The Dataset Finder: A Tool Utilizing Data Management Plans as a Key to Data Discoverability

RESEARCH PAPER

SOO-YON KIM
STEFFEN HILLEMACHER
MAX KOCHER
BERNHARD RUMPE
SANDRA GEISLER

*Author affiliations can be found in the back matter of this article

[KHK+24] S.- Y. Kim, S. Hillemacher, M. Kocher, B. Rumpe, S. Geisler:
The Dataset Finder: A Tool Utilizing Data Management Plans as a Key to Data Discoverability.
In: Data Science Journal (DSJ), Volume 23(1), pp. 1-17, Ubiquity Press, Dec. 2024.

ABSTRACT

In the past years, there has been an increased interest in sharing and reusing research data. While the importance of sharing data is urgent for enabling collaboration, many research projects are currently struggling with setting up a strategy and the right infrastructure for enabling such data-driven collaboration among the project's researchers. Through an analysis of the Cluster of Excellence project Internet of Production as a use case, we have found that to enable researchers to share and find research data, a suitable platform is needed, as well as processes that smoothly blend into existing research data management practices. We argue that leveraging data management plans from a medium of documentation to a dynamic knowledge source enhances overview and discoverability of data, while integrating easily into day-to-day workflows of researchers. We present a tool, the Dataset Finder, which is built on the basis of data management plans, and allows users to intuitively query available datasets. The current functionalities of the tools are discussed, results of a preliminary evaluation, as well as potential future features.

CORRESPONDING AUTHOR:
Soo-Yon Kim
Data Stream Management and Analysis, RWTH Aachen University, Aachen, Germany
kim@dbis.rwth-aachen.de

KEYWORDS:
Data Management Plans; Data Findability; Data Discoverability; Data Sharing; Data Reuse; Research Data Management

TO CITE THIS ARTICLE:
Kim, S.-Y., Hillemacher, S., Kocher, M., Rumpe, B. and Geisler, S. 2024 The Dataset Finder: A Tool Utilizing Data Management Plans as a Key to Data Discoverability. *Data Science Journal*, 23: 58, pp. 1–17. DOI: https://doi.org/10.5334/dsj-2024-058

1 INTRODUCTION

In the past years, there has been an increased interest in unsiloing collected data and sharing it across organizations. Large-scale research projects are an example of such settings, where it is assumed that the reproducibility and hence the quality of research, as well as interdisciplinary collaboration opportunities, will be improved through sharing and reusing research data. While sharing data is gaining importance in the scientific domain, currently, many research projects are struggling with setting up a strategy and the right infrastructure for enabling data-driven collaboration among the project's researchers.

To enable data-driven collaborative research, both data owners' and data consumers' needs with regard to sharing and reusing data, respectively, need to be addressed. Data owners need to be supported in their efforts of making their research data visible, while data consumers need to be able to find and understand other's datasets. Wilkinson et al. (2016) highlight, as one aspect of findability, that the (meta)data must be discoverable, i.e., recorded in a searchable registry. Therefore, at least the metadata of the research data should be indexed on one or multiple selected repositories. Taking large-scale research projects as an example, there are several challenges complicating the establishing of a suitable repository and accompanying structures for working with it:

> (C1) Firstly, large-scale projects are very often interdisciplinary. Between disciplines, research methods and research data types vary, such that selecting the most suitable repository or repositories in terms of discipline-specific needs can be highly tedious already.
>
> (C2) Secondly, typically, a variety of institutes work together. The introduction of new research data tools can be disruptive and may lead to duplicate work for researchers and increases the risk of inconsistencies, as their institutions already have established infrastructures and cultures for managing their research data.
>
> (C3) Thirdly, there may be a large variance in experience, knowledge, and openness towards research data management practices among researchers. Repositories and structures requiring a-priori knowledge to meaningfully work with them may hinder a large number of researchers engaging with them. This can lead to a situation where, as data owners face such challenges, many datasets of the project are not being recorded in the first place, or where data consumers are not able to locate data in a low-barrier way.
>
> (C4) Fourthly, the ratio in large-scale projects of data stewards to the researchers to be guided leads to a limitation of supporting and management capacities.

From these challenges, we have derived that a suitable solution must make sure (C1) that it is generally applicable across disciplines; (C2) that it avoids disruptions with and possibly leverages existing institutional solutions; (C3) that it is generally user-friendly and poses as few requirements to users as possible; and (C4) that it runs on a basis that requires little individual support.

To employ the above requirements, we have developed an approach that makes use of *Data Management Plans* (DMPs), and treats them as a queriable database for a large-scale project's research data's metadata. In this paper's use case, all researchers are mandated to create DMPs for their projects with the DMP tool *Research Data Management Organiser* (RDMO)[1] (Anders et al., 2024) using a specifically developed DMP template (Kim et al., 2023). The data stewards have created structures in RDMO that allow them to access all created DMPs. We aim at listing each dataset described by a DMP automatically in a registry of research data, the so-called Dataset Finder. All researchers will be able to access and search this platform, thus enabling the discoverability of research data. The approach's advantages are (C1) that the DMP questionnaire has been developed to be generally applicable across the disciplines of the project; (C2) that RDMO is a descriptive tool where no actual data is managed, hence, it is complementary to tools within institutions; (C3) that the questionnaire works with natural language, help texts are provided, and the design and user interface of the Dataset Finder can be customized, such that the barriers for researchers to work with it are low; and (C4) as

1 RDMO, https://rdmorganiser.github.io/en/.

RDMO offers an API, there is the possibility for data stewards to automatically read out entries from the DMPs, enabling the automatic creation of an accumulated list of all existing datasets, which can then be made available to the researchers. This list can be systematically analyzed, tagged, and prepared for an increased searchability. By leveraging the medium of DMPs, we therefore work towards enhancing data discoverability and data-driven collaborative research and aim to contribute towards developing a clearer understanding of the challenges in large-scale research projects and practicable approaches to address them.

In Section 2, we discuss related works with regard to sharing, finding, and reusing data. The section covers the needs of researchers, the roles of community-centered platforms, the characteristics of large-scale projects, and the medium of DMPs. In Section 3, we elaborate on the concept of our approach in the form of the Dataset Finder, and a comparison with existing platform is conducted. In Section 4, we describe the implementation and functionalities of the Dataset Finder. In Section 5, we present the findings of a preliminary evaluation. In Section 6, we discuss the next steps of our research towards enhanced data reuse in terms of transferability of our approach as well as further *Research Data Management* (RDM) services addressing not only discoverability, but also other FAIR dimensions such as accessibility. Finally, Section 7 concludes the paper.

2 SHARING, FINDING, AND REUSING RESEARCH DATA

In the current scientific domain, an extensive amount of research data exists, in large parts leaving potential in being shared and reused. Establishing the infrastructure for a data ecosystem where researchers are actively engaged in exchanging knowledge is an intensive process.

2.1 RESEARCHERS' MOTIVATION AND BARRIERS

From the perspective of data owners, while there are some concerns about intellectual property being misused or losing competitive edge (Hughes et al., 2023), many have a vested interest in their data being found, putting their datasets in repositories with the highest discoverability and reviewing the engagement with their research (Mathiak et al., 2022). Data sharing can lead to increased citations, which accounts towards one's credibility and reputation, and organizations may also provide monetary rewards as incentives for data sharing (Badewitz et al., 2020). Data owners may find it challenging though to process and provide their data in a way that enables other researchers in working with their data (Donaldson and Koepke, 2022).

