# Bridging the Data Discovery Gap: User-Centric Recommendations for Research Data Repositories

RESEARCH PAPER


MINGFANG WU
FELICITAS LÖFFLER
BRIGITTE MATHIAK
FOTIS PSOMOPOULOS
UWE SCHINDLER
AMIR ARYANI
JORDI BODERA SEMPERE
ANTICA CULINA
ANDREAS CZERNIAK
CHRIS ERDMANN
KATHLEEN GREGORY
NICK JUTY
ALLYSON LISTER
YING-HSANG LIU
SAMANTHA PEARMAN-KANZA

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



## ABSTRACT

Despite substantial investment in research data infrastructure, data discovery remains a fundamental challenge in the era of open science. The proliferation of repositories and the rapid growth of deposited data have not resulted in a corresponding improvement in data findability. Researchers continue to struggle to find data that are relevant to their work, revealing a persistent gap between data availability and data discoverability. Without rich, high-quality metadata, robust and user-centred data discovery systems, and a deeper understanding of how different researchers seek and evaluate data, much of the potential value of open data remains unrealised.

This paper presents a set of practical, evidence-based recommendations for data repositories and discovery service providers aimed at improving data discoverability for both human and machine users. These recommendations emphasise the importance of 1) understanding the search needs and contexts of data users, 2) addressing the roles that data repositories play in enhancing metadata quality to meet users' data search needs, and 3) designing discovery interfaces that support effective and diverse search behaviours. By bridging the gap between data curation practices, discovery system design, and user-centred approaches, this paper argues for a more integrated and strategic approach to data discovery.



**CORRESPONDING AUTHOR:**
**Mingfang Wu**
Australian Research Data Commons, Australia
Mingfang.Wu@ardc.edu.au



**KEYWORDS:**
Data Discovery; FAIR data; FAIR implementation

**TO CITE THIS ARTICLE:**
Wu, M., Löffler, F., Mathiak, B., Psomopoulos, F., Schindler, U., Aryani, A., Bodera Sempere, J., Culina, A., Czerniak, A., Erdmann, C., Gregory, K., Juty, N., Lister, A., Liu, Y.-H., Pearman-Kanza, S 2026 Bridging the Data Discovery Gap: User-Centric Recommendations for Research Data Repositories. *Data Science Journal*, 25: 6, pp. 1–21. DOI: https://doi.org/10.5334/dsj-2026-006

# 1. INTRODUCTION

Researchers, funding bodies, and government agencies increasingly emphasise the importance of depositing research data in recognised repositories. This reflects the broader shift toward open research practices and the critical role of research data plays in advancing research progress and ensuring research integrity. The scale in data availability has grown dramatically over the past decade. For example, DataCite (2024a) reported over 3,000 repositories and over 15 million datasets with Digital Object Identifiers (DOIs) in 2023, reflecting more than a 30 times increase in repositories and a seven times increase in datasets compared to 2014.

This rapid growth in shared data presents a significant challenge: data discoverability. Researchers face substantial hurdles in identifying data relevant to their work (Gregory *et al.*, 2020). Two key issues that underpin the challenge (Liu *et al.*, 2022): firstly, metadata quality often remains low, limiting the effectiveness of search and retrieval; secondly, existing discovery systems often fail to bridge the gap effectively between a user's data search query and the data that could meet their needs. These limitations create friction in the discovery process and degrade the user experience. Without improvements to data discovery, the vast investments made in generating and sharing research data risk being wasted, leaving valuable datasets hidden or underutilised.

The paper presents practical recommendations for repositories, broadly referred to here to include repositories, catalogues, registries, and the like that provide data discovery systems or services, with the aim of improving data discovery. The recommendations adopt a user-centric perspective, contrasting with the more common repository-centric approaches found in existing data management guidance. Foundational frameworks, such as the FAIR Guiding Principles (Findable, Accessible, Interoperable, and Re-useable) (Wilkinson *et al.*, 2016) and subsequent implementation recommendations and frameworks (e.g., Hodson, 2024), have been instrumental in advancing good metadata practices and ensuring metadata and data being FAIR. However, these recommendations primarily address the supply side of data management and are repository-centric, focusing on how data should be described, structured, and published, and they are therefore only part of the data discovery solution. They do not fully account for how researchers search for, interpret, and assess metadata (thus data) within discovery systems. For example, the Findability principle specifies that metadata should be assigned a unique identifier (F1) and data should be described with rich metadata (F2). While these principles lay an essential foundation for data discovery, they don't ensure the metadata is relevant in relation to the user's specific data search query, nor do they account for how users search, for example, by using domain-specific terminologies rather than standardised terms, or for the usability of the discovery interface. To bridge this data discovery gap, the recommendations in this paper are specifically designed around the user's experience and interaction with discovery environments, shifting the focus from simply making metadata FAIR to ensuring that data can be effectively found, evaluated, and reused in real research contexts.

## 1.1 DATA DISCOVERY JOURNEY FROM RESEARCHER PERSPECTIVE

Data discovery is not a simple, isolated activity; it is a complex and iterative process embedded in the research ecosystem and shaped by both the capabilities of the discovery service and the needs of researchers who want to find and reuse data. Figure 1 depicts five key states of a researcher's data discovery journey (Liu *et al.*, 2022), with actions and challenges described below.

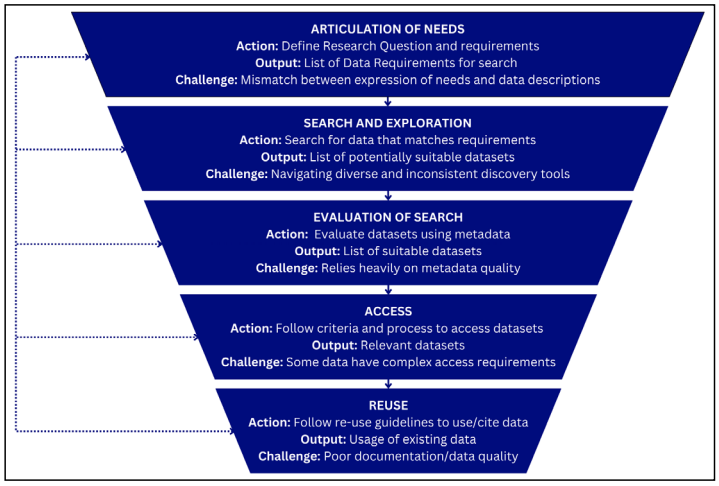


**Figure 1** Iterative data discovery journey.

1. **Articulation of needs**

   The journey begins with researchers identifying a need for data. This stage involves defining a research question, outlining methodological requirements, and specifying the characteristics of the data they need. These needs are influenced by a researcher's disciplinary context, research goals, and the intended use of the data. A common challenge here is the mismatch between how researchers express their needs and how data is described or catalogued in repositories and indexed by discovery services (Löffler *et al.*, 2021).

2. **Search and exploration**

   Once needs are identified, the researchers begin to actively search for data. This is where they engage with a variety of discovery methods and services (Koesten *et al.*, 2021). They may use general Web search engines including specialised data search tools such as Google dataset search, scholarly literature to follow data citations, or through their research networks for recommendations. They may use generalist repositories (e.g., Zenodo, Figshare, DRIAD) or discipline-specific repositories such as PANGAEA for Earth & Environmental Science (Felden *et al.*, 2023) or Inter-university Consortium for Political and Social Research (ICPSR) (Jeng *et al.*, 2017). As they explore, they develop mental models to make sense of the search system, its search result, and the available metadata (Koesten *et al.*, 2021). The efficiency of this process is heavily influenced by the availability and coverage of these services, consistent metadata quality and structure, and variability in metadata support search and interpretation. Usability factors such as interface design, the use of controlled vocabularies, and advanced filtering options that can broaden or narrow a search are also critical (Wu *et al.*, 2019).

3. **Evaluation of search result**

   After discovering potential data, researchers enter the evaluation stage to determine if the data are relevant, reliable, and fit for purpose. Koesten et al. (2017) describes three categories of selection criteria: relevance, usability, and quality. Trust is critical here (Million *et al.*, 2025; CoreTrustSeal, 2022): without adequate transparency and quality assurance, even seemingly relevant data may not be used. At this stage, researchers largely rely on metadata for an initial relevance assessment; they may access the data itself if possible, for a more in-depth assessment of its fitness for their intended purpose. The quality of metadata is even more critical for sensitive data, as it must provide sufficient metadata for evaluation without requiring direct data access, which often involves additional steps and security protocols.

4. **Access**

   If a metadata record indicates its dataset could potentially meet user needs, the next step is gaining access. As a study by Friedrich (2020) indicates, information about data accessibility is highly important, as it helps researchers decide whether to pursue a dataset. For extremely large datasets when downloading or storing data may not be trivial, researchers would like to have a preview and know statistical features of a dataset to make the assessment before they decide to download data or refine their search (Wu *et al.*, 2019). Access can be straightforward for open and small datasets but may involve additional steps for restricted or sensitive data, including application processes, approvals, or negotiation with data custodians.

5. **Reuse**

   Finally, once researchers obtain datasets, the final stage is the use of the obtained datasets. Here, the quality of documentation and data reusability directly affect the extent to which datasets can be integrated into new analyses or combined with other datasets (Liu *et al.*, 2022). Reusing others' data is not merely a technical act but also a scholarly one, requiring recognition of data creators and alignment with ethical and disciplinary standards.

This discovery journey is not linear but an interactive process, where findings at later stages may prompt users to revisit earlier ones. For example, when evaluating search results, a user may reconsider their data needs or formulate new and more effective queries to retrieve relevant datasets.

### 1.2 Key components supporting data discovery

The effectiveness of the data discovery journey relies on the interplay of three major components:

- **Users:** Throughout this paper, the term 'users' refers to researchers, automated software, or AI workflows, who actively search for data through a data discovery system. Their search context, including specific data needs, intended uses, and familiarity with both the data and the discovery system, influences their querying, navigation behaviours, and relevance assessments. The inclusion of 'machine users' signals a fundamental shift in data discovery from purely human-centric to increasingly automated and AI-driven processes.
- **Metadata:** Metadata is fundamental to data discovery, primarily referring to the descriptive records that enable users to find, assess, and reuse data. As the central component of the user's discovery journey, metadata acts as the crucial bridge between a user's search query and the relevant dataset. Consequently, poor metadata quality directly hinders data discovery usability, resulting in poor user experience, under-utilised open datasets, and loss of user trust (Kalinin and Skvortsov, 2023).

