The Open Ecosystem of e-Infrastructures for Data Discovery: A Review

Alessia Bardi, 0000-0002-1112-1292, Institute of Information Science and Technologies, National Research Council, Italy

Peter Kraker, 0000-0002-5238-4195, Open Knowledge Maps, Austria

Brigitte Mathiak, 0000-0003-1793-9615, GESIS, Germany

Heinrich Widmann, 0000-0001-9871-2687, German Climate Computing Center, Germany

Anna-Lena Flügel, 0000-0001-8218-9609, German Climate Computing Center, Germany

Antica Culina, 0000-0003-2910-8085, Ruder Boskovic Institute (IRB), Croatia & NIOO-KNAW, NL

Julien Colomb, 000-0002-3127-5520, HU Berlin

Carole Goble, 0000-0003-1219-2137, The University of Manchester, UK

Tina Heger, 0000-0002-5522-5632, Freie Universität Berlin, Institute of Biology, Germany

Valentina Hiseni, 0000-0002-1357-6127, GESIS, Germany

Navtej Juty, 0000-0002-2036-8350, The University of Manchester, UK

Abstract

Research data are among the fastest growing openly accessible scientific outputs on the web. While we have made great strides when it comes to accessibility of research data, discoverability is still one of the key challenges for open science: in many ways, we cannot cash the cheques written by this movement, if we do not increase the visibility of research outputs.

Many research data discovery services have thus emerged, often embracing the principles of openness. They aim to make data discovery more effective, address new user needs, and exploit new technologies.

This paper aims to support the conception and design of such tools by providing a descriptive framework of the current open ecosystem for research data discovery. In this framework we define the building blocks of the ecosystem (actors, roles and features of discovery services), describe how those interact with each other, and how they support the different discovery needs of researchers.

We analyse the current practices of research data discovery to identify gaps in both the infrastructure and in users' research strategy. We further analyse opportunities for innovative solutions to address the crisis of research data discoverability, improve data discovery and contribute to the evolution of the open ecosystem.

1. Introduction

FAIR research data is critical to ensure research transparency, facilitate data re-use and contribute to reproducibility and innovation. [1][2] Thanks to the Open Science movement [3], and related changes in the policies of journals, funders, and institutions, research data is now the most shared output after the traditional publication [4]. Still, many research communities report difficulties in finding relevant research products (e.g. scientific articles, research software and data) for their research activities because of the so-called “data deluge” and the dispersions of research products, which are scattered across different repositories and archives. [5][6] Kraker et al. [7] discussed the limits of “classic” literature search and identified open infrastructures as the way to overcome the discoverability crisis caused by a general lack of innovation of closed, proprietary search engines.

The same concept and reasoning can be applied to dataset discovery. Closed infrastructures are commonly recognised as a barrier to scientific progress especially during crises, like the spread of Zika and Ebola or the COVID-19 pandemic. In fact, one of the first reactions to crises was to ask for the availability of research papers and FAIR data in Open Access: openness would have sped up discoveries, quickly informed the decisions of policy makers and would have helped prevent fraudulent behaviour (for instance Surgishpere, a US health analytics company, that during the COVID-19 pandemics provided datasets about COVID-19 patients that drove healthcare best practice, but that were subsequently retracted). [8][9][10]

The barriers due to a closed infrastructure hinder effective data discovery, re-use and an equal access to knowledge and data across disciplines, geographical borders, by researchers regardless of the availability of funds and resources from their institutions and research communities.

Numerous services and e-infrastructures have been developed based on the principle of Openness to support data discovery. For example, the crowd sourced spreadsheet managed by Kramer and Bosman, counts, at the time of writing, 100 tools for discovery, 33% of which are open (of a total of about 600 tools that cover different phases of scholarly communication). [11] These infrastructures are often inter-connected and dependent on each other, thus forming an open ecosystem for research data discovery. More tools and services are expected to emerge in the years to come [12] to cover new user needs and exploit new technologies (e.g. AI, machine learning), thus bringing the level of innovation necessary to address the discoverability crisis as suggested by Kraker et al. [7]. The design of these new tools should reflect user needs and should at the same time be compatible with the current ecosystem.

The work described in this paper has been carried out by the GOFAIR Data Discovery Implementation Network (Discovery IN). After providing a collection of use cases identifying user needs [13], we set out to define the existing open ecosystem for data discovery, identifying its building blocks and its gaps. It should be noted that the Discovery IN is mainly composed of European members, and this analysis focuses on the European landscape, mirroring the bias toward users in high income countries in the ecosystem itself. [12]

Outline

Section 2 defines the open ecosystem for data discovery. Section 3 describes it in terms of actors and categories of services. Section 4 discusses how the current ecosystem addresses common discovery patterns. Section 5 presents the current practices in research data discovery and their gaps. Section 6 summarises which are the aspects that need to be clarified or improved to provide a more human-centric open ecosystem for research data discovery and support its evolution.

2. A definition of an open ecosystem for data discovery

During the workshop "Designing a FAIR Data Discovery Ecosystem" at the International FAIR Convergence Symposium (3 December 2020), organised by the Discovery IN, participants were presented with a list of e-infrastructures and services for data discovery clustered by their openness (see Table 1). The subsequent discussion revealed that there is still a debate on which e-infrastructures/services can be considered open and why (e.g. should the e-infrastructure be Open Source or is the availability of Open API enough? Whether to exclude commercial e-infrastructures or not?).

Table 1 List of Open/Closed e-infrastructures and services presented at the workshop "Designing a FAIR Data Discovery Ecosystem" at the International FAIR Convergence Symposium (3 December 2020) organised by the GOFAIR Data Discovery Implementation Network (IN)

e-infrastructure/service	Open/Closed
B2Find	Open
Google Dataset Search	Closed (e.g. no open API)
OpenAIRE	Open
Open Knowledge Maps	Open
Pangaea	Open
ResearchGate	Closed (e.g. no open API)
TextGrid Repository	Open
Zenodo	Open

In this paper, we use the term “open ecosystem for data discovery” to refer to the ecosystem of e-infrastructures, resources and services guided by research communities with open policies and open standards that support researchers at discovering research data for the purposes of research.

This definition, detailed in Table 2, is based on the concepts of the Principles of Open Scholarly Infrastructure (POSI) [14], the Global Sustainability Coalition for Open Science Services (SCOSS) [15] and previous work on data ecosystems [16]–[21], all highlighting the interdependencies among the actors of the ecosystem and the equal importance of social, technical, and policy aspects.

