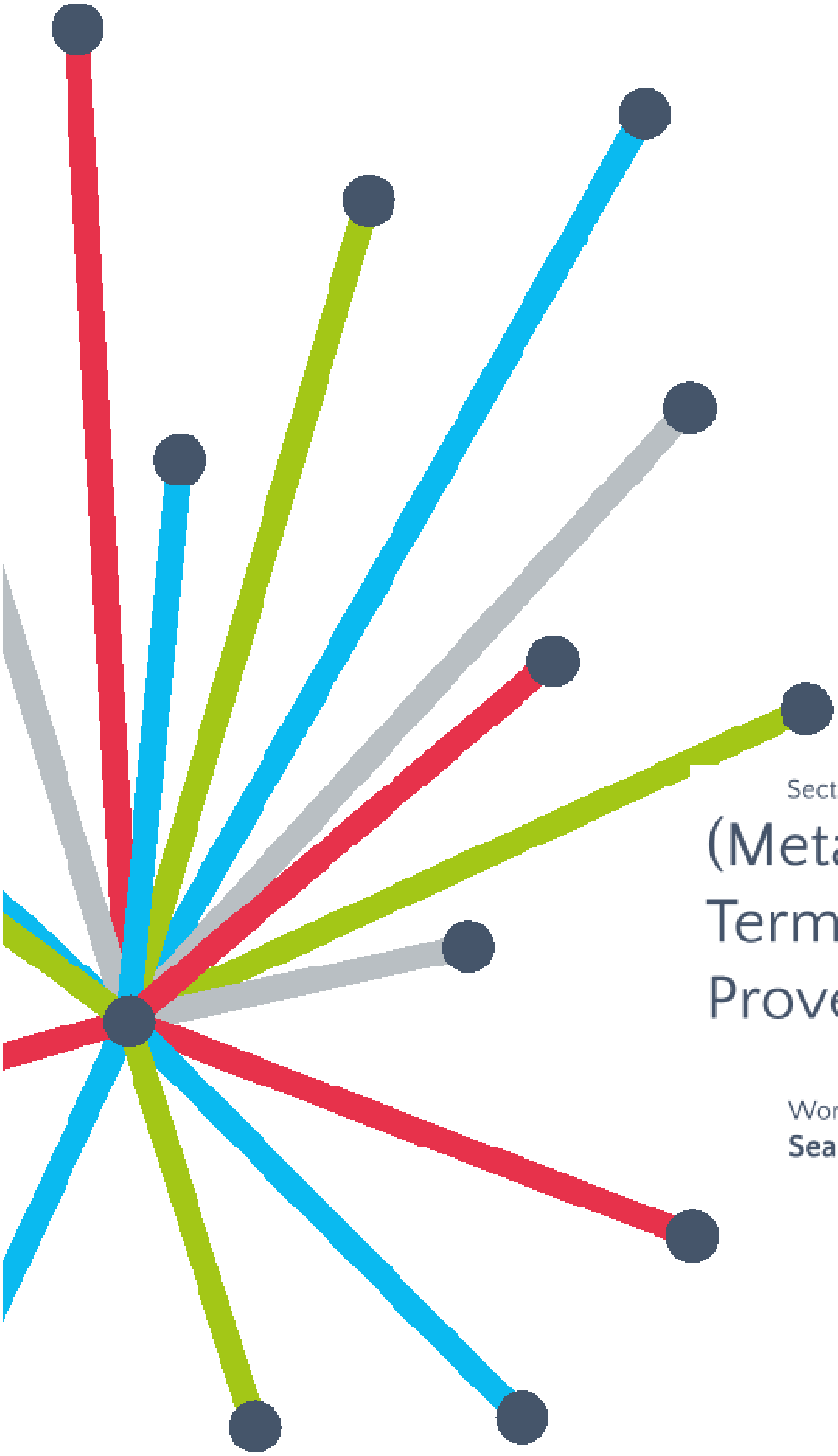
Nationale
Forschungsdaten
Infrastruktur



Section

(Meta)data, Terminology, Provenance

Working Group Charter

Search and Harvesting

Name of the working group

Search and Harvesting

Acronym

Metadata-search

Coordinators

Brigitte Mathiak, Heinrich Widmann

Authors

Brigitte Mathiak, GESIS - Leibniz-Institut für Sozialwissenschaften, 0000-0003-1793-9615

Heinrich Widmann, Deutsches Klimarechenzentrum, 0000-0001-9871-2687

Luca Ghiringhelli, Humboldt University of Berlin, 0000-0001-5099-3029

Holger Israel, Physikalisch-Technische Bundesanstalt, 0000-0002-3045-4412

Fidan Limani, Leibniz Information Centre for Economics (ZBW), 0000-0002-5835-2784

Christin Henzen, Technische Universität Dresden, 0000-0002-5181-4368

Gerhard Heyer, Sächsische Akademie der Wissenschaften (SAW), 0000-0002-0442-392X

DOI: 10.5281/zenodo.6770763

Version 1.0.0, 2022-06-21

1. Motivation

Metadata plays an important role in many use cases relating to research data and modern research infrastructures. In this working group, we will be focussing on how to extract, find, and use (meta)data, both for discovery and consumption of data by researchers (search/discovery) and for aggregation and indexing by discovery services (harvesting). For both - search and harvesting - the actors can be humans or machines.