From the perspective of data consumers, data reuse is a complex challenge for one of three reasons—absence of incentives that promote data reuse, lack of awareness of existing data, or limited metadata and documentation (Wiggins et al., 2018). The lack of awareness of existing data can be attributed to data not being shared in a manner that makes it findable, to the repository used providing limited querying functionality, or to the lack of relevant repositories in the first place. The findability of a dataset may be compromised when only discipline-specific standards are used (Musen, 2022) or when the metadata accompanying it does not facilitate discovery (Gregory et al., 2020). Repositories' functionalities may be less likely to support the discoverability of datasets as opposed to text-based academic outputs. The lack of relevant results when such repositories are queried can be discouraging (Nicholson and Bennett, 2021). On the other hand, data consumers show an interest in reusing data for pragmatic reasons, such as saving time and costs of collecting and processing new data, and gaining resources leveraging their research's objectives (Zuiderwijk and Spiers, 2019). However, it is not always seen as a resource-conserving task; potential data consumers must identify and familiarize themselves both with available repositories and datasets (Gregory et al., 2019; Krämer et al., 2021).

Both data owners and consumers state as a motivation for sharing and reusing research data their belief that demonstration of actual reuse instances can foster better relations with funding agencies (Feger et al., 2020). Better visibility of research, findability of data, and thus improved opportunities for collaboration are therefore of concern to all stakeholders of the research data ecosystem. Data discovery services should therefore be developed that address the users' specific barriers and motivations.

DOI: 10.5334/dsj-2024-058

2.2 COMMUNITY-CENTERED PLATFORMS

Users may find data through citations, repositories or personal networks. Information seeking generally begins from the web, where literature search, journals, and online databases are the access points for the initiation of data search (Krämer et al., 2021). Multiple sources have documented the use of social connections in procuring and understanding data (Gregory et al., 2019, 2020). A researcher's personal connections play an essential role in acquiring data when an initial search failed (Krämer et al., 2021) or when locating specialized datasets (Sun et al., 2023). A survey of material science engineers revealed that researchers may acquire data not just through immediate colleagues and project partners, but also through conference participation (Suhr et al., 2020). The importance of personal networks in data discovery highlights the potential of less personal, but collegial communities such as participants of a conference or members of a large-scale research project (Suhr et al., 2020), and supports the idea of building repositories designed for specific communities, with tools that provide means for users to contact the data owners.

2.3 LARGE-SCALE RESEARCH PROJECTS

Large-scale research projects present a highly collegial community, with all participants contributing towards a common research objective, carrying the potential for a more probable successful collaboration and thus an increased motivation of researchers to share and reuse research data with and of each other.

Several challenges can be derived from the characteristics of large-scale projects. The characteristics can be divided into dimensions of heterogeneity and volume. Heterogeneity-wise, many large-scale projects are naturally interdisciplinary, with discipline-specific norms not only affecting the type of data sought, but also how the data is acquired (Gregory et al., 2020). The heterogeneity within research processes and varying experience with RDM tasks at the individual level, as well as the variety of used tools and infrastructure at institute level, and finally the responsibilities and structures at project level affect the potential for successful data sharing, finding, and reuse. Volume-wise, the number of researchers and subprojects exceed one-to-one advisory and guidance.

With regard to discovering research data and seeking collaboration, therefore, while users are more willing to invest effort in searching for data as compared to literature search (Lafia et al., 2023), it is desirable to design data sharing and discovery services that are easy to access and use, and balance generic as well as specific needs of the target users.

2.3.1 Use case: The Internet of Production

The Clusters of Excellence have been established to enable internationally visible and competitive research within German universities. The Cluster of Excellence project *Internet of Production* (IoP) at RWTH Aachen University focuses on integrating data across production domains to achieve data-driven improvement of production processes and products.

The IoP is a large-scale research project, involving over 30 institutes and 200 researchers. The Cluster is subdivided into *Cluster Research Domains* (CRD) and *workstreams* specific to each CRD, with every workstream again being divided into several subprojects.

Within the IoP, there exists a vast collection of data and research assets that can form the foundation for innovative collaborations and data reuse. The number of subprojects currently comprises a total of 121 projects including 183 datasets. Heterogeneity-wise, researchers of the IoP belong to a variety of disciplines such as engineering, computer science, and social sciences. Researchers also span a wide range in terms of experience with RDM practices and the scientific domain in general, with the IoP employing researchers from all stages of their careers.

The IoP RDM protocols mandate all researchers to create a DMP, ensuring that all data within the Cluster is documented across its life cycle. The researchers in the IoP create and manage DMPs using a template provided through the RDMO tool. Using the IoP as a use case, it is possible to examine the use of DMPs as a source of information about available research data, that all project members can access and use.

2.4 THE MEDIUM OF DMPs

Data Management Plans are documents that capture a variety of information about the research outputs of a project. Several organizations include DMPs as a mandatory part of the research data lifecycle. The contents of a DMP vary based on the regulations of a project, and funding bodies, universities and other organizations have introduced templates to assist researchers in generating a well-rounded overview of their research. These templates are often structured in a question-answer format, which allows researchers to fill in the answers in a structured and intuitive way.

DMPs can be a key component towards accessing complete data lifecycle information, not only as attachment documents for proposals, but also as active artifacts for communicating research. For researchers in the role of data consumers, dataset search differs from document retrieval systems for journal articles and publications. Evaluating the relevance of a dataset requires more time compared to evaluating a literature source—researchers not only look at the dataset, but also its metadata, associated code, and publications that cite it (Kern and Mathiak, 2015). Leveraging the medium DMP may allow data owners to describe their data in an extensive and comprehensive way, and for data consumers to find a holistic summary about the data, which is essential for data consumers to make sense of potentially reusable data.

Successfully implementing DMPs as a medium in researchers' everyday work requires structural measures, as while the results of a survey that examined the outlook of the European Commission's Horizon 2020 projects concerning DMPs found that 82.2% of respondents considered DMPs to be useful (Spichtinger, 2022), the preparation of DMPs is mostly considered additional administrative work (Miksa et al., 2019). Standardizing templates, establishing submission structures, and the use of efficient tools are named as requirements for a successful realization of widely-implemented DMPs (Kim et al., 2023).

DMPs are a means to achieve FAIRness in research, and we claim that storing and indexing DMPs in suitable repositories can contribute to discoverability of data (Jones et al., 2020). More specifically, the F4 principle of FAIR[2] highlights that resources can be made discoverable not only through indexing the data itself, but also through indexing its metadata in a searchable resource. Our approach therefore is to employ DMPs as a database and developing a tool called the Dataset Finder which allows users to query the DMPs meaningfully for reusable data.

3 CONCEPT OF THE DATASET FINDER

The Dataset Finder is a tool designed to enhance the discoverability of data registered in the DMPs created in RDMO. While DMPs are instrumental in ensuring the systematic documentation of research data, their traditional use as mere archival tools presents notable drawbacks. Primarily, they contribute minimally to the findability of the actual research data they document. This is because DMPs typically focus on compliance and storage protocols rather than on the discoverability of data. To tackle these shortcomings, in Kim *et al.* (2023), DMPs were designed with a more project-specific purpose in mind. In addition to the more general information which a lot of DMPs contain within their respective research projects, the DMP template for the IoP contains project-specific building blocks which, when filled out, provide a richer source of information in the form of metadata about the documented research data. This results in the DMPs being more than just a documenting tool.

3.1 MORE THAN JUST DOCUMENTATION

In the realm of large-scale research projects, DMPs have traditionally been relegated to the role of documentation tools, designed to ensure the systematic handling, storage, and dissemination of research data. These documents, often mandated by funding bodies, delineate a structured approach to data management, encompassing the description of data types, formats, and volumes; strategies for data storage, backup, and security; protocols for access and sharing; plans for long-term preservation; and the allocation of roles and responsibilities.

