Data Discovery Paradigms:
User Requirements and Recommendations for Data Repositories

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

[1]Australia National Data Services, Melbourne, Australia
[2]Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece
[3]National Snow and Ice Data Centre, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA
[4]Office of Research Evaluation and Operations, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, USA
[5]Research Data Management Solutions, Elsevier, USA

Abstract

Purpose - The work presented in this paper is intended to help data repositories improve the findability of data in their repository.
Design/methodology/approach - Through existing resources and an online questionnaire, we collected 79 use cases describing users' needs, and the contexts of these needs, when searching for data.
Findings - By analysing these use cases, we identify nine requirements for data discovery. We propose a set of recommendations with exemplar implementations for data repositories to consider when they develop or improve their data portals.
Originality/value – This work contributes to the improvement of data repositories by enabling effective data discovery. 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. It focuses on enabling and improving the methods and tools by which users find data in these repositories.
Keywords – Data Repository, Data Findability, Metadata, Use Case, User Requirement

Introduction

A widely-endorsed statement on research data states 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 register their metadata into a data repository to make data discoverable; on the other hand, data repository operators should make the data easily indexable and searchable through the provided metadata. Over the past decade, we have seen an increasing number of public and domain specific data repositories appear. For example, re3data.org, the Registry of Research Data Repositories, had 23 repositories when it went online in 2012; the number quickly increased to over 1,200 data repositories from across the globe in three years (Pampel and Vierkant, 2015), and by September 2017, the registry had more than 1,900 repositories[1].

[1] http://www.re3data.org/, accessed on 1st of August

While it is a good thing that there are more and more data open and available through data repositories, it becomes ever more challenging for researchers to find relevant data. If researchers can't find data easily in a repository or across repositories, they may stop using it; this not only wastes time and effort put into the data repository, but also may discourage researchers from making their research data available to others. Improving data findability of a data repository will benefit all people and organisations who are involved in data ecosystem from collecting data to using data.

This paper presents work done by the Data Discovery Paradigms Interest Group (DDP IG), an Interest Group of the Research Data Alliance[2]. The goal of the presented work is to provide recommendations to enable data repositories to improve findability of data in their repository. By gathering and analysing use cases from data seekers, we identify a core set of nine functional requirements that a data repository should support. We propose a set of recommendations with exemplar implementations for data repositories to consider when they develop or improve their data portals. 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. It focuses on enabling and improving the methods and tools by which users find data on these repositories.

The rest of the paper is organised as follows: we first provide a review of current research and practices related to data discoverability, then describe the method to collect, gather and edit a series of use cases regarding data discovery and to identify requirements. We then present our recommendations and suggest how they can be adopted.

Related Work

The work presented in this paper covers multiple research and practice areas including user requirement gathering, data search analysis, data portal design, and data in the networked world. We will review related work along each area.

Requirements gathering

The W3C Data on the Web Best Practices Working Group (www.w3.org/TR/dwbp/) aimed to develop the open data ecosystem, improve communication between developers and publishers, and promote the re-use of data. The W3C group has collected use cases from 26 organisations, mostly government and public organisations, describing their current practice of managing and describing data and how these data would be used. From the collected use cases, the group identified 12 challenges and 42 requirements. The challenges include: metadata, data licenses, data provenance & quality, data versioning, data identification, data formats, data vocabularies, data access, data preservation, feedback, data enrichment, and data republication (https://www.w3.org/TR/dwbp-ucr/). The group recommends 35 best practices to support the requirements. The best practices are for data publishers to increase data reusability, processibility, interoperability, accessibility, trustiness, discoverability, likability and comprehension (Lóscio, 2017).

Gregory et al. (2017) reviewed literatures on observational data retrieval from multiple disciplines, aimed to find commonality in how users search for and evaluate data. The review is summarised along the four key stages in interactive information seeking: users and needs

[2] Research Data Alliance: Data Discovery Paradigms Interest Group https://www.rd-alliance.org/groups/data-discovery-paradigms-ig

(user contexts and data needs), user actions (the sources and search strategies used to locate research data), system (research data retrieval systems and their search functionalities), and evaluation (criteria and processes used to evaluate data for (re)use). Although all key findings from each of the above four dimensions are relevant to the study described in this paper, we would like to draw particular attention to desired system and search functionalities, which include: thematic search on maps, manipulation of time series, linkage with analytical and modelling tools, linkage data sets with citing/related publications, citation metrics for datasets, searching supporting documentations, and contacting data author, along with a few others that were specific to disciplinary data repositories.

Apart from published work, data repositories usually take a step to gather user requirements either as the first step to guide architecture design of repository or verification and usability testing of repository for further refinement. For example, JISC (https://zenodo.org/record/193011#.WgYqpluCzDB) collected use cases about research data shared services in general, DataONE (https://www.dataone.org/user-personas) developed personas to capture their user requirement, while Research Data Australia developed the repository by following best practices in the field, then conducted user interview and user experience evaluation to gather feedback to the repository and user data search needs in general. These rich resources are usually scattered on the Web, as they were primarily used for the design and development of a particular repository.

Data search analysis

In the field of information retrieval, 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 (Spärck-Jones 2006, Sanderson and Croft. 2012). Discovering data in data repositories may have some resemblance to book or literature search in a digital library or online catalogues, as both data repositories and online catalogues have structured metadata that allows implementation of rich interaction methods such as facet search and filtering. Based on an analysis of search logs from the Open Library (OL), Kim et al. (2012) compared how users find books and book-related information from within the OL (internal search) versus from web search engines (external search). Their study finds that queries sent from internal searches are shorter than external searches (2.31 vs 4.24 words), internal searches have longer duration (almost doubled) and richer interaction than external searches. Advanced search interface, faceted filtering and sorting search results by criteria are all used more frequently by internal searchers. The study recommends digital libraries or online catalogues should exploit field-level metadata for ranking, query suggestions, and other purposes.

Kern and Mathiak (2015) conducted two user studies for investigating if there were any differences in data set retrieval compared to well-known literature retrieval in the context of a requirement analysis of a data catalogue in social science domain. They find that choosing a data set is a much more important decision for a researcher than choosing a piece of literature. In particular, quantity, and quality of metadata are far more important in data set search than in literature search, where convenience is most important. On usability side, most of their participants expect the data catalogue have a Google-like input field with Boolean search, autocomplete function, search term suggestion and an auto-correct on spelling mistakes as a bare minimum.