Table 2 includes the definitions of the main terms used throughout the paper in order to base the reasoning on a common ground despite the lack of standard definitions.

Table 2 Definitions

Term	Definition
E-infrastructure	Definition from SCOSS [15]: In an Open Science context, “infrastructure” refers to the scholarly communication resources and services, including software, that we depend upon to enable the scientific and scholarly community to collect, store, organise, access, share, and assess research offered by a provider. In the context of this work, we focused on digital infrastructures, aka e-infrastructures.
E-infrastructure ecosystem	E-Infrastructures that are running autonomously but collaborating via common communication protocols and interoperability standards. The e-infrastructure ecosystem is dynamic: e-infrastructures and their resources and services may change, join, or leave the ecosystem to reflect the advancements of scholarly communication practices or new requirements from the research community.
Open e-infrastructures	E-infrastructures governed or driven by the research community with clear and established open policies, open APIs, and open licences for data, metadata and source code. [14] Open infrastructures therefore remove paywalls, avoid lock-in effects and enable community participation and outreach. They are driven solely by the requirements of the research communities and the goals of Open Science and do not pursue any other proprietary interests.
Data	Research data. Literature offers several definitions of data [22] [23] [24]. A review of the existing definition can be found in [25]. We consider data as defined in the EU directive [26] and used by the European Open Science Cloud (EOSC) [27]: “Data, not in form of scientific publication, collected or produced in the course of scientific research activities and used as evidence in the research process, or

	commonly accepted in the research community as necessary to validate research findings and results.”
Discovery	The act of finding something that had not been known before. [28] In the context of this work, the focus will be discovery services for humans, not machines. Unless otherwise specified, with the term *discovery service* we intend a data discovery service for humans, hence typically, but not necessarily, offered with a Graphical User Interface.

3. The open ecosystem for data discovery

To understand the strengths and weaknesses of the existing open ecosystem for research data discovery, we need first to understand its building blocks, how they interact with each other and how they support researchers' needs.

Several works provide overviews or descriptions of the scholarly communication ecosystem. Some focus on the actual tools and services that are used by researchers in the different phases of the research life cycle, like Kramer’s +400 tools and innovations in scholarly communication [11] and the Census of Scholarly Communication Infrastructure Providers [29]. Others analyse a specific area or application domain of the open data ecosystem to identify building block elements [16]–[21]. The work of Zuiderwijk et al. [16], despite focusing on open government data, sets the basis for our analysis, identifying three main elements of the open data ecosystem: tools and services, data providers, and data users. We expanded this definition with the concept that Jansen [21] called “ecosystem of ecosystems”, referring to the interdependency of vastly different systems managed by different stakeholders, and the notion of actors and roles.

In the next paragraphs, we detail the building blocks of the open ecosystem for data discovery: its enabling and added-value services (see Table 3 and some examples in Figure 1), and how those interact and cooperate with each other and with users (Figure 2).

Table 3 Building blocks of the open ecosystem for data discovery

Actors	**Description**	**Examples**
Researchers	Researchers use the open infrastructure to share their data, make it discoverable, and to discover data shared by others. A researcher, at a given point in time, may be affiliated to one or more organisations. Researchers may also be members of a variety of thematic research	

	communities.	
Publishers	Publishers suggest trusted repositories for the deposition of data that support published articles. Some journals, called “data journals”, are dedicated to papers that describe datasets (“data papers”); others experiment with novel publishing techniques that support new ways for readers to consume the content of a paper and its related data.	PLOS ONE https://journals.plos.org/plosone/s/recommended-repositories; Springer Nature https://www.springernature.com/gp/authors/research-data-policy; Nature Scientific Data https://www.nature.com/sdata/; “reproducible papers” by eLifeSciences.org [30]
Data repositories	Data repositories host metadata and files with research data [22]. They can be categorised along two axes: thematic and geographic coverage (see Figure 1) and offer search&browse portals for data discovery.	Thematic repositories like TextGrid or the GESIS data repository; regional repository like AfricArXiv; institutional thematic repositories like the one of NIOO; institutional, cross-discipline repository like the one of TIB.
Data aggregators	Data aggregators harvest data or metadata collected from data repositories. They can be categorised along the same two axes as data repositories (see Figure 1) and offer search&browse portals for data discovery.	Thematic aggregators like OmicsDI, Movebank and FAIDARE; regional aggregators like Canadian Federated Research Data Repository (FRDR) and cross-discipline, global aggregators like OpenAIRE, Datacite or BASE**.**
Registries of data sources	Directories of data sources that are intended to provide an organised, up-to-date and searchable collection of data sources. [22]	re3data.org fairsharing.org
PID authorities	Organisations that offer services for registering persistent and resolvable identifiers to entities.	Datacite, which issues DOI for research products; ORCID, which issues PIDs for researchers; identifiers.org, which issues PIDs for web accessible records.

Discovery services	Services with a front-end (e.g. portal) offered by a service provider that implement the discovery functionality over a set of data.	The OpenAIRE EXPLORE portal, the BASE portal, the search portal of an institutional repository.
Value-added services	Value-added services re-use content (files and metadata) originally hosted elsewhere and expand the ecosystem with innovative data discovery services that go beyond the traditional keyword based or browse searches. [31]	Open Knowledge Maps, a visual search engine ScholeXplorer, a linking service between publications and datasets.

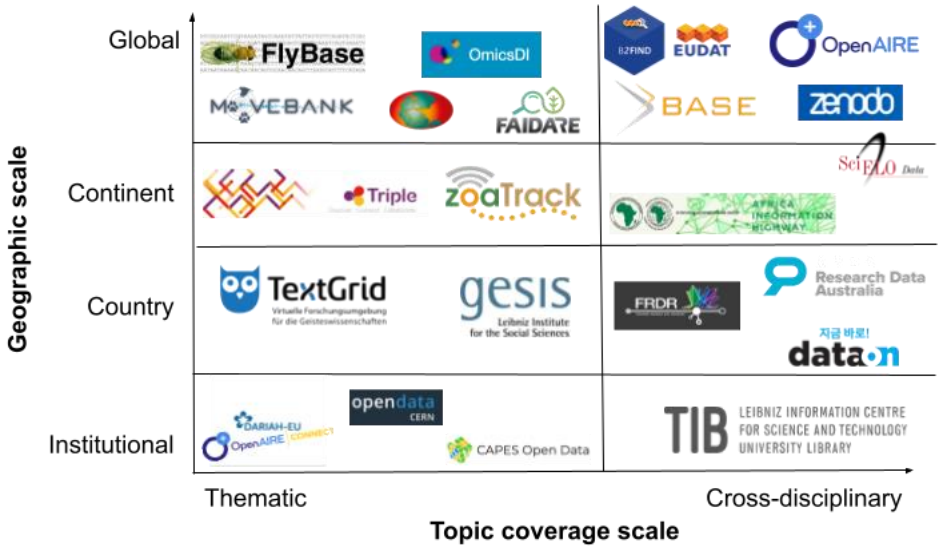

Figure 1 Two main scales along which data repositories and aggregators lie, with some examples

The possible interactions among the service providers (publishers, PID authorities, registries of data repositories, data repositories and aggregators), discovery and added-value services, and humans are depicted in Figure 2.