2 F4: (Meta)data is registered or indexed in a searchable resource, https://www.go-fair.org/fair-principles/f4-metadata-registered-indexed-searchable-resource/.

However, within the dynamic and collaborative environment of extensive research initiatives such as the IoP, the conventional function of DMPs as static documents has revealed several shortcomings. The rigidity of DMPs can lead to inflexibility, hindering their adaptation to the evolving demands of research projects. Moreover, the focus on documentation does little to enhance the discoverability of research data, posing a significant challenge for researchers seeking specific datasets. Consequently, the wealth of information contained within DMPs is often underutilized, with the emphasis placed on compliance rather than on the facilitation of research processes.

In response to these challenges, the Dataset Finder platform has been developed to transcend the traditional confines of DMPs, transforming them from mere repositories of information into dynamic, searchable sources. By leveraging the data documented within DMPs, the platform empowers researchers to locate and access research data efficiently, using keyword-based searches. This innovative approach not only augments the utility of DMPs but also promotes a more collaborative and productive research ecosystem.

In summary, while DMPs are indispensable for responsible data management, their utility in large scale research projects has been limited by their static nature and documentation-centric approach. The advent of tools like Dataset Finder marks a pivotal evolution in the role of DMPs, from passive documents to active facilitators of data discoverability, thereby enhancing the overall efficacy of research data management.

3.2 INCREASING THE FINDABILITY OF DATA

In the context of the IoP, the traditional conception of DMPs has been reimagined to address the challenges of data findability in large-scale research projects. The IoP's innovative approach has been to enrich the DMPs with additional metadata, specifically keywords that succinctly describe the research data, thereby transforming DMPs from static documents into dynamic, searchable entities.

The Dataset Finder platform, developed as part of this initiative, capitalizes on the enhanced structure of the DMPs. By incorporating descriptive keywords into the DMPs, researchers provide a more precise depiction of their data, facilitating a more intuitive search process. These keywords, while peripheral in the traditional documentation role of DMPs, become central to the functionality of the Dataset Finder, enabling researchers to efficiently locate relevant datasets within the IoP's extensive research data repository.

At its essence, the Dataset Finder operates as a sophisticated search engine that utilizes the keyword-enriched DMPs to increase the visibility and accessibility of research data. Researchers can input specific keywords into the Dataset Finder, which then queries the enriched DMPs to retrieve and present the most relevant data. This mechanism significantly enhances the findability of research data, optimizing the use of existing datasets and promoting a collaborative research environment. The strategic integration of search-friendly keywords within DMPs has improved the accessibility and utility of research data, and thus enhanced the IoP's research data management practices.

In essence, the Dataset Finder exemplifies the potential of DMPs when they are designed with an emphasis on data discoverability. This shift from passive documentation to active data facilitation reflects a broader trend in research data management towards harnessing the power of metadata to unlock the full value of research data assets.

3.3 INTEGRATING THE DATASET FINDER INTO THE PROCESS OF RDM

As discussed before, the IoP employs RDMO to facilitate the implementation of DMPs. While RDMO's capabilities are commendable for standardizing and streamlining data management processes, its generalist nature presents limitations. Notably, RDMO lacks the functionality for granular searches within DMPs, a feature that would be particularly beneficial for tailoring data management to project-specific needs. This limitation underscores the tool's orientation towards broad management tasks rather than the nuanced, dynamic requirements often essential for elevating DMPs beyond their traditional documentary role. Consequently, while RDMO contributes significantly to the general management of DMPs, it falls short in addressing

the specialized tasks that could transform DMPs into a more active and integral component of research data management. To explain the utilization of the Dataset Finder within the IoP, Figure 1 displays how it is integrated into the process of RDM.

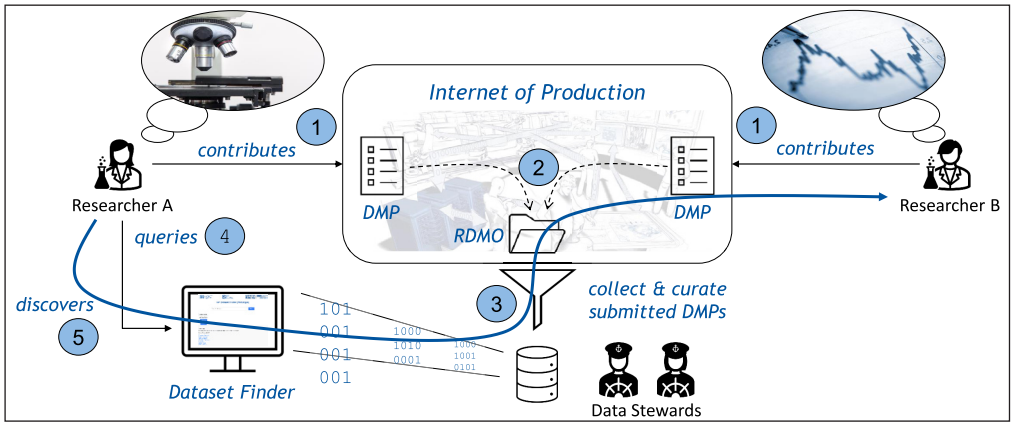

Figure 1 Integrating the Dataset Finder to the process of RDMO within the IoP.

In Figure 1, numbers ① to ⑤ mark the different steps detailing the integration of the Dataset Finder. First, marked by ①, researchers of the IoP, *Researcher A* and *Researcher B* conduct their research and eventually contribute research data to the IoP. They document their data using the IoP-specific DMP, which is managed with RDMO. Hence, displayed by ②, eventually all DMPs are managed by RDMO, turning them into the main source of information for the Dataset Finder. Next, for the integration of the Finder, marked as ③, the data stewards collect and curate the information provided by the submitted DMPs. This is necessary since RDMO does not provide an infrastructure that checks any constraints concerning the information provided by the researchers. Thus, in ③, faulty DMPs are filtered and reported back to the researchers. Afterwards, the curated information is used as the data source for the Dataset Finder. Displayed by ④, researchers are then able to query the Dataset Finder for research data they are interested in, which ultimately helps them discover research data contributed by others, marked by ⑤, and therefore improving their own research.

Although Figure 1 provides just a simplified view of the integration of the Dataset Finder, it shows how the tool improves the discoverability of research data within the IoP. Without the use of the Finder, searching through DMPs would be a cumbersome process of manually clicking through RDMO.

3.4 SYNERGIES WITH EXISTING RDM LANDSCAPE

Platforms such as Figshare,[3] Zenodo,[4] and the Open Science Framework (OSF)[5] present themselves as alternatives or possible additions, as they offer a broad spectrum of functionalities to facilitate data sharing, preservation, and discoverability.

However, within the specific context of IoP, the integration of these platforms introduces additional complexities. Each new tool or platform necessitates its own maintenance and harmonization within the existing RDM framework of the IoP. This can lead to an increased administrative burden and potential fragmentation of data management practices as researchers have to cope with several tools at once.

Conversely, the IoP Dataset Finder leverages the established infrastructure of RDMO, capitalizing on the information provided by DMPs. This strategic utilization encourages researchers to engage more actively in the documentation of their research data. Consequently, it fosters an environment where related research can be effortlessly searched and accessed within the IoP network, negating the need for data dissemination across disparate platforms. The Dataset Finder thus serves as a catalyst for transforming DMPs from static documents into dynamic tools that enhances the accessibility and utility of research data within the IoP RDM framework without adding to its complexity.

3 https://figshare.com/.

