BIBLIOGRAPHICAL DATA WORKING GROUP, DARIAH-ERIC

Workshop

Metadata-based Research in Arts, Humanities and Social Sciences

This interdisciplinary research meeting is aimed at bringing together researchers from different fields of metadata-based research to share their early stage ideas for research projects and investigations. The organisers hope that such a meeting will facilitate creation of joint research questions which could only be answered through collaboration between researchers coming from different backgrounds and possessing unique skills and competences. This two day hybrid event aims to form a new community of people interested in collaboration on large scale research projects which aim to provide a better understanding of European cultures and societies through metadata-based approaches.

March 4–5, 2025

— Na Florenci 1420/3, Prague, Upper Conference Hall + Zoom (ID: 829 3744 7315, Passcode: 377838)

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

BIBLIOGRAPHICAL DATA WORKING GROUP

Czech Literary Bibliography Research Infrastructure

Ústav pro českou literaturu AV ČR
Institute of Czech Literature of the CAS

Programme /1

Day 1 — March 4

10:00 – 11:45 / Session I – Metadata in SSH

Session Chair: *Vojtěch Malínek*

Opening words

DARIAH ERIC Bibliographical Data Working Group, Its Activities and Projects
Vojtěch Malínek *(Czech Academy of Sciences, Institute of Czech Literature, Czech Republic)* – ***Tomasz Umerle*** *(Polish Academy of Sciences, Institute of Literary Research, Poland)*

Descriptive Metadata: Investigating How It Was Made and Why That Matters
James Baker *(University of Southampton, United Kingdom)*

Network Analysis and Archival Collections
Martin Grandjean *(Université de Lausanne, Switzerland)*

Computational analyses of cultural production
Leo Lahti *(University of Turku, Finland)*

Administrative Issues

Lunch

13:30 – 14:30 / Session II – Investigating Arts Metadata

Session Chair: *Ondřej Vimr*

Exploratory Metadata Analysis and Film Historiographical Practices
Sarah-Mai Dang *(Philipps-Universität Marburg, Germany)*

Metadata in Music Search and Discovery
James Rhys Edwards *(Sinus-Institut, Germany)*

Break

15:00 – 16:00 / Session III – Registries and Tools for SSH Data Research

Session Chair: *Mikko Tolonen*

Identifying Learned Societies in Research Organisation Registry (ROR): Towards a Global Registry
Janne Pölönen *(Federation of Finnish Learned Societies, Finland)*

Exploring New Tools for SSH Research: FASCA and GRAPHIA Horizon Projects Starting in 2025
Tomasz Umerle *(Polish Academy of Sciences, Institute of Literary Research, Poland)*

Day 2 — March 5

10:00 – 11:30 / Session IV – Visualisations and Translations

Session Chair: *Tomasz Umerle*

Opening words

Metadata Enrichment and Analysis with the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) Multilingual Research Tool
Róbert Péter *(University of Szeged, Hungary)*

"Incoherent, Incomplete, and Inaccurate. Visualizing Bibliographic Data of Historical Translators"
Philipp Hofeneder *(University of Graz, Austria)*

Patterns of Translations
Péter Király *(GWDG Göttingen, Germany) –* ***András Kiséry*** *(The City College of New York, USA)*

Lunch

13:30 – 14:30 / Session V – Metadata on Book Studies

Session Chair: *Róbert Péter*

Distinguishing Works, Editions and Collections in Eighteenth-Century Books through Metadata and Text Reuse
Mikko Tolonen *(University of Helsinki, Finland)*

Exporting Children's Non-Fiction: Global Dissemination Patterns of Illustrated Czech Books for Ages 3–12
Ondřej Vimr – Charlotte Panušková *(Czech Academy of Sciences, Institute of Czech Literature, Czech Republic) –* ***Terry Klaban*** *(Charles University, Faculty of Social Sciences, Czech Republic)*

Break

15:00 – 16:00 / Session VI – Literary Metadata

Session Chair: *Péter Király*

Metadata and Derived Data in Computational Literary Studies: The GOLEM Case
Xiaoyan Yang *(University of Groningen, Netherlands)*

