MaDaBi: Building the Mannheim data bibliography with and without AI for agentic scholarly discovery

Abstract

A growing portion of research outputs resides in datasets, software, and replication packages that remain scattered across institutional, domain-specific, and general-purpose repositories. This paper presents our ongoing efforts to build the Mannheim Data Bibliography (MaDaBi) with and without AI for agentic scholarly discovery, and outlines key challenges along with our recommendations for other institutional data infrastructures.

Keywords

Data bibliography, datagraphy, FAIR, metadata extraction, metadata harvesting

1. Introduction

University libraries maintain services that manage metadata for research outputs, including publication servers and data repositories. These systems form a strong foundation for open science and research data management (RDM). However, only a small portion of the research data is deposited in institutional data repositories, limiting opportunities for RDM monitoring and findability of the data. To improve this at the University of Mannheim, we build the Mannheim data bibliography (MaDaBi) or datagraphy, i.e., a registry of metadata for datasets, software, and replication packages created or collected by the members of the University of Mannheim. We aim to improve FAIRness of data (Findable, Accessbile, Interoperable, and Reusable) and enable new metrics for data sharing and reuse. Our long-term goal is to provide our users AI-based chatbot for scholarly discovery across publications, datasets, software, and replication packages. Here, we describe our ongoing efforts.

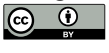
2. Building datagraphy without AI via metadata harvesting

The first challenge in building a datagraphy is defining a scope of resource types and repositories. It is a non-trivial problem at a university with many faculties and research directions. We document the desired scopes in the README-file of the MaDaBi GitHub repository https://github.com/UB-Mannheim/madabi. The metadata of the corresponding resources are harvested from the institutional **Ma**nnheim **Data** Repository MADATA as well as from the external repositories including Zenodo, GESIS Data Archive, and Harvard Dataverse. The whole workflow runs automatically on GitHub Actions every Sunday. We also developed the Mannheim data bibliography dashboard with various data metrics deployed on GitHub Pages and available at https://ub-mannheim.github.io/madabi. Our next step is to extend the range of external repositories included in the harvesting process, following the documented scopes.

3. Building datagraphy with AI via metadata extraction

We also use large language models to complement the process by extracting the metadata for datasets directly from open-access publications. Full texts of openly available publications from the **Ma**nnheim Electronic **Doc**umentation Server MADOC are processed to identify dataset mentions, availability statements, repository links, and data citations. This introduces a second opportunity to collect metadata for data beyond external harvesting and helps to capture data that may otherwise remain unlinked. It also provides insights into how researchers at Mannheim describe and cite their data, supporting

BESIDES: Beyond Search Engines-Intelligent Knowledge Discovery From Scholarly Publications. The ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) December 15-19, 2025



institutional monitoring of research data management and open-science practices. Our data collection, processing, and extraction pipelines are openly available at the MaDaBi GitHub repository.

4. Creating an agentic chatbot for metadata discovery

Separately from the described work, the Mannheim University Library is developing an agentic AI chatbot using openly accessible library webpages. Our long-term objective is to add specialized agents that operate on scholarly metadata from MADOC (publications) and MADATA (research data) as well.

As a first step, we implemented a prototype that performs retrieval-augmented answering over MADATA records. The prototype harvests OAI-PMH metadata, normalizes fields (titles, creators, subjects, and identifiers), and constructs an in-memory property graph. Nodes represent datasets, persons, and subjects. For retrieval, we combine a sparse TF–IDF index with a dense multilingual open-weight sentence-embedding model "paraphrase-multilingual-MiniLM-L12-v2". We test the command-line prototype using the "gpt-oss:20b" model running locally via the Ollama framework.

5. Challenges and Recommendations

Developing an institutional data bibliography raises both technical and organizational challenges. On the technical side, data and software often lack persistent identifiers and structured metadata. The availability statements in publications are not standardized. On the organizational side, sustainable operation requires long-term maintenance of harvesting pipelines, manual checks of metadata quality, and incentives for researchers to deposit and document their data. Institutions planning to build a data bibliography should begin by (1) defining a clear scope of resource types and repositories, (2) prioritizing metadata sources with open and documented APIs, (3) promoting the best practices of writing data/code availability statements among researchers.

6. Conclusions

Our MaDaBi efforts show how automated harvesting and AI-assisted metadata extraction can be combined to create an institutional data bibliography and a simple datagraphy-dashboard. Together with the prototype chatbot for metadata exploration, this forms a practical basis for future agentic scholarly discovery based on verifiable metadata. The MaDaBi scripts are open source.

Data and code availability statements

The MaDaBi source code is openly available at https://github.com/UB-Mannheim/madabi under the MIT license. The harvested metadata are released under the CC0 license in the same repository. The original publications used for metadata extraction are openly available at https://madoc.bib.uni-mannheim.de. All other materials in the GitHub repository are provided under the CC BY license.

Acknowledgments

This work was partially funded by the Federal Ministry of Research, Technology, and Space (BMFTR), Project TransforMA (13IHS264B): https://transfor-ma.de.

Declaration on Generative AI

The authors used ChatGPT-5 to draft content, paraphrase and rewrite, improve writing style, check grammar and spelling, and enhance content. After using this tool, the authors reviewed and edited the content and take full responsibility for the content of this publication.