4 https://zenodo.org/.

5 https://osf.io/.

3.5 COMPARISON WITH EXISTING PLATFORMS

The following section presents a comparative analysis of the IoP's Dataset Finder with three established data platforms (Table 1). Zenodo is a universal data repository operated by CERN. It provides researchers with the ability to openly archive and publish data across all scientific disciplines. EBRAINS[6] offers a suite of specialized tools and databases for neuroscientific research, supported by the European Brain Research Initiative. The Materials Genome Initiative (MGI)[7] offers a collaborative platform and a comprehensive database with the objective of accelerating the discovery and development of new materials in materials science.

Table 1 Comparison of the Dataset Finder, Zenodo, EBRAINS, and MGI.

CATEGORY	DATASET FINDER	ZENODO	EBRAINS	MGI
Domain	Production and data management.	Universal, no domain focus.	Neuroscience.	Materials science.
Timing of (Meta) Data Availability	Immediately after DMP completion.	Post-project, later in process.	Post-project, later in process.	Post-project, later in process.
Integration	Displays DMP metadata; external storage (e.g., Zenodo).	Long-term archive.	Neuroscience-specific repositories.	Materials science repositories.
Reach	Controlled cross-institute visibility.	Communities, open or restricted.	Open or restricted.	Open or restricted.
Data Owner Effort	Automated from DMPs, no manual entry.	Manual data and metadata upload.	Manual data and metadata upload.	Manual data and metadata upload.

A comparative analysis of the aforementioned services reveals their distinctive strengths and areas of focus. The comparison is based on five categories: (1) *Domain*, which reflects the platform's ability to support multiple domains; (2) *Timing of (Meta)Data Availability*, addressing the phase of the project at which the research (meta)data becomes available; (3) *Integration*, describing how data is stored and how the platform connects to other tools; (4) *Reach*, referring to the size of the target community and how easily data can be discovered; and (5) *Data Owner Effort*, describing the complexity and workload required to upload the research data.

The Dataset Finder supports a wide range of research fields related to the domains of production and data management, facilitating collaboration within a highly interdisciplinary, yet defined research area. Zenodo, with its universal design, is suitable for any domain, while EBRAINS and MGI focus on specific fields such as neuroscience and materials science.

Access to project information is provided in the Dataset Finder immediately after the DMP is completed, thus listing (meta)data in all stages of the project cycle, allowing for early-stage collaboration. Zenodo, EBRAINS, and MGI typically make research data accessible later in the project lifecycle.

The Dataset Finder displays DMP metadata from the DMP platform RDMO and integrates links to external solutions like Zenodo for data storage. Other platforms have archiving capabilities, or provide a list of suitable repositories.

The approach of the Dataset Finder defaults to a cross-institution visibility for member institutes, ensuring accessibility across the IoP, and therefore catering to and providing infrastructure specifically for the needs of the project. Although Zenodo can create similar communities, doing so requires additional effort to establish and manage these connections, which the Dataset Finder has integrated from the beginning. Zenodo, EBRAINS, and MGI provide flexible options for data visibility, ranging from open access to restricted options, allowing users to decide the level of public access.

User effort on the Dataset Finder is minimal, as it automatically utilizes data from the DMPs, while Zenodo, EBRAINS, and MGI require users to manually upload their data and metadata, which involves a higher level of user input.

6 https://www.ebrains.eu/.

7 https://www.mgi.gov/.

4 IMPLEMENTATION OF THE DATASET FINDER

With its concept in mind, the implementation of the Dataset Finder mainly consists of three parts: the data extraction, its backend, and its *Graphical User Interface* (GUI). While the process of data extraction and the building of the Finder's database can be examined separately, backend functionalities and the GUI are tightly coupled. In the following, the implementation of the Dataset Finder is presented in more detail.

4.1 UTILIZING RDMO's API

For the data extraction and the building of the Dataset Finder's database, the RDMO API provides access to DMPs of the IoP. The answers from all projects to all questions can be downloaded as JSON files. In each project, an index for each answer is stored, which links to the corresponding question. All questions belonging to a section are summarized into sets. In RDMO's internal structure, each question has an attribute that uniquely identifies it and that is referenced in the respective answers. For each set of questions, there is an additional index referring to the previous and subsequent question set. Depending on the answer given, the index can initiate a follow-up question. One example for follow-up questions would be the examination about the publication status of a dataset. In the case of a positive answer, the following question would ask for the location of the published data; in the case of a negative answer, the following question would ask for an explanation. These conditional questions are design decisions made during the creation of the IoP DMP template (Kim et al., 2023). To ensure comprehensiveness of all questions and their corresponding answers, the program runs through all question sets from all projects and extracts the questions with the corresponding identifiers. After processing all DMPs, they are stored as a JSON file and used as the database of the Finder.

4.2 THE DATASET FINDER'S CORE FUNCTIONALITIES

The Dataset Finder is a powerful tool that allows users to search for DMPs and find relevant metadata. With an intuitive user interface and various functions, the Finder offers an efficient way to obtain a comprehensive overview of all available projects and corresponding datasets.

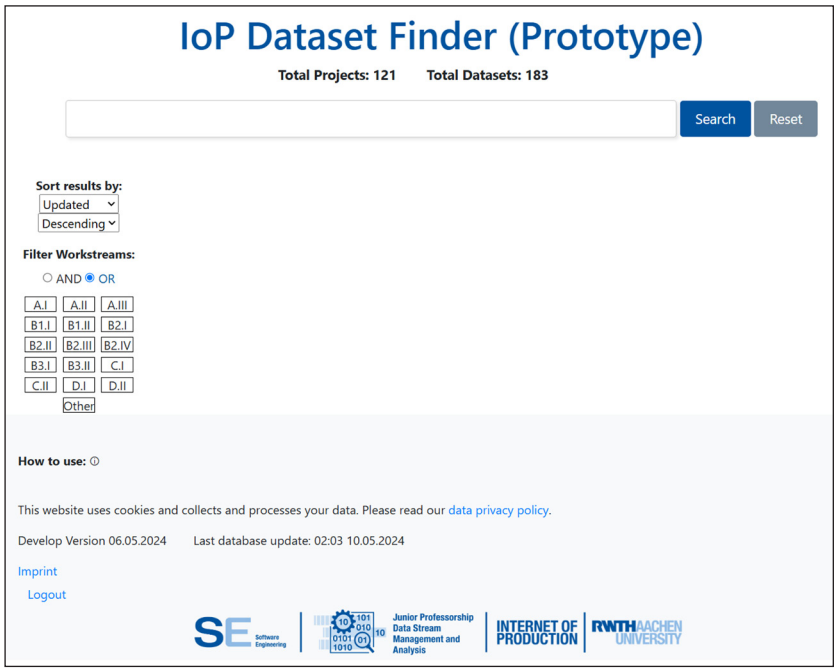

Figure 2 Start page of the Dataset Finder.

Figure 2 depicts a number of the Finder's core functionalities for user to interact with. The search function at the top allows to search for specific terms to facilitate finding relevant DMPs. For customized search result views, the sorting function on the left hand side can sort the results by creation date, modification date, or title. The filter options for workstreams offer another way to refine the search. By clicking on the corresponding fields, the desired workstreams can be selected and then linked with the logical operators AND or OR. In addition, the total number of available projects and records is displayed centrally on the platform right below the title, giving users a quick overview of the available resources.

