In addition, these pressures have prevented smaller private companies – including data analytics-related entities – from becoming or staying competitive and developing more balanced partnerships with the public stakeholders that produce scientific and cultural data.

Nevertheless, outside of this consolidation trend, alternative business models are also emerging around the open source development of library software for bibliodata production and dissemination. EBSCO's investments in the FOLIO open source system are one key example. These kinds of solutions – i.e. partnerships between public institutions and private information services to develop and provide IT support for open source systems – represent competition not only for companies like ProQuest, but also for non-open-source offerings such as OCLC's WorldShare Management Services.

Last but not least, all private operators in the research and cultural sectors face competition from big tech companies such as Microsoft and Google, which now deliver services in those spaces. These discovery services for academic publications – Microsoft Academic (MA) and Google Scholar – enjoy a solid market position and follow a different business model. While Scopus and the Clarivate's Web of Science

are subscription-based, Microsoft and Google are free to the public. However, although big tech has clearly developed some innovative solutions – key examples include MA's machine learning and natural language processing (NLP) solutions for concept detection and improved information retrieval – the sustainability of these services remains questionable. Microsoft decided to close down MA in 2021. The platform's licensing status has allowed the open science community to take over the project (OpenAlex project).

2.2.2. Understanding the production-use dimension

On the production (left-hand) side of *Figure 1*, we find services whose primary mission is the creation of bibliodata. These include national bibliography centres (e.g. Fennica), publishers, and information services which publish original works and their metadata. Various academic publishers and repositories and commercial publishers also fall in this category. Proceeding to the middle of the spectrum, we find services related to bibliodata processing and enrichment tools such as persistent identifier infrastructures (e.g. Crossref, ORCID), bibliodata publishing and linking infrastructures (e.g. OpenCitations, Open Abstracts, Wikidata/Wikicite projects), citation indexes, and metrics (e.g. Scopus, Dimensions). Further to the right we have metadata aggregators and union catalogues (e.g. OCLC, OpenAIRE, Europeana, DOAJ) that aim to harmonise and enhance externally sourced bibliodata to facilitate access to digital objects or physical holdings. Even further in that direction, there are services for the publishing of digital documents (e.g. full-text repositories like Zenodo) where bibliodata reuse and production are features offered to users and the quantity and quality of data depend largely on those users' preferences.

On the use side (i.e. the far right of the continuum), the key stakeholders are researchers, students, and personal users. These are, in other words, the entities who are the target of most other services. However users are also playing an increasingly active role in the curation of bibliodata. Moreover the research community includes a group of bibliodata researchers who work on and with bibliodata. These scholars' interactions with bibliodata extend beyond use for discovery or citation purposes.

2.2.2.1. Continuous bibliodata reuse across different services

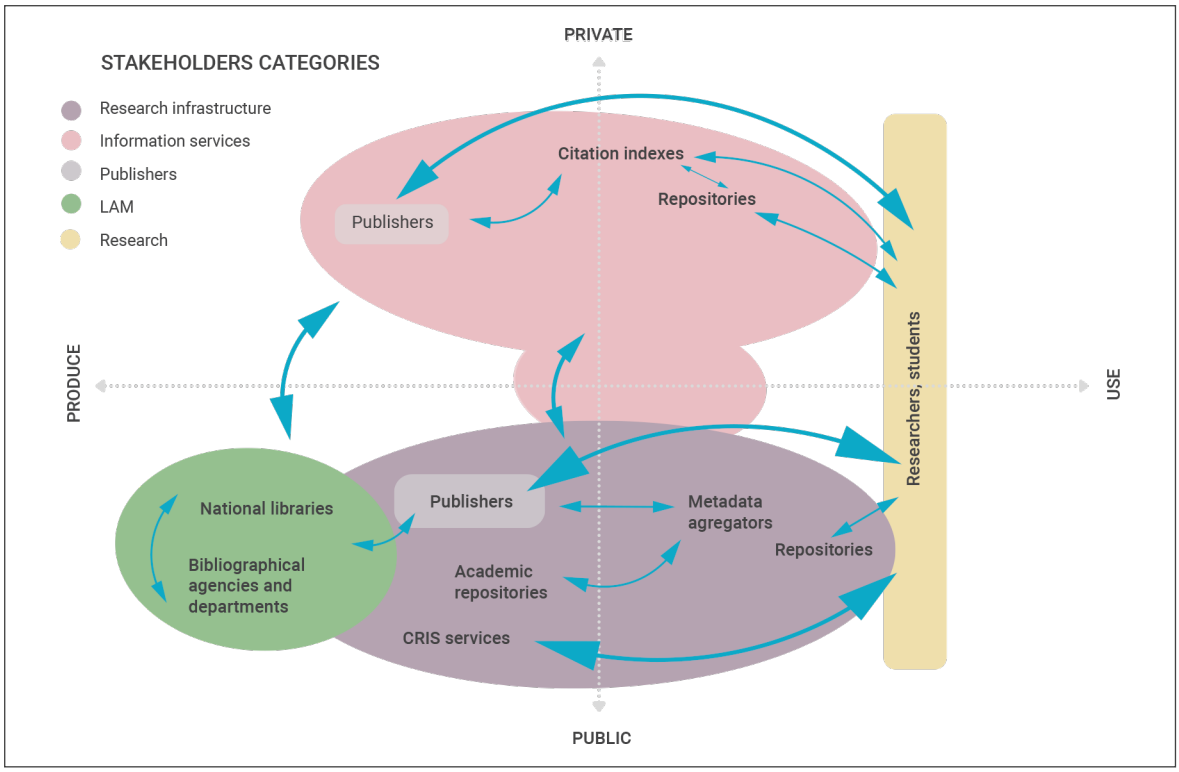


Figure 3: Continuous bibliodata reuse across different services

A defining feature of the bibliodata landscape is the **continuous bibliodata reuse across services**. In the public sector, this reuse is driven by the richness, complexity, and connectivity of the bibliodata environment and its diverse services, stakeholders, and user needs. Public policy is also a key influence, and we may note, for example, the impact of FAIR, open data, open science, and research evaluation principles.

To reuse bibliodata and meet their obligations, public stakeholders often depend on solutions from the private sector. This clearly makes for an intricate data ecosystem, and complex examples of reuse abound. Researchers may, for example, provide the metadata of their publications to different repositories, publishers, and CRIS systems, which then process, aggregate, and disseminate them. Information services like citation indexes rely on metadata from researchers and publishers while themselves also producing and enriching metadata; their output forms the basis for the bibliometrics required for scientific evaluation and further research. National bibliography centres and other public bibliodata producers use publishers' data for cataloguing (e.g. from e-ISBN services) but also provide targeted services to publishers (e.g. Cataloguing-in-Publication programmes). Similarly public research infrastructures – for example, metadata aggregators like OpenAIRE and Europeana – use externally generated

metadata, which they then normalise, enhance, and provide for further reuse and aggregation.

Libraries too often rely on union cataloguing or (re-)use the records provided by national bibliographic agencies or other institutions. They may, however, adapt those records to local standards and enrich data at any point.

As can be seen, **bibliodata production involves many stages of reconfiguration and enhancement** of the bibliodata descriptions distributed among different services. Clearly, this increased reuse of bibliodata is driving changes in the current bibliodata landscape. Among the effects are the acceleration of interoperability within the field, the creation of more spaces to develop new projects and tools, and the fostering of new data-driven methods in the humanities. In fact, contemporary bibliodata can be understood as a **collaborative effort**, a status shown by the use of shared information "entities" such as persistent identifiers, controlled vocabularies, authority controls, and collaborative cataloguing in libraries. Another key sign of such collaboration is the data enrichment, performed, for example, by metadata aggregators.