Apart from querying and search, most catalogue systems also offer faceted filtering and navigation to help their users find they search for without having to describe their need and formulate a query. Faceted filtering and navigation help users broaden or narrow a search,

which has been successfully applied by many vertical applications, such as e-commerce and digital libraries (Kong and Allan, 2014). What sets of facets to offer and in which context also affect user search experience. Niu & Hermminger (2010) investigated people's searching behaviour with two faceted library catalogues through a transaction log analysis. One studied catalogue is from a public library and the other is from a university library. Their analysis shows that people do incorporate facets when they are searching both catalogues, but, the facet usage from the public catalogue is higher than that from the university catalogue, probably due to the public catalogue supports facet browsing in addition to facet refining and has facets that describe the content better.

Data repository design and usability analysis

A data repository, as a type of digital library, should follow general design and usability guidelines of a digital library. Jooe & Lee (2011, p. 524) states that usability, in the context of digital libraries, means the ease of use, prolificacy and the extent of satisfaction it provides to its users. There are many studies on usability analysis and testing of digital libraries (e.g. Theng et al. 1999, Xie 2008, Jeng 2008, Dalkıran et al. 2014). Tsakonas and Papatheodorou (2008) explored usefulness and usability in the evaluation of open access digital libraries; a key finding from this study is that an open access digital library needs to be closely linked with users' work tasks to be self-sustainable. They also proposed a model for assessing the usefulness and usability of an open access digital library. The model attempted to demonstrate how the significant content and system attributes affect user interaction and satisfaction. Xie (2008) collected users' perceived importance of digital library, five highly ranked perceptions or criteria include interface usability, collection quality, service quality, system performance and user satisfaction. Further, the interface usability includes interface usability in general, search and browse functions, navigation, help features, view and output options and accessibility.

Data repositories should have their own special functional features from other types of digital libraries for data oriented search and archive. There is a review by Murphy and Gautier (2017) to examine what features are supported by eight repositories, features are grouped under: software features, governance/organisation, and content. Although the purpose of this activity was to improve services offered by Dataverse (https://dataverse.org/), the review would be also useful to other repositories about what are common features they should support.

Data repository in the networked world

A data repository cannot stand alone in the networked world, as Borgman (2015) argues that data exist within a knowledge infrastructure. A data repository is a node in the networked knowledge infrastructure, when all data repositories and other scholarly repositories are inter-connected, more value added services can be built. To achieve that, 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 and Web search engines. To be discoverable and be friendly to a software agent, data repositories need to follow standardised API and common vocabularies, for example, standards from 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/).

Summary

As a rapidly emerging, web-based system to support data sharing and discovery, data repositories can adapt methodologies and learnt experiences from the design of web-based information systems, and digital library in particular. Some data repositories have been following the path of user-centred system design principle - gathering use cases and requirements, involving user feedback and evaluation in the designing and development process. While some design and evaluation criteria summarised from their repository development process may apply only to their own repository, some trans-repository criteria can be generalised and may serve as guidelines for other data repositories. The study presented in this paper attempts to fill in this gap and help identify which repository criteria are of common relevance across multiple repositories.

Methods

To recommend best practices of making data more findable within data repositories, the first step is to understand what and how data repository users would search for data. For that purpose, we set out to collect use cases of what users might wish to search for data and what support they would expect a data repository to provide. We then analysed collected use cases to identify common themes and functional requirements. Specifically, we followed these steps:

Step 1. Collect existing use cases and map them to a common schema, so they could be compared.

Step 2. Conduct a survey for collecting more use cases, especially aiming for types of users who were not covered by the use cases collected in Step 1.

Step 3. Analyse all use cases to distil the large number of individual uses cases into clusters that share common themes and similar functionalities.

Step 4. Elicit functional requirements from use cases, and prioritise the requirements.

Each step is described in the following sub-sections.

Collecting existing use cases

The first step was to collect use cases. We determined that there were already several rich sources of use cases available, which were collected by different organisations using their own surveys or interviews, in the context of improving their own data search services. We aggregated use cases from the following five resources:

- Jisc Research Data Discovery Service use cases[3]
- ANDS Falling Water User Interview Responses
- BioCADDIE
- DataOne: DataOne Personas
- Spatial Data on the Web

[3] Jisc Research Data Discovery Service use cases: doi: 10.5281/zenodo.193011

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

Table 1: Examples of use cases being transformed in the open interview format.

However, these five resources describe use cases in different formats, for example, DataOne used persona, the ANDS project recorded answers to their own interview questions, while Jisc used an open interview format[4]. 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 chose the open interview format. In this open interview format, each record 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)

This format is simple, yet it records essential information: a user's background (role and discipline), needs ("I want") and purpose of a need ("so that").

From the other four resources we took those use cases that could be unambiguously transformed into this format without losing any information. For example, a use case from an interview of the ANDS Falling Water project shows that a Ph.D student, from the field of Economics, usually knows what data he wants to have; so what he wants from a (portal) homepage is a simple (clean) page with search functionality. He would like to have advanced search functionality in case he needs to refine a search. Another example of personas from DataOne[5] describes an early-career herpetologist, who is interested in finding additional data that correspond to the location of tortoise populations, and additional tortoise data so she can put her current study into perspective and perhaps find collaborators. Table 1 shows the result of transforming the above two use cases into the open interview format.

As a result of this collecting and transforming process, we had 64 use cases that can be described by the five fields from the open interview format.

Conducting survey for collecting more use cases

When reviewing the existing use cases, we found almost all of them focused on the "Researcher" role (for example: Academics, Researchers and PhD/Master students). To include more diversified roles, we conducted a survey with the above five questions. We distributed the survey to the DDP IG members and through several mailing lists representing

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

[5] Individual Personas from DataOne: Sun:Early-career herpetologist.

different communities such as ALA Scholar Communication, ACRL Science & Technology Section, NARO Physics-Astronomy-Mathematics Division, and JISC Research Data Management. Through this survey, we collected 15 additional use cases and broadened the scope of role to include librarian and funder. The survey instruments and survey data are available from Zenono (deWaard et al. 2017).

Analysing and clustering the Use-Cases

We collected 79 use cases (deWaard et al. 2017) in total from Step 1 and Step 2 of the Method Section. We first analysed the use cases to address the issue of free text mostly prevalent in the existing use-cases, especially the "*As A*" part of each record, since this exhibits a wide variance not necessarily required for the further analysis of the data. For example: the role of a researcher comes in various forms: Professor, Principal Scientist, Early Career Researcher, Student (PhD/Master). Therefore, this particular field was transformed by manually assigning each case to one of the following four major user types; "Researcher", "Research Student (PhD/Master)", "Librarian", or "Funder".