Exploring the Role of Digital Social Reading Platforms in Contemporary Literary Culture
Charlotte Panušková *(Czech Academy of Sciences, Institute of Czech Literature, Czech Republic)*

Abstracts /1

Descriptive Metadata: Investigating How It Was Made and Why That Matters

James Baker *(University of Southampton, United Kingdom)*

Collection metadata is made by people working in particular circumstances, at particular times, and in particular places. That metadata has an important role to play in research that seeks to understand and interpret collections. It can also be the subject of research itself. Knowing how our metadata was made can change how we use our metadata. In this talk, James Baker will describe his research into one early twentieth-century cataloguer, how their style was transmitted over space and time, and how this work opens up future directions for research into the plurality and multivocality of catalogue data.

Exploratory Metadata Analysis and Film Historiographical Practices

Sarah-Mai Dang *(Philipps-Universität Marburg, Germany)*

In this presentation, I explore how digital data visualizations can serve as a critical and self-reflective tool in film historical research. Drawing on feminist film theory and critical data studies, I demonstrate how metadata-driven research can foreground the situatedness of knowledge as well as archival and curatorial practices. Rather than presenting information as an objective reality, I use data visualizations to question universalization and essentialization and draw attention to the particularities, gaps and absences in data collections. Using case studies from my BMBF research group "Aesthetics of Access: Visualizing Research Data on Women in Film History" (DAVIF) (2021–2025), I will discuss both the possibilities and limitations of metadata-driven approaches in film and media studies.

Metadata in Music Search and Discovery

James Rhys Edwards *(Sinus-Institut, Germany)*

Music search and discovery – whether by humans in music libraries or by algorithmic recommender systems on streaming platforms – relies heavily on metadata. If a given musical artist's or track's metadata is not comprehensive and correct, their discoverability will inevitably suffer. The platformisation of both the commercial music sector and the music heritage sector exacerbates this risk. In the commercial sector, incorrect metadata can impede the proper payout of royalties by streaming platforms to copyright holders, whereas in the heritage sector, incorrect metadata can impede interoperability between national music information centres and international curatorial platforms like Europeana and the European Cultural Heritage Cloud. However, metadata literacy – let alone metadata management competences – are still rare among artists and other stakeholders. The research project Open Music Europe (https://openmuse.eu/ [Horizon Europe grant no. 101095295]) seeks to close this gap by developing resources and tools aimed at music stakeholders: an open-source dataspace architecture, and open-source data and metadata management tools that enable data sharing and exchange within this architecture. Our pilot project is the Slovak Comprehensive Music Database, which harmonises music

metadata held by the Slovak Music Centre and Slovak National Library, the rights management organisation SOZA, and global data systems like Wikipedia and Spotify. In this meeting of the Working Group, we will focus on the challenge of named entity disambiguation within the pilot project and the transferability of the solutions piloted to other European contexts.

Network Analysis and Archival Collections

Martin Grandjean *(Université de Lausanne, Switzerland)*

Historical network analysis is most often based on data that enable us to reconstruct the social network of individuals, their links with institutions, etc. But a network can also be extracted from the metadata of the archive itself: what can the structure of actors identified in a series of documents tell us about the bureaucracy of an organization? This presentation will show the different ways in which network analysis can be applied to historical sources, and then detail a case study which shows that the network doesn‘t necessarily have to be “social” to provide important information about the structure of an archive collection of its institution.

“Incoherent, Incomplete, and Inaccurate. Visualizing Bibliographic Data of Historical Translators”

Philipp Hofeneder *(University of Graz, Austria)*

This presentation addresses the challenges in mapping translation history, focusing on the spatial and relational complexities of translators and their works. Bibliographical data of historical translators often lacks clarity, with ambiguous information about locations and movements. I would like to explore the dynamic interplay between localization and temporal development, emphasizing relational spaces over absolute positions. Visualizations in this sense serve not as a display for results but as a starting point for further research, highlighting the subjective nature of interpretation. By examining historical geographies and migration patterns, this work seeks to redefine how translation history is visualized and understood, fostering new discussions and insights.