Even so, **it is becoming increasingly hard to understand and assess the contributions of different stakeholders at different stages of the data production process**. This is a concern since insufficient monitoring and scrutiny could lead to limits on data transmission comparable to commercial academic discovery systems' restrictions of access to citations and abstracts. These issues have been addressed retroactively by open science initiatives such as OpenCitations and Open Abstracts.

2.2.2.2. Reconfiguring traditional bibliodata production and use

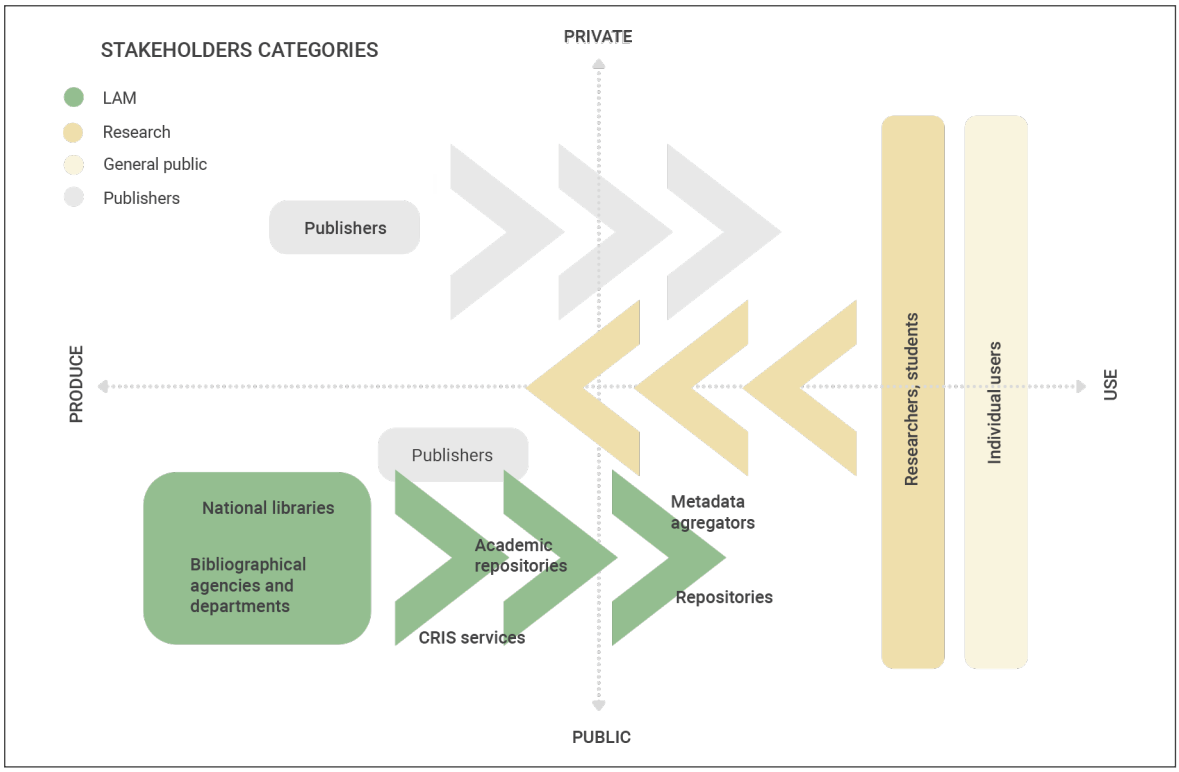


Figure 4: Reconfiguring traditional bibliodata production and use

This continuous bibliodata reuse has led not only to the emergence of more private and non-profit stakeholders, but also to a **reconfiguring of traditional bibliodata production** patterns. This especially concerns public bibliodata producers such as national and academic libraries, bibliographical agencies, and cataloguing departments. In these institutions, we may observe a **tension between traditional curatorial methods and more innovative data exchange, data mining, and research-oriented approaches**.

The more traditional approaches – which remain essential – refer to the good data production and storage practices that support the local discoverability of curators' holdings, for example, systematic cataloguing and maintaining public interfaces for local collections. In contrast, more research-oriented methods tend to view metadata as a source of cultural information which should be analysed and interpreted via bibliometrics, statistics, and other data-driven research methods. Innovative approaches focus on technological advances that support automated bibliodata production and international interoperability. Key tools here include AI, NLP, Linked Data and Semantic Web.

These trends can be seen at an organisational level, with the appearance of new departments, workflows, and job descriptions at public institutions, and at the level of new bibliodata production and processing methods. As a result, some public bibliodata producers are **expanding their curatorial practices beyond basic data production and local discovery system maintenance**. As such, they are focusing increasingly on re-using existing data while also relying on and contributing to shared knowledge through internationally recognised PIDs, authorities, and controlled vocabularies. These entities support the reuse of their own collections for purposes beyond basic data discovery. With this goal, they are drawing on the latest innovative technologies from the research and development sector.

At the same time, **users – researchers, students, and members of the general public – who traditionally approached bibliodata as a discovery tool are increasingly becoming bibliodata producers**. This is exemplified, for example, by providing metadata for the documents they submit to full-text repositories including institutional or community-driven services such as Zenodo and HAL or commercial repositories like Researchgate and Academia.edu. Others are actively building, sharing, and using bibliographical collections via information management software (e.g. Zotero, Mendeley, JabRef) or engaging in citizen science (e.g. by editing Wikipedia contents).

Finally, through their work in disciplines like **bibliometrics, cultural analytics, and book history**, bibliodata researchers are engaging in bibliodata curatorial practices including dataset creation, normalisation, and enrichment. This work enables them to use existing bibliographical datasets or produce new ones. These practices bring bibliodata researchers increasingly close to bibliodata curators since a sound knowledge of data curation processes is needed to perform high-quality research.

In sum, we are now witnessing the **converging workflows** of different stakeholders who are involved in bibliodata production and (the facilitation of) bibliodata reuse through data normalisation, enrichment, and analysis.

2.2.2.3. Inefficiencies in the reuse ecosystem

Despite this rise in data reuse in the bibliodata landscape, the process is beset by **structural limitations and glaring inefficiencies**.

The most important limitation stems from the **tension between private (commercial) services and sometimes non-profit organisations** on the one hand **and public entities** on the other. **Whereas the former may limit data access, the latter observe open science principles**. This affects both "raw" data such as citations, abstracts, and subject descriptions/keywords, and analytics like metrics, altmetrics, and indexes based on "raw" data.

Major inefficiencies in the system also arise from the **replication of bibliodata production efforts**. Under the current system, bibliodata for the same document may be produced entirely independently by different stakeholders who cannot reuse existing bibliodata transparently and efficiently in order to alleviate their workload. There is a long history of attempts to reduce such replication in both the library and publishing sectors. This includes the cooperative cataloguing tradition and Cataloguing-in-Publication programmes. However, current bibliodata output is far more dispersed, and with the rise of digital publishing, especially for research, and the surge in data reuse, a comprehensive system of cooperative data production is increasingly elusive.

The problem of **independent production of bibliodata for the same document** is exemplified by the creation of separate bibliographical descriptions by library catalogues and during digitisation processes. This is especially likely when cataloguing and digitisation happen at different institutions. The same scenario may arise if a journal article author and editorial board co-create a bibliographical record for publishing purposes, and library catalogues produce a separate record for a national bibliography.