Next we analysed use cases from two dimensions: (1) identify issues related to data discovery, and (2) identify who should take responsibility to address the data issue. We especially sought to identify those common issues related to data discovery and turn them into user requirements. For that purpose, we had one author tag each use case with an open vocabulary along the two dimensions and another author tag a second round and also check consistency across all use cases. The data issue tags resulted in 24 vocabularies. We then classified these 24 vocabularies into a top level of three major categories: Metadata, Portal Functionality, and Data. The 24 vocabularies became sub-categories. The final classification scheme derived from this bottom-up approach is shown in Figure 1.

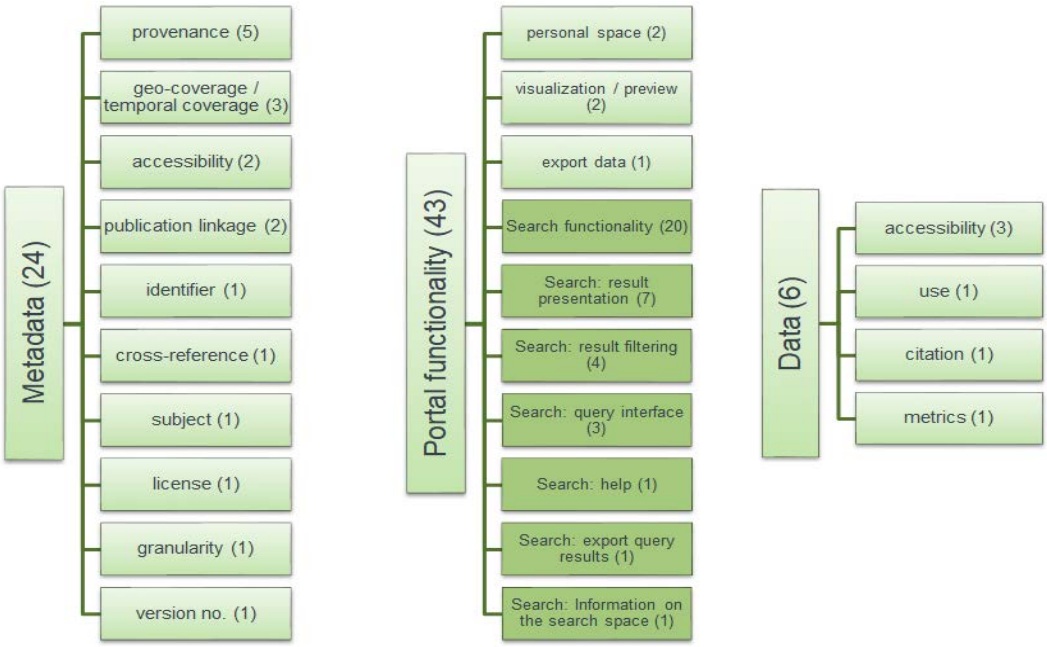

***Figure 1**: A classification scheme derived from Use Cases. The dark green color denotes a further sub-sub category of "Portal Functionality", namely "Search functionality"*[6]

[6] NB since some are counted twice, the total is not the same as the total listed in the Spreadsheet

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**	

***Table 2**: From Use Cases to Requirements*

The driving force behind this categorization was to assess potential areas in the data discovery process that were dominant, while at the same time introducing a first low-level clustering of the use-cases. Note that a use case can be in more than one category. 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 (should metadata have information accessibility information, it should be made clearly to a user). Assigning at maximum two major and two minor classes (i.e. categories and subcategories) for each use case allowed for the investigation of potential connections between the classes.

Eliciting User Requirements

The classification scheme allowed for a general overview of the missing aspects in data discovery from the perspective of the relevant technologies (i.e. Portal, Metadata and Data). However, the usefulness of these attributes can be enhanced by investigating the specific user needs that are currently being unmet. 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. Using this contextual grouping of the use cases, we identified the common aspect described by each group, and then formulated this aspect as a distinct requirement. An example of this process is shown in Table 2. Ultimately, nine individual groups (or requirements) were identified.

It is important to note that these requirements 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 address the particular requirement). We identified the following three attended audiences and assigned them to each requirement as appropriate: 1) Data Repository, 2) Data Provider and 3) Research Office/Libraries.

Finally, and in order to better understand how relevant these requirements are to the user community, we circulated a second survey, asking for a ranking of each requirement independently from 1 (Not important) to 5 (Very Significant) and including a no-opinion option. We received 31 responses to this survey, which allowed us to rank the nine requirements as

follows (in order of descending importance). The primary targeted audience in each requirement is highlighted in bold. The description is an extract from the “So that” field from the open interview format.

REQ 1. Indication of data availability

User Type: Researcher / Research Student

Target Audience: ***Data Repository****, Data Provider*

Description: *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

User Type: Funder / Researcher / Research Student

Target Audience: ***Data Repository****, Data Provider*

Description: *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)*

User Type: Researcher / Research Student

Target Audience: ***Data Provider****, Data Repository*

Description: *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)***.**

User Type: Researcher / Research Student

Target Audience: ***Data Repository****, Data Provider*

Description: *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)***.**

User Type: Researcher / Research Student

Target Audience: ***Data Provider****, Data Repository*

Description: *Having same data 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 maximise efficiency and access.*

REQ 6. Visual analytics / inspection of data / thumbnail preview

User Type: Researcher

Target Audience: ***Data Repository***

Description: *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**
User Type: Researcher / Research Student
Target Audience: ***Data Repository***
Description: *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.*

REQ 8. Accompanying educational / training material
User Type: Librarian
Target Audience: ***Research Office / Libraries****, Data Repository*
Description: *Help researchers manage and discover data in a methodical and seamless manner.*

REQ 9. Portal functionality similar to other established academic portals
User Type: Researcher
Target Audience: ***Data Repository***
Description: *For example, finding more with a subject, search by visual (i.e. draw a structure to search for), free text search, build query functionality, subscription, save lists.*

Recommendations to Data Repositories on Data Discovery

While one task force of the DDP IG was working on use cases and identifying the above set of requirements, another task force was working on identifying and recommending best practices on data discovery to data seeker, data provider, and data repository. The efforts of the two task forces were merged by matching the best practices for repositories that can meet the requirements from the use case study. This was done by taking the collected use cases and the FAIR data principles (Wilkinson, et al., 2016) as starting points, doing an environmental scan of existing data repositories and then using this information to do the matching. Drawn from research and practices as reviewed in the “Related Work” Section and our research with user search experience, we came up ten recommendations. In this section, we will discuss each recommendation and provide exemplars along with each recommendation where ever applicable.

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

Users of web search engines use few search terms (mean of 2.6 terms), and rarely use any advanced search features (Spink & Wolfram et al. 2001). 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.