4.2.1 Keyword-based search

Figure 3 illustrates an exemplary search in the RDMO catalog. The search is returning 3 of 121 projects that match the specified search parameters. Various DMP fields are queried in the search to achieve meaningful results, including project title, project or dataset description, contact persons (with associated email, ORCID, and chair), keywords, and workstreams. File size, file format, and storage medium are also part of the search criteria. The search is not limited to individual keywords; entire sentences can be searched. Multiple keywords can be separated by commas, creating a logical OR link. In this example, the keywords entered are restricted to the *Other* workstream, ensuring that only results matching this workstream are displayed. The *Searched Words* parameter lists the words from the search bar found in the project. Search matching words are highlighted in yellow. To prevent projects from being overlooked due to typing errors, the Levenshtein distance (Levenshtein et al., 1966) between all search parameters and the words entered in the search bar is calculated for each search, and sufficiently similar search results are also returned. Based on experience, a minimal Levenshtein similarity value of 85% yields satisfactory results, while limiting the number of false positives.

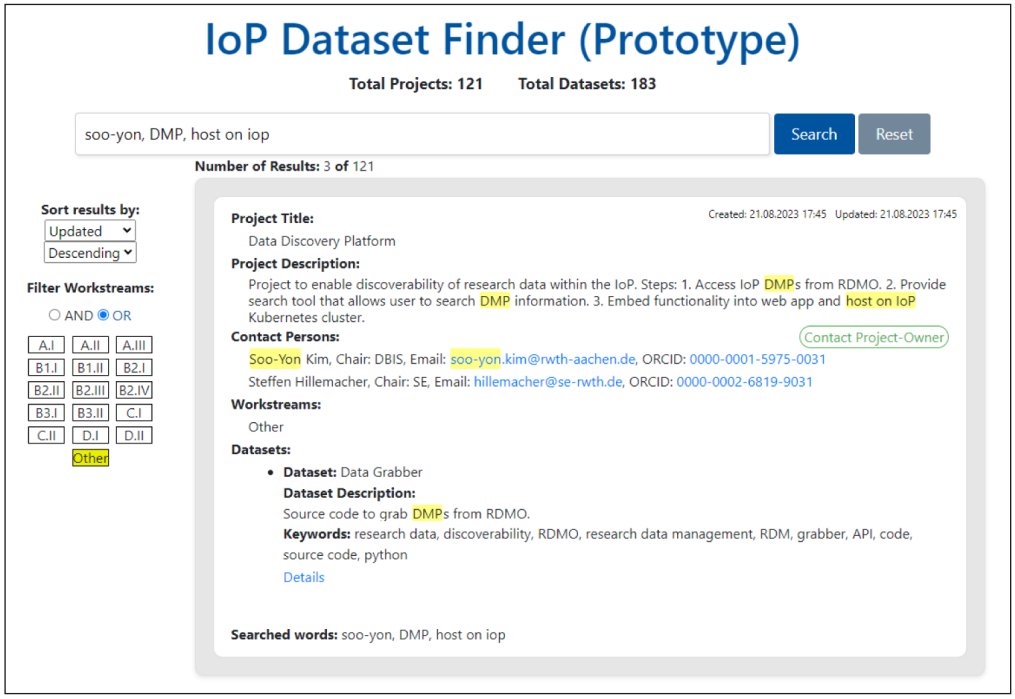

Figure 3 Search result for *soo-yon*, *DMP* and *host on IoP* with activated workstream filter in the Dataset Finder.

4.2.2 Authentication

The website is hosted within the IoP in a Kubernetes[8] cluster and is only accessible within the RWTH network. In order to further restrict access to members of the IoP only, additional authentication is required. To reduce the effort of user management, an already existing IoP portal and its user management is utilized. All members of the IoP are already registered users of the portal, thus allowing this user base to also serve as the user base of the Dataset Finder. The actual authentication process is based on the OAuth 2.0[9] protocol. In case a user attempts to access the Data Finder without authorization, the user will be redirected to the login page (Figure 4) and prompted to log in via the IoP portal. After successfully logging in, the user will be redirected back to the Dataset Finder with an attached request parameter in the URL granting access if authentication was successful. To avoid unnecessary logins, the access and refresh tokens transmitted during the authentication process are saved. When the access token expires, it is renewed using the refresh token.

8 https://kubernetes.io/.

9 https://v2.developer.constantcontact.com/docs/authentication/oauth-2.0-server-flow.html.

Figure 4 Login page to the IoP portal.

4.2.3 Contacting data owners via e-mail

To simplify communication, the Data Finder offers a built-in email client. When the user activates the *Contact the Project Owner* button in the green area next to *Contact Persons* (Figure 3), the email client opens, as shown in Figure 5 on the left hand side. In this field, the subject, message, and the user's email address can be entered. Once the email has been sent, all responsible contact persons of the corresponding project will be notified (Figure 5).

The contact form also allows communication for project owners who wish to hide their contact information within the IoP. Project owners can choose for their contact information to not be visible, such that only the *Contact the project owner* button will be shown. Users will be able to contact project owners via the contact form that does not reveal the project owner's identity. Project owners have full control over whether they want to respond or not.

This feature ensures smooth and secure communication between users and project owners. The clear separation of contact information respects the privacy of project owners while providing the possibility to communicate important requests and information to advance and support projects.

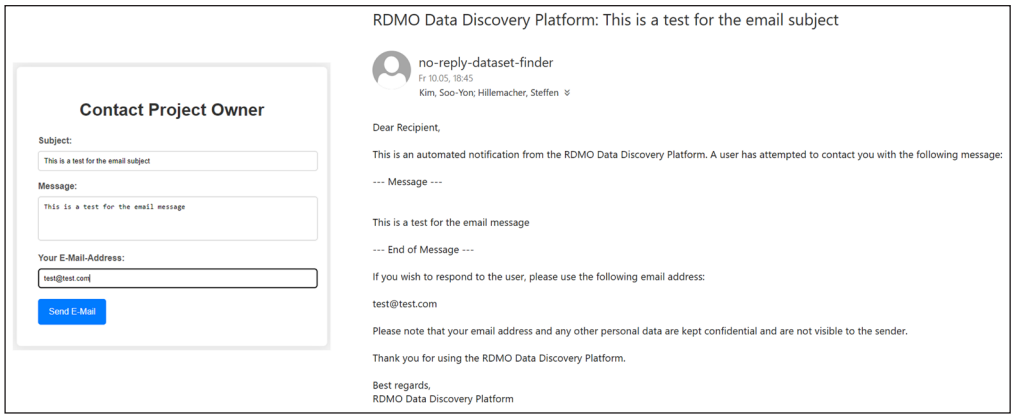

Figure 5 The image on the left depicts the user interface of the Dataset Finder email client, which enables users to contact the project owners. Upon sending a message, all specified contact persons will receive the message depicted in the image on the right.

4.2.4 Providing statistics on RDM

The standardized formatting of the data enables efficient analysis. On a dedicated statistics page, statistics are generated and dynamically updated. An illustrative example for this is the visualization to show the distribution of projects across different workstreams (Figure 6) and the file formats used in the projects (Figure 7). We can also show which responsible persons are active in how many projects and which storage media are used for data backup. We can further estimate an aggregated volume of all datasets, as well as numbers of published and planned-to-publish datasets. Keyword analyses carry the potential for identifying similar or complementary research projects and find correlations that are not obvious in the raw data.

4.3 DISPLAYING THE DATA PROVIDED BY RDMO

Currently, 17 of the 26 DMP questions are utilized, with approximately half of them being used for data retrieval in the Dataset Finder and half of them being used for statistical purposes.