At present, inefficient bibliodata reuse is widespread and reflects concerns about data quality and the lack of technological capacities to take full advantage of accessible data. This can best be illustrated by the relationship between powerful aggregators (e.g. OpenAIRE, Europeana OCLC's Worldcat) and aggregated services. These aggregators can improve the quality of aggregated data by, for example, attaching persistent identifiers and linking multiple descriptions of the same document for deduplication purposes. However, there are no systemic solutions to ensure that data providers will adopt those improvements. Such adoption might lead to more efficient workflows as original data improves and aggregators can reallocate their resources. This problem has, of course, been targeted through

the creation of dashboards for data providers such as the OpenAIRE-Provide. However, providers' ability to exploit metadata enrichment remains dependent on the interoperability of their services, know-how, and available resources.

In this context, metadata reuse faces serious limitations. Local subject classifications, for example, are often not interoperable owing to the lack of authority controls and linked data connections or because of language issues. Reference/citation mining poses similar problems since many aggregated humanities resources, especially historical texts, have not been enriched by reference/citation data. In other words, there is a complex pattern of inefficiency across the aggregation process and a lack of systemic coordination to resolve data quality issues.

Finally, the scope of bibliodata aggregation is not adequate in the humanities. A case in point is the failure to aggregate and harmonise the bibliodata produced by national libraries. This gap is particularly glaring given the development of public services like Europeana and OpenAIRE, which aggregate the bibliodata of digitised items. Moreover it speaks to the larger failure to sustain the larger European Library project. Today OCLC's Wordcat is the main service used to access the currently collected contents of library catalogues. There have, however, been persistent public efforts to harmonise the bibliodata created by early modern curators, including initiatives such as the Heritage of the Printed Book Database and the Universal Short Title Cata-logue In principle, different bibliographies from humanities disciplines should fill the gap since they aim to record documents regardless of digitisation status. However many of them have been commercialised in recent years. This is true, for example, of the *Annual Bibliography of English Language* and Literature and the *Bibliography of German Linguistics and Literature Studies*. On the other hand, the opening up of national library datasets through open protocols like Linked Open Data (LOD) should provide opportunities for new bibliodata aggregation solutions.

2.3. Overarching trends in the bibliodata ecosystem: The data influx, automation, and data-driven research

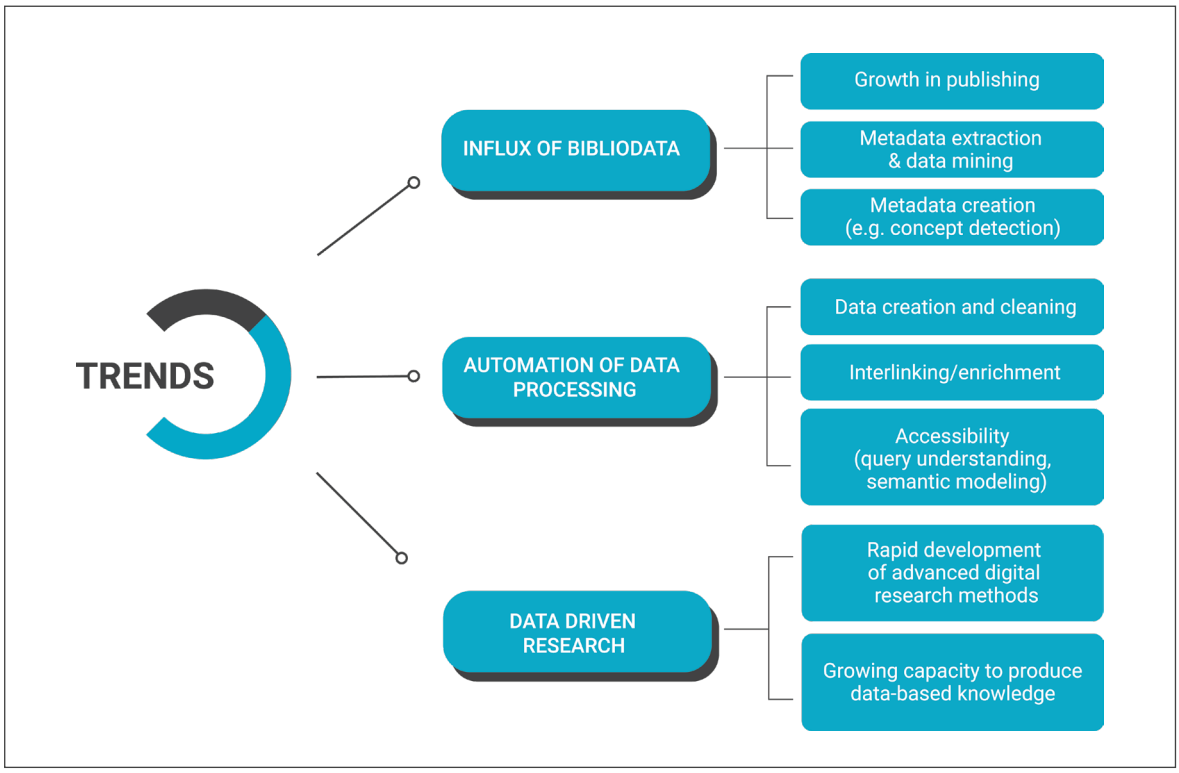


Figure 5: Overarching trends in the current bibliodata ecosystem

The bibliodata landscape is now in a phase of digital transformation characterised by the ever-growing volume of data and greater options for automated data processing. This is also a time of increased capacities to produce data-based knowledge through digital research methods.

The **influx of bibliodata** reflects the growth of publishing in general and of the digital publishing of scholarly research in particular. Other contributing factors include technical advances in areas such as metadata extraction (used for reference extraction and citation mining) and AI- and NLP-based content description (applied in automated subject classification and topic modeling). A final driver is the growing awareness among public stakeholders of the need for the open sharing of data.

The increasing accessibility of digital objects in machine-readable format also raises questions about the role of bibliodata when the full texts of documents are available. Although search engines can be used to search unstructured content effectively, bibliodata remain indispensable since they often contain expert knowledge in structured formats (e.g. reference lists and indexes), which can be hard to extract through

a full-text search. In fact, both full-text content searches and more targeted searches of structured data are useful and ultimately prove complementary.

The automation of data processing has influenced bibliodata production and use across three levels: 1) data creation and cleaning, 2) data interlinking and enrichment, and 3) access to data. Automated biblodata generation is possible whenever an item's contents are machine readable. In that case, descriptive metadata can be extracted and used as bibliodata.

Bibliodata cleaning may also be performed (semi-)automatically for taxonomy alignment, entity disambiguation, record deduplication, data normalisation, and data quality assessment. Furthermore, automation can facilitatie linking and enriching existing bibliodata sources. This can be achieved, for example, by connecting separate repositories that contain complementary information (e.g. authority records, controlled vocabularies) about the same entities or by importing external information to expand the scope of a repository. Some examples include harvesting the Web for images of authors for authority records and providing access to full-text versions of entries.

Lastly, automation can increase the accessibility of bibliodata. This area of work is still in its early stages, however there are promising developments at the level of search engines. The users of most digital bibliodata repositories rely on search and filtering tools to generate a list of matching results. In general, such searches are performed for an exact match. In other words, the process assumes that the user is highly knowledgeable and has a clear idea of what to search for and how to locate it. Allowing for fuzzier, less restrictive, semantically-based and more open search options (e.g. for fuzzy and proximity searches) could expand bibliodata's accessibility while still allowing for more advanced and precise search modalities. Such search engines would require NLP and machine-learning techniques to support query understanding, semantic modeling, and ranking systems with multiple signals.

Finally, the bibliodata landscape has been affected by the boom in advanced digital research methods, i.e. technologies, tools, and methods for advanced bibliodata research and analysis. This is tied, in turn, to increasing capacities to produce data-based knowledge. Bibliographical records provide a means not only to identify and access documents but also to transmit cultural information that yields insights into our culture, society, science, and history.