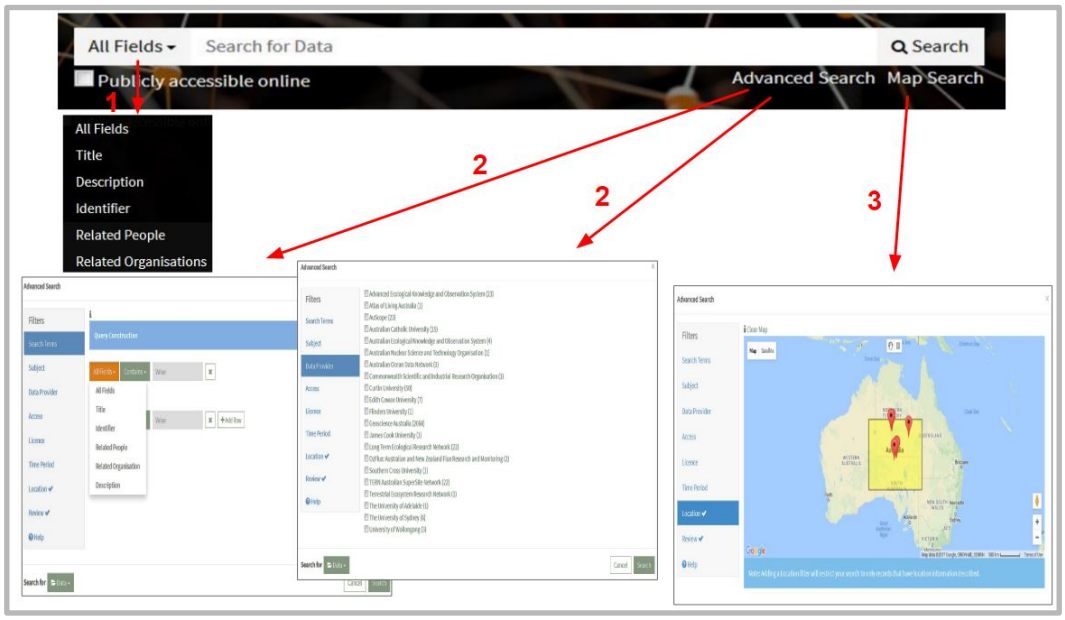

Figure 1. Query interface examples from Research Data Australia: 1) limiting search to certain field(s) and data objects, 2) advanced search form, 3) map search

- 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.)

Figure 1 shows an example query interface. It offers a simple query box where a user can post a simple query or construct an advanced query by using available search operators. 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, our environment scan finds 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.

Under this simple query box, there are two options: one is Advanced Search that allows the construction of advanced queries through a form, the other is Map Search that complements text search by enabling search via a geographic boundary with a bounding box or polygon.

A repository should investigate the most frequent search tasks from its users, 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

Figure 2. Query interface from NSIDC[7], with spatial and temporal search up front

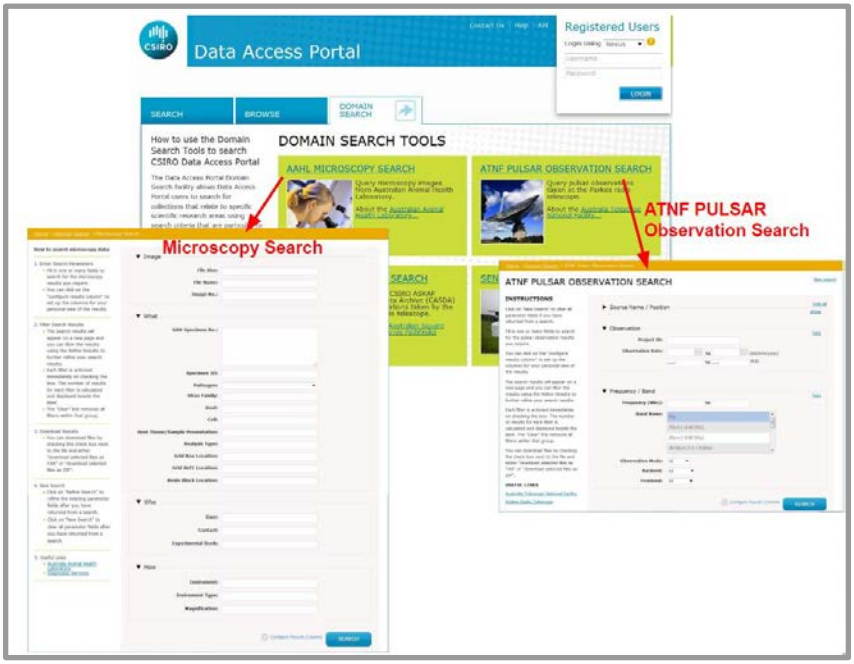

Figure 3. Query interface from CSIRO Data Access Portal[8], with advanced search customised for each research domain.

who often have clear spatial location in mind when searching for data. Figure 3 shows another example from CSIRO Data Access Portal, which provides access to data from multiple research domains, has customised query interface for each research domain.

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 and Turpin 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 a requirement that came up during the study of use cases [***ref. REQ 4***], therefore, a data search interface should support both search and browse search behaviours. One way to achieve this is by adding appropriate structures to organise

[7] National Snow & Ice Data Center: http://nsidc.org/data/search/

[8] CSIRO Data Access Portal: https://data.csiro.au

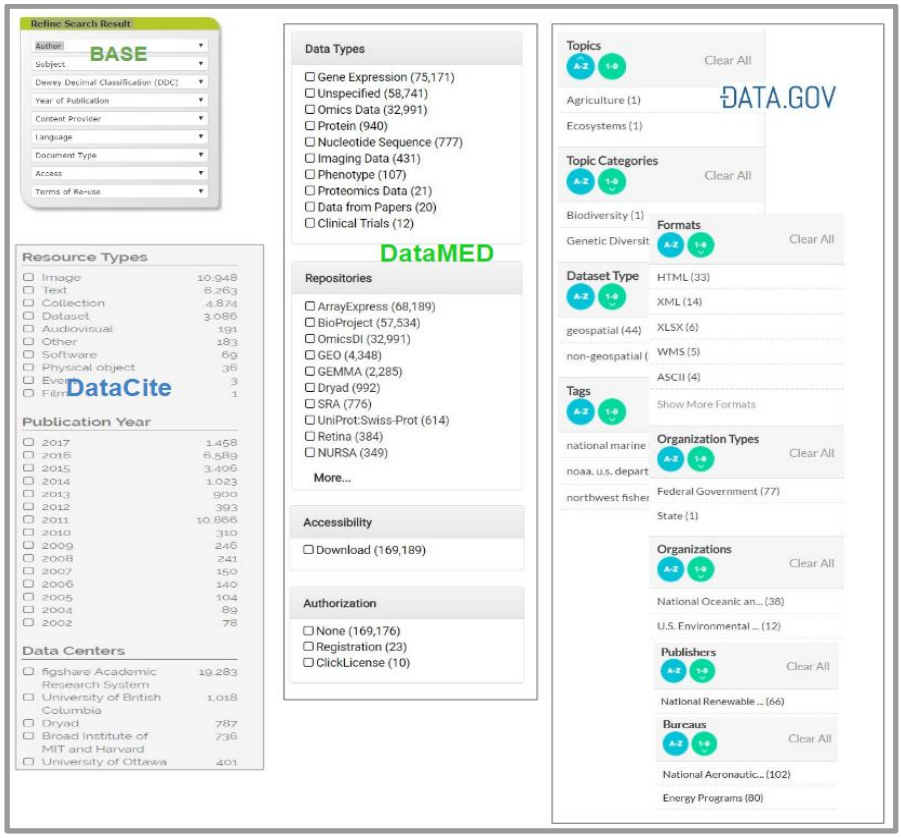

Figure 4. Examples of facet filters/browsing.