  Metadata records are typically created when researchers or data owners (also known as data providers) deposit data into a data repository or when they describe and register data within a data catalogue or a registry of repositories or catalogues. In this process, data providers and data curators generally take responsibility (largely if not solely) for ensuring the proper governance and quality of these metadata records, a process often performed during the initial data ingestion and curation phases.

  Users usually see a complete metadata record on its dedicated webpage, commonly known as a landing page, which should ideally be associated with a persistent identifier (PID). Metadata may also be harvested via APIs, indexed by web search engines, and increasingly used to train AI models.
- **Data discovery systems:** A data discovery system serves as an interface between users and metadata, thus playing a crucial role in data discovery. Such data discovery systems could be offered by a data catalogue, a data repository, a registry of repositories and catalogues, or even general web search. In this paper, we use the term 'repository' to broadly represent those (excluding web search engines) that hold metadata and/or provide a data discovery system/service. The development and maintenance of a data discovery system often involves data managers who establish metadata management policies and procedures, as well as an IT team comprising user experience designers, architects, business/data analysts, and developers who provide the necessary technical expertise and support.

Effective data discovery is shared among all stakeholders, including data discovery system providers, data curators, metadata standard bodies, data managers, etc. Improving the discovery ecosystem requires a coordinated approach where these groups work together. Within this broader context, the recommendations presented in this paper focus on the role of data repositories and actions they can take to improve research data discovery.

## 2. METHODOLOGY

We adopt participatory research methods by theoretical analysis with practitioners' practical experience (Bergold and Thomas, 2012). Specifically, we include all research partners in this collaborative knowledge creation process that brought together a diverse group of experts from the Research Data Alliance's (RDA) Data Discovery Paradigms Interest Group, including information and data scientists, data curation and management practitioners, and repository professionals.

The approach combined extensive community engagement and literature reviews. Following the principles of participatory research of democracy as a precondition, the need for a safe place, community formation and degrees of participation (Bergold and Thomas, 2012), we facilitated open dialogue and knowledge-sharing through the group's monthly meetings, RDA plenary sessions, and ongoing group discussions.

Critically, we grounded our work in the professional experience of the group's members who have direct experience in developing and operating data discovery systems. We also applied heuristic methods, a common technique in usability evaluation, to analyse real-world examples and practices from various data repositories and scholarly literature.

This iterative, evidence-informed process prioritises understanding the needs, behaviours, and expectations of users. By integrating perspectives from user-centred discovery, discovery system design, and data curation, we derived ten principles to improve data discoverability for both human and machine users, with use cases, motivation, and example-based recommendations provided for each principle. These principles also articulate shared responsibility across multiple stakeholders, including data repositories, data curators, publishers, funders, etc. (Wu *et al.*, 2024). Building directly on these ten principles, we synthesised them into four consolidated recommendations, focusing on the specific responsibilities and contributions of data repositories, with each recommendation reflecting specific groups of principles.[1]

## 3. RECOMMENDATIONS FOR IMPROVING DATA DISCOVERABILITY

This section outlines four key recommendations that address each of the three components to support a user's data discovery journey. The first recommendation reviews existing user study methods that repositories can apply to understand users' needs and search behaviours. The second recommendation focuses on integrating the repository's data discovery service with a broader data discovery ecosystem and making repositories known and discoverable by their targeted research communities. The final two recommendations address the technical foundation of discovery within a repository: metadata quality and discovery interface. While each of these four recommendations warrants extensive discussion on its theory, evaluation, and lessons learned, we provide practical guidance and highlight existing good practices. Relevant references and citations are included to offer further evidence and detailed reading for repositories that choose to pursue specific recommendations.

### RECOMMENDATION 1: USER-CENTRIC DESIGN: UNDERSTAND USERS' NEEDS AND SEARCH BEHAVIOURS

Understanding needs, behaviours, and expectations of users is paramount for the success and effectiveness of a data discovery service. This understanding directly informs critical aspects such as metadata requirements, design, implementation, and success measures. Demonstrating a clear value proposition to funders and data providers, derived from meeting user needs, is crucial for ensuring a service's long-term sustainability.

Effective data discovery services are built upon understanding their users by addressing key questions:

1. **User profiles:** Who are the intended users? What is their familiarity with the discovery system and their knowledge of the repository's subject domains?
2. **Search motivations:** Why do users seek data? What are their specific data requirements, what triggers their search within the service, and what are their intended uses for the sought data?
3. **Discovery strategies:** How do users typically find data? Do they utilise interactive portal search features, follow literature, browse, or use APIs?

Investigating these aspects enables data discovery services to address real user needs, leading to increased service engagement, and, consequently, to encouraging greater reuse of research data.

#### Recommendation 1.1: Adopt appropriate user study methods for intended purposes

Drawing from user experience research, interactive information retrieval, and human-computer interaction, a range of established methods can be employed to understand user needs and data discovery behaviours. These methods, commonly applied in digital libraries and web search, are increasingly used to study users' needs and behaviours (Sostek & Russell *et al.*, 2024). Most-used methods include surveys for broad data collection on user profiles, needs, satisfaction, and usability. Interviews, particularly using techniques like the Critical Incident Technique (Davenport, 2010; Flanagan, 1954), allow for in-depth exploration of user motivations and experiences. Interaction log analysis provides objective behavioural data on

1 For example, Principles 1 and 8 inform Recommendation 1; Principles 2 and 3 inform Recommendation 2; Principles 3, 4, 5, 6, 9, and 10 inform Recommendation 3; and Principles 4, 6, and 7 inform Recommendation 4.

search patterns and system usage. A/B testing enables the empirical comparison of design alternatives to optimize user experience. Finally, observational studies offer direct insights into usability barriers by observing users in their natural data-seeking environments. Table 1 provides a summary of the pros and cons of each method and gives examples through citations. These methods can be used individually or, ideally, in combination to provide a comprehensive understanding of users. For example, although interaction logs provide objective data on user behaviours, they offer no insight into crucial subjective factors like users' motivation, perception, and satisfaction, which are better explored through surveys, interviews, or observation studies. Integrating these complementary methods can provide a more holistic and robust understanding of user behaviour (White, 2016).

**Table 1** Pros and cons of user study methods.

| METHODS | DESCRIPTION | PROS | CONS | EXAMPLE TOOLS |
|---|---|---|---|---|
| Survey | Collect insights on user needs, functional requirements, user background, and satisfaction with a data discovery system (e.g., System Usability Score,[2] Khalsa *et al.*, 2018) | Cost-effective, scalable, can reach a large number of users, can collect both quantitative and qualitative data. | Limited to self-reported perceptions, may lack depth, response bias can be an issue. | General: SurveyMonkey, Qualtrics, Google forms, Typeforms, Microsoft Forms |
| Interview | Structured or unstructured conversation with an individual user or a group of users (focus group) (e.g., Liu *et al.*, 2023, Sostek *et al.*, 2024). | Provide in-depth insights into user motivations for a data search and experience with a discovery service, allow for clarification and follow-up questions, useful for exploratory research. | Time-consuming, resource-intensive, findings may not be easily generalizable, requires skilled interviewers and interview data analysis. | Self-developed cheatsheet with interview steps and questions. AI tools can be used to transcript interview recordings and analyse interview transcripts (Wollin-Giering *et al.*, 2024) that need to be verified by researchers. |
| Interaction log analysis | Analysis of search logs that captures user interactions with a system, allowing the analysis of search patterns, quality of relevance ranking, query and click behaviours, etc. (e.g., Kacprzak *et al.*, 2018; Sharifpour *et al.*, 2023) | Captures actual user behaviour beyond self-reporting, reveals data search patterns and system usage, can be tailored to specific research questions. | Lacks contextual explanations for behaviour, requires technical expertise for analysis, may not capture user motivations or frustrations directly. | Google analytics or Matomo (Quintel and Wilson, 2020) for general web traffic, most visited pages, user activities (e.g., page view, position of clicked search result).<br>Self-coding for advance analysis of targeted investigation/research questions. |
| A/B testing | A/B testing compares two design alternatives to measure user preference and impact. This can be done with low-fidelity wireframes or fully functional systems (Vega-Gorgojo *et al.*, 2016; Löffler *et al.*, 2023). | Provides empirical evidence of design effectiveness, allows for direct comparison of alternatives, and can reach a large number of users. | Requires careful experiment design to ensure validity and avoid bias, can be time and cost-intensive, may not explain *why* one design performs better. | By randomly directing real users to alternative sites and analyse logs, or in a controlled setting where recruited users testing different designs and provide feedback[3] (tools like Crazy Egg and Hotjar can record heatmaps and scroll maps). |
| Observational study | Observing users as they search for data, either in controlled environments or their natural workflow (e.g., Thomas *et al.*, 2021) | Provides direct insights into challenges and pain points reveals unexpected behaviours and pain points. | Resource-intensive (time, personnel), potential for observer bias, user behaviour may be influenced by observation, findings may not be easily generalizable. | Can utilise screen and session recording tools, e.g., Lookback, Silverback, Hotjar, and Crazy Egg (for heatmaps). |

### Recommendation 1.2: Adopt appropriate user study methods at different stage of data discovery service development

The choice of method(s) should be guided by the repository's maturity and available resources. A structured, phased approach to user engagement is crucial for continuous improvement. Below and in Figure 2, we outline three broad stages of repository development by grouping 7 phases of an information system development life cycle (Pressman and Maxim, 2015) and suggest suitable study methods for each:

1. **System ideation and design:** This initial stage focuses on user requirement gathering, data source identification, and system architecture design. Methods such as surveys, interviews, and focus groups are ideal here for these activities.