The main technological advances now driving bibliodata-based knowledge production in the humanities relate to semantic publishing. They are encapsulated in knowledge graphs implementations by research output aggregators (Microsoft Academic,

OpenAIRE, Semantic Scholar) and linked data services in the national libraries sector. These efforts to re-model bibliographical data seek to facilitate a more detailed and context-sensitive approach to publishing data that translates into heightened capacities to produce and visualise knowledge. For its part, the research community has reacted positively to these developments and there is growing interest in bibliodata--driven research in disciplines such as the digital humanities, prosopography, book history, bibliometrics, scientometrics, and cultural analytics.

2.4. Bibliography

1. Aspesi, C., Allen, N. S., Crow, R., Daugherty, S., Joseph, H., McArthur, J. T., & Shockey, N. (2019, March 28). *SPARC landscape analysis: The changing academic publishing industry – implications for academic institutions* (ver. 2). doi.org/10.31229/osf.io/58yhb
2. Aspesi, C., Allen, N., Crow, R., Hollister, V., Joseph, H., McArthur, J., Shockey, N., & Steen, K. (2021). *2021 Update: SPARC landscape analysis and roadmap for action*. www.sparcopen.org/wp-content/uploads/2021/10/2021-Landscape-Analysis-101421.pdf
3. Beaudiquez, M. (1998). *National bibliographic services at the dawn of the 21st century: Evolution and revolution*. ICNBS Copenhagen 25–27 November 1998. Proceedings of the International Conference on National Bibliographic Services. www.ifla.org/wp-content/uploads/2019/05/assets/bibliography/publications/beam-e.pdf
4. Blom, H.M.C.W, Jagersma, R., Reboul, J.M. (2020). Printed Private Library Catalogues as a Source for the History of Reading in 17th–18th century Europe. In M. Hammond (Ed.). *The Edinburgh History of Reading 1: Early Readers, 2020, Edinburgh University Press*, Edinburgh, 249–269.
5. Breeding, M. (2016). Smarter Libraries through Technology: Library Technology via Nonprofits. *Smart Libraries Newsletter*, 36(3), 1–2. librarytechnology.org/document/21779
6. Breeding, M. (2020). 2020 Library Systems Report. *American Libraries*, May, 30–41. www.americanlibrariesmagazine.org/2020/05/01/2020-library-systems-report
7. Bryant, R. (2022). Convening the OCLC RLP around bibliometrics and research impact (BRI). www.hangingtogether.org/convening-the-oclc-rlp-around-bibliometrics-and-research-impact-bri/
8. Chodacki, J., Fenner, M., & Lowenberg, D. (2020, November 7). Open Metrics Require Open Infrastructure. *Make Data Count*. www.makedatacount.org/2020/07/10/open-metrics-require-open-infrastructure
9. Colavizza, G., Peroni, S., & Romanello, M. (2021). The case for the Humanities Citation Index (HuCI): a citation index by the humanities, for the humanities. *arXiv*, abs/2110.00307. arxiv.org/abs/2110.00307

10. Commission Regulation (EU) No 651/2014 of 17 June 2014 declaring certain categories of aid compatible with the internal market in application of Articles 107 and 108 of the Treaty. (2014). *Official Journal of the European Union*. eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0651&from=PL
11. Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., Mayr, P., Romanello, M., & Zumstein, P. (2020). The OpenCitations Data Model. In: *The Semantic Web – ISWC 2020*. Lecture Notes in Computer Science, vol. 12507. Springer, Cham, 447–463. doi.org/10.1007/978-3-030-62466-8_28
12. ESFRI. (n.d.). Research Infrastructure (RI). In *Glossary*. Retrieved December 7, 2021 from www.esfri.eu/glossary
13. Fyfe, A., Coate, K., Curry, S., Lawson, S., Moxham, N., & Røstvik, C. M. (2017). Untangling academic publishing: A history of the relationship between commercial interests, academic prestige and the circulation of research. *Zenodo*. doi.org/10.5281/zenodo.546100
14. Gasparini, A., Kautonen H. (2017). Understanding Artificial Intelligence in Research Libraries – Extensive Literature Review. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 32(1). doi.org/10.53377/lq.10934
15. Haslhofer, B., Isaac, A., & Simon, R. (2018). Knowledge graphs in the libraries and Digital Humanities domain. In S. Sakr & A. Zomaya (Eds.), *Encyclopedia of Big Data technologies*. Springer, Cham. doi.org/10.1007/978-3-319-63962-8_291-1
16. IFLA Bibliography Standing Committee. (2021). *Common practices for national bibliographies in the electronic age*. www.ifla.org/wp-content/uploads/2019/05/assets/bibliography/common_practices_for_national_bibliographies_2021-01.pdf
17. Jaradeh, M. Y., Auer, S., Prinz, M., Kovtun, V., Kismihók, G., & Stocker, M. (2019). Open Research Knowledge Graph: Towards machine actionability in scholarly communication. *Semantic Scholar*. semanticscholar.org/paper/Open-Research-Knowledge-Graph%3A-Towards-Machine-in-Jaradeh-Auer/a97a18b630032b42c1a1c19db64f3fbf09a1c30b
18. Király, P. (2019). Validating 126 million MARC records. *DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. 161–168. doi.org/10.1145/3322905.3322929
19. Kokash, N., Romanello, M., Suyver, E., & Colavizza, G. (2022). From Books to Knowledge Graphs. *arXiv preprint*, 2204.10766. doi.org/10.48550/arXiv.2204.10766
20. Koundouri, P., Chatzistamoulou, N., Dávila, O. G., Giannouli, A., Kourogenis, N., Xepapadeas, A., & Xepapadeas, P. (2021). Open access in scientific information: Sustainability model and business plan for the infrastructure and organization of OpenAIRE. *Journal of Benefit-Cost Analysis*, 12(1), 170–198. doi.org/10.1017/bca.2020.26
21. Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic Data Science and the History of the Book (c. 1500–1800). *Cataloging & Classification Quarterly*, 57(1), 5–23. doi.org/10.1080/01639374.2018.1543747

22. Larivière V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLoS ONE*, 10(6), e0127502. doi.org/10.1371/journal.pone.0127502
23. Lawlor, B. (2019). An overview of the NFAIS 2019 Annual Conference: Creating strategic solutions in a technology-driven marketplace. *Information Services & Use, 39*(3), 127–165. doi.org/10.3233/ISU-190051
24. Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177. doi.org/10.1016/j.joi.2018.09.002
25. Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & López-Cózar, E. D. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126, 871–906. doi.org/10.1007/s11192-020-03690-4
26. Montoya, A.C. (2021). Enlightenment? What Enlightenment? Reflections on Half a Million Books (British, French, and Dutch Private Libraries, 1665–1830). *Eighteenth-Century Studies*, 54(4), 909–934. doi:10.1353/ecs.2021.0097
27. Péter, R., Szántó, Z., Seres, J., Bilicki, V., & Berend, G. (2020). AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts. In G. Berend, G. Gosztolya, V. Vincze (Eds.), XVI. *Magyar Számítógépes Nyelvészeti Konferencia,* 43–55.
28. Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications, 9*(12), 1–59. doi.org/10.3390/publications9010012
29. Siwecka, D. (2018). Knowledge Organization Systems Used in European National Libraries Towards Interoperability of the Semantic Web. In F. Ribeiro, M. E. Carveira (Eds.) *Challenges and Opportunities for Knowledge Organization in the Digital Age*. Ergon Verlag, Baden-Baden, 633–643.
30. Stegaeva, M. V. (2016). Cooperative cataloging: History and the current state. *Scientific and Technical Information Processing*, 43, 28–35. doi.org/10.3103/S0147688216010056
31. Schonfeld, R. C. (2017). When is a Publisher not a Publisher? Cobbling Together the Pieces to Build a Workflow Business. *The Scholarly Kitchen*. scholarlykitchen.sspnet.org/2017/02/09/cobbling-together-workflow-businesses
32. Schonfeld, R. C. (2021). Clarivate to Acquire ProQuest. *The Scholarly Kitchen*. scholarlykitchen.sspnet.org/2021/05/18/clarivate-to-acquire-proquest
33. Tay, A., Martín-Martín, A., & Hug, S. E. (2021, May 27). Goodbye, Microsoft Academic – Hello, open research infrastructure? *LSE Impact Blog*. www.blogs.lse.ac.uk/impactofsocialsciences/2021/05/27/goodbye-microsoft-academic-hello-open-research-infrastructure
34. Tolonen, M., Hill, M. J., Ijaz, A.Z., Vaara, V., & Lahti, L. (2021). Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production. In I. Baird (Ed.), *Data Visualization in Enlightenment Literature and Culture*. Palgrave Macmillan, 63–119. doi.org/10.1007/978-3-030-54913-8_3