Researchers (data producers) *deposit data in* a data repository, which possibly *requests a PID for the data* to a PID authority. Researchers may identify the appropriate repository for deposition among those *suggested* by a publisher or using the discovery service offered by a registry of data repositories. The latter *lists* trusted repositories that can be used for research purposes. If the repository has a curation policy in place, data curators *curate* the deposited data and ensure the quality of (meta)data. Examples are data experts annotating datasets in collaboration with the researchers in large research consortia, and domain experts curating thematic collections or databases (e.g. Flybase).

Once deposited, data is made discoverable via the discovery service of the repository. Data aggregators can then *collect its metadata*, making the data discoverable from additional discovery services. In this way datasets become more visible to a wider set of possible data consumers.

Furthermore, value-added services can *use the content* (metadata and data, according to the access rights and licence) to offer innovative discovery services that go beyond the traditional keyword based or browse searches [32]. Our landscape study identified three main typologies of value-added services focusing on data discovery:

- *Visual search engines* like Open Knowledge Maps, which executes live queries on repositories and/or aggregators and presents the results in innovative visual ways.
- *Scholix hubs* [31] like ScholeXplorer, which aggregates links between research literature and data to support the navigation from data to publications and vice versa.
- *Virtual research environments (VREs)* provide web user interfaces and tools for scientists to collaborate or process/manipulate data. [22] The discovery phase is just the initial interaction of the users with the VREs: once the datasets of interest are discovered, the VRE supports the users with its analysis and processing (e.g. proposing algorithms, models, pipelines that can be executed on the selected dataset). Examples include the Galaxy project (https://galaxyproject.org/) [33] and the VREs provided by D4Science.org [34] to the biodiversity, fishery and aquaculture research communities (https://www.d4science.org/integrated-data-catalogue).

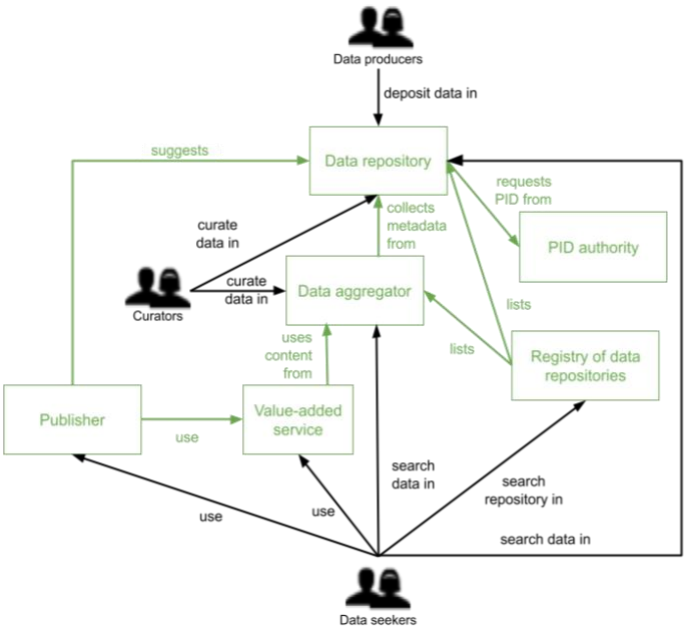

Figure 2 Building blocks of the open ecosystem for data discovery and their interactions

The variety of discovery services helps at serving different research communities (e.g. researchers of a specific domain, researchers in scientometrics and science of science) and personas of researchers (e.g. different levels of digital skills, different level of knowledge of the

domain, different motivations). Finally, because the data discovery ecosystem is dynamic and because metadata can be reused (thanks to open licences and open protocols) new discovery systems can join the open ecosystem for data discovery to address new requirements and needs of researchers.

4. Current practices in data discovery

Data consumers select the discovery service of preference based on their discovery patterns and habits. Zuiderwijk et al. [16] define the term “search” as an interaction between the tools and the users: “A non-linear process containing many feedback loops.” No single platform or service can be seen as independent, as they mirror each other in terms of search technologies, metadata and tools. Users jump from one system to the next with apparent ease, but also out of necessity, as no single resource has everything they need neither in the open nor in the closed infrastructure (despite several high-profile efforts, including Google’s [35], [36]). Instead, small local archives, often community-driven, thrive as increasing numbers of researchers are willing to share and re-use data [4].

Research looking at the “socio-technical practice” of data search [37], or observation of data searchers [38] reveals the strong interconnectivity of these activities with literature search, web search and personal networks, but also among the different services that provide data and data services. This is also the first point of “The Principles of Open Scholarly Infrastructure” [14]: “...research transcends disciplines, geography, institutions and stakeholders. The infrastructure that supports it needs to do the same.” Knowledge infrastructure plays an important role in this process because it "mediates exchanges between creators and consumers by both enabling and restricting the use of that data". [39]

In this section we discuss search patterns and problems as they occur in practice, based on the available literature, the use cases as reported by end-users in surveys conducted by the GOFAIR Discovery Implementation Network [13] and the feedback collected during the workshop “Designing a FAIR Data Discovery Ecosystem” at the International FAIR Convergence Symposium (3 December 2020) organised by the GOFAIR Data Discovery Implementation Network.

Data discovery patterns

While there are no comprehensive studies on the data discovery process, interviews [37], [40]–[42], observations [38] and surveys [13], [43] indicate that the most common strategies to launch data discovery are the following: people (contacting colleagues, visiting conferences), use of generic and domain agnostic search engines, querying domain data repositories, and literature review.(as illustrated in Figure 3). These strategies are not mutually exclusive, and many users seem to follow multiple strategies, depending on the context.