2 System Usability Score: https://measuringu.com/sus/.

3 NN/g: A/B testing 101 https://www.nngroup.com/articles/ab-testing/.

2. **New service development and a minimal viable product (MVP) launch:** During this stage, the Minimum Viable Product is developed in consultation with stakeholders, leading to its first public release. Methods like surveys, A/B testing, and observation are appropriate for gathering iterative input and testing a product's usability and effectiveness before its full launch. This allows for rapid adjustments and improvements based on early user interactions.
3. **Service refinement and evolution:** As the service matures and receives feedback from its user base, existing features are refined, and new features are added based on user feedback and the availability of new technology. Continuous use of A/B testing, observation, and surveys, complemented by interaction log analysis, is crucial for evidence-based refinement.

The progression of user study methods across development phases highlights an iterative, agile approach to developing a data discovery system. This continuous feedback loop ensures that the discovery system remains responsive to evolving user needs and technological advancements, rather than being a static product.

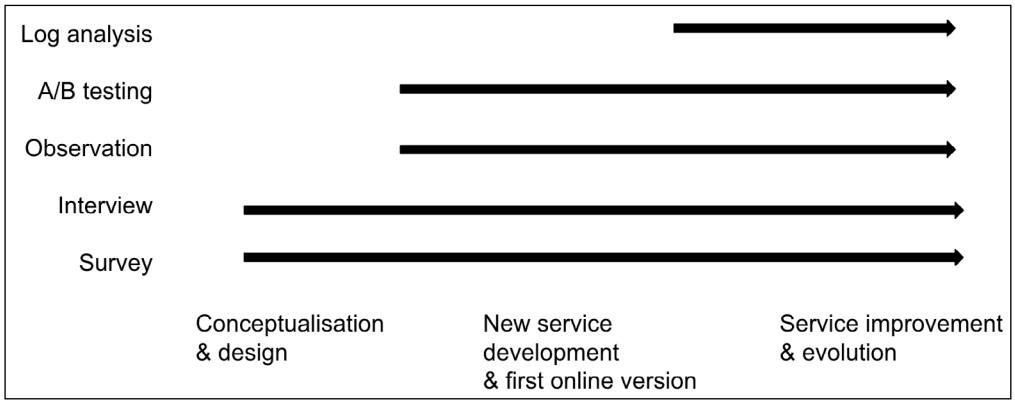


**Figure 2** Recommended user study methods at different stages of data discovery service development.

## RECOMMENDATION 2: LEVERAGING EXTERNAL ECOSYSTEMS FOR DATA DISCOVERY

To ensure research data is discoverable, repositories must cater to a variety of search behaviours by supporting multiple access points and discovery paths. This involves enhancing a repository's presence not only on the web but also within academic literature and research communities, as studies (Gregory *et al.*, 2020; Liu *et al.*, 2022) reveal that web search, following literature, and research networks remain common methods among researchers.

### Recommendation 2.1: Enhance web presence

Web search engines play an important role for researchers to discover data; Gregory et al. (2020) found that about 59% of surveyed researchers find data via web search. Repositories also report a similar trend. For example, Research Data Australia, a national data catalogue, reports that the majority of their users arrive via web searches (Wu, 2022); similarly, PANGAEA, a data repository for Earth & Environmental Science, finds that about 31% of users come through web search engines. To improve web discoverability, repositories should:

- **Register repositories:** Registering repositories with international open-access resource registries, such as OpenAIRE,[4] FAIRsharing[5] or re3data,[6] enhances their exposure to web search engines. These registries act as a form of reputable online directories that search engines such as Google regularly crawl and index.[7] Reputable registries often have good domain authority; when they link to a repository, it can provide a valuable backlink that signals to web search engines that the repository is a trustworthy data resource. A discoverable and authoritative repository will boost its data discoverability.[8]

4 OpenAIRE: https://www.openaire.eu/.

5 FAIRsharing registry of standards, databases and policies: https://fairsharing.org/.

6 Registry of research data repositories: https://www.re3data.org/.

7 Overview of crawling and index: https://developers.google.com/search/docs/crawling-indexing.

8 A guide to Google Search Engine Ranking Systems: https://developers.google.com/search/docs/appearance/ranking-systems-guide.

- **Apply SEO best practices:** Utilising Search Engine Optimization (SEO) techniques can help web search engines understand a repository and assist users in finding the site.[9] It is beneficial to leverage existing reputable domains, such as university domains, for linking to repositories, as these often have a high ranking. Regularly reviewing search logs or web tracking data (see Recommendation 1), if available, can help in identifying and addressing problems as they arise.
- **Embed machine-readable metadata:** It is important to embed structured, machine-readable metadata on all metadata webpages using vocabulary like schema.org[10] (Wu *et al.*, 2021). This not only improves a repository's ranking in search results but also facilitates seamless interoperability with other metadata aggregators (e.g., https://oceaninfohub.org/) and data discovery systems.

### Recommendation 2.2. Appear in literature and research communities

Researchers also frequently discover data through academic literature and their professional networks (Gregory *et al.*, 2020; Liu *et al.*, 2022). To tap into these discovery pathways, repositories should:

- **Provide downloadable data citations:** Repositories should offer clear citation guidance and downloadable citation formats on all metadata landing pages, for example, platforms like Zenodo exemplify this approach by exporting citations in major styles (e.g. APA, Harvard, MLA, etc.) as required by publishers (Stall *et al.*, 2023), this encourages and simplifies the process for researchers to cite data in publications, therefore enhancing the data discoverable through citation indexes.
- **Link metadata to publications and related research objects:** Repositories should adopt metadata schemas that enable data description to be linked to other related research objects, such as models, software, and publications, that are either used to analyse data or generated from it. Some metadata schemas already support such relational linking and description. For example, the DataCite schema (2024b) provides a suite of controlled vocabularies for describing 'relationType' (e.g., isDerivedFrom, isCitedBy/Cites), many of which are bidirectional. Establishing linking and bidirectional links not only enhances data discoverability but also supports the development of linked open data and knowledge graphs, e.g., Scholix (Burton *et al.*, 2017) and the OpenAIRE Graph (Manghi *et al.*, 2019); therefore, enabling more structured, machine-driven discovery queries.
- **Engage with research communities:** Beyond citations, repositories should actively engage with research/user communities, for example, by joining academic societies, organising data focused workshops at conferences, and hosting data challenge events. Such proactive involvement with communities that generate and use data increases repository visibility and, in return, the discoverability of its data. It also fosters a culture of data sharing and contributes to the growth of the repository's data holding, thereby supporting its long-term sustainability as a data infrastructure (Cooper and Springer, 2019).

## RECOMMENDATION 3: ENSURE HIGH-QUALITY, STRUCTURED, INTEROPERABLE, AND DISCOVERABLE METADATA

Metadata is fundamental to data discovery. As discussed in the user's discovery journey, metadata acts as the bridge between a user's search query and relevant data, and is crucial for relevance assessment, data access and reuse. The effectiveness of this bridge depends on metadata quality, structure, and interoperability. While FAIR (meta)data principles (Wilkinson *et al.*, 2016) provide a high-level roadmap, this recommendation offers practical guidance for repositories on how to implement them from the user's perspective. The following sub-recommendations are structured to guide repositories from foundational strategies like interoperability and semantic richness to granular description and robust quality assurance processes.

9 SEO Starter guide: https://developers.google.com/search/docs/fundamentals/seo-starter-guide.

10 Schema.org Schemas is a standard vocabulary that enables consistent, machine-readable description of resources on the web: https://schema.org/docs/schemas.html.

### Recommendation 3.1: Enhance coverage and interoperability

Repositories should balance generalist and discipline-specific collections by prioritising metadata interoperability and broad disciplinary coverage to support researchers from both disciplinary and multidisciplinary backgrounds (Smith, 2020). This requires a harmonised yet flexible approach where a common set of metadata attributes is adopted across all disciplines, while also allowing for discipline-specific extensions, to support a spectrum of exploratory and granular searches by users with different backgrounds and search strategies (Marchionini, 2006).

- **Improve metadata coverage and exchange through strategic partnerships:** To enhance data coverage and ensure broad representation, especially in underrepresented disciplines, repositories should regularly exchange metadata with other relevant disciplinary and generalist repositories and engage with targeted disciplinary communities (see Recommendation 2.2) for promoting data sharing. Many repositories provide a metadata feed for other repositories to harvest. For example, the data.gov platform not only harvests metadata from all participating agencies but also offers its complete metadata catalogue via RESTful API, allowing other repositories to easily ingest. Beyond metadata exchange, discoverability can also be improved by integrating the search functionality of specialised partners, such as by recommending a search result from a targeted repository's API within the primary repository's search results.
- **Use standardised metadata schemas and vocabularies:** Improving metadata interoperability through standardised schemas, structured vocabularies and ontologies, as well as machine-actionable encoding, is essential for metadata exchange and aggregation with other repositories. The WorldFAIR project (Gregory *et al.*, 2024) has been investigating the cross-domain interoperability framework (CDIF) by analysing use cases from eleven disciplines and provided guidance by recommending vocabularies from Schema.org and DCAT for data discoverability, and DDI-CDI[11] for data structure, and SKOS[12]/XKOS[13] and OWL[14] for semantics. Resources such as FAIRsharing (Lister and Sansone, 2023) provide a searchable registry of data and metadata standards that can help repositories identify relevant schemas and vocabularies.

### Recommendation 3.2: Support flexible searching across broad and specific terms

A discovery system should support searching across both broad and specific terms to meet diverse needs of both human users and machine agents. Broad terms enable exploratory discovery and topic navigation when human users are unfamiliar with precise terminology (Lafia *et al.*, 2023; Sharifpour *et al.*, 2023), while specific terms allow domain experts and automated workflows to retrieve highly relevant data. Supporting both ensures more comprehensive discovery and improves search success for all users.

To achieve this, repositories should leverage multiple classification systems simultaneously. For example, this Health Data Australia (HDA)[15] portal supports both the general ANZSRC Field of Research[16] classification and domain-specific vocabularies such as MeSH or ANZCTR[17] Conditions for clinical data. Additionally, repositories can enable vocabulary mapping to connect data described using different terminologies, such as between MeSH and the International Statistical Classification of Diseases and Related Health Problems (National Library of Medicine, 2021) and the mapping of vegetation classification systems across different countries (Sun *et al.*, 1997). Including mapped terms in indexes enables richer search filters and query expansion support for human users (see Recommendation 4) (Smith, 2020). For machine users, well-structured

11 DDI CDI: Cross-Domain Integration: https://ddialliance.org/Specification/ddi-cdi.

12 SKOS Simple Knowledge Organisation System is a standard model for encoding controlled vocabularies for interoperable metadata: https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html.

