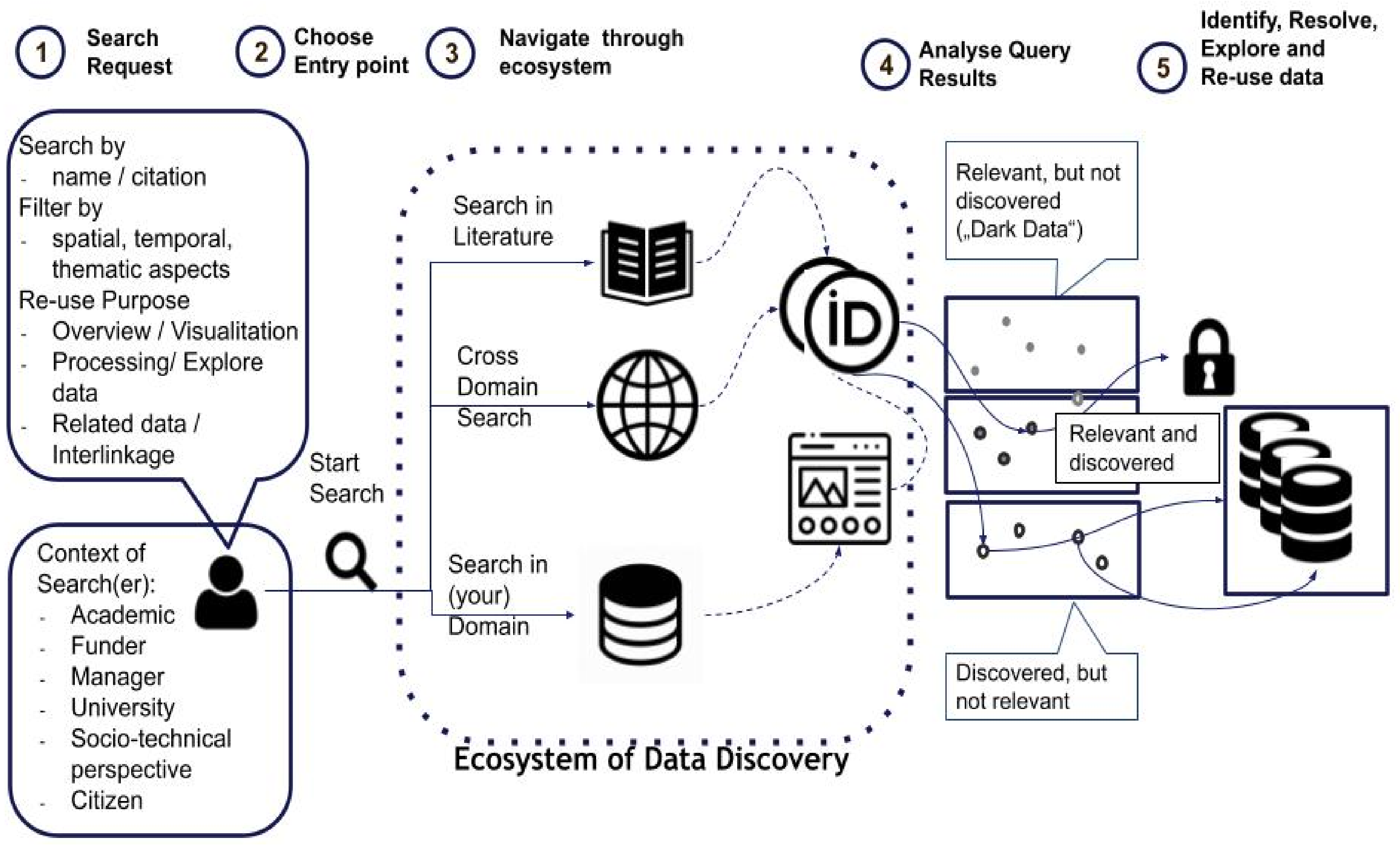
Search and Harvesting across NFDI Consortia - Gaps and Challenges

Brigitte Mathiak [1], Heinrich Widmann [2], Gerhard Heyer [3], Christin Henzen [4], Andreas Czerniak[5]

Search and harvesting use cases on harmonized metadata play an important role in several NFDI consortia (see Meta(data), Terminology and Provenance section concept). The working group Search and Harvesting works on a common understanding of user requirements (for search) and service requirements (for harvesting), analysis of the data sources landscape, and recommendations - with respect to common and specific needs, e.g., for spatial or sensitive data. On this poster, we present as our first outcome an overview on identified and structured search and harvesting gaps and challenges across NFDI consortia, which fosters a common understanding of a multidisciplinary vision for search & harvesting solutions.

Data Search

To identify search and harvesting gaps and challenges, we analyzed the search process. The users' search begins by finding and choosing an appropriate entry point (Figure 1, No. 2) based on the given search request (No. 1). They need to decide on using a non-disciplinary, a multidisciplinary, or a disciplinary search service or starting from a literature review. Regardless the chosen entry point and navigation through the search ecosystem (No. 3), the users should end up on a landing page of a dataset, which allows to assess the dataset's relevance based on the provided metadata, and, in a best-case scenario, provides a download option (No. 4 and 5).