search results. Figure 4 shows examples from a few repositories, where users browse search results by applying appropriate faceted filters or start browsing and then search from there.

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

After a user gets a search result s/he will make assessment of which items from candidate lists are relevant to his/her 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 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 and Fuller et al., 2001, Turpin & Scholer et al., 2009). It is recommended that search systems:

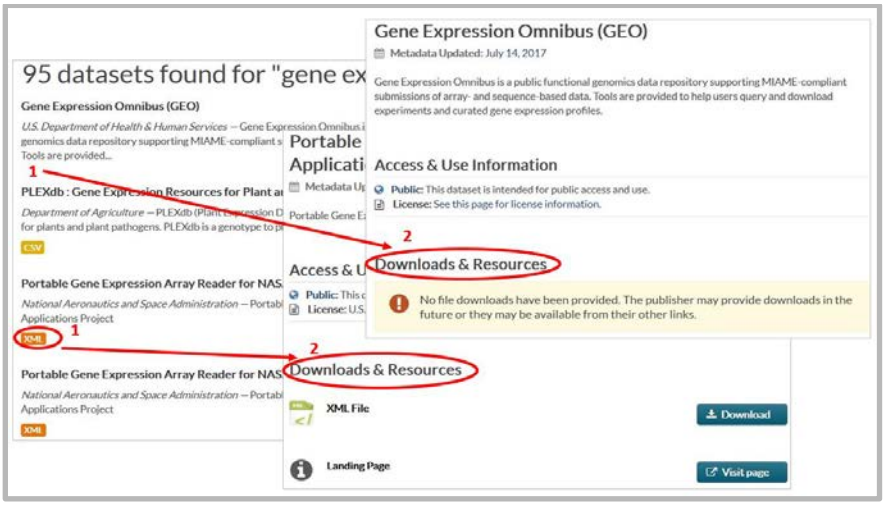

Figure 5. An example from data.gov that clearly labels data accessibility and access method from search result (1) and at record level (2)

- **Highlight query terms in search results**
  This is an almost a default practice for all search systems. 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 when users search for data, they most care about data accessibility. If a data object is not accessible, they may discontinue their search. Figure 5 shows an example from data.gov that clearly labels data accessibility as a part of snippet of a record (area 1) and at a record level (area 2). If nothing is shown in the area 1, that means no files are there for download; first time users may have to learn this rule only when they click and land on metadata record page, so it is recommended to make the rule explicit, for example, having a label "No file for download yet" or similar labels with the same effect.
- **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. If a data provider hasn't provided a licence to a data, displaying "No License available" would help users assess the data's reusability. Figure 5 shows a good example from data.gov.
- **Provide preview or statistics of a data set**
  In many cases, users would like to have a preview and know statistical features of a data collection [***ref. REQ 6***] in order to decide if the data fits their research purpose before they proceed to download data, especially if a collection is big and includes large sets of files. Figure 6 shows an example from Elsevier Datasearch where a user can click anywhere in the area of a snippet to preview data. Figure 7 shows an example of an advanced data preview from Australian Plant Phonemics Facility[9] collection portal, where a click on a collection triggers an application which enables a user to explore various statistical features of selected data points and variables (Sainsbury and Berget et al., 2017).

[9] Australian Plant Phenomics Facility: http://www.plantphenomics.org.au/

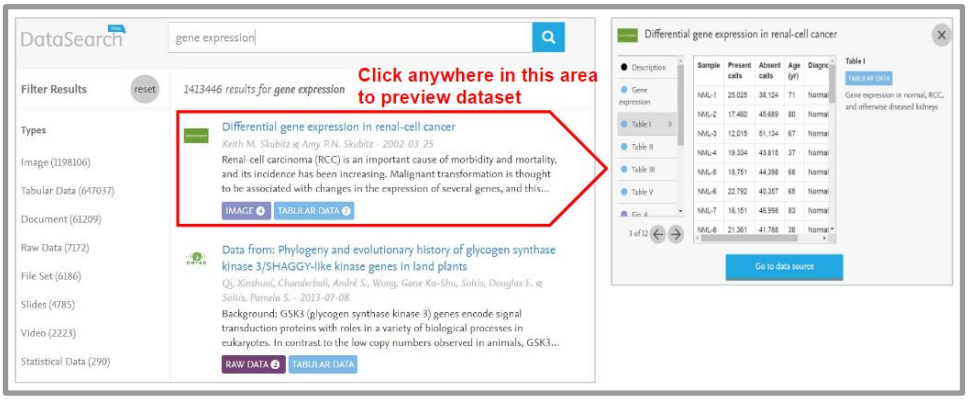

Figure 6. An example of preview data from Elsevier Datasearch[10]

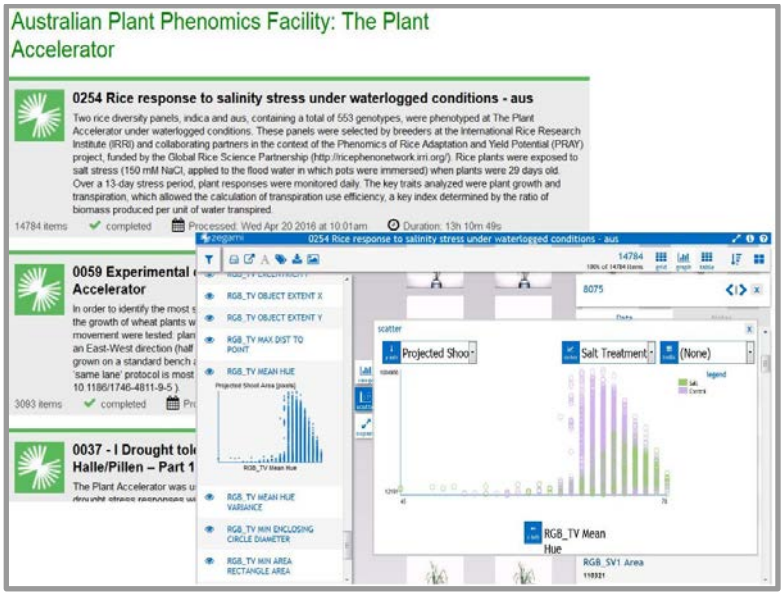

Figure 7. Examples of advanced data preview from Australian Plant Accelerator

- **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 fields clearly and unambiguously, and make actionable 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 method and/or software have been used to collect data, and where data are derived from. Figure 8 shows such an example. This provenance information will help users to