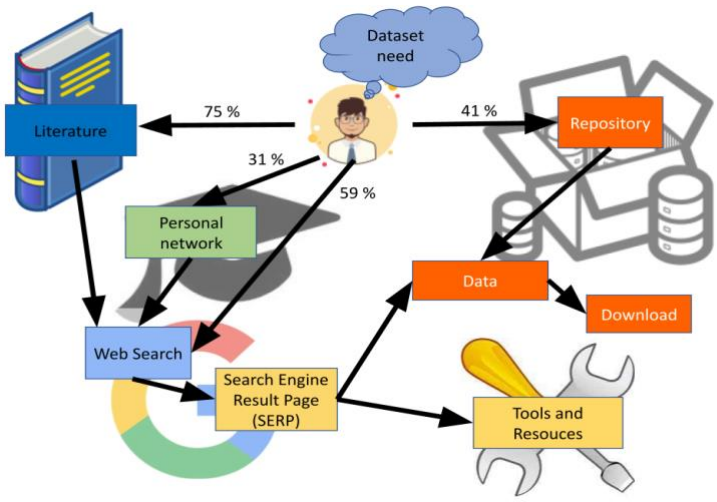

Figure 3 Illustration of researchers' data search strategies (globally distributed, multidisciplinary survey) Percentages from Gregory et al, 2020 [40]

Based on the globally distributed multidisciplinary survey with nearly 1,700 responding researchers conducted by Gregory et al. in 2020 [40], most researchers search for the targeted data via literature (75%), a common web search engine like Google (59%), or directly in a repository or discovery service they already know (41%), rather than first identifying the discovery service that best matches their use case.

Thus, it is crucial for a discovery service to become a part of the general knowledge of the community. Some services have achieved this in a larger user community (e.g. Zenodo and OSF), or their domain-specific community (e.g. Pangaea [41] for georeferenced data from earth system research). Importantly, the general knowledge of the community must be updated and nurtured. Services like re3data.org and fairsharing.org for the discovery of data repositories and data discovery services are building blocks of the open ecosystem, together with other services that build on top of them to provide aggregated views of available research data or suggest thematic repositories based on specific features, like the Data Deposit Recommendation Service of DARIAH-EU (https://ddrs-dev.dariah.eu/ddrs/).

Regarding the types of searches performed by users, log analysis of data search portals [44] confirms that exploratory searches are not so common. In the majority of cases, users search for a specific dataset they know about, and they try to find it by entering the informal name of the dataset directly in the search form of the discovery service. This pattern suggests the need for curated and updated metadata about research data. Using informal names is common due to the generic slow innovation in research data discovery approaches, when compared to literature discovery where we can observe some major advances to facilitate researchers in the so-called "literature deluge", like visual overviews of research topics with Open Knowledge Maps, proposition of new papers based on the current paper read, or a library of papers previously read or published by the user.

The described discovery patterns and search strategies (summarised in Figure 4) are not scalable: the number of available research data is in continuous increase (Zenodo reported 10K

new depositions every week in 2021 [44]). In addition, researchers' expectations about discovery solutions options are increasing higher due to the constant progress of innovations in the fields of artificial intelligence and recommendation systems [45]. The open ecosystem is a good terrain where such innovative practices can be fertilised, piloted and promoted for wider uptake, thanks to its the openness features that characterises it (both at the policy and technical levels) and the direct involvement of research communities.

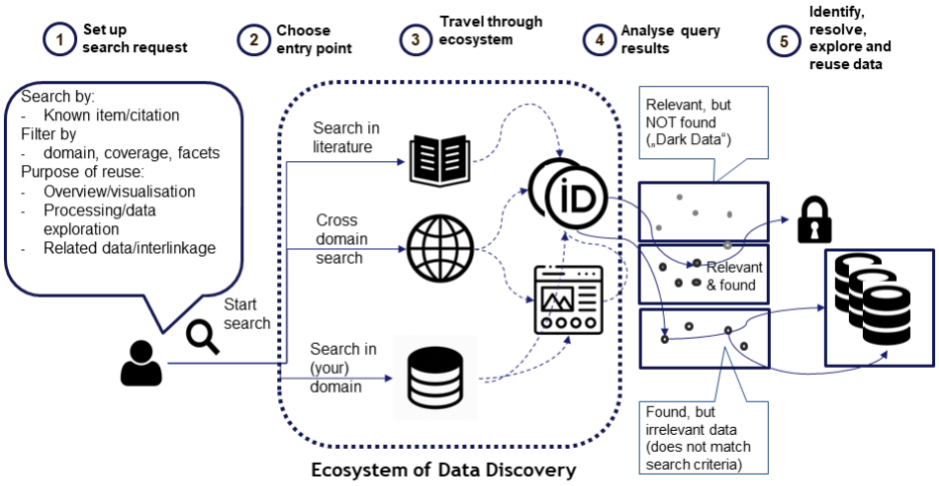

Figure 4 Phases and Stages during the Data Discovery Journey through the Ecosystem

5. Data discovery services: costs and benefits for data consumers

In Section 3 we have identified the set of building blocks of the open ecosystem of data discovery and described how they interact and offer data discovery services.

Not all discovery services support data discovery in the same way and with the same effectiveness. The costs and benefits of the discovery process for data consumers can be expressed in a) time needed to locate (all) relevant datasets; and b) the completeness and the relevance of the search results (i.e. have all the relevant datasets been identified? Are the identified datasets relevant?). Needed time and completeness of results, in turn, depend on consumers' specific need(s), the service coverage (according to its specificity for topic and region) and the service discovery features (metadata quality and search functionality).