1 Search Request

2 Choose Entry point

3 Navigate through ecosystem

4 Analyse Query Results

5 Identify, Resolve, Explore and Re-use data

Search by
- name / citation

Filter by
- spatial, temporal, thematic aspects

Re-use Purpose
- Overview / Visualitation
- Processing/ Explore data
- Related data / Interlinkage

Context of Search(er):
- Academic
- Funder
- Manager
- University
- Socio-technical perspective
- Citizen

Start Search

Search in Literature

Cross Domain Search

Search in (your) Domain

Ecosystem of Data Discovery

Relevant, but not discovered („Dark Data")

Relevant and discovered

Discovered, but not relevant

Figure 1: Search steps. Adapted from https://doi.org/10.5281/zenodo.7468089

Gaps and Challenges

We collected gaps and challenges based on our own experiences and derived from the literature. By summarizing and structuring, we derived three main categories: users, metadata, and technology.

Users

- **Inexperienced users:** For literature search, researchers usually get a foundational introduction during their education. This basic knowledge is then deepened during the research career through frequent use. However, for data search, there is still a lack of proper education and training materials.
- **Inappropriate and invisible entry points:** Unlike for literature, users are often not aware of relevant entry points for their data search. Instead, they often rely on 1) general web search or 2) literature review to find names of datasets and then try to find them on the Web. As a result, datasets don't get found, especially when they are published in disciplinary data repositories with a small number of provided datasets. However, common Web-based search engines lack specific search and filter options. To increase the visibility and the discoverability, entry points should be offered with faceted search filters as well as with search user interfaces adapted to the research domain and needs, for example filtering spatial coverage via map-based UI for georeferenced data.
- **Missing feedback mechanism and detailed user statistics:** Providers of search and discovery services often don't know much about their users, users' needs and issues that occur by using their services. This is due the lack of easy-to-use or automated feedback mechanisms for both - for the relevance/suitability of discovered data and for the provided functionality, e.g., where users report missing/needed features. Especially information on suitable data could be used to improve the provided search functionality, but can hardly be collected systematically for existing services.

CoRDI
Conference on Research Data Infrastructure

NFDI WG Search & Harvesting
in NFDI section (Meta)data, Terminology, Provenance
Mailing list: section-metadata-wg-search@lists.nfdi.de
charter: https://doi.org/10.5281/zenodo.6770763

Metadata

- **Missing links:** While approaches on linking different resource types on a metadata level, e.g. datasets and related Web services, are well-known, search and& harvesting services of the overall ecosystem are often not linked to each other or links are not visible/usable in the user interfaces. However, research data needs context information for re-use and machine-actionable processing, e.g. from documentation, relevant research articles or tools. Thus, missing links address interoperability and adequate UI design challenges. An approach to overcome this issue would be a persistent identifier (PID) graph service that allows interlinking resources across disciplines and data sources by using PIDs.
- **Granularity issues:** "Common" metadata properties such as *title, description, author*, etc. (see for instance the Dublin Core schema) are not sufficient for many disciplines, and the granularity level of metadata for a dataset is likewise not sufficient for an efficient search. For some disciplines, therefore, several metadata schemas have been developed that allow to reference specific information, and to describe fine-grained data in a discipline-specific hierarchy of metadata. In the context of search and harvesting, however, it still needs to be investigated to what extent such fine-grained metadata can be merged, or inherited. As it still needs major efforts to provide harmonized metadata across the NFDI consortia, common definitions for resource types or the description of spatial or temporal information can be used as a starting point to develop strategies that foster search and harvesting for fine-grained/detailed metadata.
- **Provenance:** The provision of provenance information within metadata is essential for the reuse of data and the evaluation of the data quality. However, it is still often lacking or partly lacking as some metadata schemas only offer options to describe one part of the provenance chain, such as one processing step. Moreover, when input data and final data products are published in different sources, we still lack services that can gather, summarize and visualize provenance information across those different services.

Technology

- **Missing information retrieval concepts for multidisciplinary datasets:** Unlike for literature, there is not much research on how information retrieval for dataset discovery works for multidisciplinary use cases. This is despite the fact that multidisciplinary research data retrieval is complex and needs context information that is typically distributed across different sources.
- **Harmonization / mapping of results is missing:** When harvesting dataset metadata from different sources, a dataset that is registered in several repositories should only be listed once in the result list. Mapping/harmonization can be performed easily, if the dataset is registered with the same identifier or name. However, that is not always the case, in particular for related datasets that are only linked from provenance graphs.
- **Open questions on the technology level include:** How can we rank datasets for multidisciplinary use cases? How can we show links to additional materials to provide context in an effective and efficient manner addressing multidisciplinary user needs? How can we effectively remove duplicates when harvesting different sources? How can we measure that provided search and harvesting services address user needs? How can we get more information on how many datasets are underutilized and how often searches fail and why?

Conclusion

Our working group identified gaps and challenges for search and harvesting. That helps us to identify future activities and target groups. To improve search and harvesting concepts and services across NFDI consortia, we need comprehensive strategies for data providers to provide high-quality metadata for data users to provide their needs, to give feedback on search results and evaluated data, and for repository providers to provide suitable functionalities across repositories. As a starting point for the next activities, we envision to provide recommendations for the above-mentioned roles.

[1] GESIS - Leibniz Institute for the Social Sciences [2] DKRZ - Deutsches Klimarechenzentrum [3] SAW - Sächsische Akademie der Wissenschaft [4] Technische Universität Dresden [5] Bielefeld University Library, Germany