35. Vimr, O. (2020). Big Translation History and the Use of Data Mining and Big Data Approaches: Panel Report and Observations (EST Congress 2019 in Stellenbosch). *Chronotopos – A Journal of Translation History*, 1(2), 192–195. doi.org/10.25365/cts-2019-1-2-11
36. Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1), 20–41. doi.org/10.1162/qss_a_00112
37. Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Yuxiao, D., Qian, J., Kanakia, A., Chen, A., & Rogahn, R. (2019). A review of Microsoft Academic Services for science of science studies. *Frontiers in Big Data*, 2(45). doi.org/10.3389/fdata.2019.00045

3. KEY CHALLENGES AND OPPORTUNITIES FOR PUBLIC BIBLIODATA STAKEHOLDERS

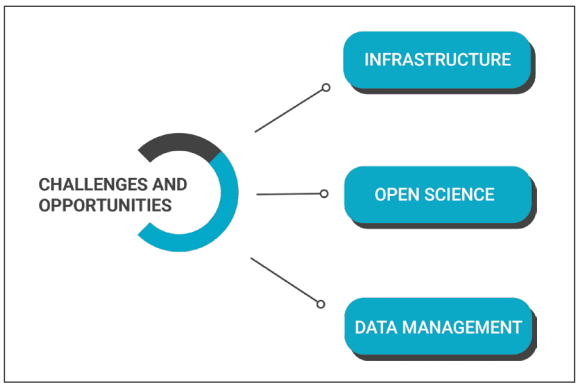


Figure 6: Main challenges and opportunities for the public bibliodata stakeholders

Our map of the current landscape in Chapter 2 enables us to identify a number of challenges and untapped opportunities that should be targeted by future bibliodata-related efforts. In this chapter, we explore these challenges and opportunities across three key areas: 1) infrastructure; 2) open science, and 3) bibliodata management.

3.1. Infrastructure

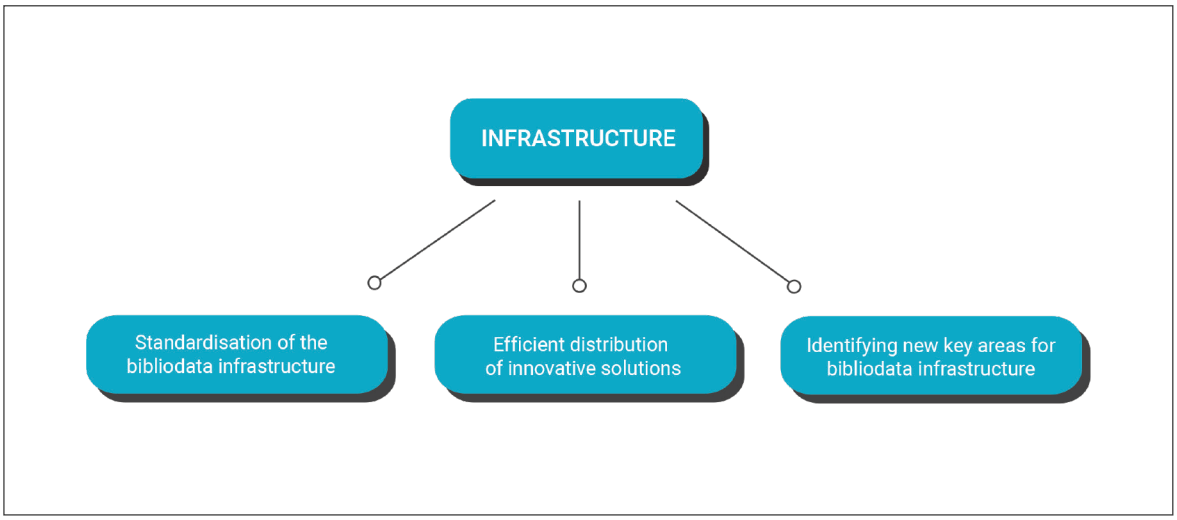


Figure 7: Bibliodata infrastructure

The main components of bibliodata infrastructure include services and tools to process, transform, and enrich bibliodata; services to assign and resolve persistent identifiers for publications, scholars, organisations, and other objects; support for bibliodata management; and materials and activities to promote bibliodata literacy.

We identify three main tasks for stakeholders related to bibliodata infrastructure: its standardisation, disseminating innovative solutions, and identifying new investments.

The standardisation of bibliodata infrastructure remains a challenge since despite the existence of efficient solutions such as international aggregators, access to stan-

dard formats and protocols for bibliodata varies widely across the public sector. The divisions in the bibliodata landscape outlined in the previous chapter also create roadblocks to standardisation. In this context, we would distinguish at least three large data ecosystems: a LAM ecosystem (led by libraries as the main stakeholder); a research-oriented public ecosystem (for research infrastructures); and a private sector ecosystem for services that process research outputs (for commercial information services).

Stakeholders are now increasingly rejecting a focus on any single data format or standard or centralised data exchange solution in favour of approaches like semantic publishing that enable cross-institutional and international interoperability. This is itself a challenge since it demands a clear set of common standardisation objectives. Here the end goal is not to adopt any single (elusive) format or standard, but to apply general standards such as FAIR principles more flexibly to local collections and services while also enriching, mapping, and converting these resources to enhance interoperability.

Standardisation promises to create major opportunities in the bibliodata landscape since it would enable a profound connectivity among stakeholders that could foster reuse and integration and minimise costs. The availability of standardised bibliodata would enable the seamless integration of services that produce or use those data. It would also facilitate innovative solutions for reuse and exploitation.

Efficient sharing of innovations in bibliodata production and use is another challenge that calls for extensive resources and expertise. These innovative solutions include the technological advances of the semantic web (semantic graphs, linked data access points, etc.) and artificial intelligence (machine learning techniques for bibliodata generation, enrichment, mining, interlinking, and searches). Adopters of these solutions tend to be entities with vast resources such as big tech services, commercial information services, more robust metadata aggregators, and larger public stakeholders like national libraries. Sector specialists like Semantic Scholar and GROBID are also involved in some of these pioneering efforts.

