DATA SCIENCE JOURNAL

Wu, M, et al. 2019. Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. *Data Science Journal,* 18: 3, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2019-003

RESEARCH PAPER

Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories

Mingfang Wu[1], Fotis Psomopoulos[2,3], Siri Jodha Khalsa[4] and Anita de Waard[5]

[1] Australian Research Data Commons, AU

[2] Institute of Applied Biosciences, Centre for Research and Technology, Hellas, Thessaloniki, GR

[3] Dept of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, SE

[4] National Snow and Ice Data Center, University of Colorado, Boulder, US

[5] Research Data Collaborations, Elsevier, US

Corresponding author: Mingfang Wu (mingfang.wu@ardc.edu.au)

As data repositories make more data openly available it becomes challenging for researchers to find what they need either from a repository or through web search engines. This study attempts to investigate data users' requirements and the role that data repositories can play in supporting data discoverability by meeting those requirements. We collected 79 data discovery use cases (or data search scenarios), from which we derived nine functional requirements for data repositories through qualitative analysis. We then applied usability heuristic evaluation and expert review methods to identify best practices that data repositories can implement to meet each functional requirement. We propose the following ten recommendations for data repository operators to consider for improving data discoverability and user's data search experience:

1. Provide a range of query interfaces to accommodate various data search behaviours.
2. Provide multiple access points to find data.
3. Make it easier for researchers to judge relevance, accessibility and reusability of a data collection from a search summary.
4. Make individual metadata records readable and analysable.
5. Enable sharing and downloading of bibliographic references.
6. Expose data usage statistics.
7. Strive for consistency with other repositories.
8. Identify and aggregate metadata records that describe the same data object.
9. Make metadata records easily indexed and searchable by major web search engines.
10. Follow API search standards and community adopted vocabularies for interoperability.

Keywords: Data discovery; Usability; Data repository; Requirements and recommendations; FAIR data

1. Introduction

A widely-endorsed statement on research data asserts that data should be FAIR: '*Findable, Accessible, Interpretable and Reusable*' (Wilkinson et al., 2016). The FAIR Guiding Principles further specify four criteria for making data findable, one of them is '*F4. (meta)data are registered or indexed in a searchable resource.*' On one hand, this requires data owners or providers to create metadata and register it to a data repository in order to make data discoverable; on the other hand, data repository operators need to index the metadata and make it easily discoverable. With more data open and available through data repositories, it becomes challenging for researchers to find relevant data and to assess their fitness for intended use. Improving data discoverability will benefit all people and organizations who are involved in the data lifecycle, from data production to eventual data applications.

Providing easy data discovery and an overall user-friendly experience is a key service to all data repositories (Iosifescu and Plattner et al., 2018), as Joo & Lee (2011, p. 524) states that usability of digital libraries means the ease of use, prolificacy and the extent of satisfaction it provides to its users. Data repositories have been following their own path to develop their portal with support to data discovery (Murphy and Gautier 2017). For example, DataONE followed user-centric system design principles – gathering use cases and requirements, involving user feedback and evaluation in the designing and development process, so that the developed system can meet its user and stakeholder's need (Michener and Allard et al., 2012). Yet, there still lacks a common understanding of what data repositories should offer to support data discoverability. This study attempts to fill in this gap by examining trans-repository criteria and identifying requirements and practices that are of common relevance across repositories.

With the goal of enabling data repositories to improve discoverability of their data holdings, we can research and draw lessons from several fields. In particular, considering the fields of information retrieval and digital libraries, there has been substantial research on why and how users search for information and how search algorithms and systems should model and support user search behaviour and search task (Spärck Jones 2006, Sanderson and Croft. 2012, Niu and Hermminger, 2010, Kim and Field et al., 2012). We will further review the research findings and recommendations from previous work in the relevant sections below. In this study, we first adopt the case study methodology (Soy, 2018) as used by previous work of similar kind such as the W3C Data on the Web Best Practices Working Group (www.w3.org/TR/dwbp) (Lóscio, 2017): we collected use cases or data search scenarios, and then we performed a qualitative analysis of the use cases to infer user's data search needs, from which we identified a core set of nine functional requirements. We propose a set of recommendations with exemplar implementations for data repositories to consider when they develop or improve their data portals. The recommendations focus on enabling and improving the methods and tools by which users find data in these repositories. This paper is aimed at developers, project and product managers of data repositories, and researchers who are involved in developing data repositories, community platforms, or interfaces to data collections.

The rest of the paper is organized as follows: we first describe in detail the methodology we used to collect use cases and identify requirements. We then present our recommendations and conclude with discussion of contributions and future work.

2. Case Study Methodology

To recommend best practices for making data more findable within data repositories, first we need to understand why, what and how data repository users search for data. We adopted the case study methodology (Soy 2018) by gathering data discovery needs from representative users, then categorised needs, elicited functional requirements and made generalisations. We followed the following steps:

Step 1. Collect use cases.
Step 2. Analyse use cases to identify common themes and similar functionalities.
Step 3. Elicit functional requirements and prioritise the requirements.

Each step is further detailed in the following sections.

2.1. Collecting use cases

We used two methods for collecting use cases. In first method, we collected existing use cases by different organisations in the context of improving their own data search services of their data repositories. In the second method, we conducted a survey to collect more use cases in order to cover wider user representation.