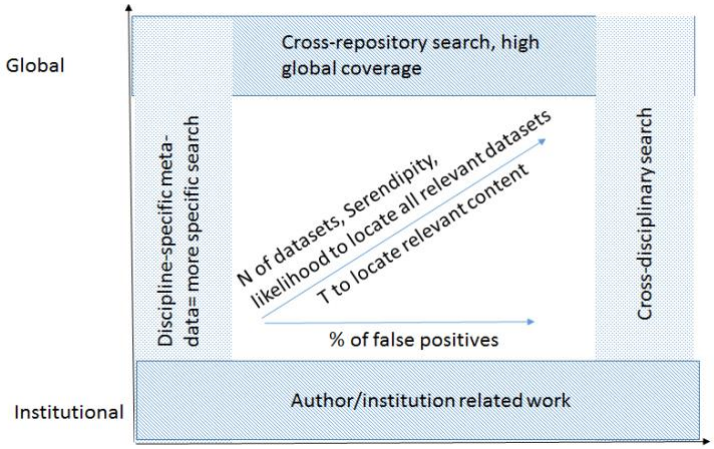

Figure 5 Benefits and limitations of discovery services based on their 'geographic scale' ('institutional' through 'global'; y-axis) and domain scope ('thematic' through to 'cross-discipline'; x-axis)

As summarised in Figure 5, the time to locate relevant datasets increases with the increase of the geographical and topic coverage of the discovery service. On the one hand, global, cross-disciplinary discovery services (like DataCite[1] or OpenAIRE[2]) increase the likelihood to locate datasets of interest with one single query (i.e. the searcher does not need to run the same query on several smaller discovery services and combine their results). On the other hand, their result lists typically return a higher number of false positives, due to technical challenges related to the implementation of effective ranking features and also due to the language in which the queries are expressed by users, especially when searching by terms that are used in different domains with different meanings [46] (e.g. think about "Mouse" in biology and computer science).

Appropriate filters and advanced functionalities such as visualisation services can help users find true positives more effectively and efficiently. In addition, using data discovery services with large geographic and topic-coverage scale is the best option for users who are specifically looking for cross-domain data and wish to maximise the possibility of finding relevant datasets that are useful for them, but originally devised for other domains.

Thematic discovery services, instead, offer the possibility, typically missing in cross-disciplinary discovery services, to perform more advanced searches thanks to a domain specific metadata model. Examples are FlyBase[3] for genetic and molecular data on flies (drosophilidae), or Pangaea[4] for geoscience. In an environment that is highly fractured into many small repositories, discovery services offered by global thematic aggregators can help users to avoid having to locate and search through multiple small databases independently. This way, costs of invested time versus benefits in terms of completeness of results can be maximised.

Institutional services (that can be geographically or domain limited) are not necessarily small or irrelevant. They shine with high specificity, allowing for fast searches, curated and trusted

[1] https://search.datacite.org/
[2] https://explore.openaire.eu/
[3] https://flybase.org/
[4] https://www.pangaea.de/

metadata and very often a community that offers additional services around the data, such as tools to analyse the data, or additional documentation. An example is the CERN Open Data Portal[5], that offers discovery functionality on the data produced by the LHC experiments, together with training material and services for data visualisation and analysis.

6. Gaps in research data discovery

In practice, researchers often report that research data discovery is difficult [13], [32], [47]. While travelling through the pathways of the data lake labyrinth, researchers encounter diverse obstacles, gaps and dead ends depending on their search use-case. For instance, a manager who needs an overview about coarse metadata of project-related repository entries might fare better than a data modeller who needs open data they can download or process via an API. The search outcome may vary: the user may successfully locate all or some relevant data, the discovery journey may lead to data that was not searched for at all, or the search yields no useful datasets at all.

In this section, we discuss known issues with data discovery. We group the identified gaps in problems originating from the user's seeking approach and issues due to shortcomings in the open ecosystem for data discovery.

In [13] the GOFAIR Discovery IN collected and analysed over 100 data discovery use cases and categorised them in twelve clusters based on the required and common discoverability functionalities. In the following, we map these clusters to gaps and obstacles in data discovery. While the cluster 'Metadata for Discovery' is addressed in the gap 'Missing and Misleading Information', the use cases in the cluster 'Data Citation' point to the lack of services that enable data citation search. The use cases in the clusters 'Overview' and 'Discoverability', in turn, indicate that the knowledge about existing discovery services improves scholarly work in terms of the completeness and correctness of the search results.

Table 4 Overview of challenges, deficiencies and gaps in data discovery

Gap	**Search strategy deficiency**	**Data infrastructure deficiency**
1 - Unstructured or missing search strategies	Missing overview of data discovery ecosystem, lack of data search literacy, missing search strategy, imprecise search terms on researcher's side.	Missing metadata, coarse data granularity, missing search facets/filter possibilities.
2 - Inadequate user interfaces		Missing user involvement, lack of innovative features,

[5] https://opendata.cern.ch/

		proprietary licences prohibiting reuse of software and services
3 - Lack of interoperable and interconnected discovery ecosystem		Lack of technical interoperability, organisational cooperation and metadata interoperability between repositories and aggregators, closed and proprietary indexes
4 - Low recall or low precision	Unfitting repository/search engine, imprecise search terms	Suboptimally configured search engine, missing metadata, missing filter options
5 - Problems with identification, access and reuse of data	Misinterpretation or misuse of data found	Missing provenance, missing information of licences and restrictions

Unstructured or missing search strategies

The user's search journey ideally begins by making informed decisions from the start. First, users have to discover the appropriate entry point as discussed already above in the context of Figure 4. They need to decide on whether to search at a cross domain level, or directly go to a domain specific search portal.

The ignorance of suitable entry points is mostly due to a lack of data search literacy. While for literature search, researchers usually get a foundational introduction during their education, which is then deepened during the PhD through constant use, data search is a much rarer occurrence [47].

As such, the most common starting points are usually not a metadata collection for research data, such as would be the norm for literature, but literature search or web search [40], which leads to known item searches and to repositories that can then be queried. Unfortunately, this introduces a strong bias against services that do not have the original copy of the data. Pure metadata collections are not on this search path, nor are value-adding services. And as such many users will never encounter them and thus never learn of their benefits.

Another even more strategic approach is to use repository registries and identify a fitting data repository for the user's search query, e.g. via the Registry of Research Repositories (http://re3data.org) or FAIRSharing (https://fairsharing.org) as a starting point.

For optimised discoverability, the challenge for the user is to design the search query in such a way that it is at the same time sufficiently general to achieve the greatest possible recall and sufficiently specific to achieve the best possible precision. At first glance, formulating the search query seems like a trivial task. But lack of prior knowledge of the existing search functionalities and imprecise setup of the search terms, result in low/bad quality and relevance of the results. The initial search criteria can always be readjusted and refined later on of course. However, the clearer and tighter the search target is specified from the beginning, the more likely dead ends and detours will be avoided during the discovery journey. A good balance must be achieved with respect to the granularity level as well: if a user wants to find a fine-grained, small dataset or even a single file in a big data collection, they can often not find it by searching for it directly. Instead, as a first step the package which comprises the data must be found, and then they can find out whether the specific data (e.g. for a variable in high temporal resolution) is in there. In addition, for some use cases it's not that easy to specify the search request, e.g. if you are looking for sensitive data like patient data from medicinal studies.