13 XKOS extends SKOS and encodes statistical classifications for interoperable metadata: https://linked-statistics.github.io/xkos/xkos-best-practices.html.

14 Using OWL and SKOS: https://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html.

15 Health Data Australia: https://researchdata.edu.au/health/.

16 The Australian and New Zealand Standard Research Classification (ANZSRC-FoR): https://www.arc.gov.au/manage-your-grant/classification-codes-rfcd-seo-and-anzsic-codes.

17 Australian New Zealand Clinical Trials Registry (ANZCTR): https://www.who.int/clinical-trials-registry-platform/network/primary-registries/australian-new-zealand-clinical-trials-registry-(anzctr).

hierarchical vocabularies and mappings are critical for enabling more sophisticated, automated query expansion and semantic search capabilities (Allahim *et al.*, 2025; Nasir *et al.*, 2019).

Implementing and maintaining these capabilities requires ongoing curation effort, especially since researchers or depositors often provide minimal or inconsistent metadata during self-deposit. Repositories can provide proactive deposit support by offering structured metadata templates and clear guidelines during the deposit process to encourage the use of controlled vocabularies. Repositories can apply post-deposit enrichment either by repository staff, community curators, or through automated workflow utilising NLP tools (e.g., BioPortal[18] and AgroPortal[19] annotators) to enrich metadata and maintain vocabulary mapping over time.

### Recommendation 3.3: Describe data at different granular levels (e.g., collection, item, variable)

**Differentiate granular levels:** Repositories should implement data annotation strategies that differentiate between collection-level and item-level, or between study-level and variable-level metadata, depending on disciplinary conventions. For example, in a national election survey, the study-level metadata describes the overall survey design, whereas variable-level metadata details specific questions or variables (e.g., age group or suburb), which can be reused in other analyses. Enriching higher-lever metadata with key terms from more granular levels can improve search engine ranking for item-level records (Foulonneau *et al.*, 2005), and support both within disciplinary and across disciplinary data searches. The PANGAEA repository, for instance, linked sub-level dataset references to parent collections, as shown in Figure 3. Collection-level metadata summarises the metadata of sub-level datasets, which may differ in structure but are interconnected, such as in bundled publications (Felden *et al.*, 2023). This allows users to find (and cite) the whole bundle, but still navigate to the sub-level datasets for a more detailed discovery.

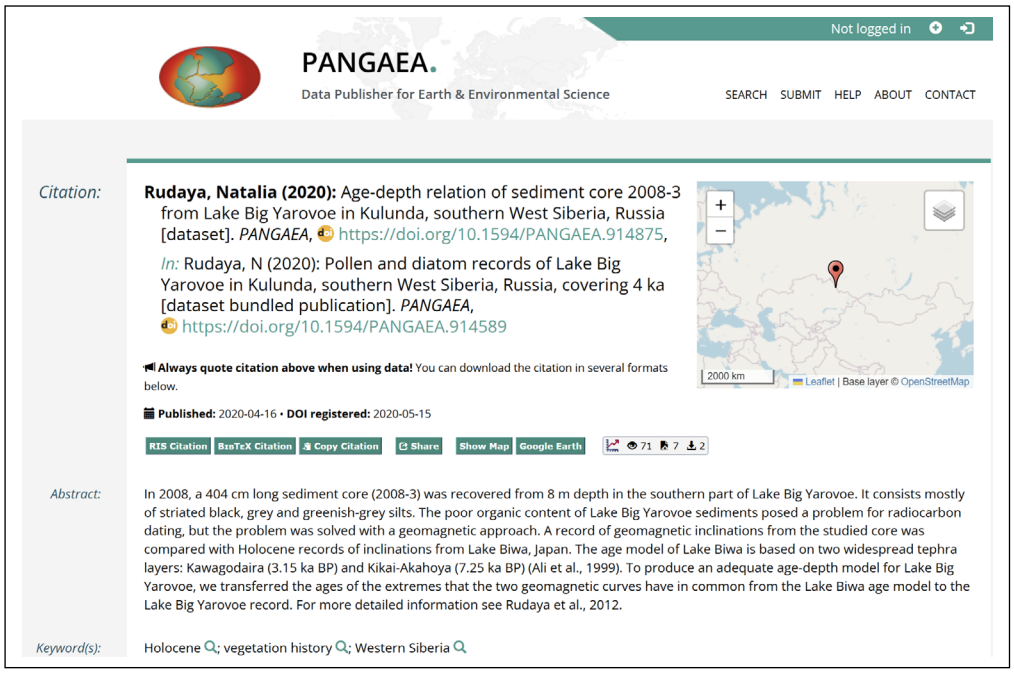


**Figure 3** An example of a sub-level dataset which refers to a collection.

Determining the right granularity requires close collaboration with research communities to ensure annotations support discovery and align with shared standards. Repositories must also be cautious about privacy issues (Slavković and Seeman, 2023), as granular metadata (e.g., postcode, ethnicity, or health conditions) may contain sensitive information requiring specific protocols and access controls to protect privacy.

**Use structural metadata to link resources:** Repositories are encouraged to adopt schemas with vocabularies for structural metadata, which describes the relationships between resources or their parts, such as sequence or hierarchy (NISO, 2004). Relation vocabularies can include general terms (e.g., 'partOf', 'isVersionOf' from Schema.org, or concept hierarchy 'skos:broader'

18 National Center for Biomedical Ontology Bioportal: https://bioportal.bioontology.org/annotator.

19 AgroPortal: https://agroportal.lirmm.fr/.

and 'skos:narrower' from SKOS) or domain-specific ones (e.g., 'mutualism', 'commensalism', 'parasitism', and 'competition' between species) (Lang and Benbow, 2013). Structural metadata helps organise related resources, enhances research result ranking (e.g., boosts datasets linked to many resources), guides exploratory navigation, and enables linked data for both human and machine users (Taniguchi and Hashizume, 2023).

### Recommendation 3.4: Implement robust metadata quality assurance

Improving metadata quality means ensuring metadata is fit for purpose. Metadata quality is inherently multidimensional, encompassing aspects such as completeness, accuracy, consistency, timeliness, and relevance (Wang and Strong, 1996). For supporting effective data discovery, metadata should be structured and described in ways that support how intended users search and interpret underlying data, as metadata quality is critical for a data search system to be able to retrieve metadata records that are relevant to a user's query, and users also rely heavily on metadata for initial relevance assessment and determination of fitness for their intended purpose.

- **Define and enforce core metadata attributes:** While FAIRness assessments provide a valuable foundation, metadata quality extends beyond FAIR to include the richness, accuracy, and usability of metadata from both human and machine perspectives (Peng *et al.*, 2024).

  Because "quality" is inherently context-dependent, and studies report that metadata currently poorly reflecting user needs is the biggest obstacle in retrieving relevant data (Löffler *et al.*, 2021), repositories should work with their user communities to identify which metadata elements, controlled vocabularies, and levels of descriptive detail are most critical for successful discovery and assessment in their field.

  Based on user input, repositories should define a core set of essential 'discovery' metadata attributes that are mandatory. This core set must include not only domain-independent metadata fields but also domain-specific elements crucial for contextual understanding (e.g., biodiversity domain environments, species, materials, and chemicals for biodiversity research) (Löffler *et al.*, 2021). Repositories should also prioritise fields that add semantic richness (e.g., linking to relevant software PIDs or providing detailed provenance) to improve both machine readability and human interpretation (Alencar *et al.*, 2024). This ensures the metadata provides not just descriptive but also contextual information for determining fitness for purpose.

- **Integrating proactive support and metadata quality assessment**

  Effective metadata quality assurance requires a dual approach: proactive support for data providers and holistic validation of the resulting metadata quality.

  First, help support providers in generating good metadata by providing guidance, tools, and automated checks (e.g., metadata quality assessment tools such as https://data.europa.eu/mqa/methodology) (Wentzel *et al.*, 2023). This upfront support helps data providers create records that accurately and completely represent the data within the schema, mitigating poor-quality metadata like poorly written descriptions without considering users' familiarity with disciplinary terminologies, missing vital information for intended use, metadata elements without using standard vocabularies (e.g., subject headings, variable names from a community-endorsed dictionary), or inconsistent naming conventions.

  Secondly, validate and assess metadata FAIRness and quality by establishing a baseline using assessment tools like the FAIRness assessment and maturity metrics (Candela *et al.*, 2024; *Krans et al.*, 2022). These assessments should be complemented with broader quality frameworks (Lacagnina *et al.*, 2023) and user-driven evaluations to ensure metadata truly meets discovery and reuse needs. For example, organisations like NASA improve Earth observation data discovery by assessing metadata quality focused on correctness, completeness, and consistency (Bugbee *et al.*, 2021), demonstrating the importance of going beyond FAIR to evaluate and improve metadata in practice.

### Recommendation 3.5: Manage duplicated metadata records

Duplication is a top repository data quality issue as reported in this interview study (Liu *et al.*, 2023), mirroring challenges found in large-scale bibliography systems like WorldCat (Weitz, 2020), where institutions like OCLC provide clear instructions for new record entry.[20] For data repositories, duplication commonly occurs in two main scenarios (Wu *et al.*, 2024): 1) Metadata cross-aggregation: repositories cross-aggregating metadata from multiple sources often end up with identical or augmented records that refer to the same underlying data and 2) Institutional requirements: researchers or data creators are asked by their respective institutions to register the same data to their local institutional repositories, leading to multiple copies of the metadata across different systems. These duplicate metadata records clutter search results, wasting user time and degrading the user experience.