At present, innovative solutions are often seen as crowning achievements for mature collections or services. However, in many cases, there is enormous potential for their application to smaller stakeholders and less processed resources. Many non-digitised bibliographical resources, particularly printed book-length bibliographies, could benefit from machine-learning approaches to retrospective conversion (in contrast, current

prospects of manual digitisation often nip retroconversion plans in the bud). Similarly, journal publishers could employ metadata extraction to improve data quality. If provided with enough know-how and resources, the curators of smaller bibliographical services could boost the interoperability of their resources by using targeted forms of content enrichment such as PIDs and linked data or some level of international authority control. This could be an alternative to overhauling curatorial approaches (e.g. by changing data formats or software). If the entire landscape is to flourish, then open solutions need to be developed and geared to smaller stakeholders, who should be offered concrete tools. Ultimately, larger players that are able to implement innovations rely on smaller stakeholders' resources. To break this cycle, a targeted effort is needed to disseminate innovations more equally.

A final challenge concerns the **identification of new areas for infrastructure investment**. These choices require a good understanding of the landscape and a well-established pattern of cooperation among diverse stakeholders. This common approach is needed to identify oversights in existing resources such as the aforementioned gap around the aggregation of national libraries' bibliodata. Such knowledge-sharing can also avoid the duplication of efforts and inefficient spending, which is exemplified in the dependence of many public stakeholders on private/commercial services.

By making joint efforts to identify new investment opportunities in bibliodata infrastructure, stakeholders can create pathways to better and more efficient services for end users. While these public entities have their own bibliodata interests and approaches, there is much common ground to be explored in standardisation, automation, IT tools, and beyond. Finding gaps and inefficiencies in the bibliodata landscape should be a priority for all entities and not only commercial service providers.

This better alignment of public stakeholders' infrastructure needs could also lead to clearer boundaries between public and private stakeholders. This might, in turn, reduce tensions and foster more cooperative relationships between these groups.

3.2. Open science

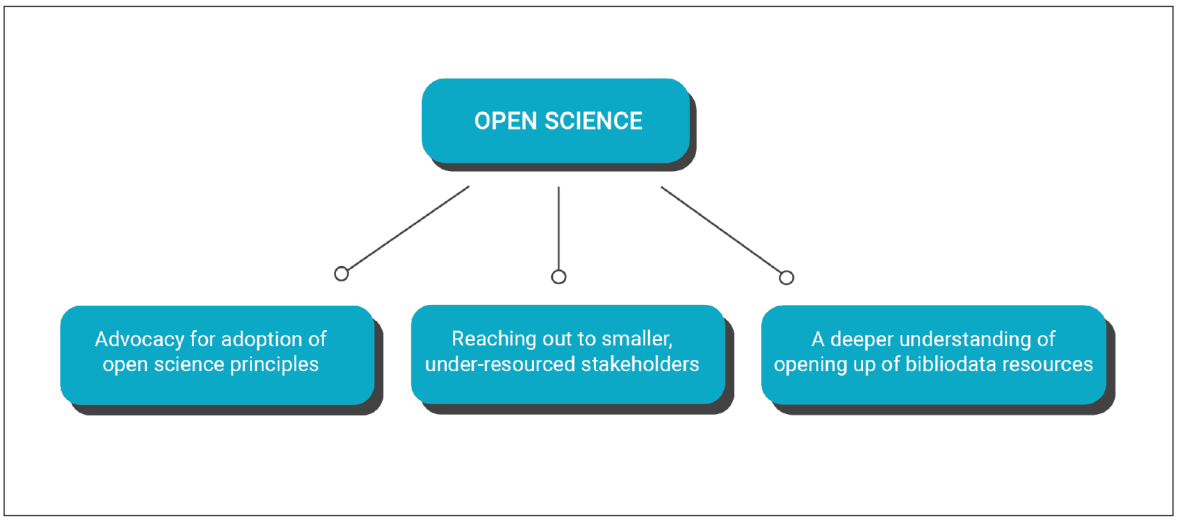


Figure 8: Bibliodata and Open Science

Open science refers to a broad set of technical and cultural issues related to access to information. Such information includes (digital) materials such as data, methods, and reports, and the concept also extends to the openness of infrastructures, collaboration models, and evaluation metrics.

We identify three main challenges for stakeholders seeking open access to bibliodata: ensuring that open science principles are widely adopted; directing advocacy at smaller stakeholders; and expanding understandings of "openness".

The widespread adoption of open science principles remains the greatest challenge for participants in the bibliodata landscape. This is due to the persistent division of data ecosystems into open and closed systems, a division closely aligned with the one between public and private stakeholders. Private sector business models may allow for the exclusive ownership of data, tools, and services. The growth of commercial services is, thus, a key factor in this landscape. As reliance on these services increases, access to metadata collections may decline since users are often restricted to a given number of entries or a certain level of information. These restrictions may complicate or even completely disable access to bibliodata along with their potential (re-)use. Those most affected tend to be stakeholders/users outside the core academic community who are not affiliated with research institutions and cannot benefit from subscription contracts.

Another major barrier to open accessibility lies in the technological constraints affecting different stakeholders. These may result in issues around interoperability

and the quality of data and methods. Adopting common open science standards around licensing, data presentation, and workflow documentation (see also Section 3.1 above on infrastructure) would allow for the interoperability of cross-institutional and international data ecosystems. In the case of data presentation, such open science standards should clearly include FAIR principles.

As both the OpenCitations and Open Abstracts initiatives have shown, there are many opportunities for open science initiatives in the bibliodata landscape. Those projects properly identify the limitation of access to elements of bibliodata during the ongoing reuse cycle, and they have succeeded in "regaining" access via advocacy, collaboration, and infrastructure building. This work provides a model for other open science advocacy, which could apply these strategies locally and/or extrapolate from them conceptually, for example, by replacing citations or abstracts with subject descriptions or keywords. Another prospect for promoting open science exists at the public policy level and includes the positions of national and international grant providers and governmental bodies. Infrastructures for open research resources such as the European Open Science Cloud (EOSC) and Social Sciences & Humanities Open Cloud (SSHOC) may also use their influence to broaden open science policy and integrate open data principles into current research workflows.

Despite these promising developments, there are still significant obstacles to applying open science standards across the bibliodata landscape. One stumbling block relates to the **relatively small size of some public sector participants involved in data creation and sharing**. Open access initiatives may be limited by a lack of know-how if resources are practically and formally accessible but not legally (i.e. in licensing terms) or technologically ready for future reuse. In the previous section, we outlined several issues related to the unequal distribution of technological innovations and called on larger public stakeholders to reach out to smaller ones with infrastructure solutions for the sake of the entire landscape. (Here a key issue is the reliance of powerful services such as metadata aggregators and discovery services on local collection quality.) Similarly, the licensing of bibliodata resources often depends on local institutional know-how and curators' appreciation of the impact of this legal issue on reuse options. Open science advocates need to advise smaller stakeholders about the importance of licensing as well as its costs and benefits. Above all, they can help clarify the expectations of those smaller entities by offering common recommendations.

Another challenge concerns the need for a **deeper understanding of open science in the context of bibliodata resources**. In particular, it should be understood that "open science" covers not only open access to data, but also openness of the workflows, methods, and tools used in data processing. As such, it concerns the input of both service curators and researchers. On the first count, this includes data harmonisation, reconciliation, and linking methods as well as algorithms for metadata extraction, analysis, and reproducible reporting. On the second, it incorporates both research data and the (pre-)processing and modelling methods of bibliodata researchers.

Of course, as we have seen in other areas of data curation and research, a lack of knowledge, know-how, and resources may limit the scope of what is effectively "opened up" (i.e. published, documented, and disseminated) for different stakeholders. However, we contend that the wider sharing of data processing methods and workflows can lead to exceptional gains for all participants both in terms of efficiency and the overall quality of data curation and research.

