# Fostering Collaboration to Enable Bibliodata-driven Research in the Humanities

**Malínek, Vojtěch**
malinek@ucl.cas.cz
Institute of Czech Literature, Czech Academy of Sciences, Czech Republic

**Umerle, Tomasz**
tomasz.umerle@ibl.waw.pl
Institute of Literary Research, Polish Academy of Sciences, Poland

**Tolonen, Mikko**
mikko.tolonen@helsinki.fi
University of Helsinki, Finland

**Karlińska, Agnieszka**
agnieszka.karlinska@ibl.waw.pl
NASK National Research Institute, Warszaw

**Romanello, Matteo**
matteo.romanello@unil.ch
Université de Lausanne, Switzerland

**Colavizza, Giovanni**
g.colavizza@uva.nl
University of Amsterdam, Netherlands

**Peroni, Silvio**
silvio.peroni@unibo.it
University of Bologna, Italy

**Siwecka, Dorota**
dorota.siwecka@uwr.edu.pl
University of Wroclaw, Poland

**Łubocki, Jakub**
jakub.lubocki@mnwr.pl
National Museum of Wrocław, Poland

**Rißler-Pipka, Nanette**
nanette.rissler-pipka@gwdg.de
GWDG, Göttingen, Germany

**Lindemann, David**
david.lindemann@ehu.eus
University of Basque Country, Spain

**Labropoulou, Penny**
penny@athenarc.gr
ATHENA Research Centre, Greece

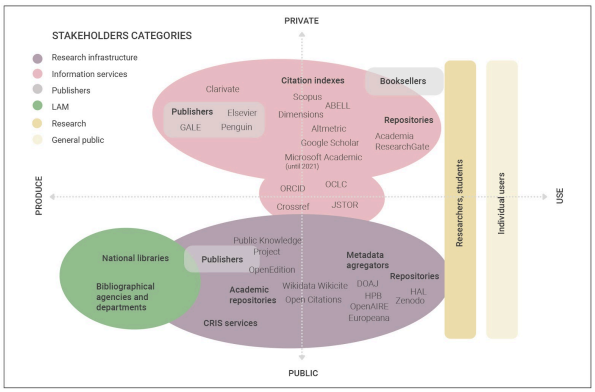
**Klaes, Christiane**
c-klaes@tu.brauschweig.de
Technische Universität Braunschweig, Germany

## Panel Overview

**Bibliodata Landscape in the Humanities**
Vojtěch Malínek, Institute of Czech Literature, Czech Academy of Sciences

Tomasz Umerle, Institute of Literary Research, Polish Academy of Sciences

Bibliographical data (bibliodata) are one of the most important types of data for data-driven research and curation in the humanities. Bibliodata has solid foundations in terms of data production (constant flow of data produced by different stakeholders) and an established user base. The current bibliographical data ecosystem is shaped by constant interactions between various public and private stakeholders which engage in production and (re)use of bibliodata (Umerle et al. 2022).




This complex and rich data landscape is responsible for registering cultural, societal and research phenomena through national libraries (i.e. BNF), union catalogues (i. e. ESTC) and specialized bibliographies (i.e. BOSLIT), academic and public repositories (i.e. Zenodo) and CRIS systems, citation indexes (i.e. Scopus, WoS), metadata aggregators ( OpenAIRE, Europeana, TRIPLE), social networks (e. g. Goodreads) or information services (i.e. Dimensions, Altmetric).

These sources and services are used by the general public, researchers and curatorial institutions to identify, describe and make available different information resources, but at the same time there is a long lasting and fast growing trend to produce knowledge based on bibliographical data (Aspesi et al. 2021).

This trend is followed by various bibliodata stakeholders including researchers in the field of bibliometrics, cultural analytics, book history and service providers such as public e-infrastructures (OpenAIRE), or private companies like Dimensions that increasingly engage in providing research data for transforming information into knowledge.

Unfortunately, despite large amounts of data, diverse interest in data-based research and relative accessibility to this type of resources (in comparison e.g. to full text collections), bibliodata-based research has not reached its full potential to provide relevant knowledge on the social and cultural, contemporary and historical processes.

This is due to the number of challenges in the field, especially dispersion of data resources amongst divergent stakeholders, fragmentation of the stakeholders, unbalanced distribution of innovative solutions and best practices, uncoordinated data curation standards in different fields and disciplinary barriers between established research methodologies (e.g. bibliometrics, book history, cultural analytics) (Király 2019).