[10] Elsevier Datasearch: https://datasearch.elsevier.com/#/

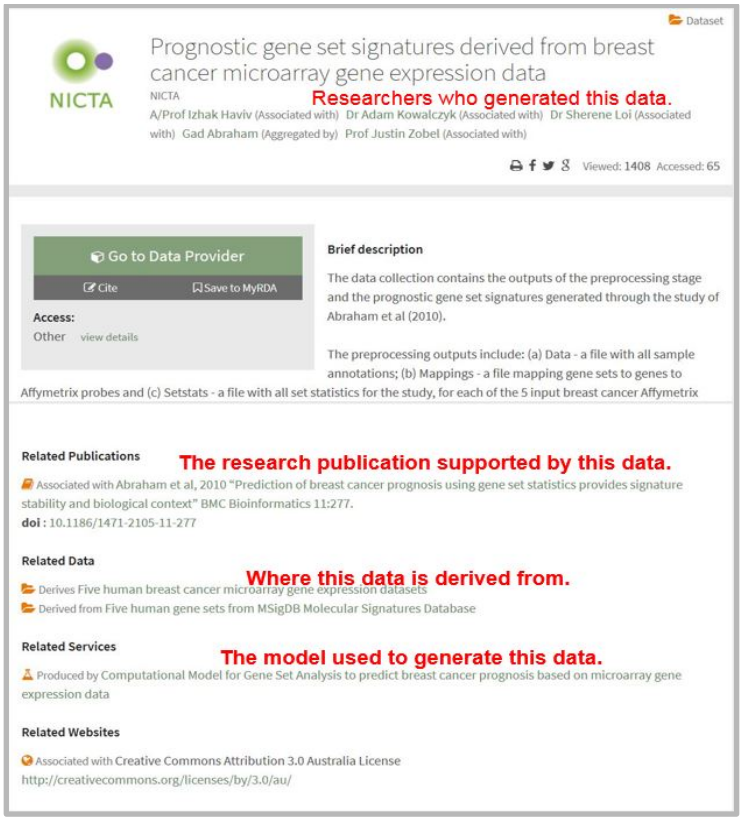

Figure 8. An example of displaying provenance information[11]

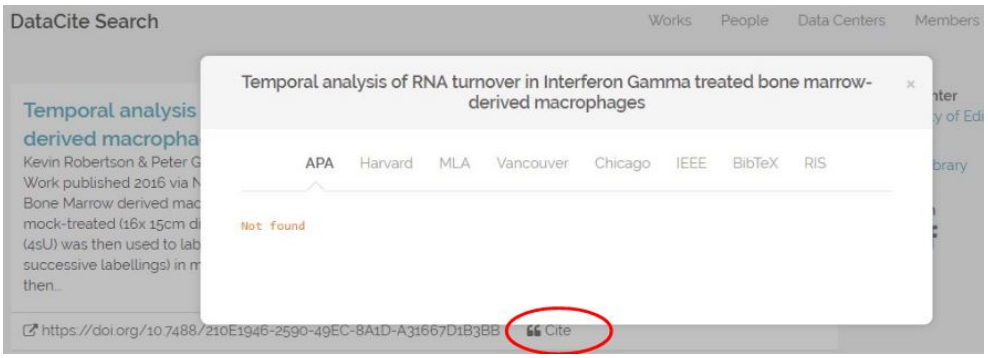

Figure 9 . An example of supporting for data citation (from DataCite)

assess data accountability and ideally for data reproducibility (Wu and Treloar, 2015). [***ref. REQ 2 & 3***].

REC 5. Be able to share and output bibliographic references

It is important that references can be exported to popular formats (e.g. Evernote, Bibtex, etc.) so that a researcher can save this information for later reference or sharing it with his/her colleagues [***ref. REQ 7***]. This feature often comes with other academic portals such as library reference systems and research paper publishers, it is recommend to have this functionality from a data repository as well [***ref. REQ 9***]. Figure 9 shows an example of this feature from DataCite Search. When the link Cite

[11] An example of displaying provenance information: https://researchdata.ands.org.au/11572/

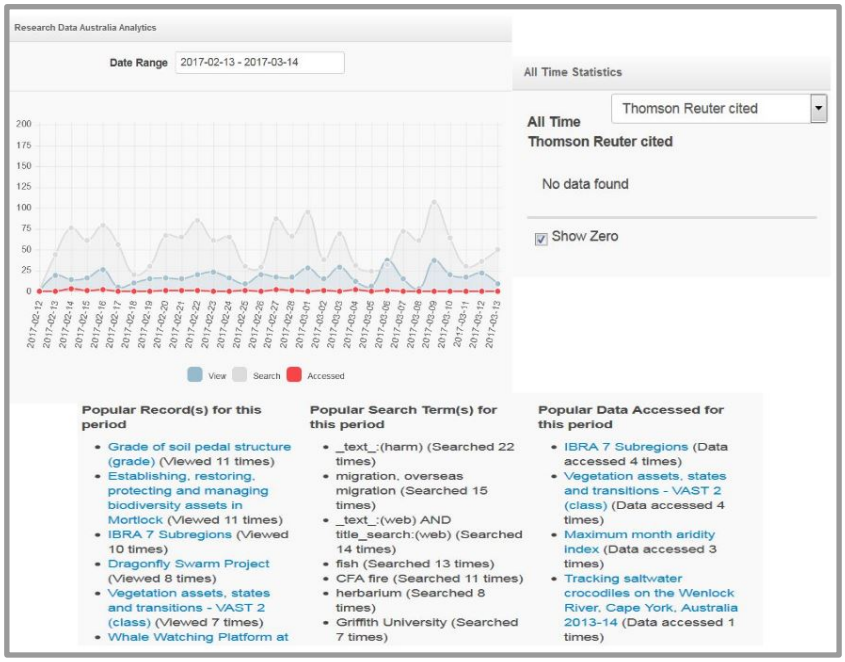

Figure 10. An example of data access analytics aggregated per data source (from Research Data Australia)

is clicked, a floating window opens where a user can choose from a few supported citation format options. Being able to output a data citation in popular publication acceptable styles will encourage data users to cite the data properly, which will make the data even more discoverable.