In the first method, we aggregated use cases from the following five resources:

1. JISC Research Data Discovery Service use cases (Ferguson, 2016)
2. ANDS User Interview Responses[1]
3. BioCADDIE[2]
4. DataONE: DataONE Personas[3]
5. Spatial Data on the Web[4]

[1] ANDS User Interview Responses: https://goo.gl/CDw1gp.
[2] BioCADDIE Working Group 4 (Use Cases and Testing Benchmarks): https://biocaddie.org/sites/default/files/d7/project/430/wg4_final_report.pdf.
[3] DataONE Personas: https://www.dataone.org/user-personas.
[4] W3C Spatial Data on the Web Use Cases & Requirements: https://www.dataone.org/user-personas.

This method enabled us to cover a variety of disciplinary backgrounds (e.g. biomedical and healthcare, earth science, economy and humanity), and thus a wide representation of disciplines and user groups. However, the five resources describe use cases in different formats as a result of adopting different use case development methods; for example, DataONE used persona and the ANDS project recorded answers to their own interview questions, while JISC used an open interview format.[5] We needed to adapt these use cases into a single framework/schema for cross analysis and summary. After a review of the structure from the five sources, we adapted the description from the open interview format to re-write existing use cases into single format. In this description format, each use case has the following fields:

1. '***As a***' (i.e. role)
2. '***Theme***' (i.e. scientific domain/discipline)
3. '***I want***' (i.e. requirement, missing feature, supported function)
4. '***So that***' (i.e. the user need that is addressed)
5. '***Comments***' (anything that are not covered by the above four fields)

We then used the above description of five fields to re-write the existing use cases, keeping only those use cases that could be unambiguously re-written in this new format without any loss of information.

For example, a use case from the ANDS user interviews showed that a Ph.D student, from the field of Economics, usually knows what data they want to have; so what they want from a (portal) homepage is a simple page with search box. They would like to have advanced search in case they need to refine a search. Another example of personas from DataONE[6] describes an early-career herpetologist, who is interested in finding tortoise data and the location of tortoise populations, so they can put their study into perspective and perhaps find collaborators. **Table 1** shows the result of re-writing the above two use cases into the new description format. Ultimately, we collected 64 use cases as a result of the re-writing process.

We found almost the entirety of the 64 use cases focused on the 'Researcher' role. To include more diversified roles such as data librarians, we collected additional use cases ourselves by turning the above five fields into questions. We invited participants representing different communities, such as ALA Scholar Communication, ACRL Science & Technology Section, NARO Physics-Astronomy-Mathematics Division, to complete the survey. As a result, we collected 15 additional use cases and broadened the scope of role to include librarian and funder.

In total, we collected 79 use cases for further analysis. We have made the data from the collected use cases openly accessible through Zenodo (de Waard, et al., 2017).

2.2. Analysing and clustering use cases

Next, we analysed the 79 use cases along two dimensions: (1) identify issues related to data discovery, and (2) identify intended audience who may take responsibility to address each issue. We especially sought to identify those common issues related to data discovery and turn them into user requirements. For that purpose, we first normalised various users' backgrounds as captured from the field ('*As A*') into the following four user types: 'Researcher', 'Research Student (PhD/Master)', 'Librarian', or 'Funder'. We then applied an initial open coding method to label each use cases (Charmaz, 2006): we had one author label each use case with maximum of two open vocabulary terms along the two dimensions and another author label a second round while also checking for consistency across all use cases. The open coding activity resulted in 24 vocabulary terms, which are provisional, comparative, and grounded in the use cases. We then applied an axial coding method to identify relationships among the 24 terms (Charmaz, 2006). As a result, we classified these 24 terms into three groups, labelled as: Metadata, Portal Functionality, and Data. **Figure 1** shows the three

Table 1: Examples of use cases being re-written.

As A	Theme	I want	So that
Ph.D Candidate	Economics	To have advanced search functionality	So he can refine a search when needed
Researcher	Herpetology	To find more data to correlate with the locations of her tortoise populations	So she can put her research into perspective and identify collaborators

[5] User stories as purposed for the agile methodology: https://www.scrumalliance.org/community/articles/2013/september/agile-user-stories.

[6] DataOne: Sun:Early-career herpetologist: https://www.dataone.org/personas/sun-early-career-herpetologist.

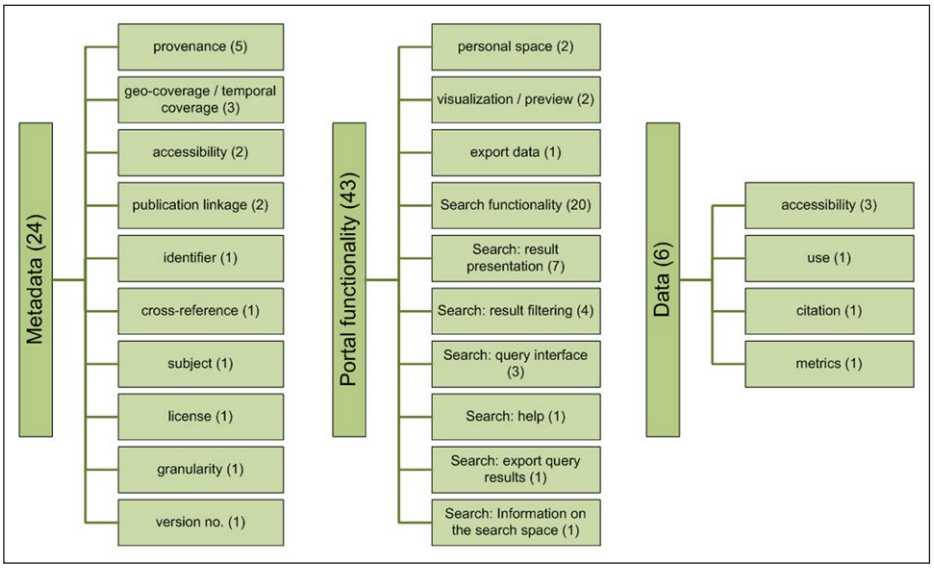

Figure 1: Two layered grouping of Use Cases. The first and second layer are from axial coding and open coding respectively.