For search and harvesting queries to be efficient, i.e. for them to deliver the optimal result, some requirements on metadata have to be met. In general the metadata should be as FAIR as possible, at a minimum, data identifiers (PID/DOI) must be provided, and - depending on the use case - the metadata should have the highest possible degree of correctness, completeness, and interoperability that can be reached within the interested communities. However, what “complete” or “correct” means depends on the specific research questions in the related scientific communities, which cannot be known a priori. Hence, we need to strive for an optimum in flexibility on the one hand, while ensuring efficient search and harvesting of metadata on the other.

When considering the consortia’s requirements, we need to distinguish between the requirements from the content side (Comprehensive and rich data description, which scientific tasks need to be supported) and requirements from the IT side (Allow re-use of metadata for interoperable search and harvesting in distributed repositories).

The group’s work is important for ensuring a FAIR compliant usage of research data in modern research infrastructures while keeping the “great pictures” set out by the consortia and NFDI e.V.

2. Objectives

- Collection of search & harvesting related requirements from different consortia/institutions to give feedback to basic services
- Collection of best practices from consortia/institutions to derive transferrable “lessons learned”
- Collection of data sources and the technologies that are used for harvesting
- Development and provision of recommendations to harmonize harvesting services based on common standards
- Network with relevant groups and coordinate efforts (see adoption plan)

3. Work Plan

WP 1 Requirements for Search and Harvesting from both Service and User Perspective

Depending on the discipline-specific and multi-disciplinary challenges of the search-and-harvesting ecosystem, there are different requirements from within the different consortia concerning any NFDI-wide infrastructure. The goal of this work package is to collect these requirements and consolidate them. This will serve as a basis for the following work packages and potential basic services.

Task 1 Collect requirements from within the working group

The Search & Harvesting working group represents a wide array of consortia (see Initial Membership List and Adoption Plan). Our first step is therefore to collect and consolidate the requirements we know first-hand. This refers to both requirements that come from the communities as well as requirements that arise with the challenges of running search & harvesting services within a consortium.

Task 2 Collect and evaluate state-of-the-art

There are already many resources and groups that address search & harvesting related topics. In collaboration with the efforts of WP 5, we will collect and systematize these additional resources and add them to the requirements of Task 1.

Milestone 1: Report on requirements (an ever evolving draft can be found here: *https://drive.google.com/drive/folders/1w1gHe9k_J7bmZyCsgng0lfqeUcS6K1DR?usp=sharing*)

WP 2 Best Practices for Search and Harvesting

Based on WP 1, we will discuss Best Practices within our group, especially with regards to practices, which we have first-hand knowledge of. These will serve as a basis for recommendations made in WP 4.

Milestone 2: Collection on Best Practices (internal)

WP 3 Data Sources

Harvesting relies on knowing what is there and how to access it. Compiling a list of available resources is not only helpful to gain a better understanding of the current landscape of NFDI data sources, it also improves the visibility of these resources, by opening a dialogue with harvesters both within NFDI and outside of it as part of WP 5.

Task 1 Acquire information from the consortia

Many consortia have already compiled lists of their data sources, e.g. the Forschungsdatenzentren of KonsortSWD; others may need additional help to generate such lists. In both cases, the lists will then need to be merged and information relevant for harvesting, such as the used or not yet available but needed APIs or metadata standards and content, added.

Milestone 3: Snapshot of the current data sources landscape of NFDI consortia

WP 4 Harvesting recommendations

Based on the report of WP 1, the best practices of WP 2, and the snapshot of WP 3, we will derive recommendations for harvesting to be distributed among the data sources, and active harvesters of the consortia and beyond.

Milestone 4: Publish recommendations for harvesting

WP 5 Networking
Many other working groups in the sections have topics that directly and indirectly touch upon the topic of Search & Harvesting. We have identified some of them in the Adoption Plan below. We intend to invite speakers from these groups or visit them to promote collaboration. We will also distribute outcomes from this group via section channels and beyond.