- **Use PIDs or link to the primary PIDs:** Having a PID for the same data and linking metadata records of the same data helps prevent and identify duplication in the first place. This requires repositories to follow a similar procedure of allowing data depositors to register data using an existing PID (e.g., DOI) for repositories to retrieve metadata from the PID (e.g., DataCite API) instead of creating yet another metadata of the same data. If direct PID registration isn't feasible, repositories should allow linking duplicated or nearly identical metadata records through the relation, such as 'IsIdenticalTo' (DataCite, 2024b).
- **Version-augmented metadata records:** A repository should set up a policy and process of how to handle augmented metadata records to make it easy to detect augmented metadata records for the same dataset. Zenodo provides a good example here;[21] It handles metadata versions by allowing creators to 1) make editorial changes without changing its DOI, and 2) create new distinct versions of a metadata record with a new DOI when having significant updates to data files. If an augmented metadata record is treated as a new version, link it to its previous version, e.g., isVersionOf (or isNewVersionOf, isPreviousVersionOf), as well as with an indication that the new version is about the metadata record, not the data (Klump *et al.*, 2021).
- **Utilise technology for detection and management:** When duplicate or augmented records exist despite best practices, technology can support duplicate detection and management. Lightweight tools like OpenRefine (Miller and Vielfaure, 2022) can help repository managers to identify and reconcile duplicates, while more advanced indexing solutions like Elasticsearch, Apache Solr, or OpenSearch support features such as 'result grouping' and 'top docs aggregation', provided the underlying schema/mapping supports these functionalities. In practice, repositories may combine automated approaches (e.g., natural language processing to assess record similarity) with community or user feedback mechanisms to flag duplicates. This process directly supports the presentation of cleaner search results (see Recommendation 4.2).

### Recommendation 3.6: Apply latest AI technologies to enrich metadata

Cutting-edge AI technologies, particularly large language models (LLMs), offer opportunities to enhance metadata by automating tasks like subject annotation and indexing (Chae and Davidson, 2025; Wu *et al.*, 2023). For instance, natural language processing models can extract variable-level metadata from unstructured text data by identifying key entities like dates, locations, measurements, and objects in images, etc., thus providing structured information for improving search and analysis. The OSCARS AI-SCOPE project exemplifies this by introducing a sophisticated AI analysis tool for surface scattering experiments and simultaneously generating rich metadata annotations.[22] Adopting these technologies is particularly valuable for machine users, as it creates machine-actionable metadata that supports more advanced, automated discovery and integration workflows, thereby improving a repository's 'AI readiness'.

20 OCLC: When to input a new record.

21 Zenodo Documentation: Manage versions: https://help.zenodo.org/docs/deposit/manage-versions/.

22 OSCARS AI-SCOPE project: https://oscars-project.eu/projects/ai-scope-ai-driven-enhancement-surface-scattering-data-open-science-platforms-across.

However, a cautious approach is essential. Since LLMs can produce plausible and inaccurate results, they should not be blindly relied upon. Repositories should implement a 'human-in-the-loop' workflow, where curators validate, correct, and supplement automated outputs, ensuring that the scale of AI is balanced with the accuracy and domain expertise of human oversight.

## RECOMMENDATION 4: SUPPORT FLEXIBLE SEARCH AND USER-FRIENDLY INTERFACE

A user-friendly search system is the critical mediator between users and the repository's metadata records, playing a central role in successful data discovery. Even when a repository dedicates substantial effort to collecting comprehensive and high-quality metadata, a poor search system, which lacks flexible search functionality and a user-friendly interface, will effectively negate that investment, driving users away and rendering the valuable data undiscoverable. Therefore, the search system's design and performance are important to unlock the value of shared research data.

### Recommendation 4.1: Support a flexible search interface with multiple search pathways

To cater to varying user needs and search behaviours, data repositories should offer a range of search functions and support multiple search pathways. This includes expanding beyond simple text queries and Boolean queries to offer advanced filters and specialised search tools.

- **Provide versatile search options:** A data discovery system should support general and structured search and browse, including simple keyword queries, complex Boolean queries, and structured advanced queries, as well as advanced filters ranging from broad classifications (e.g., research discipline, data type) to narrow fields (e.g., specific variables, instrument) (Wu *et al.*, 2019).

  Search functionality should also accommodate different data types, recognising that searching text-based data differs from searching numeric or structured data. Versatile search options empower users to refine queries and uncover hidden insights. For example, the OGC API EDR Standard[23] provides a simple web interface for retrieving specific spatio-temporal subsets (e.g., weather forecasts) by position, area, trajectory, or corridor, delivering only the needed data and abstracting backend complexity.

- **Support query expansion and refinement:** Support query expansion or refinement to include broad or narrow terms (e.g., suggesting broader terms if a specific search yields no results). This enhances the likelihood of finding relevant data and is highly valued by users (Löffler *et al.*, 2023). The expanded query terms can be derived from existing subject classification schemas, terminology services, user community input, or analysis of past search logs (Terolli *et al.*, 2020). Figure 4 shows GFBio's search interface,[24] offering a classical keyword search and a semantic search. In the semantic search, the search result is expanded with synonyms obtained from a terminology service.[25]

- **Integrate domain-specific tools:** The complexity of offering rich search pathways increases when a repository handles datasets from multidisciplinary datasets. To effectively address user needs in repositories with diverse data types, the search interface must adapt to the data's inherent properties. For example, for data with bounding boxes or coordinates, repositories should implement map-based search interfaces as standard. While this is easier in a discipline-specific repository where all data share this geospatial attribute, multidisciplinary repositories should utilise standards like the OGC API – Environmental Data Retrieval (EDR) standards to enable complex spatial and temporal queries only for relevant datasets. This ensures that users searching for geospatial data are given appropriate map tools to account for users from different disciplinary backgrounds or search behaviours.

23 OGC API Environmental Data Retrieval (EDR) Standard for retrieving targeted environmental data subsets: https://ogcapi.ogc.org/edr/.

24 GFBio's search interface: https://search.gfbio.org/.

25 GFBio's terminology service: https://terminologies.gfbio.org/.

- **Enhance data discovery with LLMs:** Further technical improvements for retrieving relevant data can be achieved through integration of LLMs into data discovery systems. By combining classic statistical search algorithms and deep-learning techniques, LLMs enable users to express their search needs in natural language rather than relying solely on a few chosen keywords or Boolean expressions. This shift allows for more intuitive, conversational, and semantically rich interactions with data repositories, helping users articulate complex data needs more effectively (Silva and Barbosa, 2024; Gao *et al.*, 2023).

  Moreover, search systems incorporating Retrieval Augmented Generation (RAG) can not only identify relevant datasets but also generate contextualised summaries or explanations drawn from multiple sources. Such systems return full-text answers accompanied by links to datasets that directly address the user's query, thereby bridging the gap between research question-based search and structured data retrieval (Amugongo *et al.*, 2025; Wang *et al.*, 2024). As these technologies mature, they have the potential to significantly enhance both the precision and accessibility of data discovery, particularly for users unfamiliar with technical metadata structures or domain-specific vocabularies.

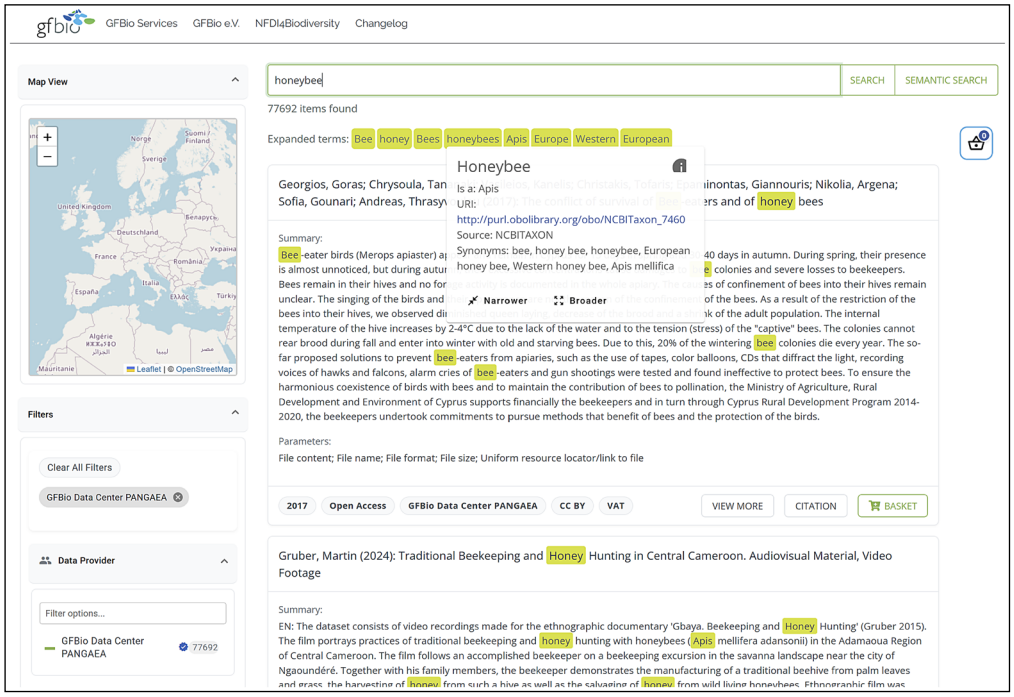


**Figure 4** Screenshot of GFBio's search interface with an expanded search including synonyms obtained from a terminology service.

## Recommendation 4.2: Present search results in a user-friendly and accessible way

Flexible search interfaces allow researchers to explore data at various levels of granularity, as Shneiderman et al. (2016) advocated 'overview first, details on demand'. Consider presenting search results to allow exploring data attributes or distributions through visualisation (Wu *et al.*, 2019; Dixit *et al.*, 2018).

- **Group metadata records for the same data or different versions:** Displaying multiple metadata records for duplicates or different data versions separately can clutter search results and waste the user's time as they sift through repetitive entries and keep track of the duplications mentally. Once duplicates are identified (see Recommendation 3.5), repositories can aggregate them and present them as a single entry in a search result, like Google Dataset Search (see Figure 5).[26] A straightforward method is to display a representative record with a link like 'More records like this' or 'More sources/versions' to group similar records of the same data. Careful interface design is crucial when displaying aggregate results from multiple sources (Sostek *et al.*, 2024), and repositories should adopt appropriate user study methods (see Recommendation 1) to iterate and verify the design.