Table 2: From Use Cases to Requirements.

As A	Theme	I want	So that
Researcher	Social Science	To see what data is available right now	Make a forecast
Researcher	Social Science	Cares about Access Conditions	
Researcher	Physical Science	wants a very prominent Download button	
Researcher	Computer Science	see (data) publish date or available date	
Researcher	Health Science	data	
Requirement		**Indication of Data availability**	

groups at the first layer and the distribution of the 24 vocabulary terms among the groups at the second layer. Note that a use case can be in more than one category (but we allowed no more than two categories). For example, a need from a use case is '*care about data access condition*' (if a data is not available, that user would not bother with it, nor click further), this is a metadata issue (to code accessibility of data) but also portal functionality issue (to clearly display data accessibility if provided, or display "unknown" otherwise).

2.3. Eliciting user requirements

The classification resulting from the above qualitative analysis allowed for a general overview of the missing aspects in data discovery from the perspective of the relevant technologies (i.e. Portal functionality, Metadata and Data). However, the usefulness of these attributes can be enhanced by investigating the specific user data search needs. As such, the final step in the process was to infer the user requirements from the use cases. In order to do this, we grouped all 79 use cases based on the context of the '*I want*' field, i.e. the specific data discovery need, and identified the common aspect described by each group, and then formulated this aspect as a distinct requirement using the vocabulary terms identified earlier as a guiding principle. An example of this grouping is shown in **Table 2**. Ultimately, nine individual groups (or requirements) were identified.

The nine requirements, as described in **Table 3**, capture the user perspective in the data discovery process, and therefore each requirement has a distinct target audience (i.e. the community that needs to take responsibility to address the particular requirement). We identified the following three intended audiences and assigned them to each requirement as appropriate: 1) Data Repository, 2) Data Provider and 3) Research Office/Libraries.

Table 3: Nine user requirements elicited from use cases.

User Requirements	User Type	Actors who can meet requirement	Description (extracts from the 'so that' field)
REQ 1. Indication of data availability	Researcher/ Research Student	Data Repository Operator, Data Provider	If there is no clear indication of data availability, the search is usually dropped within the first 2 minutes. A 'sort by availability' function could also reveal potential data embargo. Ideally should have an evident big button for 'Download'.
REQ 2. Connection of data with person/institution/ paper/citations/grants	Funder/ Researcher/ Research Student	Data Repository Operator, Data Provider	This allows for ranking of datasets, the connection of the information displayed with personal details as well as accountability. Also, this information can be used for grant application as well as for comparative studies (datasets underpinned several papers). Finally, allow for the upload of manuscript for direct connection.
REQ 3. Fully annotated data *(including granularity, origin, licensing, provenance, and method of production, times downloaded)*	Researcher/ Research Student	Data Provider, Data Repository Operator	This information will validate the use of a dataset in a particular study, as well as remove the step of having to read the corresponding manuscript to understand the data. To judge validity, need to know where and when the data was measured, and the basic experimental and instrumental parameters. These are more important than e.g. who created the data. To assess the validity of the data, look at repository/paper, then look at the data first to see if it makes sense.
REQ 4. Filtering of data based on specific criteria on multiple fields at the same time *(such a release date, geo coverage, text content, date range, specific events).*	Researcher/ Research Student	Data Repository Operator, Data Provider	Support targeted studies (e.g. find global temperature records for volcanic eruptions in the last century; find articles on bronze age in Britain).
REQ 5. Cross-referencing of data *(same or different repositories).*	Researcher/ Research Student	Data Provider, Data Repository Operator	Having the same with different identifiers is not sufficiently convenient for studies. Also, there are multiple instances/versions and reproducibility necessitates specific uses every time. Finally, cross-referencing will avoid duplication and maximize efficiency and access.
REQ 6. Visual analytics/ inspection of data/ thumbnail preview	Researcher	Data Repository Operator	Decide if this data set is right for a research purpose. Also allows quick visual filtering from a results set.
REQ 7. Sharing data *(either whole dataset, particular records, or bibliographic information)* **in a collaborative environment**	Researcher/ Research Student	Data Repository Operator	Make sure that there is a common space of keeping both data and their versions across time – alleviate the need to rerun at the last minute to check nothing has been published since last study/search, or to share bibliographic information about data.

(Contd.)

User Requirements	User Type	Actors who can meet requirement	Description (extracts from the 'so that' field)
REQ 8. Accompanying educational/training material	Librarian	Research Office/ Libraries, Data Repository	Help researchers manage and discover data in a methodical and seamless manner.
REQ 9. Portal functionality similar to other established academic portals	Researcher	Data Repository Operator	For example, finding more within a subject, search by visual (i.e. draw a structure to search for), free text search, build query functionality, subscription, save lists.