To leverage the full potential of bibliodata-driven research in the humanities we need a multi-faceted approach to tackle those challenges. In this panel we propose a renewed framework for scaling up bibliodata-driven research through collaborative efforts:

1) propel creation of the scalable and multidisciplinary bibliodata workflows for bibliographical data science, especially through collaboration of metadata-based research with the AI and textual analysis communities,

2) adjust open data standards and implementations to the needs of digital humanities, especially focusing on solutions diverse actors relevant for the humanities – not only research ones, but GLAM ones,

3) share and implement interoperability standards and best practices,

4) create and implement standardised and detailed documentation of bibliographical resources.

**Opportunities and Challenges for Bibliographical Data Science**

Mikko Tolonen, University of Helsinki
Agnieszka Karlińska, NASK National Research Institute, Warszaw

**Bibliographical data science (BDS)** is an approach targeted at enabling the use of bibliographical metadata as a research object, deriving from the more generic paradigms of book history, open science and data science (Lahti et al. 2019). Since bibliographical metadata is connected to all aspects of public discourse, BDS potential to influence scholarly practices is enormous. Today BDS is facing two critical challenges: 1) **creation of enrichment and harmonisation workflows** to facilitate large-scale bibliodata-driven research, 2) the need to grow as a community and embrace the developments at the **intersections of metadata processing and textual analysis** by developing collaborations with the NLP and ML experts.

Wider reach of bibliographical data research calls for **usage of interdisciplinary, multilingual** collections originating from diverse sources. Independently maintained collections demand automated harmonisation enhancing their readiness for quantitative analysis (Tolonen et al. 2021). The critical challenge in this regard is the creation of scalable workflows for producing knowledge that cover all stages of the bibliodata research. On the one hand, there is a need to genuinely take an international perspective. On the other hand, we need to overcome disciplinary silos and bring together diverse experts.

At the same time, **bibliodata-driven research** cannot focus solely on the metadata and **needs to grow as a community** and create a framework **combining metadata-based workflows with the NLP methods** aimed at data mining and information extraction (Péter et al. 2020). Together these approaches produce larger workflows critical for the humanities research that should be collaboratively developed and openly shared.

**Building the Humanities Citation Index: a Case in Point for Open Bibliodata**

Matteo Romanello, University of Lausanne
Giovanni Colavizza, University of Amsterdam
Silvio Peroni, University of Bologna

Building a citation index for the Arts & Humanities is a typical example of an endeavour that, in order to be realised, would require a **radical change** in the way GLAM institutions represent, organise and exchange bibliographical data (Stone 1982). The implementation of such a Humanities Citation Index (HuCI) not only raises some substantial technical challenges (Colavizza et al. 2022), but also requires **close collaboration** between various stakeholders responsible for producing, publishing or consuming bibliodata (Martín-Martín et al. 2021). These include researchers, libraries, archives, publishers, and learned societies. In this contribution, we will discuss the following aspects which are key to the realisation of HuCI:

**Licensing and documentation:** bibliodata ought to come with explicitly defined open licenses (Ficarra et al. 2020), as well as with thorough documentation in order to maximise their reusability.

**Shared bibliodata formats:** to enable their use in citation mining pipelines and processes, bibliodata must be represented in a shared, concise, and "easy-to-process" format. In the absence of widely adopted common formats, we will require services for mapping existing formats into a common one.

**Provision of persistent identifiers:** any source to be indexed by HuCI needs a unique, persistent identifier. This raises some issues in terms of coverage and granularity, and requires to find economically viable solutions to minting unique identifiers for legacy publications.

**Data and service APIs:** the exchange of bibliodata between GLAM institutions and any potential user must happen without need for *any* human intervention, by means of APIs and regular data dumps.

**Documentation of Bibliodata Resources**

Dorota Siwecka, University of Wroclaw
Jakub Łubocki, National Museum of Wroc#aw
Nanette Rißler-Pipka, GWDG, Göttingen

Creating, maintaining and providing access to resources of a bibliographical nature is proving to be insufficient for the average user. Even the best-constructed bibliographical database, using international metadata standards, open licenses and API or similar open interfaces, turns out to be moderately useful when its description is missing (Bilder et al. 2020). A lack of proper user-friendly documentation containing basic information on the methodology of creating bibliographical datasets affects many factors, i.e. users' awareness of the level of relevancy and completeness of the obtained results. In the context of data reuse depicting the contents of the database seems crucial in order to draw well reflected conclusions from bibliodata analysis.