3.3. Data management lifecycle: Creation and documentation

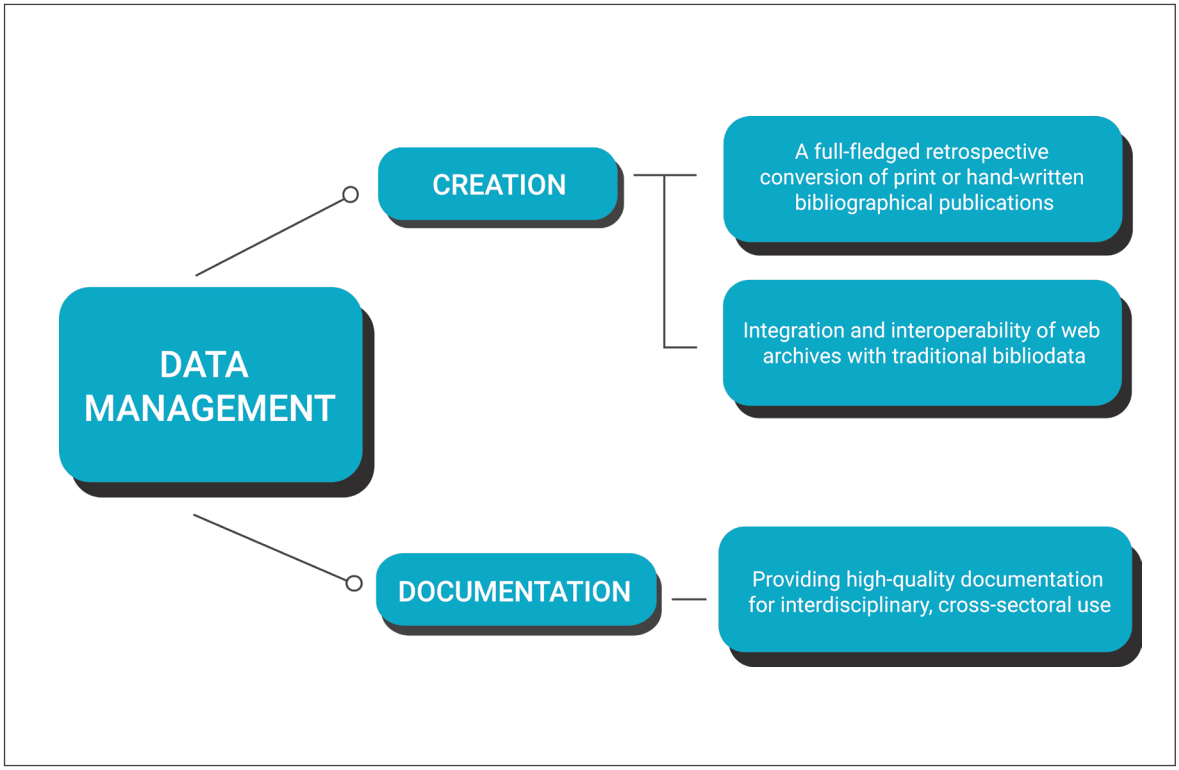


Figure 9: Bibliodata management: main challenges

Data management refers to any practices involving data, including their conceptualisation, storage, and dissemination. Data management lifecycle is a term used to

describe the series of processes included in data management. In this section, we focus on two stages of this lifecycle: data creation and documentation. In this regard, we identify two main tasks for bibliodata management: continually and proactively adjusting the scope of the bibliodata created and rigorously documenting existing datasets.

Each stakeholder has its own data creation plan. Nevertheless while the bibliodata landscape is expansive, complex, and varied, an **overview of its entirety can help identify gaps that have not been sufficiently addressed**. Such an overview reveals the failure to commit to a **fully-fledged retrospective conversion of print or hand-written bibliographical publications** (e.g. printed enumerative bibliographies, inventories, hand-written listings, classification schemes) into structured, machine-readable data formats. While library catalogues have been systematically digitised, for the publications themselves, conversion calls for a more organised approach. The challenges here include copyright and technological issues, however these must be offset against this extraordinary opportunity to produce and publish large quantities of highly valuable data.

From a legal standpoint, it is significant that many of these bibliographical publications are the work of public institutions. This alone increases their retroconversion prospects. Even if some items cannot be converted into databases for legal reasons, the scope of many bibliographies overlaps and knowledge of existing resources could still lead to great gains. On the technical side, advances in Handwritten Text Recognition (HTR), Optical Character Recognition (OCR), Optical Layout Recognition (OLR), metadata extraction, and data structuring and parsing tools may all be exploited when taking up this challenge (see for example GROBID applications). These technologies are already being used for bibliodata processing but they have not yet been systematically applied to the retroconversion of bibliographical publications. Stakeholders that invest in these methods will profit especially from being able to prove and showcase the benefits of retrospective conversion; the process will result in the creation of large new datasets and related innovations in data processing methods (AI, machine learning, data parsing and cleaning, etc.). It may also generate opportunities for cooperation between the bibliodata and natural language processing communities, who may apply knowledge from both fields in their work on semi-structured data.

A second challenge related to current inefficient bibliodata creation concerns the organising of knowledge about the massive volume of new online content being produced daily. These items include blogs, artists' websites, podcasts, social media feeds, digital art, e-literature, etc. Many of those documents have cultural

significance, however the bibliographical control over them is limited compared with that over printed documents (i.e. via national bibliographies, library catalogues, etc.) and digital research output. Some online documents have been registered in public web archives (mostly maintained by national libraries) and/or by non-profit organisations such as Internet Archive. In some cases, traditional standards such as the ISSN are being used to record these online materials.

Nevertheless, the quality of the metadata (e.g. subject and type classifications, granularity of description, etc.) attached to online documents remains incomparable to the standards for "traditional" bibliodata such as printed books in libraries or digital research output. Moreover, these online resources are not connected or aggregated by bibliodata services. If bibliodata services are truly to represent the documents produced by a given society, then they must include the metadata of online documents at least to some extent. Significantly, we are already seeing a surge of interest in data-driven research which compares "traditional" bibliodata (e.g. from scientific literature) to the metadata of online documents. This includes, for example, studies that compare the popularity of certain topics in published scholarship and on social media. There is, thus, an opportunity for bibliodata services to meet the needs of researchers who are now collecting and organising online documents in order to compare them to the "traditional" bibliodata accessible through bibliodata services.

As regards **bibliodata documentation**, the most significant challenge remains the creation of records that extend beyond a formal description of the data. Such records should include a **content description and methodological background about data processing and the relationships between relevant datasets**, that is, information that would allow users to assess the data's representativeness. This task needs to be addressed through close cooperation between curators and researchers, but the current bibliodata landscape does not support the creation of stable cross-sectional teams that could invest time in documenting datasets. This is, however, an essential step for future research. The challenge is, thus, not only to produce comprehensive documentation but to create an environment where that process can happen. Providing **high-quality documentation** for existing datasets would **foster cooperation in the field**, support interdisciplinary initiatives, and increase data use. This is an opportunity for both data curators and users, and researchers especially may benefit from partnering with curatorial institutions to develop documentation and work on data collection.