Finally, and in order to better understand how relevant these requirements are to the intended communities, we circulated a second survey, asking for a ranking of each requirement independently, ranging from 1 (Not important) to 5 (Very Significant) and including a no-opinion option. In order to ensure that we capture as many of the different scientific disciplines possible, the survey was circulated through both official RDA mailing lists (such as the Data Discovery Paradigms, the FAIRSharing and the Research Data Management groups) as well as through targeted networks (such as DANS, NIH and NBDC Japan). Ultimately, we received 31 anonymous responses, which allowed us to rank the nine requirements as listed in **Table 3** (in the order of descending importance). Please note that the survey itself didn't capture participants' background; however, people from the above mailing lists are mostly data providers, data infrastructure operators, data librarians and researchers.

2.4. Summary

Through analysis of collected use cases, we produce a classification scheme leading to a set of core requirements in supporting data discovery. The classification offers a more comprehensive view upfront, which can be used by various stakeholders for different purposes: for example, when data managers selecting a metadata schema to describe data, they may take the Metadata and the Data classes as basic requirement of discovery metadata; data repository developers may check if their repository supports requirements from the Portal functionality class.

The set of core requirements is at a more abstract level. For people who would like to examine further what use cases are behind each requirement, we provide the mapping between the requirements and use cases in (de Waard, et al., 2017). The requirements can be used as a scaffold for verifying best practices or providing better services for the various audiences such as data providers, data managers, data repository operators.

In this paper we focus on the role of data repository in making data more discoverable. We expect that data repository operators can use the requirements for the following (but not limited to) purposes:

- As a checklist for designing and implementing a data service portal.
- For existing data discovery services, the list of requirements can be used as guidelines for heuristic evaluation of a specific data discovery service (Nielsen, 1995), and therefore plan for future improvements when necessary.
- In the era of big data, research on data discovery paradigms is at an all-time high. A user's perspective provides a strong foundation on which to construct the paradigms of the future.

3. Recommendations to data repositories on data discovery

By taking the requirements and the FAIR data principles (Wilkinson, et al., 2016) as starting points, as well as drawing from research and practices as reviewed and referenced in this section, we came up with ten recommendations for repositories to meet the requirements through heuristic evaluation and expert review method (Nielsen, 1993 & 1995). Note that when we summarised recommendations, we recognised the relationship between recommendations and requirements is not one to one, but one to many. Therefore, we will first discuss each recommendation and the requirement(s) it relates to, provide exemplars whenever applicable, then map recommendations to requirements. Note that, for reference purposes only, we will number recommendations, however, the numbers do not indicate priority over each other.

REC 1. Provide a range of query interfaces to accommodate various data search behaviours

Spink & Wolfram et al. (2001) found that users of web search engines rarely used any advanced search features. However, there are differences between discovering data from a repository and searching information on the Web. These include:

- Metadata from a repository are well-structured, which provide more search options, such as field operators and advanced search interfaces.
- Studies show that structured queries that exploit document structures provide more precise answers than those from unstructured queries (Mihajlovic & Hiemstra et al., 2006).
- Users of data repositories may be more aware of advanced search features, possibly having worked with other search systems such as bibliographic search and geographical information search engines. This leads to the requirement that users would like a repository to offer similar search interfaces and search experiences to systems they are familiar with [***ref. REQ 9***].

Overall, we recommend that a repository offer the following three query interfaces:
- Simple search box
- Advanced search
- Map search (if data in a repository is of geospatial in nature.)

A repository may provide a set of search operators or query modifiers for advanced searchers; if so, the repository should keep its search operators as consistent to others' as possible, otherwise users have to learn and remember these operators per repository. For example, we find three repositories offering three different syntax for the 'title' field search operator: *tit:query terms*, *query terms[title]*, *title:query terms*; it would be good if all follow a same syntax.

A repository should investigate the most frequent search tasks from its user and configure its query interface to support these search tasks. For example, the initial data search interface (**Figure 2**) from NSIDC (National Snow & Ice Data Centre) puts spatial and temporal search parameters up-front, as its users are mostly geoscientists who often have clear spatial location in mind when searching for data. The three search functionalities can also be mutually connected as in EnviDat (Iosifescu and Plattner et al., 2018).

REC 2. Provide multiple access points to find data (e.g. search, subject browse, faceted browse/filtering).

Users have different intents when searching for data. Some users may look for a specific data collection and are able to describe the data they are after, while others may not have a clear search target but would like to explore repositories to find any available data (Wu, et al., 2010, Niu & Hemminger, 2010). In many cases, users may need to go through several iterations of search and browse to learn about resources and refine their search to get what they are after (Hearst & Elliott et al., 2002). This is also an identified requirement [***ref. REQ 4***], therefore, a data search interface should support both search and browse search behaviours. A way to achieve this may include: providing subject browse, adding appropriate structures to organise search results, applying appropriate faceted filters. Assante and Candela et al. (2016) reviewed five repositories (Dryad, Figshare, Zenodo, CSIRO DAP and 3TU.Datacentrum); they found all of them offer keyword-based search, facet search and facet filtering.

Facets are usually derived from controlled vocabularies (e.g. subject, data type, file format etc). Data repositories and data providers should work together and adopt community accepted vocabularies, this will give users a consistent search experience across repositories. The tenth recommendation below will discuss using community adopted vocabularies for making machine-to-machine search interoperable.