26 Google Dataset Search: https://datasetsearch.research.google.com/.

- **Support data reuse assessment:** The main purpose for data search is data reuse. Scholars look for suitable data to compare novel findings with leverage data or to integrate data from various sources for data synthesis to explore novel hypotheses. Thus, search results need to provide relevant information for data reuse, such as the licence, data format, citation information, and publication date (Wu *et al.*, 2019). Furthermore, information on data type, domain-specific categories, or statistics on data reuse is also helpful for a scholar's decision to reuse a dataset.
- **Accessibility:** For users with special needs, it is essential to ensure an accessible user interface.[27] For example, not-text content needs to be provided with alternative labels, content must be adaptable (responsive design, different devices), and all information presented in the search interface needs to be accessible by keyboard.

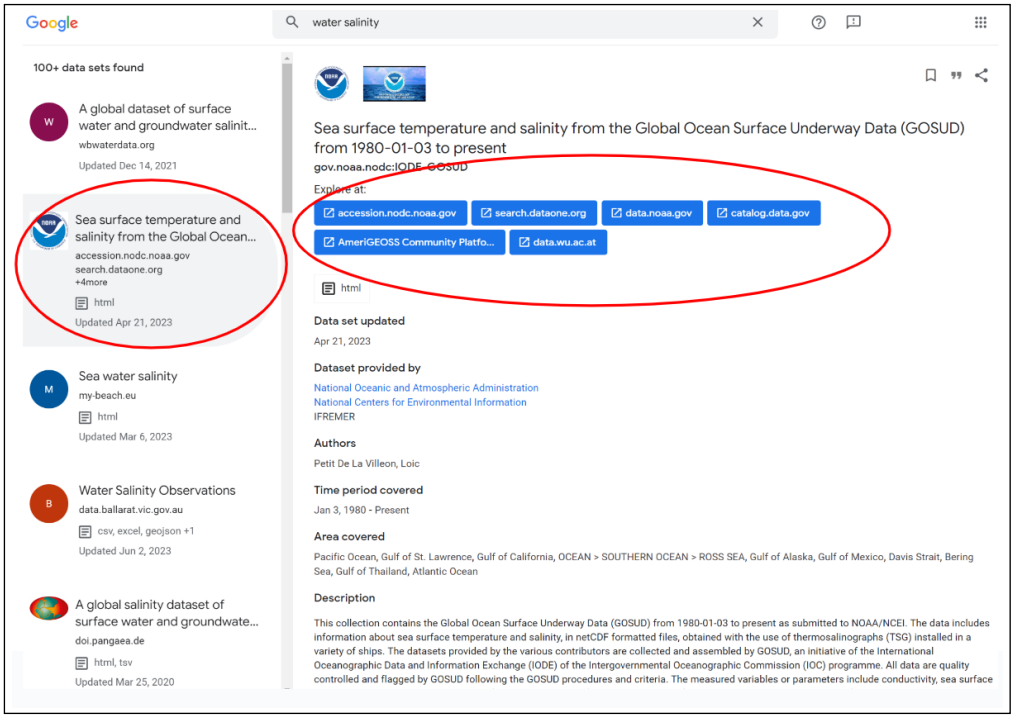


**Figure 5** An example of aggregating duplicated metadata records from several data catalogues.

### Recommendation 4.3: Implement continuous search system performance assessment

Repositories should establish and maintain a continuous, user-centric process for assessing and tuning the performance of their data discovery system. Recommendation 1 provides a set of user studies and their use cases for evaluating the discovery system's performance, including quantitative metrics (from log analysis and Google Analytics) to monitor search success rate (e.g., percentage of search queries that have zero hits or high-exit searches without any click on a search result), qualitative metrics through interviews and surveys and usability testing, and regular tuning and iteration of search parameters of underline search engines as new metadata records are added into the repository. A continuous assessment ensures the search system remains an effective mediator between users and metadata.

## 4. SUMMARY

This paper addresses the challenge of research data discovery, a fundamental issue in the era of open research where the rapid growth in data availability hasn't translated into a corresponding improvement in findability. To bridge the gap, the paper presents recommendations for data repositories, arguing for an integrated approach that places the needs of users at its core. The recommendations are structured around key pillars:

a. A foundational commitment to user-centric design (Recommendation 1): The paper argues that a deep understanding of users' needs, motivations, and search behaviours is paramount for any effective discovery service. The recommendation provides repositories

27 W3C web accessibility: https://www.w3.org/TR/WCAG22/.

with a toolkit of user study methods and guidance on how to apply them at different stages of a data discovery service's development lifecycle.

b. Strategic integration with the broader discovery ecosystem (Recommendation 2): Recognising that researchers discover data through multiple pathways, the recommendations advise repositories to look beyond their own discovery platforms.

c. A rigorous focus on high-quality metadata (Recommendation 3): As the bridge between user queries and relevant data, metadata is central to discovery. The recommendations offer detailed guidance on improving metadata quality from the users' perspectives.

d. The development of user-friendly and accessible interfaces (Recommendation 4): The final recommendation addresses the point of interaction between the user and the system. It advocates for flexible search interfaces that support multiple discovery pathways, from simple keyword searches to advanced semantic queries.

Recognising that implementing these recommendations may be challenging, repositories are encouraged to adopt a strategic, phased approach. The first step is to assess their status using suitable user study methods to understand their users' data discovery experiences and challenges and to prioritise evidence-based actions based on their available resources and specific user community needs.

While the recommendations provide user-centred recommendations for improving data discovery, they may not fully address certain challenges such as discovery data across multiple heterogeneous repositories, which is particularly important for interdisciplinary research. Additionally, although user-centred methods are recommended, the paper doesn't cover all possible evaluation metrics or approaches for assessing discovery effectiveness across diverse user communities.

The recommendations also highlight the potential of emerging AI technologies, such as Large Language Models and Retrieval-Augmented Generation, providing a forward-looking perspective for increasingly automated, AI-driven discovery. Any AI application must be carefully validated through a 'human-in-the-loop' workflow to ensure accuracy and reliability. Given that AI in data discovery is both rapidly evolving and still emerging, the implications for repositories to integrate and govern AI tools remain under discussion; this warrants dedicated research and future recommendation once both discussions and proven practices mature.

In conclusion, improving data discovery requires a coordinated and proactive effort from all stakeholders. By adopting the user-centred recommendations outlined in this paper, data repositories can move beyond simply making data available to genuinely making it discoverable. This will significantly enhance the value of their services and help the research community unlock the full potential of its shared research assets.

## ACKNOWLEDGEMENTS

This work was developed as part of the Research Data Alliance (RDA) Interest Group entitled 'Data Discovery Paradigms', and we acknowledge the support provided by the RDA community and structures. We would like to thank members of the group for their support and their thoughtful discussions through plenary sessions and regular monthly calls.

We would like to thank: Leyla Jael Castro, Chris Hunter, Jeffrey Grethe, Maggie Hellström, Christin Henzen, Live Kvale, Bénédicte Madon, Andrea Medina-Smith, Graham Parton, Andrea Pörsch, Emanuel Söding, Dimitri Szabo, Lucas van der Meer, Nina Weisweiler, Heinrich Widmann, CJ Woodford; who contributed the conceptualisation discussion, review, editing, or commenting on the RDA supporting output "Ten principles to improve dataset discoverability", which informed the writing up of this manuscript.

## FUNDING INFORMATION

Ying-Hsang Liu has been supported by the DFG project Intentional Forgetting and Changes in Work Processes: A Process-Conditional Approach in the Administrative and IT Context under project number 427257555.

Brigitte Mathiak has been supported by the DFG project KonsortSWD under project number: 442494171.

Antica Čulina was supported by the Croatian Science Foundation under the project number HRZZ-IP-2022-10-2872.

Andreas Czerniak was funded by the German Research Foundation (DFG) – 506475377.

Samantha Pearman-Kanza has been supported by the Careers and Skills for Data-driven Research (CaSDaR) Network+ under the UKRI grant UKRI739, and the Physical Sciences Data Infrastructure (PSDI) via the EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1 and the AI for Chemistry: AIchemy Hub via the EPSRC Grants EP/Y028775/1 and EP/Y028759/1.

Allyson Lister acknowledges contributions from the ELIXIR Interoperability Platform, where FAIRsharing is an adopted service delivered by the ELIXIR-UK Node.

Amir Aryani's contribution was partially supported by the Australian Government through the Australian Research Council's Industrial Transformation Training Centre for Information Resilience (CIRES) project number IC200100022.

## COMPETING INTERESTS

Kathleen Gregory is a member of the DSJ editorial board.

The authors have no competing interests to declare.

## AUTHOR AFFILIATIONS

**Mingfang Wu**
Australian Research Data Commons, Australia

**Felicitas Löffler** orcid.org/0000-0001-6423-7427
Thuringian Ministry for Digitization and Infrastructure, Germany

**Brigitte Mathiak** orcid.org/0000-0003-1793-9615
GESIS – Leibniz institute for the Social Sciences, Germany

**Fotis Psomopoulos** orcid.org/0000-0002-0222-4273
Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece

**Uwe Schindler** orcid.org/0000-0002-1900-4162
PANGAEA, MARUM, University of Bremen, Germany

**Amir Aryani** orcid.org/0000-0002-7180-0246
Swinburne University, Australia

**Jordi Bodera Sempere** orcid.org/0000-0002-0388-015X
European Synchrotron Radiation Facility, France

**Antica Culina** orcid.org/0000-0003-2910-8085
Ruder Boskovic Institute, Croatia

**Andreas Czerniak** orcid.org/0000-0003-3883-4169
Library, Bielefeld University, Bielefeld, Germany

**Chris Erdmann** orcid.org/0000-0003-2554-180X
SciLifeLab, Sweden

**Kathleen Gregory** orcid.org/0000-0001-5475-8632
Leiden University, The Netherlands

**Nick Juty** orcid.org/0000-0002-2036-8350
The University of Manchester, United Kingdom

**Allyson Lister** orcid.org/0000-0002-7702-4495
Oxford e-Research Centre, Department of Engineering Science, University of Oxford, United Kingdom