REC 6. Provide feedback about data usage statistics

This usage statistics includes metrics such as metadata viewed, data viewed, data downloaded, data cited, etc. This information can be useful to different groups of people for different purposes:

- Repository managers may want to see this information to better manage and promote their data, and improve their data discovery services. Figure 10 shows an example of aggregating data access and usage statistics per data source. Repository managers can see the number of records viewed, searched and accessed for a selected time period; most popular records, search terms and data accessed. Other features (not shown in the figure) include the number of data records that have minted DOI's and the number of records that meet different levels of metadata quality.
- Data authors or 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 may be one of the factors to influence if they would use a data collection [***ref. REQ 3***].

REC 7. Be consistent 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 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 sensitive, it should be labelled clearly.
- Secondly, keep consistency 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 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 institutional or domain 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 metadata records describe the same object and the two records replicate each other.
>
> ***Parallel metadata records***: Two records describe the same object but they 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***: A record is regarded as an augment of another if the two records describe the same object, and 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; or make them abandon a search [***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 Ntarmos et al., 2011; Weissman and Ayhan 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's web interfaces as well as by web search engines, as many users search for data through web search engines

like Google. 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 we recommend that repositories:

1. **Make metadata records easily indexed by web search engines.** Web search engines build their indexes by sending their web crawler (or spider) to browse the Web from seed URLs, following link to link recursively, and fetching available web pages that are pointed to. For a repository that generates a metadata page in html via the repository API, it is recommended to have a sitemap[12] that lists unambiguous URLs of landing pages for each data object. Having a persistent identifier such as DOI for a data object will help when it is published to DataCite and/or cited in publications, that will link back to the data object's landing page. This type of linkage gets landing pages a higher chance to be crawled, but also increases their ranking in a search result because of the PageRank algorithm (Page and Brin et al. 1999, Fenner and Crosas et al. 2016).
2. **Make metadata understandable by web search engines**. When a metadata record is fetched and indexed by web search engines, the metadata should also be described in a way that is understandable by web search engines. A Data Citation Roadmap [Fenner and Crosas et al. 2016] recommends to encode 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 may bring two advantages:
    a. It may make a difference in determining ranking of the metadata in a search result as search engines may take into account where a search term occurs, e.g. title or description[13].
    b. If a search engine understands more about a metadata record, it may create rich search features[14] and snippet for users[15].

       Detailed guidance on how to improve the ability of search engines to find and understand datasets are provided on a google developer's website (https://developers.google.com/search/docs/data-types/datasets).

REC 10. Follow API search standards and community adopted vocabularies for interoperability

Interoperability is one of the four FAIR data principles. What language and vocabulary are used to describe data does affect the construction of a search interface and user's mental model of data objects to be searched. Recommendations to data repositories are to use community adopted vocabularies and follow API search standards. For many services that

[12] Learn about sitemaps: https://support.google.com/webmasters/answer/156184?hl=en

[13] How search works: https://www.google.com/intl/en_us/insidesearch/howsearchworks/crawling-indexing.html

[14] About Schema.org: http://schema.org/docs/faq.html?visit_id=1-636249696842738521-3472474488&hl=en&rd=1

[15] Schema.org and Google for Local Discovery: Some Key Takeaways: http://blogs.library.duke.edu/bitstreams/2014/03/27/schema-org-and-google-for-local-discovery-some-key-takeaways/

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[16] or SRU- Search and Retrieval by URL[17] (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).

Requirements vs. Recommendations

Now that we have described how the Requirements were gathered and discussed Recommendations for data repositories in the two sections above, we provide a mapping of how the Recommendations support the Requirements in Table 3. As Table 3 shows each

	REQ1: Data availability	REQ2: Connection of data	REQ3: Annotations	REQ4: Filtering	REQ5: Cross-referencing	REQ6: Inspection of data	REQ7: Collaborative environment	REQ9: Similarity across portals	REQ8: Training material
REC 1: Query interfaces				✓		✓		✓	Ten simple rules for finding data
REC 2: Multiple access points		✓		✓		✓		✓	
REC 3: Summarize search results	✓		✓			✓			
REC 4: Metadata records readable		✓	✓						
REC 5: Bibliographic references							✓		
REC 6: Usage statistics			✓						
REC 7: Consistentcy								✓	
REC 8: Identify duplicates		✓			✓				
REC 9: Findability from web SEs	Support data searches from web search engines								
REC 10: Interoperability	The Fair Data Principles								

Table 3: Mapping between the Requirements and Recommendations.

Requirement is supported by at least one Recommendation, except for the Requirement 8 “Accompanying education/training material”. Although data repositories play a role in satisfying this requirement by providing a “Help” page to describe a set of search operators and how the site is organised, the primary responsibility here may rest with on librarians and research offices. Thus we map this requirement to the “Ten Simple Rules for Data Seekers”[18], which librarians and research offices can present to their users.

In Table 3, the Recommendations 9 and 10 are not directly mapped to any requirements as requirements were inferred from use cases from human users, nevertheless these two

[16] OpenSearch: http://www.opensearch.org
[17] Search and Retrieval via URL: http://www.loc.gov/standards/sru/
[18] Ten simple rules for data seekers: https://docs.google.com/document/d/1SVxYxa052a_NP-W15Z6tg9E8bFKXyARRNmpR0GosMDc/edit?usp=sharing

requirements are important that: the Recommendation 9 addresses a common behaviour that many researchers are using web search engines as their primary tool to search for publication and data, the Recommendation 10 is to support 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 requirement.

Conclusions and Recommendations

We documented and collected a set of use cases related to data discovery. Through analysis of collected use cases, we produce a classification 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 functionalities 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 the Requirement tab of this spreadsheet (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 data repository operators can use the requirements for the following (but not limited to) purposes:

- As a checklist for designing and implementing a data services portal.
  By addressing each one of these requirements, one can ensure that the data services portal will provide the state-of-the-art in the data discovery process, from the users' perspective.
- For existing data discovery services, the list of requirements can be used as a guidelines for heuristic evaluation of a specific data discovery service [Nielsen 95], and therefore plan for future improvements when necessary.
- As an indication of the missing aspects in the current data discovery landscape.
  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.

The requirements address issues related to data, metadata and portal functionality. Clearly, improving data discovery paradigms requires a collective effort by data collectors, data providers, data repositories, data librarians and research trainers to optimize the data discovery process for researchers. Although there is no single best route to building an optimal data discovery portal, we hope that in collecting these use cases and recommendations we have provided some pointers to improve data search features on domain and institutional repositories and portals.