3.4. Bibliography

1. Bilder, G., Lin, J., & Neylon, C. (2020). *The principles of Open Scholarly Infrastructure*. doi.org/10.24343/C34W2H
2. Bourrier, K., Thelwall M. (2020). The Social Lives of Books: Reading Victorian Literature on Goodreads. *Journal of Cultural Analytics* 5 (1). doi.org/10.22148/001c.12049
3. Chambers, S. (2020). Web-archives for Open Science: How FAIR can we go? [video]. *WARCnet kickoff meeting, 4–6 May 2020*. www.cc.au.dk/en/warcnet/presentations/kickoff-meeting-2020
4. Dooley, J. & Bowers, K. (2018). Descriptive metadata for web archiving: Recommendations of the OCLC research library partnership web archiving metadata working group. *OCLC Research*. www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata/recommendations.html
5. ESFRI. (2021). *Strategy Report on Research Infrastructures Roadmap 2021. Public Guide*. www.esfri.eu/esfri-roadmap-2021
6. Faniel, I. M. & Zimmerman, A. (2011). Beyond the data deluge: A research agenda for large-scale data sharing and reuse. *International Journal of Digital Curation, 6*(1), 58–69. doi.org/10.2218/ijdc.v6i1.172
7. Fecher, B. & Friesike, S. (2014). Open Science: One term, five schools of thought. In S. Bartling & S. Friesike (Eds.), *Opening Science*. Springer, Cham. doi.org/10.1007/978-3-319-00026-8_2
8. Ficarra, V., Fosci, M., Chiarelli, A., Kramer, B., & Proudman, V. (2020). Scoping the Open Science Infrastructure landscape in Europe. *Zenodo*. doi.org/10.5281/zenodo.4159838
9. Goudarzi, S., Pugh, K., Rhinesmith, V., Staines, H., & Thaney, K. (2021). Designing a preparedness model for the future of Open Scholarship. *Zenodo*. doi.org/10.5281/zenodo.5218968
10. Lindemann, D., Khemakhem, M., Romary, L. (2018). Retro-digitizing and Automatically Structuring a Large Bibliography Collection. *European Association for Digital Humanities (EADH) Conference*. hal-01941534
11. Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). Open Data Metrics: Lighting the fire (Version 1) [Computer software]. *Zenodo*. doi.org/10.5281/zenodo.3525349
12. Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., & Nevalainen, T. (2020). Wrangling with non-standard data. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference: Riga, Latvia, October 21–23, 2020. *CEUR Workshop Proceedings*, 2612. www.ceur-ws.org/Vol-2612/paper6.pdf
13. Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J., & Mechant, P. (2021). Web-archiving and social media: An exploratory analysis. *International Journal of Digital Humanities*. doi.org/10.1007/s42803-021-00036-1

4. CONCLUSIONS: TOWARDS JOINT AGENDAS FOR PUBLIC BIBLIODATA STAKEHOLDERS

As an overview of bibliodata ecosystems in the humanities, this report may serve as a conversation starter among entities – especially those in the public sector – who recognise the need to take advantage of the potential of bibliodata, better allocate limited public resources, and collaborate to meet common goals. This document may also provide a basis for the creation of the **joint agendas** among these stakeholders – whether in the form of further reports, case studies, and/or grant proposals.

Below we offer recommendations for future work in the hope that they may facilitate further discussions among interested parties. These recommendations may also provide a roadmap for productive and successful cooperation around public stakeholders' joint agendas.

- Focus on the unique issues around bibliodata

Bibliodata-related issues are often overlooked by powerful stakeholders whose focus is on large-scale digitisation (LAM services), access to full-text materials (information services, public metadata aggregators), and the development of IT tools (RIs). Bibliodata landscape stakeholders are uniquely placed to provide a bibliodata-based perspective on the humanities and cultural heritage. Such an analysis through a bibliodata lens can undoubtedly benefit the community as a whole. Potential gains may include the identification of untapped partnerships between public stakeholders with similar obligations (e.g. libraries and academic publishers, two key bibliodata producers in the humanities) that could lead to the development of shared infrastructure and the harmonising of standards for further data reuse. Identifying common goals is particularly crucial for parties undertaking major challenges that call for large-scale resources and strong leadership. These challenges include the retrospective conversion of bibliodata, bibliodata extraction and mining, and implementing machine learning and NLP for metadata production and query improvement.

- Pursue advocacy and education

When addressing the unique issues around bibliodata, advocates should explain the stakes involved in bibliodata services development. This includes highlighting how investing in bibliodata can translate into improved organisation, access, and knowledge sharing as well as the overall development of the humanities.

Advocates and educators should highlight bibliodata issues to specific humanities and cultural heritage stakeholders. This can be done in many ways, including by

sharing "success stories" from the bibliodata field, supporting and enhancing data literacy, and facilitating open access to bibliographical data science methods and workflows. Public relations initiatives should also inform the humanities community about the nature of bibliographical work.

- Emphasise cooperation between diverse stakeholders

Our ability to meet the most pressing challenges in the bibliodata landscape depends on the development of cross-institutional, cross-sectional cooperation. Collaborative efforts must include not only work across different public sectors – especially LAM and research – but also continuing efforts to build partnerships with the private sector through initiatives like OpenCitations and Open Abstracts.

Ideally this cooperation should be issue-based and tackle challenges that are critical to wide-ranging stakeholders. Those challenges may include the automatic classification and description of documents, analysis and mining of online content, and support for data reuse in the humanities.

- Take advantage of the open science movement

Open science principles are the default standard for public bibliodata stakeholders. Open science and surrounding policies (e.g. FAIR principles) can galvanise joint efforts to standardise bibliodata, document and license existing datasets, and engage smaller stakeholders in the movement. Public policies should be adapted and contextualised so that they fit bibliodata needs, and the agendas of open science initiatives should be similarly enhanced (a good point of comparison can be found in projects such as Invest in Open Infrastructure and Make Data Count). In all these activities, stakeholders in the humanities should treat data originating from the research and LAM sectors as equally important for the development of the field.

- Share innovations in bibliodata curation and research

To date there has been great progress in bibliodata curation and research. This includes advances in semantic publishing, which can be used to enrich data and improve interoperability; new automation technologies (e.g. NLP, AI, and machine learning) to classify documents and control data quality; and data science methods which produce new knowledge. Some public stakeholders have managed to co-produce or take advantage of these developments, but there is no guaranteed "trickle-down" effect.

It is critical for innovation-leading public stakeholders – whether in the curation or research spaces – to provide smaller actors models of workflows, methods, and tools that are openly accessible and easily understandable. Ultimately, a more efficient distribution of existing and newly developed solutions will lead to more effective public spending. This will, in turn, improve the overall quality of the bibliodata available.

Acronyms

AI – Artificial Intelligence
API – Application Programming Interface
BIBO – The BIBliographic Ontology
BIBFRAME – BIBliographic FRAMEwork
CRIS – Current Research Information System
DOI – Digital Object Identifier
EOSC – European Open Science Cloud
ESFRI – European Strategy Forum on Research Infrastructure
FAIR – Findability, Accessibility, Interoperability, and Reusability
HTR – Handwritten Text Recognition
ILS – Integrated Library System
ISBD – International Standard Bibliographic Description
ISSN – International Standard Serial Number
ISBN – International Standard Book Number
LAM – Libraries, Archives and Museums
MARC – MAchine Readable Cataloguing
MODS – Metadata Object Description Schema
NLP – Natural Language Processing
OCLC – Online Computer Library Center
OCR – Optical Character Recognition
OJS – Open Journal System
OLR – Optical Layout Recognition
ONIX– ONline Information EXchange
OPAC – Online Public Access Catalogue
ORCID – Open Researcher and Contributor ID
PID – Persistent Identifier
PKP – Public Knowledge Project
RDA – Resource Description and Access
RDF – Resource Description Framework
SSHOC – Social Sciences & Humanities Open Cloud
VIAF – Virtual International Authority File

This was partly prepared as a part of Czech Literary Bibliography research infrastructure project (LM2018136), funded by the Ministry of Education, Youth and Sports of the Czech Republic within its activities in support of research infrastructures.

This text was proofread within the project CZ.02.2.69/0.0/0.0/18_054/0014701 Development of Research and Popularisation Resources of the Institute of Czech Literature of the CAS, co-funded by the EU's European Structural and Investment Funds within the operational programme Research, Development and Education.