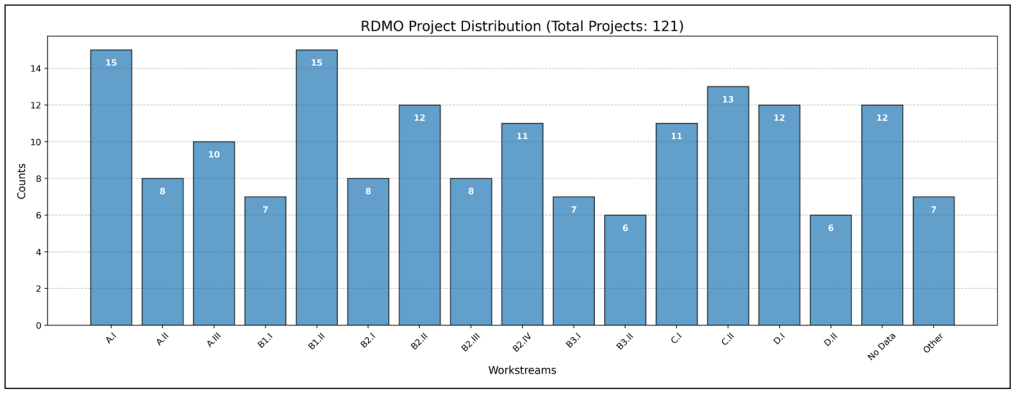

Figure 6 Distribution of the RDMO projects to the workstreams.

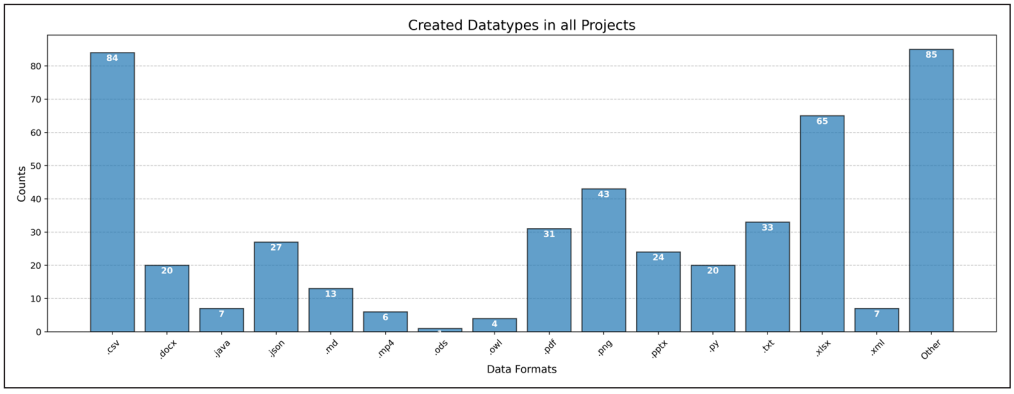

Figure 7 Distribution of the created data types within the projects.

Questions asking for potential keywords or responsible persons for instance have a standardized format for their respective answers (Kim et al., 2023). This facilitates the processing of the answers by the Dataset Finder and they can be used to improve the findability of projects. In contrast, questions that can be answered by free text are not standardized and, as a consequence, are considerably more complex to process. Thus, including them in the search criteria would significantly increase the implementation effort, and would require a more extensive analysis on how they could be utilized to enhance the findability of the data, which is the focus of this research. They could be of great use for data consumers to gain a more holistic understanding of a dataset once they have discovered it in the Finder. Handling of these questions will be a focus of future project advancements.

5 PRELIMINARY EVALUATION

To evaluate the functionality and relevance of the Dataset Finder, we have formulated the anticipated needs of researchers in an overview below (Table 2). For each need, we specify which features of the Dataset Finder address the need, and discuss the maturity of the current solution. The maturity is graded following a three-star scheme: ⋆ means that the need is not met; ⋆⋆ means that the need is partly met; ⋆⋆⋆ means that the need is fully or almost fully met.

Table 2 Overview of anticipated needs, features, and maturity of the current solution.

ANTICIPATED NEEDS	IMPLEMENTED FEATURES	MATURITY OF SOLUTION
Single point of reference	Automatic extraction from RDMO, coverage of all IoP institutes, transforming DMPs into a unified data catalog	⋆⋆⋆
Support of heterogeneous search behavior	Large coverage of DMP attributes, fuzzy search, multi-query support	⋆⋆
Intuitive GUI	Search bar, filtering and sorting, design and operation modes closely aligned with conventional search engines	⋆⋆
Comprehensive depiction of results	Result highlighting, sorting, “Details” page	⋆⋆
Quality of results	–	⋆
Actionability	Button to contact data owner, if provided: link to dataset	⋆⋆

A key anticipated need for researchers in large-scale projects is a **single point of reference** to identify existing datasets within the project. Previously, no such resource existed, and discovering reusable data or collaboration opportunities within the project required highly proactive and labor-intensive communication with individual institutes and researchers. By leveraging the project-wide coverage of DMPs created in RDMO and processing the DMPs through an automatic pipeline, a comprehensive data catalog has been created for researchers to be queried or browsed.

Users may exhibit profoundly **heterogeneous search behaviors**, such as searching purposefully versus browsing exploratively, or searching for highly diverse concepts, such as names of authors, domain-specific terms, or technical attributes. To address this, the platform enables browsing through all entries, and covers most of the DMP fields as the basis for targeted searches. A fuzzy search mechanism enhances robustness by accounting for typographical errors, and multi-term query support allows users to search for multiple concepts simultaneously. While these features enable a flexible operation of the Dataset Finder, future enhancements could support even more search behaviors by including an extension of covered DMP fields, and enhancement of browsing functions, e.g., by suggesting keywords.

To address the need for an **intuitive graphical user interface (GUI)**, the platform is developed to have a familiar design and user-friendly navigation. A central search bar facilitates the initiation of queries, while filtering and sorting options on the side bar allow for precise refinement of search results. The interface is designed on principles commonly found in conventional search engines, which aim to make the system more familiar and easier to use. While these features contribute to enhanced usability, further optimization could be achieved through surveying and incorporating user feedback.

A further anticipated need is the ability of researchers to cognitively process and understand search results. The Dataset Finder includes several features addressing this need for a **comprehensive depiction of results**: search terms are highlighted in the results, the results can be sorted by criteria such as timeliness (e.g., last updated or created), and the results are presented in a compact format where only key information is displayed initially, with more detailed information accessible via a clickable "Details" page. While these features enhance comprehensibility, some researchers may still find the information overwhelming, while others may prefer details currently hidden on the "Details" page to appear upfront for certain datasets. Future improvements could include a customized display of information to show on the main results page rather than the details page depending on dataset types, collapsible sections or layout adjustments to balance accessibility with information density, and a definition of a "relevance" metric to sort results by.

A critical topic of interest of researchers using the Dataset Finder we anticipate is the **quality of the results**, specifically whether the searching and browsing functions are able to provide results relevant for the researcher, and whether the provided information in each entry is sufficient for a researcher to determine if a dataset is relevant for reuse or collaboration. Addressing the latter is challenging, as the current Dataset Finder lists all datasets described in RDMO, regardless of the level of detail provided by the data owner. For instance, while image datasets may be of interest to researchers based on attributes such as resolution, this information can be missing. Since such details cannot be automatically generated, the quality of the entries depends heavily on the completeness of data descriptions. Enhancing this quality could involve making specific fields in the DMP mandatory, but this introduces a trade-off: While it could improve dataset descriptions, it might also reduce reporting of datasets due to the added effort required. The integration of automated tools to infer and supplement missing metadata, thereby improving the utility of search results without relying on the completeness of information provided by data owners, might be a field for subsequent work. With regard to the relevance of the results, future enhancements could explore semantic search to recommend similar projects or to create thematic collections.

