RESEARCH DATA ALLIANCE

Data Discovery Paradigms Interest Group

April 5, 2017

RDA 9th Plenary Meeting, Barcelona, Spain

research data sharing without barriers

rd-alliance.org

Anita de Waard,
Siri Jodha Singh Khalsa
Fotis Psomopoulis
Mingfang Wu

Charter Data Discovery Paradigms Interest Group:

- Motivation:
  - For data to be Findable, we need a data infrastructure that supports users in discovering research data regardless of the manner in which it is stored, described and exposed.
  - This interest group aims to explore common elements and shared issues that those who search for data, and who build systems that enable data search, share.
- Use cases:
  - Builders of data search engines are interested in developing components and practices to connect resources and results
  - Data repositories are interested in improving and expanding search on their platforms
  - Users are interested in better interfaces and fewer places to look for data
  - Data creators are interested in a shared set of data metrics for all search engines
  - Data search builders are interested in sharing knowledge and tools about ranking, relevance and content enrichment.
- Goals:
  - Provide a forum where representatives across the spectrum of stakeholders and roles pertaining to data search can discuss issues related to improving data discovery.
  - Identify concrete deliverables such as a registry of data search engines, common test datasets, usage metrics, and a collection of use cases and competency questions.

Timeline:

- Apr 16 (RDAP7): Held BoF on Datasearch, planned IG
- Sep 16 (RDAP8): Held kickoff meeting at RDA 8: established topics (long list, to be narrowed down)
- Oct 16: Established web presence, mailing list, did poll of potential Task Force topics
- Dec 16: Identified set of Task Forces & got to work!
- Mar 17: Preliminary Task Force Outputs Distributed
- **Apr 17 (RDAP9): Discuss outputs Task Forces, plan next steps and new Task Forces.**
- Sep17 (RDAP10): Summarize & Distribute Outputs, possible joint session, discuss testbed.

Ranking of Topics From Survey:

Topic	Nr Points	Rank
Guidelines for making data findable	194	1
Use cases, prototyping tools and test collections	263	2
Metadata enrichment	232	2
Relevancy ranking	255	3
Cataloging common API's	255	3
Data Citation practices and metrics	272	4
Granularity, domain-specific cross-domain issues	312	5
De-duplication of search results	293	5
Using upper-level ontologies	320	6
Search personalisation	348	7

Agenda Today:

16:00- 16:10: Introduction of the group's goals and progress (= *this!)*

16:10 -16:55: Overview of each of the 3 active Task Forces:

- Relevancy Ranking Task Force
- Use Cases, Prototyping Tools and Test Collections Task Force
- Best Practices for Making Data Findable Task Force

17:00 Discuss new Task Forces and Coordination :

- Volunteers to lead other top-5: Common API's, Metadata Enrichment?
- What other topics to start?
- Do we want to turn Task Forces into Working Groups?
- Collaborations with other Working/Interest Groups?