The search portal or catalogue provider has to, at the same time, provide high precision search criteria (via high-granularity metadata and data) to return relevant and precise results to the users. Providing fine-grained data for search and access is expensive and not always necessary, but meeting the contextual needs of data consumers usually results in fine-grained (meta)data descriptions. In addition, search guides, well-designed GUIs and search facets targeted the portal's user needs will help with a successful data retrieval (see also the next point).

Inadequate user interfaces

A fundamental issue when it comes to poor discoverability of research data is the lack of adequate user interfaces for data discovery [48]. Often, existing interface concepts for publications are extended to datasets, meaning that they do not take into account different characteristics and challenges of datasets. These include different types of metadata, a wide variety in aggregation (from whole databases to individual files), and different types of access (from downloadable files to databases with their own specific search facilities).

In addition, many innovative features such as visualisations, recommender systems, semantics, content mining, annotation and responsible metrics are not yet widely available for research data discovery. Many frontends are designed from the systems' rather than the users' perspective and fail to cover use cases and requirements of researchers and other stakeholders of research.

It will therefore be important to increase usability and usefulness of data discovery solutions, which can only be done through user involvement and participatory design. Furthermore, to realise the full potential of open research data, users beyond academia need to be taken into account, involving all stakeholders of research data. Furthermore, facilitating reuse of interfaces and user-facing services enables continued innovation in this area. We therefore need FAIR and open infrastructures, which are the prerequisite for this practice.

Lack of interoperable and interconnected discovery ecosystem

From a data provider's view, a lack of technical interoperability between repositories and discovery portals and the (sometimes) insufficient compatibility of metadata standards often result in information loss and thus less metadata for the user to be discovered. Without interoperability, repositories do not form a discovery ecosystem, but standalone data silos that users have trouble finding and/or accessing.

From the researcher's view, a lack of connections between data repositories and discovery services means a significantly lower recall of found datasets. Having to access many different repositories instead of being able to search in one place is time-consuming. In addition, each repository might have different access policies, e.g. only allowing access for institute members, which could impose further hurdles for the data seeker. In the worst case, the user gives up in frustration or settles for a meagre result.

More joined cooperation between the discovery service providers and a top-level federated search portal would improve the situation.

Low recall or low precision

In the worst case where no data is found at all, users may be presented with less relevant or 'irrelevant' results, i.e. referring to data which does not match their search criteria (see gap 1). This issue can be split into three components, related to the metrics recall and precision [22]:

a. **Incomplete results or low recall**: only a small fraction of objects relevant to the query were found, due to lack of identifying relevant knowledge from unfamiliar research domains with their own terminology, semantics and publication culture. This issue is related to 'Dark data', data which are stored and available in data repositories, but never re-used, because they are unlikely to be discovered.

b. **Invalid results or low precision**: only a small fraction of found objects were relevant to the query, due to not being able to filter out only relevant findings from the totality of all found data.

c. **Missing Information, resources and tools** to process the data, although they would be useful to the researcher, because they were not on the search path.

Problems with identification, access and reuse of data

Even when data is found, there are still open challenges. As mentioned above, metadata, documentation and relevant tools to handle the data might not be known to the researcher. For example, a researcher re-uses copyleft data in an improper way by ignoring CC-BY-SA licence. In other cases, open access is not possible at all or at least complicated due to restrictions like embargoes or the need of registration to the data repository before access is granted.

In addition, many datasets and databases are so complex or poorly documented that navigation and interpretation can be difficult. [49] In the worst-case-scenarios, data could be misinterpreted and lead to counter-factual science.

All these challenges pose significant problems for researchers, leading to delays, unnecessarily duplicated work, and unusable results.

Observing these shortcomings of discovery workflows confirms that discoverability itself is in crisis as stated by Kraker et al. [7].

7. Outlook/Conclusion

The open ecosystem for data discovery has rapidly evolved over the last decades and additional innovative tools are expected to be developed to respond to the evolving discovery needs of the research community, and to exploit new technologies to address the data discoverability crisis.

The GOFAIR Data Discovery Implementation Network aims to facilitate the evolution of this ecosystem, through suggesting improvements for existing data discovery services and identifying and releasing new ones. This cannot happen without a clear knowledge about the status quo of the current ecosystem, user needs, discovery patterns and identifying existing deficiencies and gaps. In this paper, we have highlighted the main actors and roles that form the open ecosystem for data discovery, highlighting the connections that exist among them, as well as those that are possible through the adoption of open policies and technologies.

Yet, many gaps for data discovery remain. Based on our analysis of use cases and previous studies, those gaps and their main cause have been identified. Our study resulted in the identification of 5 main gaps, whose causes are both related to deficiencies of the search strategy of the users and of the ecosystem itself: (i) unstructured or missing search strategies; (ii) inadequate user interfaces; (iii) Lack of interoperable and interconnected discovery services; (iv) low recall or low precision; (v) problems with the identification, access, and re-use of data.

It is worth observing that some of the gaps are already being addressed by some initiatives. For example, recent activities in the context of the European Open Science Cloud include establishing an interoperability framework, and developing an overarching Open Science Graph indexing resources within and beyond EOSC. Other examples are the alignment of guidelines for data providers, e.g. by OpenAIRE and EUDAT-B2FIND, community-driven interoperability efforts like the endorsement of using schema.org markup throughout the life science community by initiatives such as bioschema.org, allowing for better indexability by search engines of Life Science resources. ScholeXplorer (http://scholexplorer.openaire.eu) is an example of added value service that supports data discoverability exploiting the links between scientific literature and data made available according to the Scholix framework. In terms of user interfaces, Open Knowledge Maps is for example actively working on adapting their visual discovery approach to take into account the unique characteristics of datasets.

Several new entrants to the data discovery market, however, are following a closed and proprietary model, which means that these services and interfaces cannot be reused, preventing innovation and possibly causing high costs and new paywalls down the line. There is therefore a danger of repeating the same mistakes that were made when it comes to the digital infrastructure for publications, a scenario that should definitely be avoided.

In any case, no single organisation can solve all the problems identified, and no single discovery service suits any researchers' discovery needs.

An important role of the GoFAIR Discovery Implementation Network is to raise more awareness of the gaps that still exist, and to network with stakeholders involved in the open ecosystem in order to come up with sustainable solutions that improve data discovery and evolve the ecosystem for research data discovery towards the fulfilment of researchers' needs.