We proposed ten recommendations to address the requirements. 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. We would like to work with data repositories in the future to validate the requirements, evaluate and refine the recommendations.

Acknowledgement:

We would like to thank the RDA DDP IG members, especially Dom Fripp, 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.

References:

Borgman, C. L. (2015), *Big data, little data, no data: Scholarship in the networked world*, *MIT press*.

Dalkıran, Ö., Aker, İ., Öztemiz, S., Taşkın, Z. and Tunç, S. K., "Usability testing of digital libraries: The experience of EPrints", In Procedia - Social and Behavioural Sciences, Volume 147, 2014, Pages 535-543, ISSN 1877-0428, DOI:10.1016/j.sbspro.2014.07.155.

Fenner M, Crosas M, Grethe J, Kennedy D, Hermjakob H, Roca-Serra P, Berjon R, Martone M, Clark T. (2016), " A data citation roadmap for scholarly data repositories", In BioArxiv. DOI: 10.1101/097196.

Gregory, K., Groth, P., Cousijn, H. and Wyatt, S. (2017). "Search data: a review of observational data retrieval practices", Accessed from https://arxiv.org/abs/1707.06937 on 7 Nov. 2017.

Hammond T. (2010), "OpenSearch: A case study in OpenSearch and SRU integration", D-Lib Magazine. Volume 16, Number July/August 2010. DOI:10.1045/july2010-hammond

Hearst M, Elliott A., English J., Sinha Rashmi, Swearingen K, Yee K. (2002), "Finding the flow in Web site search". In Communications of the ACM, Vol 45(9), Sept. 2002, Pages 42-49. DOI:10.1145/567498.567525

Jeng, J. (2008). "Usability of academic digital libraries: Effectiveness, Efficiency, Satisfaction, and Learnability", Libri, 55(2-3), pp. 96-121. Retrieved 11 Nov. 2017, from doi:10.1515/LIBR.2005.96.

Joo, S., & Lee, J. Y. (2011), "Measuring the usability of academic digital libraries: Instrument development and validation", Electronic Library, 29(4), 523–537.

Kern, D. and Mathiak, B. (2015), "Are there any differences in data set retrieval compared to well-known literature retrieval?", In S. Kapidakis et al. (Eds.): Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, LNCS 9316, pp.197-208, DOI:10.1007/978-3-319-24592-8_15

Kim, J. Y., Field, H. and Cartright, M. A. (2012), "Understand book search Behaviour on the Web", In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM'12)*, pp. 744-753. Oct 29-Nov. 2, 2012, Maui, HI, USA.

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), New York, NY, USA, pp.433-442. DOI: https://doi.org/10.1145/2063576.2063643

Kong, W. and Allan J. (2014), “Extending faceted search to the general Web”, Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM), Nov. 2014, pp.839-848. doi:10.1145/2661829.2661964.

Lóscio, B. F., Burle, C. and Calegari, N. (editors) (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.

Mihajlovic V., Hiemstra D., Blok H.E., Apers P.M.G. (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 B. M. (2010), “Beyond text querying and ranking list: How people are searching through faceted catalogues in two library environments”, In Proceedings of the Association for Information Science and Technology, Vol 47(1). November/December 2010, pp.1-9.

Pampel, H. and Vierkant, P. (2015), “Current status and future plans of re3data.org - Registry of research data repositories”, In: Wagner J, Elger K, (editors). GeoBerlin2015: Dynamic Earth from Alfred Wegener to today and beyond; Abstracts, Annual Meeting of DGGV and DMG. Berlin, Germany; 2015. p. 287—288. Available from: http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1369620.

Page L, Brin S, Motwani R. and Winograd T. (1999), “The PageRank Citation Ranking: Bringing Order to the Web”, Technical Report from Stanford InfoLab, available at: http://ilpubs.stanford.edu:8090/422/.

Sanderson, M. and Croft, W.B. (2012), “The history of information retrieval research”, In: Proceedings of the IEEE 100. Special Centenial Issue, pp. 1444 –1451. Issn:0018-9219. doi:10.1109/JPROC.2012.2189916.

Sainsbury G., Berger B., Garnett T., and Conway S. (2017), “ Zegami: A visual discovery and analytics platform”, Presented at Monthly TechTalk, March 2017. Slides is available at https://drive.google.com/file/d/0B3urY1fwzAEcWmNGTnpsUjVUQWM/view

Shearer, K., Rodrigues, E., Walk, P. and Perakakis, P. (2016), “Next Generation Repositories” Accessed at 1st Dec. 2017, from Zenodo. DOI:10.5281/zenodo.163264

Spärck-Jones, K. (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, S. K. (Eds.). ACM, New York, NY, USA, Article 1, 7 pages. DOI=http://dx.doi.org/10.1145/1364742.1364743

Spink S., Wolfram D., Jansen, B. J., and Saracevic T. (2001). “Searching the Web: The public and their queries”, Journal of the American Society for Information Science, 52(3):226-234.

Theng Y.L., Duncker E., Mohd-Nasir N., Buchanan G., Thimbleby H. (1999), “Design guidelines and user-centred digital libraries”, In: Abiteboul S., Vercoustre AM. (eds) Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science, VOL1696. Springer, Berlin, Heidelberg.

Tsakonas, G., & Papatheodorou, C. (2008), “Exploring usefulness and usability in the evaluation of open access digital libraries”, Information Processing and Management, 44: 1234-1250.

Turpin A., Scholer F., Jarvelin K., Wu M. and Culpepper J. S. (2009), “Including Summaries in System Evaluations”, Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, July 2009, pp.508-515.

Weissman, S., Ayhan, S., Bradley, J. and Lin, J. (2015), “Duplicate and contradictory information in Wikipedia”, In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15). ACM, New York, NY, USA, 57-60. DOI:10.1145/2756406.2756947.

Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. (2016), “The FAIR Guiding Principles for scientific data management and stewardship”, Sci. Data 3:160018 doi: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). ACM, New York, NY, USA, 375-381. DOI:10.1145/383952.384028.

Wu, M., Turpin, A., Puglisi, S. J., Scholer, F. and Thom, J. A. (2010), “Presenting query aspects to support exploratory search”, Proceedings of the Eleventh Australasian Conference on User Interface, Vol. 106, pp.23-32.

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

de Waard, A., Khalsa, SJ, Psomopoulos, F. and Wu, M. (2017). “RDA IG Data Discovery Paradigms IG: Use Cases data [Data set]”. Zenodo. DOI:10.5281/zenodo.1050976.

Xie, H. I. (2008), “Users’ evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment”, In Information Processing & Management, VOL 44, Issue 3, 2008, pp.1346-1373, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2007.10.003.