As we expect that researchers want to follow up on discoveries in the Dataset Finder relevant to them, the platform's ability to enable **actionability** beyond its core functionality of displaying information presents an essential need. Several features have been integrated to support user interaction and data utilization. These include a contact function, enabling users to directly reach out to dataset owners, and the possibility to access research data directly via a link in the

Dataset Finder if the data has been referenced in the DMP with a link or persistent identifier. In future work, extending the pipeline to integrate data management platforms and repositories with actual data should be prioritized to present a more holistic solution for researchers.

6 NEXT STEPS FOR THE DATASET FINDER

In addition to improving the Dataset Finder's capabilities of processing the information provided by DMPs, two parts are essential for the next steps concerning the development of the Dataset Finder. First, the Dataset Finder is research software. Consequently, as part of research software engineering it is important to make it or parts of its implementation available for reuse in other research projects or more general application scenarios. Hence, the transferability of the approach plays an important part in the future work. Second, currently, the Dataset Finder only displays metadata of the research data provided by the IoP DMP. For the future, it is planned to display the referencing of the actual data and where it is stored in the Finder. Extensions and interfaces with data storage could also be considered.

6.1 TRANSFERABILITY OF THE DATASET FINDER APPROACH

When looking at the transferability of the Dataset Finder to other applications scenarios, its architecture plays a critical role. Basically, it is bifurcated into two pivotal components: the DMP data extraction and the search logic including the GUI. The extraction process leverages the RDMO API to procure DMPs pertinent to the IoP, subsequently curating a specialized database that underpins the Finder's operational capabilities. The search logic, augmented by a user-friendly GUI, incorporates bespoke functionalities such as authentication protocols, contact forms, and advanced filtering mechanisms.

While these features are meticulously tailored to accommodate the unique structure and requirements of the IoP, they exhibit varying degrees of transferability to other research endeavors. The custom curation of data and project-specific authentication and filtering processes may pose challenges when attempting to adapt the Dataset Finder to alternative contexts. These elements are intrinsically designed to align with the IoP 's infrastructure, and the structuring of the IoP DMP, potentially necessitating significant modifications to ensure compatibility with other projects.

Conversely, the Dataset Finder also encompasses a suite of generic functionalities that hold promise for broader applicability. The utilization of the RDMO API for data extraction and the foundational aspects of the database's creation are not inherently project-specific. These components offer a versatile framework that can be readily integrated into other research projects, facilitating a seamless transition of the Dataset Finder's core principles. This duality of bespoke and generic elements within the Dataset Finder's design underscores its potential as both a specialized tool for the IoP and a transferable asset for the wider research community, provided that necessary adjustments are made to cater to the distinct demands of each new research landscape.

6.2 DATA STORAGE AND INTEGRATION OF RESEARCH DATA

As the Dataset Finder evolves within IoP, a pivotal enhancement on the horizon is the integration of metadata with actual data repositories. Currently, the RDMO serves as the backbone for the Dataset Finder DMPs, but not extending to the storage of research data itself. This delineation restricts the Dataset Finder to providing only metadata, leaving the actual data unlinked and inaccessible through the platform.

Addressing this limitation, future developments aim to forge a symbiotic relationship between the DMPs within the IoP and the *Collaborative Scientific Integration Environment* (Coscine)[10] (Lang et al., 2024; Politze et al., 2023) developed at RWTH Aachen University. By embedding direct links to Coscine 's repositories within the IoP DMPs, the Dataset Finder will be poised to offer a more holistic service. Researchers will be able to utilize the metadata to locate relevant research data and seamlessly navigate to the corresponding Coscine repositories to access the data, subject to permissions.

10 https://coscine.rwth-aachen.de/.

This strategic linkage promises to streamline the research process, eliminating the need for additional platforms and reducing the fragmentation of data management tools. The integration will not only enhance the functionality of the Dataset Finder but also reinforce the existing RDM framework of the IoP. By capitalizing on the established infrastructure of RDMO and Coscine, the Dataset Finder is set to transition from a tool that catalogues metadata to one that serves as a gateway to tangible research data, thereby enriching the research data management ecosystem. Alternatively, platforms like NFDI4Ing's[11] Jarves (Hamann and Werheid, 2023) might provide a promising way for integrating the Dataset Finder, RDMO, and Coscine.

7 CONCLUSION

The Dataset Finder improves upon existing frameworks and tooling of RDM, particularly within the context of large-scale research projects like the IoP. By transforming DMPs from static documents into dynamic, searchable databases, the Dataset Finder greatly enhances the discoverability and accessibility of research data. It not only facilitates interdisciplinary collaboration by allowing researchers to efficiently locate and utilize each other's research data but also streamlines the RDM process by integrating with existing RDMO infrastructure. The Dataset Finder's innovative approach serves as a prototype for future RDM tools, demonstrating the potential of leveraging DMPs beyond compliance, towards active facilitation of research processes and data sharing within the scientific community.

FUNDING INFORMATION

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production –390621612.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Soo-Yon Kim, Steffen Hillemacher.

Conceptualization, Methodology, Writing.

Max Kocher.

Software implementation, Writing.

Bernhard Rumpe, Sandra Geisler.

Revision.

AUTHOR AFFILIATIONS

Soo-Yon Kim orcid.org/0000-0001-5975-0031
Data Stream Management and Analysis, RWTH Aachen University, Aachen, Germany

Steffen Hillemacher orcid.org/0000-0002-6819-9031
Software Engineering, RWTH Aachen University, Aachen, Germany

Max Kocher orcid.org/0000-0001-9801-0300
Data Stream Management and Analysis, RWTH Aachen University, Aachen, Germany

Bernhard Rumpe orcid.org/0000-0002-2147-1966
Software Engineering, RWTH Aachen University, Aachen, Germany

Sandra Geisler orcid.org/0000-0002-8970-6282
Data Stream Management and Analysis, RWTH Aachen University, Aachen, Germany

11 https://nfdi4ing.de/.

REFERENCES

Anders, I., Enke, H., Hausen, D.A., Henzen, C., Jagusch, G., Lanza, G., Michaelis, O., Peters-von Gehlen, K., Rathmann, T., Rohrwild, J., Schönau, S., Wedlich-Zachodin, K.V. and **Windeck, J.** (2024) 'The research data management organiser (rdmo) – a strong community behind an established software for dmps and much more', *Data Science Journal*, 23(1), p. 28. Available at: https://doi.org/10.5334/dsj-2024-028 (Accessed 15 May 2024).

Badewitz, W., Kloker, S. and **Weinhardt, C.** (2020) 'The data provision game: Researching revenue sharing in collaborative data networks', *in 2020 IEEE 22nd Conference on Business Informatics (CBI)*, Vol. 1, pp. 191–200. Available at: https://doi.org/10.1109/CBI49978.2020.00028 (Accessed 15 May 2024).

Donaldson, D.R. and **Koepke, J.W.** (2022) 'A focus groups study on data sharing and research data management', *Scientific Data*, 9(1), p. 345. Available at: https://www.nature.com/articles/s41597-022-01428-w; https://doi.org/10.1038/s41597-022-01428-w (Accessed 15 May 2024).

Feger, S. S., Wozniak, P. W., Lischke, L. & **Schmidt, A.** (2020), ''Yes, I comply!': Motivations and Practices around Research Data Management and Reuse across Scientific Fields', *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW2), pp. 141:1–141:26. Available at: https://doi.org/10.1145/3415212 (Accessed 15 May 2024).