Acknowledgment

Heger was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number HE 5893/8-1); Flügel was supported by European Commission Horizon 2020 programme under DICE (101017207); Bardi was supported by European Commission Horizon 2020 programme under OpenAIRE Nexus (No. 101017452); Kraker was supported by European Commission Horizon 2020 programme under TRIPLE (grant agreement no. 863420); Widmann was supported by European Commission Horizon Europe programme under FAIRCORE4EOSC (grant code 101057264); Hiseni was supported by German Research Foundation, grant_number 217852844: Smart Harvesting 2 and grant_number 189200501: InFoLiS II - Integration of research literature and data; Colomb was supported by German Research Foundation, grant_number 327654276: SFB 1315: Mechanisms and disturbances in memory consolidation: From synapses to systems; Mathiak was supported by German Research Foundation, grant_number 442494171: KonsortSWD; Culina was supported by the NWO VENI personal grant (grant no. 016.Veni.181.054).

Author contributions

Conceptualization: Bardi, Kraker, and Mathiak. *Methodology*: Bardi, Mathiak, Widmann, and Flügel. *Project administration:* Kraker. *Supervision*: Bardi. *Visualization*: Bardi, Kraker, Mathiak, Culina, and Widmann. *Writing - original draft:* Bardi, Kraker, Mathiak, Widmann, and Flügel. *Writing - review & editing*: Bardi, Kraker, Juty, Mathiak, Culina, Widmann, Goble, Colomb, Flügel, Hiseni, and Heger.

References

[1] J. M. Jeschke, S. Lokatis, I. Bartram, and K. Tockner, 'Knowledge in the dark: scientific challenges and ways forward', *FACETS*, Aug. 2019, doi: 10.1139/facets-2019-0007.

[2] S. Bechhofer *et al.*, 'Why linked data is not enough for scientists', *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, Feb. 2013, doi: 10.1016/j.future.2011.08.004.

[3] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Sci Data*, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.

[4] C. Tenopir *et al.*, 'Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide', *PLoS ONE*, vol. 15, no. 3, p. e0229003, Mar. 2020, doi: 10.1371/journal.pone.0229003.

[5] M. Baglioni *et al.*, 'The OpenAIRE Research Community Dashboard: On Blending Scientific Workflows and Scientific Publishing', in *Digital Libraries for Open Knowledge*, Cham, 2019, pp. 56–69. doi: 10.1007/978-3-030-30760-8_5.

[6] A. Bardi, M. Manunta, E. Toth-Czifra, T. Vergoulis, P. Manghi, and M. Baglioni, 'D7.3 – Interoperability with Research Infrastructures', Sep. 2019, doi: 10.5281/zenodo.3701394.

[7] P. Kraker, M. Schramm, and C. Kittel, 'Discoverability in (a) Crisis', *ABI Technik*, vol. 41, no. 1, pp. 3–12, Feb. 2021, doi: 10.1515/abitech-2021-0003.

[8] 'Coronavirus (COVID-19): sharing research data', *Wellcome*. https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak (accessed Feb. 08, 2022).

[9] L. Besançon *et al.*, 'Open science saves lives: lessons from the COVID-19 pandemic', *BMC Medical Research Methodology*, vol. 21, no. 1, p. 117, Jun. 2021, doi: 10.1186/s12874-021-01304-y.

[10] R. C.-19 W. Group, 'RDA COVID-19 Recommendations and Guidelines on Data Sharing', Jun. 2020, doi: 10.15497/rda00052.

[11] B. Kramer and J. Bosman, '400+ Tools and innovations in scholarly communication'. [Online]. Available: http://bit.ly/innoscholcomm-list

[12] L. Bezuidenhout and J. Havemann, 'The varying openness of digital open science tools'. F1000Research, May 17, 2021. doi: 10.12688/f1000research.26615.2.

[13] B. Mathiak, N. Juty, A. Bardi, J. Colomb, and P. Kraker, 'Discoverability Use Cases to help define Requirements for Research Data Discovery Tools', Zenodo, Dec. 2021. doi: 10.5281/zenodo.5771603.

[14] G. Bilder, J. Lin, and C. Neylon, 'The Principles of Open Scholarly Infrastructure', *The Principles of Open Scholarly Infrastructure*, 2020. https://openscholarlyinfrastructure.org/ (accessed Feb. 08, 2022).

[15] 'Defining open infrastructure – SCOSS – The Global Sustainability Coalition for Open Science Services'. https://scoss.org/what-is-scoss/defining-open-infrastructure/ (accessed Feb. 08, 2022).

[16] A. M. G. Zuiderwijk, M. F. W. H. A. Janssen, and C. B. Davis, 'Innovation with open data: Essential elements of open data ecosystems', *Information Polity, 19 (1-2), 2014*, 2014, Accessed: Feb. 28, 2022. [Online]. Available:

https://repository.tudelft.nl/islandora/object/uuid%3A288c0ce7-70cf-42b9-90a9-05408edc33a8

[17]R. Pollock, 'Building the (Open) Data Ecosystem', *Open Knowledge Foundation blog*, Mar. 31, 2011. https://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/ (accessed Feb. 28, 2022).

[18]A. Jaime, M. A. Osorio-Sanabria, T. Alcántara-Concepción, and P. L. Barreto, 'Mapping the open access ecosystem', *The Journal of Academic Librarianship*, vol. 47, no. 5, p. 102436, Sep. 2021, doi: 10.1016/j.acalib.2021.102436.

[19]R. Adner, 'Ecosystem as Structure: An Actionable Construct for Strategy', *Journal of Management*, vol. 43, no. 1, pp. 39–58, Jan. 2017, doi: 10.1177/0149206316678451.

[20]M. A. Phillips and P. Ritala, 'A complex adaptive systems agenda for ecosystem research methodology', *Technological Forecasting and Social Change*, vol. 148, p. 119739, Nov. 2019, doi: 10.1016/j.techfore.2019.119739.

[21]S. Jansen, 'A focus area maturity model for software ecosystem governance', *Information and Software Technology*, vol. 118, p. 106219, Feb. 2020, doi: 10.1016/j.infsof.2019.106219.

[22]A. Culina, M. Baglioni, T. W. Crowther, M. E. Visser, S. Woutersen-Windhouwer, and P. Manghi, 'Navigating the unfolding open data landscape in ecology and evolution', *Nat Ecol Evol*, vol. 2, no. 3, Art. no. 3, Mar. 2018, doi: 10.1038/s41559-017-0458-2.