**Ying-Hsang Liu** orcid.org/0000-0001-6504-4598
Chemnitz University of Technology, Germany

**Samantha Pearman-Kanza** orcid.org/0000-0002-4831-9489
University of Southampton, United Kingdom

## REFERENCES

**Alencar, V., Kohwalter, T., Braganhole, V., da Silva, J.** and **Murta, L.** (2024) 'Prov-Dominoes: An approach for knowledge discovery from provenance data', *Expert Systems with Applications*, 245. Available at: https://doi.org/10.1016/j.eswa.2023.123030

**Allahim, A., Shamsuddin, S.M.** and **Meulien, J.** (2025) 'Semantic approaches for query expansion: Taxonomy, challenges, and future research directions', *PeerJ Computer Science*. Available at: https://doi.org/10.7717/peerj-cs.2664

**Amugongo, L.M., Mascheroni P., Brooks S., Doering S.** and **Seidel J.** (2025) Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6), e0000877. Available at: https://doi.org/10.1371/journal.pdig.0000877

**Bergold, J.** and **Thomas, S.** (2012) 'Participatory research methods: A methodological approach in motion', *Historical Social Research/Historische Sozialforschung*, 37(4), pp. 191–222. Available at: https://www.jstor.org/stable/41756482

**Bugbee, K., le Roux, J., Sisco, A., Kaulfus, A., Staton, P., Woods, C., Dixon, V., Lynnes, C.** and **Ramachandran, R.** (2021) 'Improving discovery and use of NASA's Earth observation data through metadata quality assessments', *Data Science Journal*, 20(1), p. 17. Available at: https://doi.org/10.5334/dsj-2021-017

**Burton, A., Aryani, A., Koers, H., Manghi, P., Bruzzo, S.L., Stocker, M., Diepenbroek, M., Schindler, U.** and **Fenner, M.** (2017) 'The Scholix framework for interoperability in data-literature information exchange', *D-Lib Magazine*, 23(1/2), pp. 1–20. Available at: https://doi.org/10.1045/january2017-burton

**Candela, L., Mangione, D.** and **Pavone, G.** (2024) 'The FAIR assessment conundrum: Reflections on tools and metrics', *Data Science Journal*, 23(1), p. 33. Available at: https://doi.org/10.5334/dsj-2024-033

**Chae, Y.** and **Davidson, T.** (2025) 'Large language models for text classification: From zero-shot learning to instruction-tuning', *Sociological Methods & Research*. Available at: https://doi.org/10.1177/00491241251325243

**Cooper, D.M.** and **Springer, R.** (2019) 'Data communities: A new model for supporting STEM data sharing', *Ithaka S+R*, 13 May. Available at: https://doi.org/10.18665/sr.311396

**CoreTrustSeal Standards and Certification Board** (2022) *CoreTrustSeal requirements 2023–2025 (V01.00)*. Available at: https://doi.org/10.5281/zenodo.7051012

**DataCite** (2024a) *DataCite thriving communities: 3000 repositories and counting*. Available at: https://doi.org/10.5438/63qf-5740 (Accessed: 17 April 2024).

**DataCite Metadata Working Group** (2024b) *DataCite metadata schema for the publication and citation of research data and other research outputs*. Version 4.5. DataCite e.V. Available at: https://doi.org/10.14454/znvd-6q68

**Davenport, E.** (2010) 'Confessional methods and everyday life information seeking', *Annual Review of Information Science and Technology*, 44(1), pp. 533–562. Available at: https://doi.org/10.1002/aris.2010.1440440119

**Dixit, R., Rogith, D., Narayana, V., Salimi, M., Gururaj, A., Ohno-Machado, L., Xu, H.** and **Johnson, T.R.** (2018) 'User needs analysis and usability assessment of DataMed – a biomedical data discovery index', *Journal of the American Medical Informatics Association*, 25(3), pp. 337–344. Available at: https://doi.org/10.1093/jamia/ocx134

**Felden, J., Möller, L., Schindler, U., Huber, R., Schumacher, S., Koppe, R., Diepenbroek, M.** and **Glöckner, F.O.** (2023) 'PANGAEA – Data publisher for earth & environmental Science', *Scientific Data*, 10, p. 347. Available at: https://doi.org/10.1038/s41597-023-02269-x

**Flanagan, J.C.** (1954) 'The critical incident technique', *Psychological Bulletin*, 51(4), pp. 327–359. Available at: https://doi.org/10.1037/h0061470

**Foulonneau, M., Cole, T.W., Habing, T.G.** and **Shreeves, S.L.** (2005) 'Using collection descriptions to enhance an aggregation of harvested item-level metadata', in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*. Denver: Association for Computer Machinery, pp. 32–42. Available at: https://doi.org/10.1145/1065385.1065393

**Friedrich, T.** (2020). *Looking for data: Information seeking behaviour of survey data users* (Doctoral dissertation). Humboldt-Universität zu Berlin. Available at: https://doi.org/10.18452/22173

**Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M.** and **Wang, H.** (2023) 'Retrieval-augmented generation for large language models: A surve', arXiv:2312.10997v5 [cs.CL]. Available at: https://doi.org/10.48550/arXiv.2312.10997

**Gregory, A., Bell, D., Brickley, D., Buttigieg, P.L., Cox, S., Edwards, M., Doug, F., Gonzalez Morales, L.G., Heus, P., Hodson, S., Kanjala, C., Le Franc, Y., Maxwell, L., Molloy, L., Richard, S., Rizzolo, F., Winstanley, P.** and **Wyborn, L.** (2024) *WorldFAIR (D2.3) (version 1)*. Available at: https://doi.org/10.5281/zenodo.11236871

**Gregory, K., Groth, P., Scharnhorst, A.** and **Wyatt, S.** (2020) 'Lost or found? Discovering data needed for research', *Harvard Data Science Review*, 2(2). Available at: https://doi.org/10.1162/99608f92.e38165eb

**Hodson, S.** (2024) WorldFAIR (D2.2) *WorldFAIR's experience with FIPs (second set of FAIR implementation profiles for each case study) (version 1)*. Available at: https://doi.org/10.5281/zenodo.11236094

**Jeng, W., He, D.** and **Chi, Y.** (2017) 'Social science data repositories in data deluge: A case study of ICPSR's workflow and practices', *The Electronic Library*, 35(4), pp. 626–649. Available at: https://doi.org/10.1108/EL-11-2016-0243

**Kacprzak, E., Koesten, L., Tennison, J.** and **Simperl, E.** (2018) 'Characterising dataset search queries', in *WWW '18: Companion Proceedings of the Web Conference 2018*. *Geneva: International World Wide Web Conferences Steering Committee*, pp. 1485–1488. Available at: https://doi.org/10.1145/3184558.3191597

**Kalinin, N.A.** and **Skvortsov, N.A.** (2023) 'Difficulties of FAIR principles implementation in cross-domain research infrastructures', *Lobachevskii Journal of Math*, 44, pp. 147–156. Available at: https://doi.org/10.1134/S199508022301016X

**Koesten, L., Gregory, K., Groth, P.** and **Simperl, E.** (2021) 'Talking datasets—Understanding data sense-making behaviours', *International Journal of Human-Computer Studies*, 146, 102562. Available at: https://doi.org/10.1016/j.ijhcs.2020.102562

**Koesten, L.M., Kacprzak, E., Tennison, J.F.** and **Simperl, E.** (2017). '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* (pp. 1277–1289). Available at: https://doi.org/10.1145/3025453.3025838

**Khalsa, S., Cotroneo, P.** and **Wu, M.** (2018) 'A survey of current practices in data search services', *Mendeley Data*, V1. Available at: https://doi.org/10.17632/7j43z6n22z.1

**Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R.R.** and **Asmi, A.** (2021) 'Versioning data is about more than revisions: A conceptual framework and proposed principles', *Data Science Journal*, 20(1), p. 12. Available at: https://doi.org/10.5334/dsj-2021-012

**Krans, N.A., Ammar, A., Nymark, P., Willighagen, E.L., Bakker, M.I.** and **Quik, J.T.K.** (2022). 'FAIR assessment tools: Evaluating use and performance', *NanoImpact*, 27, p. 100402. Available at: https://doi.org/10.1016/j.impact.2022.100402

**Lacagnina, C., David, R., Nikiforova, A., Kuusniemi, M.E., Cappiello, C., Biehlmaier, O., Wright, L., Schubert, C., Bertino, A., Thiemann, H.** and **Dennis, R.** (2023) *Towards a data quality framework for EOSC (1.0.0)*. Available at: https://doi.org/10.5281/zenodo.7515816

**Lafia, S., Million, A.J.** 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)*. New York: Association for Computing Machinery, pp. 128–136. Available at: https://doi.org/10.1145/3576840.3578275

**Lang, J.M.** and **Benbow, M.E.** (2013) 'Species interactions and competition', *Nature Education Knowledge*, 4(4), p. 8. Available at: https://www.nature.com/scitable/knowledge/library/species-interactions-and-competition-102131429/

**Lister, A.** and **Sansone, A.** (2023, July 28) *FAIRsharing in a nutshell*. Available at: https://doi.org/10.5281/zenodo.8191958

**Liu, Y.-H., Wu, M., Power, M.** and **Burton, A.** (2022) *Elicitation of data discovery contexts: An interview study (1.0)*. Available at: https://doi.org/10.5281/zenodo.7179526

**Liu, Y.-H., Wu, M., Power, M.** and **Burton, A.** (2023) *Elicitation of contexts for discovering clinical trials and related health data: An interview study (V1.0)*. Available at: https://doi.org/10.5281/zenodo.7839282.