Creating, sharing and disseminating best practices in documentation for users of bibliographical datasets, is what we propose as a solution to improve the quality of bibliodata resources, as well as increase their reuse (Faniel and Zimmermann 2011). A collaboration between divergent stakeholders is needed to find and compare existing recommendations from different countries (Vlassenroot et al. 2021). On this basis, international guidelines for bibliodata providers should be created taking into account the user's point of view (concerning scope and coverage of database, sources of information, structure, metadata model, license, etc.). Developed recommendations should also include a template of such dataset description. It will allow comparison of this kind of resources and make it easier to select the reliable datasets for research. A de-

scription template can be then used e.g. for creating a register of European bibliographical databases for Humanities.

**Bibliodata LOD-ification using free software**

David Lindemann, University of Basque Country
Penny Labropoulou, ATHENA Research Centre
Christiane Klaes, Technische Universität Braunschweig

This contribution discusses workflows for the conversion of bibliodata into Linked Data, using free software. Under "bibliodata" we consider publication metadata from bibliographic catalogues, citation relations and content-describing subject headings. For the discussion, we present three use cases, namely LexBib (Lindemann et al. 2018; Lindemann 2021), a bibliography of Lexicography and Dictionary Research, Inguma (Erriondo 2006), a collection of scientific written production in the Basque language, and CLB-LOD, a dataset derived from the electronic Czech Literary Bibliography. Final goal in all the use-cases on hand is federation with or integration in Wikidata (Vrandečić / Krötzsch 2014; Van Veen 2019), a large free Knowledge Graph, while their sources deploy different formats and models: (1) Zotero collection, (2) SQL database with custom (non-standard) data model, and (3) MARC21 (XML version), a standard widely used in library catalogues.

To that end, we deploy instances of the Wikibase software (i.e., the same software that underlies Wikidata, cf. Lindemann (2022)). For interaction with Wikibase, e.g. data upload, synchronization, and entity linking (reconciliation of literal values against authority files such as ( VIAF or Wikidata), we use our own scripts, and OpenRefine.

We present advantages and drawbacks of the chosen software and tool pipeline, taking into account benefits of including bibliographical data in the Wikidata Knowledge Graph, such as their visibility through related bibliometrics tools like Scholia (Nielsen et al. 2017), and how our approach could be adopted by a larger community.

# Bibliography

**Aspesi, Claudio** / **Allen, Nicole** / **Crow, Raym** / **Hollister, Valorie** / **Joseph, Heather** / **McArthur, Joseph** / **Shockey, Nick** / **Steen, Katie** (2021): *2021 Update: SPARC landscape analysis and roadmap for action*. https://www.sparcopen.org/wp-content/uploads/2021/10/2021-Landscape-Analysis-101421.pdf [30.10.2022].

**Bilder, G.** / **Lin, J.** / **Neylon, C.** (2020): *The principles of Open Scholarly Infrastructure*. DOI: 10.24343/C34W2H.

**Colavizza, Giovanni** / **Peroni, Silvio** / **Romanello, Matteo** (2022): "The case for the Humanities Citation Index (HuCI): a citation index by the humanities, for the humanities", in: *International Journal on Digital Libraries* . DOI: 10.1007/s00799-022-00327-0 .

**Erriondo, Lorre** (2006): "Soziolinguistika eta UEU (Udako Euskal Unibertsitatea)", in: *Bat: Soziolinguistika aldizkaria* 61: 71–84.

**Faniel, Ixchel M.** / **Zimmerman, Ann** (2011): "Beyond the data deluge: A research agenda for large-scale data sharing and reuse", in: *International Journal of Digital Curation* 6, 1: 58–69. DOI: 10.2218/ijdc.v6i1.172.

**Ficarra, Victoria** / **Fosci, Matiia** / **Chiarelli, Andrea** / **Kramer, Bianca** / **Proudman, Vanessa** (2020): *Scoping the Open Science Infrastructure landscape in Europe*. Zenodo. DOI: 10.5281/zenodo.4159838.