Patterns of Translations

Péter Király *(GWDG Göttingen, Germany) –*
András Kiséry *(The City College of New York, USA)*

What Hungarian literary works were translated, into what languages, and when? What changes, what tendencies, what patterns can we see over these 210 years? What do these patterns reveal about Hungarian literature—and, perhaps most importantly: What do the metadata of translations of Hungarian works suggest about the world system of translations? These questions led to a qualitative research project about the general patterns of translation. Our project aims to analyze data about the circulation of Hungarian literature beyond its original linguistic and cultural context, and reveal the various “translationscapes” of Hungarian literature, that is: show what Hungarian literature looks like, what authors are visible and what texts are important from the perspectives of various languages. But our ambition goes beyond this. We want to know

not only what works were successful, when, and where: we also want to figure out what this reveals about how literature—not just Hungarian literature, but any translated literature—circulates among languages, countries, and political systems. Hungarian is a small literature, both in terms of the sheer size of the corpus of texts and in terms of its reach, however, it passed through channels and mechanisms similar to, or identical with, the channels and mechanisms through which any literary translation circulates. One of our aims is to trace the historical evolution of the world system of literary translations.
In the presentation I will show some results based on Hungarian and international data, and I also talk about how to detect translations in a library catalogues.

Computational analyses of cultural production

Leo Lahti *(University of Turku, Finland)*

Quantitative analyses of cultural production can benefit from computational integration of interlinked metadata collections. Data streams from statistical authorities, surveys, and data observatories, are increasingly integrated into reproducible research workflows. The complexity of the data and algorithms in data-intensive mixed methods research, and the tensions between open and closed resources are challenging transparency and reuse, however. This talk will discuss solutions based on collaborative open source development in our recent studies on data-driven analyses of cultural production of literature and music, building on open digital resources retrieved from European and national statistical authorities and memory organizations.

Exploring the Role of Digital Social Reading Platforms in Contemporary Literary Culture

Charlotte Panušková *(Czech Academy of Sciences, Institute of Czech Literature, Czech Republic)*

This presentation introduces a dataset acquired from a Czech digital social reading platform, a local alternative to the more well-known Goodreads. Digital social reading platforms have been a rich yet often overlooked source of data on reading communities. They offer unique insights into reading habits that were previously studied using limited groups through questionnaires and interviews. Existing research on Goodreads and other digital social reading sites has demonstrated the potential of these platforms for analyzing reading preferences, genre eclecticism, story-world absorption through reviews, and the correlation between comments and ratings. This presentation will discuss the dataset's structure, its potential applications in literary and social research, and the broader implications of digital reading platforms in understanding contemporary reading behavior.

Metadata Enrichment and Analysis with the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) Multilingual Research Tool

Róbert Péter *(University of Szeged, Hungary)*

This presentation introduces the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) multilingual research tool that enables researchers to critically analyse bibliographic data and texts at scale using data-driven methods supported by Natural Language Processing (NLP) techniques. AVOBMAT offers a range of dynamic text and data mining tasks with interactive parameter tuning and control, from preprocessing to analysis. Its analytical and visualisation tools facilitate both close and distant reading of texts and bibliographic data. Currently, it supports 178 metadata fields.
The presentation focuses on how AVOBMAT preprocesses, enriches, analyses, and visualises metadata. Through case studies, it demonstrates how the tool helps users critically engage with texts and metadata. AVOBMAT's key features include: (i) the use of transformer language models on a scalable, cloud-based infrastructure that allows researchers to preprocess and analyse texts and metadata at scale; (ii) it combines metadata and textual analysis, enabling users to ask complex research questions in one integrated, interactive, and user-friendly web application; (iii) it analyses and enriches texts and metadata in 16 languages; and (iv) private databases can be made public.

Identifying Learned Societies in Research Organisation Registry (ROR): Towards a Global Registry

Janne Pölönen *(Federation of Finnish Learned Societies, Finland)*

Learned societies are key to advancing academic disciplines yet remain underrepresented in open infrastructures like ROR. This presentation introduces a novel methodology for identifying learned societies within ROR based on their publishing activities and organizational names within ROR, addressing a critical gap in research metadata (Kulczycki et al., 2024). We propose construction of a global registry of learned societies using ROR identifiers to improve visibility and inclusivity in research metadata, fostering open science. Such a registry would provide a publicly accessible and editable resource for identifying learned societies for research and practical purposes.