Data Discovery Paradigms IG

Relevancy Ranking Task Force

research data sharing without barriers

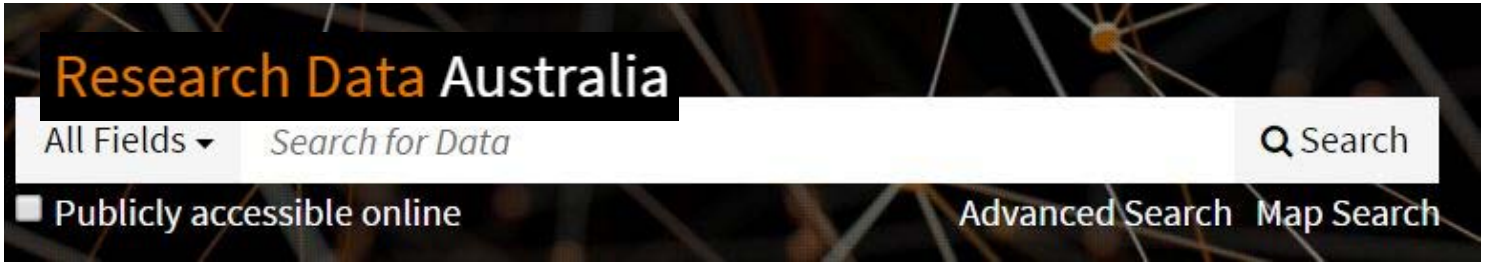
rd-alliance.org

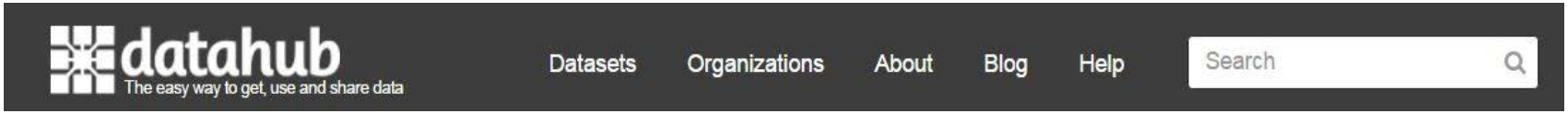
The FAIR Guiding Principles

- **F**indable
  - F1. (meta)data are assigned a globally unique and persistent identifier
  - F2. data are described with rich metadata
  - F3. metadata clearly and explicitly include the identifier of the data it describes
  - F4. (meta)data are registered or indexed in a searchable resource
- **A**ccessible
- **I**nteroperable
- **R**eusable

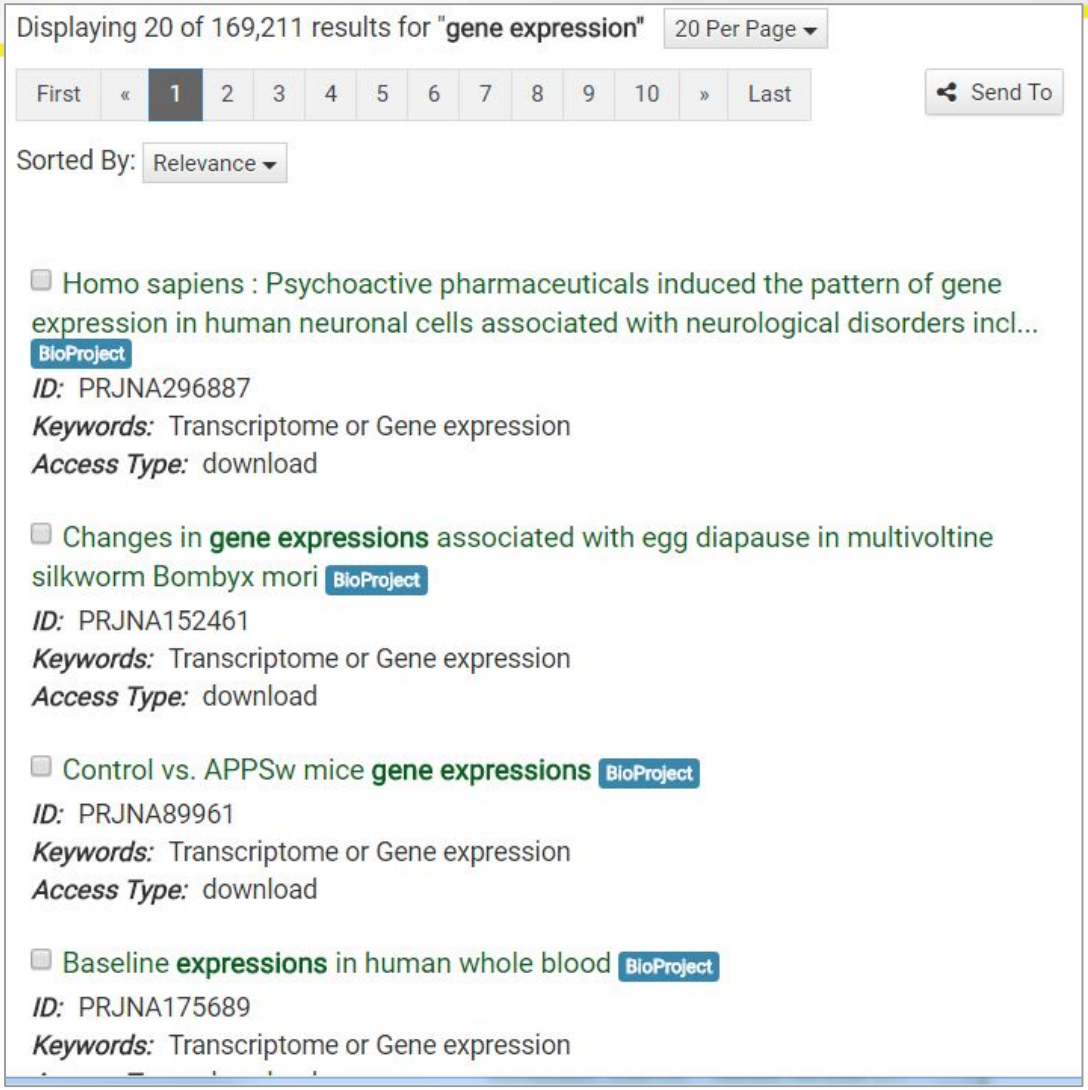
Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi:10.1038/sdata.2016.18 (2016).