4. Membership List

	Name	Institution	NFDI consortia
1	*Brigitte Mathiak (Co-Speaker)*	*GESIS*	*KonsortSWD*
2	*Heinrich Widmann (Co-Speaker)*	*DKRZ*	*NFDI4Earth*
3	*Gerhard Heyer*	*SAW*	*Text+*
4	*Jürgen Kett*	*DNB*	*Text+, 4Culture*
5	*Gerald Steilen*	*Verbundzentrale GBV*	*NFDI4Objects*
6	*Alexander Behr*	*TU Dortmund*	*NFDI4Cat*
7	*Swantje Dogunke*	*FSU Jena / ThULB*	
8	*Rachit Khare*	*TU Munich*	*NFDI4Cat*
9	*Akhil Patil*	*DLR-TS*	*NFDI4Ing*
10	*Fidan Limani*	*ZBW*	*KonsortSWD, BERD@NFDI, NFDI4DS*
11	*Luca Ghiringhelli*	*Humboldt Berlin*	*FAIRmat*
12	*Dirk Weisbrod*	*DIPF/ VerbundFDB*	*Konsort SWD*
13	*Dorothea Iglezakis*	*Uni Stuttgart*	*NFDI4Ing, MaRDI*
14	*Sandra Göller*	*FIZ Karlsruhe*	*NFDI4Culture, NFDI4Chem*
15	*Holger Israel*	*PTB*	*(NFDI4Phys), Punch4NFDI, Daphne4NFDI, NFDI-MatWerk*
16	*Alexander Reis*	*Dt. Schifffahrtsmuseum*	
17	*Mehtap Özaslan*	*TU Braunschweig*	*NFDI4Cat*
18	*Oliver Bothe*	*Hereon*	
19	*Noriko Cassman*	*FSU Jena*	*NFDI4Microbiota*
20	*Peter Mutschke*	*GESIS*	*NFDI4DS, BERD, KonsortSWD*
21	*Stefan Dietze*	*GESIS & HHU*	*NFDI4DS, BERD*
22	*Johannes Darms*	*ZB MED*	*NFDI4Health*
23	*Atif Latif*	*ZBW*	*BERD@NFDI*
24	*Doris Jaeger*	*KIT*	
25	*Sonja Schimmler*	*Fraunhofer FOKUS*	*4Cat, 4DS*
26	*Rainer Stotzka*	*Karlsruhe Institute of Technology*	*NFDI4Ing, NFDI-MatWerk*
27	*Nadiia Huskova*	*HLRS*	*NFDI4Cat*
28	*Christin Henzen*	*TU Dresden*	*NFDI4Earth*
29	*Björn Schembera*	*University of Stuttgart*	*MaRDI*

30	*Ziyad*	*ZB Med*	*NFDI4Microbiota*
31	*Stephan Hachinger*	*Leibniz Supercomputing Centre (LRZ)*	*NFDI4Earth, NFDI4Ing, etc.*

5. Adoption Plan

The following consortia are represented in the group, directly contributing to the outputs and benefiting from them:

- *KonsortSWD*
- *NFDI4Earth*
- *Text+*
- *NFDI4Culture*
- *NFDI4Cat*
- *NFDI4ing*
- *BERD@NFDI*
- *FAIRmat*
- *MaRDI*
- *NFDI4Chem*
- *NFDI4Phys*
- *Punch4NFDI*
- *Daphne4NFDI*
- *NFDI-MatWerk*
- *NFDI4Microbiota*
- *NFDI4DataScience*
- *NFDI4Health*

6. Collaboration Plan

Several aspects - cross-cutting through different NFDI projects - affect the results of this WG. In an effort to adopt such relevant practices in our approach, we plan to establish, maintain, and update as necessary a collaboration process with relevant groups both within and outside of the NFDI. Since the relevance and relationship between our WG and these groups is mutual, this process will enable us to adopt recognized standards and practices set across NFDI WGs or external (to the NFDI) organizations, as well as provide findings from this WG to them. We next present the initial candidates for such a collaboration from NFDI projects and external initiatives.

Cross-links to other NFDI WGs:

- Research Software Engineering WG (CI)
- Data integration (CI)
- Ontology Harmonization and mapping (Meta)
- Semantic Interoperability & Terminology Services (Meta)
- Knowledge Graphs (Meta)
- Cookbook(s), Guidance and Best Practices (Meta)

Groups outside of the NFDI:

- RDA Data Discovery Paradigms IG – (B. Mathiak, H. Widmann member)
- RDA Data Granularity WG - B. Mathiak (Co-chair)

- RDA IG FAIR Digital Object Fabric - R. Stotzka (Co-chair)
- GoFAIR Discovery IN - B. Mathiak (Co-chair), H. Widmann member
- EOSC Taskforce Semantic Interoperability - (H. Widmann member)