Distinguishing Works, Editions and Collections in Eighteenth-Century Books through Metadata and Text Reuse

Mikko Tolonen *(University of Helsinki, Finland)*

This work-in-progress highlights an approach for integrating metadata-driven analyses with large-scale text reuse to differentiate between "works", "editions" and "collections" in eighteenth-century bibliodata. Building on our earlier algorithmic grouping of English Short Title Catalogue (ESTC) mainly by title and author, we now incorporate text-overlap information from Eighteenth Century Collections Online (ECCO). By systematically detecting overlapping passages—particularly

in cases where multiple works by the same (or different) author are compiled into new collections—we generate enriched metadata that clarifies how texts of different size were bundled, reprinted or revised. This enhanced framework addresses known shortcomings of purely metadata-based methods and opens new avenues for historical research, enabling scholars to switch between these refined categories when studying publishing practices, authorship and readership in eighteenth-century Britain. The idea is not to create metadata that is necessarily fit for cataloguing purposes but something that researchers can put to practice.

Exploring New Tools for SSH Research: FASCA and GRAPHIA Horizon Projects Starting in 2025

Tomasz Umerle *(Polish Academy of Sciences, Institute of Literary Research, Poland)*

This presentation will introduce you to new tools for SSH research developed within the FASCA and GRAPHIA projects. FASCA focuses on enhancing the GoTriple platform by developing data science support tools for in-depth analysis and interpretation of research data. GRAPHIA, on the other hand, is building an advanced SSH knowledge graph (SSH KG), powered by AI tools, next-generation research instruments, and an SSH citation index. With FASCA, you'll be able to take advantage of improved GoTriple features, such as an advanced API query system, an extended version of the Pundit annotation tool, and a data science workspace that supports FAIR-compliant research publishing. GRAPHIA will provide access to a comprehensive set of APIs, enabling integration with analytical applications and custom research workflows. We invite you to get involved by joining FASCA's open calls for pilot projects, which offer support from a data science team. GRAPHIA also welcomes researchers to contribute to the development of the SSH knowledge graph (SSH KG) and test new AI-driven tools through workshops and innovation prototyping labs (e.g. IPLs). Your participation will help shape the future of SSH research by improving access to, understanding of, and reuse of research data.

Exporting Children's Non-Fiction: Global Dissemination Patterns of Illustrated Czech Books for Ages 3–12

Ondřej Vimr – Charlotte Panušková *(Czech Academy of Sciences, Institute of Czech Literature, Czech Republic) –* ***Terry Klaban*** *(Charles University, Faculty of Social Sciences, Czech Republic)*

This paper introduces a collaborative research project at the intersection of bibliographical data science, publishing, and translation studies. Partnering with an industry stakeholder, the project examines the global sales and dissemination patterns of Czech children's non-fiction books in 2014–2024. The industry partner provides comprehensive data while seeking deeper insights into market trends. Key questions include: What specific insights is the industry partner seeking? Are their expectations aligned with the realities of bibliographical, publishing, and translation studies? What analytical approaches and research methodologies (quantitative, qualitative, or mixed methods) can this dataset support?

Rather than providing definitive answers, this paper aims to open a discussion on the potential and challenges of such an inquiry.

Metadata and Derived Data in Computational Literary Studies: The GOLEM Case

Xiaoyan Yang *(University of Groningen, Netherlands)*

This presentation introduces the GOLEM project ("Graphs and Ontologies for Literary Evolution Models"), a large-scale initiative to create a graph database of online fiction corpora in six languages (English, Spanish, Italian, Indonesian, Korean, and Chinese). GOLEM focuses on capturing textual features such as narrative elements, stylistic traits, and reader responses (e.g., likes, comments) to enable comparative analysis without accessing full texts. The project leverages programmable corpora and SPARQL endpoints to facilitate reusable analytical pipelines, ensuring compatibility with existing ontologies like Wikidata and MiMoText. By analyzing metadata and derived data from millions of fanfiction stories published between 2000 and 2022, GOLEM aims to provide insights into the evolution of fiction writing and reader engagement.