REC 3. Make it easier for researchers to judge relevance, accessibility and reusability of a data collection from a search summary

After a user gets a search result they will make assessment of which items from candidate lists are relevant to their data search task. The current standard operation of search systems requires users to view summaries of search results; users only proceed to examine a full metadata record itself (as

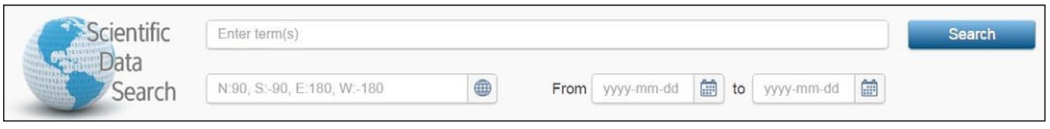

Figure 2: Query interface from National Snow & Ice Data Center (http://nsidc.org/data/search/), with spatial and temporal search up front.

presented in a form of web page) if they find its summary appealing. Summaries of search results do affect how users relate research results to their search topic and their search success (Wu et al., 2001, Turpin & Scholer et al., 2009). It is recommended that search systems:

- **Highlight query terms in search results**
  Highlighting query terms make it clear to data searchers why an item is in search result.
- **Make it clear if data are accessible**
  The first requirement [***ref. REQ 1***] indicates that users care most about data accessibility. The accessibility should be made clear at search summary page and display of an individual data record.
- **Make the data license clear**
  It should be clear what conditions apply for re-using data [***ref. REQ 3***]. If data is associated with an open license, this should be displayed clearly. Even when data provider hasn't provided a licence to a data, displaying 'No License available' would be helpful.
- **Provide preview or statistics of a data set**
  Users search for data in order to find data that can be used for a (research) purpose. Assessing fitness of data for the purpose is an important part of data discovery process. Users would like to have a preview and know statistical features of a data collection [***ref. REQ 6***] in order to make the assessment before they decide to download data or further refine their search. **Figure 3** shows an example from Elsevier Datasearch where a user can click anywhere in the area to preview data.
- **Mark data coverage on a map**
  For spatial search, displaying results on a map will provide a quick summary of search results, and guide users to focus on data from relevant geospatial areas. Furthermore, if a search is of both spatial and temporal features, search results can be displayed on a map with a time slider or layer to help narrow down to a relevant subset.

REC 4. Make Individual metadata records readable and analysable

The presentation structure of a metadata record should have information from most important fields on top of a page, label each field clearly and unambiguously, and make clickable links and buttons recognisable.

Whenever possible, a metadata record should include and clearly display provenance information, for example who collected data, who owns the data, what methods and/or software have been used to collect and process data, and where data are derived from. This provenance information will help users to assess data accountability and, ideally, reproducibility (Wu and Treloar, 2015). [***ref. REQ 2 & 3***].

REC 5. Enable sharing and downloading of bibliographic references

Exporting a data reference to popular formats (e.g. Evernote, Bibtex, etc.) can help a researcher manage references or share the reference with colleagues [***ref. REQ 7***]. This feature often comes with academic portals such as library reference systems and research paper publishers; it is recommended

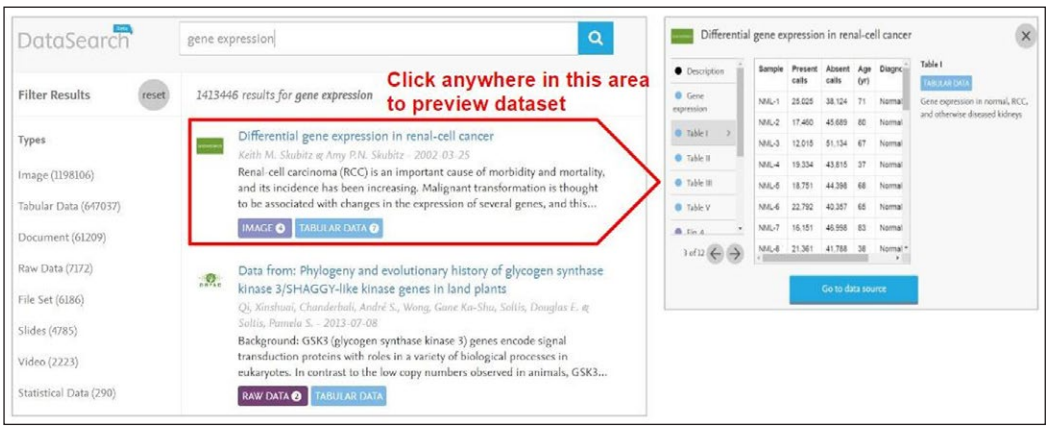

Figure 3: An example of preview data from Elsevier Datasearch (https://datasearch.elsevier.com/#/).

to have this functionality from a data repository as well [***ref. REQ 9***]. Being able to output a data citation in popular publication-acceptable styles (e.g. APA, MLA, etc.) will also encourage data users to cite the data properly.

REC 6. Expose data usage statistics

Usage statistics includes metrics such as metadata viewed, data viewed, data downloaded, data cited, etc. This information can be useful for different purposes:

- Repository managers may want to see this information to better manage and promote their data and improve their data discovery services.
- Data providers want to see their most cited data and to see who cited and viewed their data
- Data users may use data access statistics to gauge if a data collection is widely used by their research community. This information is one of the factors to influence if they would use a data collection [***ref. REQ 3***].

REC 7. Strive for consistency with other repositories.

Consistency is one of the most basic usability principles (Nielsen, 1993). Our users also require portal functionality reminiscent of other established academic portals [**ref. REQ 9**]. It is recommended that a data repository realises consistency at two levels:

- First, a data repository should keep visual appeal, site design, vocabulary and labels and functionality consistent within its own repository. Same action should result in the same effect. If facets are used as filters to search result, use the sets of facets consistently; if a set of facets is sensitive to query and search result, it should be labelled clearly.
- Second, be consistent with other repositories and established academic portals. Research what functionalities are provided by popular repositories and academic portals, very likely your users would demand the same, as indicated by REQ 9. The consistency between repositories can go beyond functionality, it may include the same or similar labels and vocabulary for facets.