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A. and **Wyatt, S.** (2019) 'Searching data: A review of observational data retrieval practices in selected disciplines', *Journal of the Association for Information Science and Technology*, 70(5), pp. 419–432. Available at: https://doi.org/10.1002/asi.24165 (Accessed 15 May 2024).

Gregory, K.M., Cousijn, H., Groth, P., Scharnhorst, A. and **Wyatt, S.** (2020) 'Understanding data search as a socio-technical practice', *Journal of Information Science*, 46(4), pp. 459–475. Available at: https://doi.org/10.1177/0165551519837182 (Accessed 15 May 2024).

Hamann, T. and **Werheid, J.** (2023) 'Jarves: CoRDI 2023', *1st Conference on Research Data Infrastructure (CoRDI 2023)*. Karlsruhe, Germany, 12–14 September. Available at: https://doi.org/10.5281/zenodo.8315363 (Accessed 15 May 2024).

Hughes, L.D., Tsueng, G., DiGiovanna, J., Horvath, T.D., Rasmussen, L.V., Savidge, T.C., Stoeger, T., Turkarslan, S., Wu, Q., Wu, C., Su, A.I. and **Pache, L.** (2023) 'Addressing barriers in FAIR data practices for biomedical data', *Scientific Data*, 10(1), p. 98. Available at: https://www.nature.com/articles/s41597-023-01969-8; https://doi.org/10.1038/s41597-023-01969-8 (Accessed 15 May 2024).

Jones, S., Pergl, R., Hooft, R., Miksa, T., Samors, R., Ungvari, J., Davis, R.I. and **Lee, T.** (2020) 'Data management planning: How requirements and solutions are beginning to converge', *Data Intelligence*, 2(1–2), pp. 208–219. Available at: https://doi.org/10.1162/dint_a_00043 (Accessed 15 May 2024).

Kern, D. and **Mathiak, B.** (2015) 'Are there any differences in data set retrieval compared to well-known literature retrieval?, in S. Kapidakis, C. Mazurek and M. Werla (eds.) *Research and Advanced Technology for Digital Libraries*. Lecture Cham: Notes in Computer Science, Springer International Publishing, pp. 197–208. Available at: https://doi.org/10.1007/978-3-319-24592-8_15 (Accessed 15 May 2024).

Kim, S.-Y., Hillemacher, S., Decker, S., Rumpe, B. and **Geisler, S.** (2023) 'Designing and implementing practicable data management plans in large-scale projects', *Bausteine Forschungsdatenmanagement*, 2023(3), pp. 1–12. Available at: https://www.se-rwth.de/publications/Designing-and-Implementing-Practicable-Data-Management-Plans-in-Large-Scale-Projects.pdf (Accessed 15 May 2024).

Krämer, T., Papenmeier, A., Carevic, Z., Kern, D. and **Mathiak, B.** (2021) 'Data-seeking behaviour in the social sciences', *International Journal on Digital Libraries*, 22(2), pp. 175–195. Available at: https://doi.org/10.1007/s00799-021-00303-0 (Accessed 15 May 2024).

Lafia, S., Million, A. and **Hemphill, L.** (2023) 'Direct, orienting, and scenic paths: How users navigate search in a research data archive', in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, Association for Computing Machinery, New York, NY, USA, pp. 128–136. Available at: https://dl.acm.org/doi/10.1145/3576840.3578275; https://doi.org/10.1145/3576840.3578275 (Accessed 15 May 2024).

Lang, I., Nellesen, M. and **Politze, M.** (2024) 'Rdm platform coscine-fair play integrated right from the start', *ing.grid*, 1(2). Available at: https://doi.org/10.48694/inggrid.3952 (Accessed 15 May 2024).

Levenshtein, V. I. *et al.* (1966) 'Binary codes capable of correcting deletions, insertions, and reversals', in *Soviet physics doklady*, Vol. 10, Soviet Union, pp. 707–710.

Mathiak, B., Juty, N., Bardi, A., Colomb, J. and **Kraker, P.** (2022) 'Discoverability Use Cases to help define Requirements for Research Data Discovery Tools', Zenodo. Available at: https://doi.org/10.5281/zenodo.5833952 (Accessed 15 May 2024).

Miksa, T., Simms, S., Mietchen, D. and **Jones, S.** (2019) 'Ten principles for machine-actionable data management plans', *PLOS Computational Biology*, 15(3), p. e1006750. Publisher: Public Library of Science. Available at: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006750; https://doi.org/10.1371/journal.pcbi.1006750 (Accessed 15 May 2024).

Musen, M.A. (2022), 'Without appropriate metadata, data-sharing mandates are pointless', *Nature*, 609(7926), p. 222. Available at: https://www.nature.com/articles/d41586-022-02820-7; https://doi.org/10.1038/d41586-022-02820-7 (Accessed 15 May 2024).

Nicholson, S.W. and **Bennett, T.B.** (2021) 'Do institutional repository deposit guidelines deter data discovery?', *Evidence Based Library and Information Pratice*, 16(3), pp. 2–17. Available at: https://doi.org/10.18438/eblip29913 (Accessed 15 May 2024).

Politze, M., Lang, I. and **Jansen, K.** (2023) 'Coscine. nrw landesweite basisversorgung zur verwaltung von forschungsdaten im open source modell', in *Proceedings of the Conference on Research Data Infrastructure*, 1. Available at: https://doi.org/10.52825/cordi.v1i.235 (Accessed 15 May 2024).

Spichtinger, D. (2022) 'Data management plans in horizon 2020: what beneficiaries think and what we can learn from their experience [version 2; peer review: 2 approved, 1 approved with reservations]', *Open Research Europe*, 1(42). Available at: https://doi.org/10.12688/openreseurope.13342.2 (Accessed 15 May 2024).

Suhr, B., Dungl, J. and **Stocker, A.** (2020) 'Search, reuse and sharing of research data in materials science and engineering-A qualitative interview study', *PLOS ONE*, 15(9), p. e0239216. Available at: https://doi.org/10.1371/journal.pone.0239216 (Accessed 15 May 2024).

Sun, G., Friedrich, T., Gregory, K. and **Mathiak, B.** (2023) 'Supporting data discovery: A meta-synthesis comparing perspectives of support specialists and researchers', arXiv:2209.14655 [cs]. Available at: http://arxiv.org/abs/2209.14655; https://doi.org/10.5334/dsj-2024-048 (Accessed 15 May 2024).

Wiggins, A., Young, A. and **Kenney, M.A.** (2018) 'Exploring visual representations to support data re-use for interdisciplinary science', *Proceedings of the Association for Information Science and Technology*, 55(1), pp. 554–563. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2018.14505501060; https://doi.org/10.1002/pra2.2018.14505501060 (Accessed 15 May 2024).

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and **Mons, B.** (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1), p. 160018. Available at: https://www.nature.com/articles/sdata201618; https://doi.org/10.1038/sdata.2016.18 (Accessed 15 May 2024).

Zuiderwijk, A. and **Spiers, H.** (2019) 'Sharing and re-using open data: A case study of motivations in astrophysics', *International Journal of Information Management*, 49, pp. 228–241. Available at: https://doi.org/10.1016/j.ijinfomgt.2019.05.024 (Accessed 15 May 2024).

TO CITE THIS ARTICLE:
Kim, S.-Y., Hillemacher, S., Kocher, M., Rumpe, B. and Geisler, S. 2024 The Dataset Finder: A Tool Utilizing Data Management Plans as a Key to Data Discoverability. *Data Science Journal*, 23: 58, pp. 1–17. DOI: https://doi.org/10.5334/dsj-2024-058

Submitted: 15 May 2024
Accepted: 04 December 2024
Published: 31 December 2024

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.