[23]'Research data – CASRAI'. https://casrai-test.evision.ca/glossary-term/research-data/ (accessed Feb. 08, 2022).

[24]'Open access - H2020 Online Manual'. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm (accessed Feb. 08, 2022).

[25]A. H. Renear, S. Sacchi, and K. M. Wickett, 'Definitions of dataset in the scientific and technical literature', *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–4, 2010, doi: 10.1002/meet.14504701240.

[26]'EUR-Lex - 32019L1024 - EN - EUR-Lex'. https://eur-lex.europa.eu/eli/dir/2019/1024/oj (accessed Feb. 08, 2022).

[27]'EOSC Glossary', *EOSC Portal*, Mar. 26, 2019. https://eosc-portal.eu/glossary (accessed Feb. 08, 2022).

[28]K. Achenbach *et al.*, 'Defining discovery: Is Google Scholar a discovery platform? An essay on the need for a new approach to scholarly discovery [version 1; peer review: 1 approved, 1 approved with reservations]', *Open Research Europe*, vol. 2, no. 28, 2022, doi: 10.12688/openreseurope.14318.1.

[29]K. Skinner, *Mapping the Scholarly Communication Landscape – 2019 Census*. Atlanta, Georgia: Educopia Institute, 2019. Accessed: Mar. 02, 2022. [Online]. Available: https://educopia.org/2019-census/

[30]‘Introducing eLife’s first computationally reproducible article’, *eLife*, Feb. 20, 2019. https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article (accessed Feb. 08, 2022).

[31]A. Burton *et al.*, ‘The Scholix Framework for Interoperability in Data-Literature Information Exchange’, *D-Lib Magazine*, vol. 23, no. 1/2, Jan. 2017, doi: 10.1045/january2017-burton.

[32]A. Chapman *et al.*, ‘Dataset search: a survey’, *The VLDB Journal*, vol. 29, no. 1, pp. 251–272, Jan. 2020, doi: 10.1007/s00778-019-00564-x.

[33]E. Afgan *et al.*, ‘The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update’, *Nucleic Acids Research*, vol. 46, no. W1, pp. W537–W544, Luglio 2018, doi: 10.1093/nar/gky379.

[34]M. Assante *et al.*, ‘Enacting open science by D4Science’, *Future Generation Computer Systems*, vol. 101, pp. 555–563, Dicembre 2019, doi: 10.1016/j.future.2019.05.063.

[35]O. Benjelloun, S. Chen, and N. Noy, ‘Google Dataset Search by the Numbers’, in *The Semantic Web – ISWC 2020*, Cham, 2020, pp. 667–682. doi: 10.1007/978-3-030-62466-8_41.

[36]D. Brickley, M. Burgess, and N. Noy, ‘Google Dataset Search: Building a search engine for datasets in an open Web ecosystem’, in *The World Wide Web Conference*, New York, NY, USA, Mai 2019, pp. 1365–1375. doi: 10.1145/3308558.3313685.

[37]K. M. Gregory, H. Cousijn, P. Groth, A. Scharnhorst, and S. Wyatt, ‘Understanding data search as a socio-technical practice’, *Journal of Information Science*, vol. 46, no. 4, pp. 459–475, Aug. 2020, doi: 10.1177/0165551519837182.

[38]T. Krämer, A. Papenmeier, Z. Carevic, D. Kern, and B. Mathiak, ‘Data-Seeking Behaviour in the Social Sciences’, *Int J Digit Libr*, vol. 22, no. 2, pp. 175–195, Jun. 2021, doi: 10.1007/s00799-021-00303-0.

[39]C. L. Borgman, P. T. Darch, I. V. Pasquetto, and M. F. Wofford, ‘Our knowledge of knowledge infrastructures: Lessons learned and future directions’, Jun. 2020, Accessed: Feb. 23, 2022. [Online]. Available: https://escholarship.org/uc/item/9rm6b7d4

[40]K. Gregory, P. Groth, A. Scharnhorst, and S. Wyatt, ‘Lost or Found? Discovering Data Needed for Research’, *Harvard Data Science Review*, vol. 2, no. 2, Apr. 2020, doi: 10.1162/99608f92.e38165eb.

[41]C. L. Palmer, ‘Scholarly work and the shaping of digital access’, *Journal of the American Society for Information Science and Technology*, vol. 56, no. 11, pp. 1140–1153, 2005, doi: 10.1002/asi.20204.

[42]L. M. Koesten, E. Kacprzak, J. F. A. Tennison, and E. Simperl, ‘The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour’, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Mai 2017, pp. 1277–1289. doi: 10.1145/3025453.3025838.

[43]T. Friedrich, ‘Looking for data’, Nov. 2020, doi: 10.18452/22173.

[44]A. Ioannidis, ‘Hardening our service’, *Zenodo Blog*, Dec. 07, 2021. https://blog.zenodo.org/2021/12/07/2021-12-07-hardening-our-service/ (accessed Mar. 16, 2022).

[45]P. Kraker, M. Schramm, and C. Kittel, ‘Open Knowledge Maps: Visual Discovery Based on the Principles of Open Science’, *Communications of the Association of Austrian Librarians*, vol. 72, no. 2, Art. no. 2, Oct. 2019, doi: 10.31263/voebm.v72i2.3202.

[46]T. Alrashed, D. Paparas, O. Benjelloun, Y. Sheng, and N. Noy, ‘Dataset or Not? A Study on the Veracity of Semantic Markup for Dataset Pages’, in *The Semantic Web – ISWC 2021*, Cham, 2021, pp. 338–356. doi: 10.1007/978-3-030-88361-4_20.

[47]D. Kern and B. Mathiak, ‘Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?’, in *Research and Advanced Technology for Digital Libraries*, Cham, 2015, pp. 197–208. doi: 10.1007/978-3-319-24592-8_15.

[48]G. Peng *et al.*, ‘Global Community Guidelines for Documenting, Sharing, and Reusing Quality Information of Individual Digital Datasets’, *Data Science Journal*, vol. 21, no. 1, Art. no. 1, Mar. 2022, doi: 10.5334/dsj-2022-008.

[49]GO FAIR Discovery IN, ‘Manifesto of the Discovery GO FAIR Implementation Network: Open User Interfaces for Increased Visibility of Research Results’. 2019. Accessed: May 12, 2022. [Online]. Available: https://www.go-fair.org/wp-content/uploads/2019/02/GO-FAIR-Manifesto_-Discovery-final.pdf