REC 8. Identify and aggregate metadata records that describe the same data object

There are cases where: either the metadata of a data collection is published to multiple repositories because of co-ownership. In certain cases, each repository assigns their own Persistent Identifier (PID) to the data collection; or some data repositories cross-harvest each other's metadata records. This may result in:

> ***Duplicate metadata records:*** Two records replicate each other.
>
> ***Parallel metadata records:*** Two records don't completely overlap with each other. Parallel records can be from different organisations (as a result of collaborative work) or different (cataloguing) languages.
>
> ***Augmented metadata records:*** One record has the other record's content as a subset.

This can lead to several copies of metadata records of these types being retrieved. Displaying multiple records of the same data collection may confuse users and waste their time [***ref. REQ 5***].

It would be easier to detect the above types of metadata records if a data collection has a consistent PID across multiple metadata records. However, if this is not possible, a data repository may attempt to use metadata fields such as title, authors, description and linked publication etc. to identify duplicate, parallel and augmented metadata records (Koloniari and Ntarnos, et al., 2011; Weissman, et al., 2015). Users will be helped by the repository aggregating these metadata records and displaying them in a way to make it clear that these records are for the same data collection.

REC 9. Make metadata records easily indexed and searchable by major web search engines

It is important to make data searchable via a data repository as well as by web search engines, as many users search for data through web search engines. Also, researchers who make their research data open would like to have their data searchable through web search engines thus providing wider exposure of their research. To assist in this approach, we recommend that repositories:

1. **Make metadata records easily indexed by web search engines.**
   For a repository that generates a metadata page in html via the repository API, it is recommended to have a sitemap that lists unambiguous URLs of landing pages for each data object.

2. **Make metadata understandable by web search engines.**
   When a metadata record is indexed by web data search tools such as Google dataset search,[7] the metadata should be described in a way that is understandable by web search engines. A Data Citation Roadmap (Fenner and Crosas et al., 2016) recommends encoding Dublin Core metadata etc. in HTML meta tags and/or annotate landing page with schema.org in JSON-LD format to represent schema.org metadata. This structured way of describing data can help to improve data discovery on the Web by enabling web data search tools to link structured metadata to scientific publications, authors, or even knowledge graph (Noy and Brickley, 2017).

REC10. Follow API search standards and community adopted vocabularies for interoperability

A data repository is a node in the networked knowledge infrastructure (Borgman, 2015); when all data repositories and other scholarly repositories are inter-connected, more added-value services can be built. To achieve this, the next generation of repositories need to achieve new levels of web-centric interoperability (Shearer et al., 2016). In this networked environment, it is important for a data repository to provide services that support both human users and software agents. It needs be findable by data repository aggregators and applications such as Google Scholar, Web search engines and Web dataset search tools. To be discoverable and friendly to a software agent, data repositories should use community adopted vocabularies for example the W3C standards for describing semantics of Web resources and linked data (https://www.w3.org/standards/semanticweb/), the W3C Data Catalogue Vocabulary (https://www.w3.org/TR/vocab-dcat/) and schema.org (http://schema.org/).

Data repositories should also follow API search standards. For many services that aggregate search results from multiple repositories, repositories syndicating search results or recommending similar data collections from other repositories, using a community adopted search API, such as OpenSearch[8] or SRU- Search and Retrieval by URL[9] (Hammond, 2010), and community adopted (machine readable) vocabulary will enable interoperability between various starting points and offer greater flexibility and processability for data consumers (Lóscio, Burle and Calegari, 2017). Improvement in interoperability will enable greater data discoverability across repositories.

Table 4 shows a mapping between recommendations and requirements: each Requirement is supported by at least one Recommendation, except for the Requirement 8 '*Accompanying education/training material*'. Although data repositories can play a role in satisfying this requirement by providing a 'Help' page, the primary responsibility here may rest with on libraries and research offices. Thus, we map this requirement to a publication '*Eleven quick tips for finding research data'* (Gregory et al., 2018) which librarians and research offices can present to their users.

In **Table 4**, Recommendations 9 and 10 are not directly mapped to any requirements as requirements were inferred from use cases from human users. Nevertheless, these two requirements are important in that Recommendation 9 addresses a common behaviour that many researchers are using web search engines as their primary tool to search for publications and data. Recommendation 10 supports one of the four FAIR data principles -interoperability – that will not only benefit software agents but also enable the consistency as discussed in the seventh recommendation.

4. Discussion and Conclusion

In order for data repositories to better support users' data discovery activities, we need first to understand the users' requirements: why users search for data, what they use data for and how users would search for data. In this study, we collected and documented 79 use cases and clustered them into three broad categories. While the use cases indicate requirements for a range of purposes, such as for data providers to consider what information should be provided in metadata, we focus on use cases that lead to functional requirements that data repository operators can consider or implement when developing their data portal.

[7] Google dataset search: https://toolbox.google.com/datasetsearch.
[8] OpenSearch: http://www.opensearch.org.
[9] Search and Retrieval via URL: http://www.loc.gov/standards/sru/.

Table 4: Matching requirements to recommendations.