**Löffler, F., Wesp, V., König-Ries, B.** and **Klan, F.** (2021) 'Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?' *PLoS ONE*, 16(3), e0246099. Available at: https://doi.org/10.1371/journal.pone.0246099

**Löffler, F., Shafiei, F., Witte, R., König-Ries, B.** and **Klan, F.** (2023) 'Semantic search for biological datasets: A usability study on modes of querying and explaining search results', *20th Conference on Database Systems for Business, Technology and Web, BTW 2023*. Dresden, Germany, 6–10 March. Available at: https://doi.org/10.18420/BTW2023-56

**Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J.** and **Principe, P.** (2019) *The OpenAIRE research graph data model*. Available at: https://doi.org/10.5281/zenodo.2643199

**Marchionini, G.** (2006) 'Exploratory search: from finding to understanding', *Communications of the ACM*, 49(4), pp. 41–42. Available at: https://doi.org/10.1145/1121949.1121979

**Million, A.J., York, J., Lafia, S.** and **Hemphill, L.** (2025) 'Data, not documents: Moving beyond theories of information-seeking behavior to advance data discovery', *Journal of the Association for Information Science and Technology*, 76(4), pp. 649–664. Available at: https://doi.org/10.1002/asi.24962

**Miller, M.** and **Vielfaure, N.** (2022) 'OpenRefine: An approachable open tool to clean research data', *Bulletin – Association of Canadian Map Libraries and Archives (ACMLA)*, 170. Available at: https://doi.org/10.15353/acmla.n170.4873

**Nasir, J.A., Varlamis, I.** and **Ishfaq, S.** (2019) 'A knowledge-based semantic framework for query expansion', *Information Processing & Management*, 56(5), pp. 1605–1617. Available at: https://doi.org/10.1016/j.ipm.2019.04.007

**National Library of Medicine** (2021) *SNOMED CT to ICD-10-CM map*. Available at: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html.

**NISO (National Information Standards Organization)** (2004) *Understanding metadata*. Bethesda: NISO Press. Available at: https://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

**Peng, G., Berg-Cross, G., Wu, M., Downs, R.R., Shrestha, S.R., Wyborn, L., Ritchey, N., Ramapriyan, H.K., Clark, S.J., Wood, J., Liu, Z.** and **Marouane, A.** (2024) 'Harmonizing quality measures of FAIRness assessment towards machine-actionable quality information', *International Journal of Digital Earth*, *17*(1). Available at: https://doi.org/10.1080/17538947.2024.2390431

**Pressman, R.S.** and **Maxim, B.R.** (2015). *Software Engineering: A Practitioner's Approach* (8th ed.). McGraw-Hill Education.

**Quintel, D.** and **Wilson, R.** (2020) 'Analytics and privacy: Using Matomo in EBSCO's discovery service', *Information Technology and Libraries*, 39(3). Available at: https://doi.org/10.6017/ital.v39i3.12219

**Sharifpour, R., Wu, M.** and **Zhang, X.** (2023) 'Large-scale analysis of query logs to profile users for dataset search', *Journal of Documentation*, 79(1), pp. 66–85. Available at: https://doi.org/10.1108/JD-12-2021-0245

**Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N.** and **Diakopoulos, N.** (2016) *Designing the user interface: Strategies for effective human-computer interaction*. 6th ed. Boston: Pearson.

**Silva, L.** and **Barbosa, L.** (2024) 'Improving dense retrieval models with LLM augmented data for dataset search', *Knowledge-Based Systems*, 294, 111740. Available at: https://doi.org/10.1016/j.knosys.2024.111740

**Slavković, A.** and **Seeman, J.** (2023) 'Statistical data privacy: A song of privacy and Uutility', *Annual Review of Statistics and Its Application*, 10, pp. 189–218. Available at: https://doi.org/10.1146/annurev-statistics-033121-112921

**Smith, L.C.** (2020) 'Interdisciplinary searching as a use case for vocabulary mapping', in M. Lykke, T. Svarre, M. Skov and D. Martínez-Ávila (eds.) *Knowledge organization at the interface: Proceedings of the sixteenth international ISKO conference, 2020 Aalborg, Denmark*. vol. 17. Baden-Baden: Ergon-Verlag, pp. 428–435. Available at: https://doi.org/10.5771/9783956507762-428

**Sostek, K., Russell, D.M., Goyal, N., Alrashed, T., Dugall, S.** and **Noy, N.** (2024) 'Discovering datasets on the web scale: Challenges and recommendations for Google dataset search', *Harvard Data Science Review*, special issue 4. Available at: https://doi.org/10.1162/99608f92.4c3e11ca

**Stall, S., Bilder, G., Cannon, M., Hong N.C., Edmunds, S., Erdmann, C.C., Evans, M., Farmer, R., Feeney, P., Friedman, M., Giampoala, M., Hanson, R.B., Harrison, M., Karaiskos, D., Katz, D.S., Letizia, V., Lizzi, V., MacCallum, C., Meunch, A., Perry, K., Ratner, H., Schindler, U., Sedora, B., Stockhause, M., Townsend, R., Yeston, J.** and **Clark, T.** (2023) 'Journal production guidance for software and data citations', *Scientific Data*, 10, 656. Available at: https://doi.org/10.1038/s41597-023-02491-7

**Sun, D., Hnatiuk, R.J.** and **Neldner, V.J.** (1997) 'Review of vegetation classification and mapping systems undertaken by major forested land management agencies in Australia', *Australian Journal of Botany*, 45(6), pp. 929–948. Available at: https://doi.org/10.1071/BT96121

**Taniguchi, S.** and **Hashizume, A.** (2023) 'Transforming metadata content guidelines and instructions to linked data', *Journal of Documentation*, 51(4). Available at: https://doi.org/10.1177/01655515221142428

**Thomas, K., 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.

**Terolli, E., Ernst, P.** and **Weikum, G.** (2020) 'Focused query expansion with entity cores for patient-centric health search', in J.Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal (eds.) *The semantic web – ISWC 2020*. Cham: Springer, pp. 547–564. Available at: https://doi.org/10.1007/978-3-030-62419-4_31

**Vega-Gorgojo, G., Slaughter, L., Giese, M., Heggestoyl, S., Soylu, A.** and **Waaler, A.** (2016) 'Visual query interfaces for semantic datasets: An evaluation study', *Journal of Web Semantics*, 39(C), pp. 81–96. Available at: https://doi.org/10.2139/ssrn.3199241

**Wang, R.Y.** and **Strong, D.M.** (1996). 'Beyond Accuracy: What Data Quality Means to Data Consumers', *Journal of Management Information Systems*, 12(4), pp. 5–33. Available at: https://doi.org/10.1080/07421222.1996.11518099

**Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Y., Xu Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X.** and **Huang, X.** (2024) 'Searching for best practices in retrieval-augmented generation', in Y. Al-Onaizan, M. Bansal and Y.-N. Chen (eds.) *Proceedings of the 2024 conference on empirical methods in natural language processing*. Miami: Association for Computational Linguistics, pp. 17716–17736. Available at:: https://doi.org/10.18653/v1/2024.emnlp-main.981

**Weitz, J.** (2020) 'Improving WorldCat quality: Resolving to reduce duplicates', *Organizacija znanja*, 25 (1–2), 2025003. Available at: https://doi.org/10.3359/oz2025003

**Wentzel, B., Kirstein, F., Jastrow, T., Sturm, R., Peters, M.** and **Schimmler, S.** (2023) 'An extensive methodology and framework for quality assessment of DCAT-AP datasets', in I. Lindgren, C. Csáki, E. Kalampokis, M. Janssen, G.V. Pereira, S. Virkar, E. Tambouris and A. Zuiderwijk (eds.) *Electronic Government*. Cham: Springer, pp. 262–278. Available at: https://doi.org/10.1007/978-3-031-41138-0_17

**White, R.W.** (2016) *Interactions with search systems*. Cambridge: Cambridge University Press.

**Wilkinson, M., Dumontier, M., Aalbersberg, I., 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., 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, 160018. Available at: https://doi.org/10.1038/sdata.2016.18

**Wollin-Giering, S., Hoffmann, M., Höfting, J.** and **Ventzke, C.** (2024) 'Automatic transcription of English and German qualitative interviews', *Forum Qualitative Sozialforschung Forum: Qualitative Social Research*, 25(1). Available at: https://doi.org/10.17169/fqs-25.1.4129

**Wu, M.** (2022) *ARDC Project: Eliciting data search context*. Available at: https://doi.org/10.5281/zenodo.6819787

**Wu, M., Juty, N., RDA Research Metadata Schemas WG, Collins, J., Duerr, R., Ridsdale, C., Shepherd, A., Verhey, C.** and **Castro, L.J.** (2021) *Guidelines for publishing structured metadata on the web (3.1)*. Available at: https://doi.org/10.15497/RDA00066

**Wu, M., Psomopoulos, F., Khalsa, S.J.** and **de Waard, A.** (2019) 'Data discovery paradigms: User requirements and recommendations for data repositories', *Data Science Journal*, 18(1), p. 3. Available at: https://doi.org/10.5334/dsj-2019-003

**Wu, M., Brandhorst, H., Marinescu, M., Lopez, J.M., Hlava, M.** and **Busch, J.** (2023) 'Automated metadata annotation: What is and is not possible with machine learning', *Data Intelligence*, 5(1), pp. 122–138. Available at: https://doi.org/10.1162/dint_a_00162

**Wu, M., Gregory, K., Löffler, F., Mathiak, B., Psomopoulos, F., Schindler, U., Aryani, A., Bodera, J., Castro, L.J., Culina, A., Czerniak, A., Erdmann, C., Grethe, J., Hellström, M., Henzen, C., Hunter, C., Juty, N., Kvale, L., Lister, A., Liu, Y.-H., Madon, B., Medina-Smith, A., Parton, G., Pearman-Kanza, S., Pörsch, A., Söding, E., Szabo, D., van der Meer, L., Weisweiler, N., Widmann, H.** and **Woodford, C.J.** (2024) *Ten principles to improve dataset discoverability (1.0)*. Available at: https://doi.org/10.15497/rda/00120

**TO CITE THIS ARTICLE:**
Wu, M., Löffler, F., Mathiak, B., Psomopoulos, F., Schindler, U., Aryani, A., Bodera Sempere, J., Culina, A., Czerniak, A., Erdmann, C., Gregory, K., Juty, N., Lister, A., Liu, Y.-H., Pearman-Kanza, S 2026 Bridging the Data Discovery Gap: User-Centric Recommendations for Research Data Repositories. *Data Science Journal*, 25: 6, pp. 1–21. DOI: https://doi.org/10.5334/dsj-2026-006

**Submitted:** 19 July 2025
**Accepted:** 06 January 2026
**Published:** 12 February 2026



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