**Király, Péter** (2019): „Validating 126 million MARC records", in: *DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, Brussels, Belgium, May 2019: 161–168. DOI: 10.1145/3322905.3322929.

**Lahti, Leo** / **Marjanen, Jani** / **Roivainen, Hege** / **Tolonen, Mikko** (2019): "Bibliographic Data Science and the History of the Book (c. 1500–1800)", in: *Cataloging & Classification Quarterly* 57, 1: 5–23. DOI: 10.1080/01639374.2018.1543747.

**Lindemann, David** (2022): "LOD-Ification of Bibliographical Data Using Free Software: CLB-LOD Wikibase". *Mutual Learning Workshop for Improving Cultural Heritage Bibliographical Data*, Prague, Czech Republic, October 2022. DOI: 10.5281/zenodo.7250730.

**Lindemann, David** (2021): "Zotero to Elexifinder: Collection, Curation, and Migration of Bibliographical Data", in: *SiKDD 21 Slovenian KDD Conference*. Ljubljana, Slovenia, October 2021 https://ailab.ijs.si/dunja/SiKDD2021/Papers/LindemannDavid.pdf [30.10.2022].

**Lindemann, David** / **Kliche, Fritz** / **Heid, Ulrich** (2018): "LexBib: A Corpus and Bibliography of Metalexicographical Publications", in: *Proceedings of the XVIII EURALEX International Congress*. Ljubljana, Slovenia, August 2018: 699–712. http://euralex.org/publications/lexbib-a-corpus-and-bibliography-of-metalexicographical-publications/ [30.10.2022].

**Martín-Martín, Alberto** / **Thelwall, Mike** / **Orduna-Malea, Enrique** / **Delgado López-Cózar, Emilio** (2021): "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations", in: *Scientometrics* 126, 1: 871–906. DOI: 10.1007/s11192-020-03690-4.

**Nielsen, Finn Årup** / **Mietchen, Daniel** / **Willighagen, Egon** (2017): „Scholia, Scientometrics and Wikidata", in Blomqvist, E. et al. (eds.): *The Semantic Web: ESWC 2017 Satellite Events*. Cham: Springer International Publishing: 237–259. DOI: 10.1007/978-3-319-70407-4_36.

**Péter, Róbert** / **Szántó, Zsolt** / **Seres, József** / **Bilicki, Vilmos** / **Berend, Gábor** (2020): "AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts", in Berend, Gábor / Gosztolya, Gábor / Vincze, Veronika (eds.): *XVI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, JATEpress: 43–55.

**Stone, Sue** (1982): "Humanities scholars: Information needs and uses", in: *Journal of Documentation* 38, 4: 292–313. DOI: 10.1108/eb026734.

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

**Umerle, Tomasz** / **Colavizza, Giovanni** / **Herden, Elźbieta** / **Jagersma, Rindert** / **Király, Péter** / **Koper, Beata** / **Lahti, Leo** / **Lindemann, David** / **Łubocki, Jakub Maciej** / **Malínek, Vojtěch** / **Milanova, Alexandra** / **Péter, Róbert** / **Rißler-Pipka, Nanette** / **Romanello, Matteo** / **Roszkowski, Marcin** / **Siwecka, Dorota** / **Tolonen, Mikko** / **Vimr, Ondřej** (2022): *An Analysis of the Current Bibliographical Data Landscape in the Humanities. A Case for the Joint Bibliodata Agendas of Public Stakeholders.* Zenodo. DOI: 10.5281/zenodo.6559857 .

**Van Veen, Theo** (2019): „Wikidata: From "an" Identifier to "the" Identifier", in: *Information Technology and Libraries* 38, 2: 72–81. DOI: 10.6017/ital.v38i2.10886.

**Vlassenroot, Eveline** / **Chambers, Sally** / **Lieber, Sven** / **Michel, Alejandra** / **Geeraert, Friedel** / **Pranger, Jessica** / **Birkholz, Julie** / **Mechant, Peter** (2021): "Web-archiving and social

media: An exploratory analysis", in: *International Journal of Digital Humanities* 2: 107–128. DOI: 10.1007/s42803-021-00036-1.

**Vrandečić, Denny** / **Krötzsch, Markus** (2014): "Wikidata: A Free Collaborative Knowledge Base", in *Communications of the ACM* 57: 78–85 https://research.google/pubs/pub42240/ [30.10.2022].