Requirement	Recommendations
REQ1: data availability	REC3 Assessable search result
REQ2: Connection of data	REC2 Multiple access points REC4 Readable metadata records REC8 Identifiable duplicates
REQ3: Annotations	REC3 Assessable search result REC4 Readable and analysable metadata records REC6 Available data usage statistics
REQ4: Filtering with single or multiple criteria	REC1 Multiple query interfaces REC2 Multiple access points
REQ5: Cross-reference	REC8 Identifiable duplicates
REQ6: Inspection of data	REC1 Multiple query interfaces REC2 Multiple access points REC3 Assessable search result
REQ7: Collaborative environment	REC5 Available bibliographic references
REQ8: Training material	Eleven quick tips for finding research data
REQ9: Similarity across portals	REC1 Multiple query interfaces REC2 Multiple access points REC7 Consistent interface
Support data searchers from web search engines	REC9 Findable from web search engines
The Fair Data Principles – interoperability	REC10 Interoperability with other repositories

Through qualitative analysis of the use cases, we derived nine requirements that can be applied to data repositories. Note that some requirements are not special to data repositories but have been applicable to information discovery systems in general; requirements as such include REQ4 (Be able to filter data based on specific criteria), REQ 8 (Provide educational/training material), and REQ 9 (Have similar search functionalities and interfaces to other established academic portals). The remaining five requirements, REQ 1. (Indication of data availability), REQ 2 (connection to related resources), REQ 3 (fully annotated data), REQ 5 (cross-referencing of data), REQ 7 (sharing data in a collaborative environment) are more data discovery oriented. While the use cases cover wide range of research disciplines and several roles (e.g. researchers, librarians and funders), each data repository is unique in one way or another. When developing a data portal, a data repository can consult their end users with these nine general requirements and prioritise them. They may also elicit new requirements specific to data they hold and their user community.

We presented a set of recommendations and discussed how these requirements can be supported through recommendations. Some recommendations, such as REC1 (multiple query interfaces), REC2 (multiple access points) and REC3 (assessable search result), were drawn from studies of information discovery systems in general and academic digital libraries in particular. Nevertheless, we discussed each of these recommendations for their applicability to data repositories. Recommendations that are specific to data repositories include REC4 (readable and analysable metadata records), REC8 (identifiable duplicate, parallel and augmented metadata records), REC9 (findable from web search engines) and REC10 (interoperable with other repositories). Data repositories can take the ten recommendations as guidelines when implementing a new repository or as a checklist when conducting heuristic evaluation of an existing repository. Data repositories can implement all or prioritise their implementation based on their user needs and available resources.

Clearly, improving data discovery paradigms requires a collective effort by data collectors, data providers, data repositories, data librarians and research trainers. Although there is no single best route to building an optimal data discovery portal, we hope that the use cases, requirements and recommendations provide a starting pointer to improve data search features. In the future, we would like to work with data repositories to validate the requirements and evaluate and refine the recommendations.

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, especially Dom Fripp, Jennie Larkin, William Michener, Natalia Atkins, Beth Huffer, Jens Klump, Andrea Perego, Kathleen Fontaine, Kathleen Gregory, Antica Culina, Anusuriya Devaraju and Tim Clark who contributed to the Use Cases and the Best Practices Task Forces. This paper was supported by the RDA Europe 4.0 project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777388

Competing Interests

The authors have no competing interests to declare.

References

Assante, M, Candela, L, Castelli, D and **Tani, A.** 2016. Are Scientific Data Repositories Coping with Research Data Publishing? *Data Science Journal*, 15: 6. DOI: https://doi.org/10.5334/dsj-2016-006

Borgman, CL. 2015. Big data, little data, no data: Scholarship in the networked world. MIT press.

Charmaz, K. 2006. Constructing grounded theory: A practical guide through qualitative analysis. London: Sage Publications.

de Waard, A, Khalsa, SJ, Psomopoulos, F and **Wu, M.** 2017. RDA IG Data Discovery Paradigms IG: Use Cases data [Data set]. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.1050976

Fenner, M, Crosas, M, Grethe, J, Kennedy, D, Hermjakob, H, Roca-Serra, P, Berjon, R, Martone, M and **Clark, T.** 2016. A Data Citation Roadmap for Scholarly Data Repositories. In: *BioArxiv*. DOI: https://doi.org/10.1101/097196

Ferguson, N. 2016. Jisc Research Data Shared Service metadata focus group use cases [Data set]. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.193011

Gregory, K, Khalsa, SJ, Michener, WK, Psomopoulos, FE, de Waard, A and **Wu, M.** 2018. Eleven quick tips for finding research data. *PLoS Computational Biology*, 14(4): e1006038. DOI: https://doi.org/10.1371/journal.pcbi.1006038

Hammond, T. 2010. nature.com OpenSearch: A Case Study in OpenSearch and SRU integration. *D-Lib Magazine*. Volume 16, Number 7/8, July/August 2010. DOI: https://doi.org/10.1045/july2010-hammond

Hearst, M, Elliott, A, English, J, Sinha, R, Swearingen, K and **Yee, K.** 2002. Finding the Flow in Web Site Search. In: *Communications of the ACM*, 45(9): 42–49. Sept. 2002. DOI: https://doi.org/10.1145/567498.567525

Iosifescu Enescu, I, Plattner, G-K, Espona Pernas, L, Haas-Artho, D, Bischof, S, Lehning, M and **Steffen, K.** 2018. The EnviDat Concept for an Institutional Environmental Data Portal. *Data Science Journal*, 17: 28. DOI: https://doi.org/10.5334/dsj-2018-028

Joo, S and **Lee, JY.** 2011. Measuring the usability of academic digital libraries: Instrument development and validation. *Electronic Library*, 29(4): 523–537. DOI: https://doi.org/10.1108/02640471111156777

Kim, JY, Feild, H and **Cartright, MA.** 2012. Understanding Book Search Behavior on the Web. In: *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM'12)*, 744–753. Oct 29–Nov. 2, 2012, Maui, HI, USA. DOI: https://doi.org/10.1145/2396761.2396856

Koloniari, G, Ntarmos, N, Pitoura, E and **Souravlias, D.** 2011. One is enough: Distributed filtering for duplicate elimination. In: *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*, Berendt, B, de Vries, A, Fan, W, Macdonald, C, Ounis, I and Ruthven, I (eds.), 433–442. ACM, New York, NY, USA. DOI: https://doi.org/10.1145/2063576.2063643

Lóscio, BF, Burle, C and **Calegari, N.** (eds.) 2017. Data on the Web Best Practices. (W3C Recommendation 31 January 2017). Available at: https://www.w3.org/TR/dwbp/ Accessed on 9 August 2017.

Michener, WK, Allard, S, Budden, A, Cook, RB, Douglass, K, Frame, M, Kelling, S, Koskela, R, Tenopir, C and **Vieglais, DA.** 2012. Participatory Design of DataONE Enabling Cyberinfrastructure for the Biological and Environmental Sciences. United States. DOI: https://doi.org/10.1016/j.ecoinf.2011.08.007

Mihajlovic, V, Hiemstra, D, Blok, HE and **Apers, PMG.** 2006. Exploiting Query Structure and Document Structure to Improve Document Retrieval Effectiveness. Report from Centre for Telematics and Information Technology, University of Twente, is Available at: https://core.ac.uk/display/11470014.

Murphy, D and **Gautier, J.** 2017. A comparative review of various data repositories. Accessed from: https://dataverse.org/blog/comparative-review-various-data-repositories on 29 Nov. 2017.

Nielsen, J. 1993. Usability Engineering. Publisher: Morgan Kaufmann.

Nielsen, J. 1995. How to Conduct a Heuristic Evaluation. Available from: https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/ accessed on 31 Oct. 2017.

Niu, X and **Hemminger, BM.** 2010. Beyond text querying and ranking list: How people are searching through faceted catalogs in two library environments. In: *Proceedings of the Association for Information Science and Technology*, 47(1): 1–9. November/December 2010. DOI: https://doi.org/10.1002/meet.14504701294

Noy, N and **Brickley, D.** 2017. Facilitating the discovery of public datasets. Retrieved on 25 Oct. 2018 from Google AI Blog: https://ai.googleblog.com/2017/01/facilitating-discovery-of-public.html.

Sanderson, M and **Croft, WB.** 2012. The History of Information Retrieval Research. In: *Proceedings of the IEEE 100*. Special Centenial Issue, 1444–1451. ISSN: 0018-9219. DOI: https://doi.org/10.1109/JPROC.2012.2189916

Shearer, K, Rodrigues, E, Walk, P and **Perakakis, P.** (2016, October). Next Generation Repositories. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.163264

Soy, SK. 2018. The Case Study as a Research Method. Accessed 26/10/2018 from: http://www.gslis.utexas.edu/~ssoy/usesusers/l391d1b.htm.

Spärck, JK. 2006. Information retrieval and digital libraries: Lessons of research. In: *Proceedings of the 2006 international workshop on Research issues in digital libraries (IWRIDL '06)*, Majumder, P, Mitra, M and Parui, SK (eds.), Article 1, 7 pages. ACM, New York, NY, USA. DOI: https://doi.org/10.1145/1364742.1364743

Spink, S, Wolfram, D, Jansen, BJ and **Saracevic, T.** 2001. Searching the Web: The public and their queries. *Journal of the American Society for Information Science*, 52(3): 226–234. DOI: https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R

Turpin, A, Scholer, F, Jarvelin, K, Wu, M and **Culpepper, JS.** 2009. *I*ncluding Summaries in System Evaluations. *Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 508–515. Boston, MA. July 2009. DOI: https://doi.org/10.1145/1571941.1572029

Weissman, S, Ayhan, S, Bradley, J and **Lin, J.** 2015. Identifying Duplicate and Contradictory Information in Wikipedia. In: *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15)*, 57–60. ACM, New York, NY, USA. DOI: https://doi.org/10.1145/2756406.2756947

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, 3: 160018 DOI: https://doi.org/10.1038/sdata.2016.18

Wu, M, Fuller, M and **Wilkinson, R.** 2001. Searcher performance in question answering. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*, 375–381. ACM, New York, NY, USA. DOI: https://doi.org/10.1145/383952.384028

Wu, M and **Treloar, A.** 2015. Metadata in Research Data Australia and the Open Provenance Model: A Proposed Mapping. In: Weber, T, McPhee, MJ and Anderssen, RS (eds.), *MODSIM2015, 21st International Congress on Modelling and Simulation*, 641–647. Modelling and Simulation Society of Australia and New Zealand, December 2015, ISBN: 978-0-9872143-5-5. http://www.mssanz.org.au/modsim2015/C4/wu.pdf.

Wu, M, Turpin, A, Puglisi, SJ, Scholer, F and **Thom, JA.** 2010. Presenting query aspects to support exploratory search. *Proceedings of the Eleventh Australasian Conference on User Interface*, 106: 23–32.

How to cite this article: Wu, M, Psomopoulos, F, Khalsa, S and de Waard, A. 2019. Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. *Data Science Journal,* 18: 3, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2019-003

Submitted: 22 August 2018 **Accepted:** 13 December 2018 **Published:** 08 January 2019

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

OPEN ACCESS