Relevancy Ranking - from a search

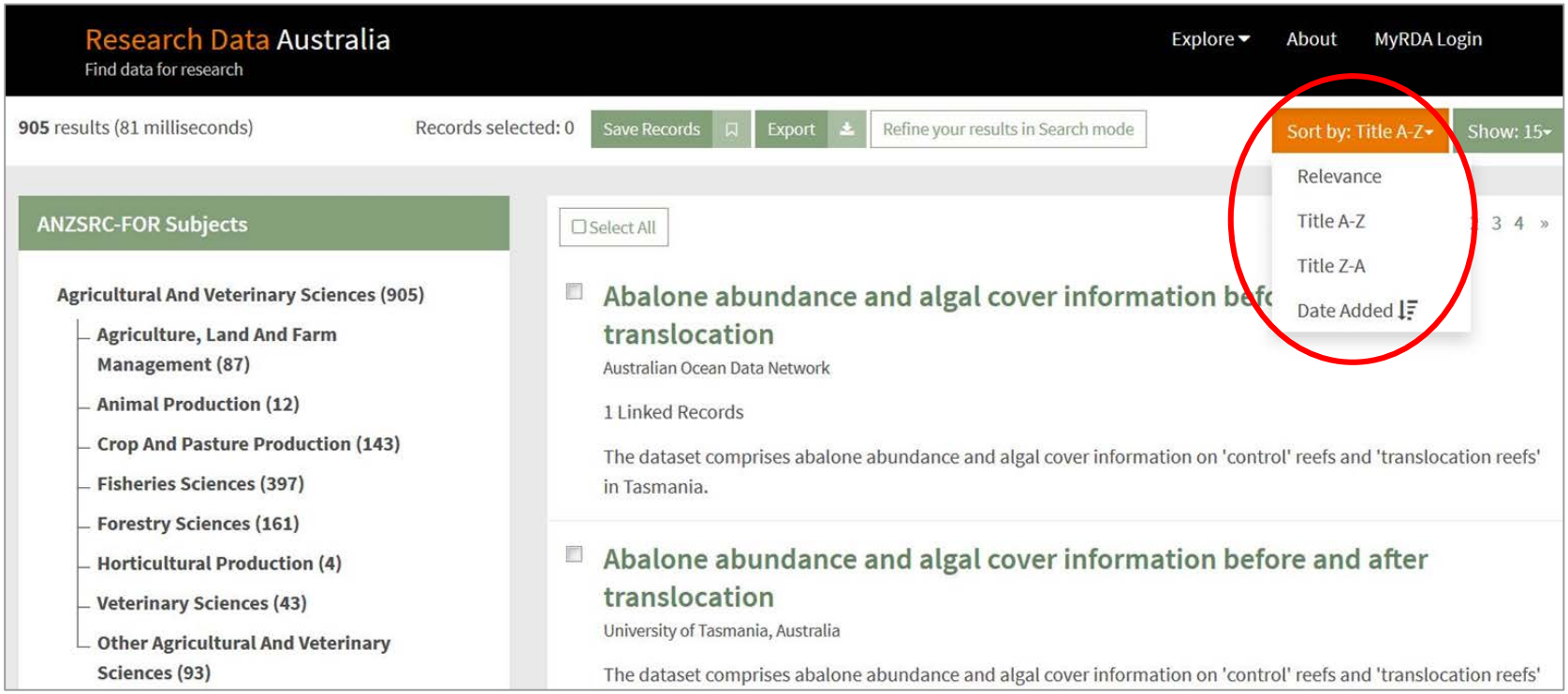




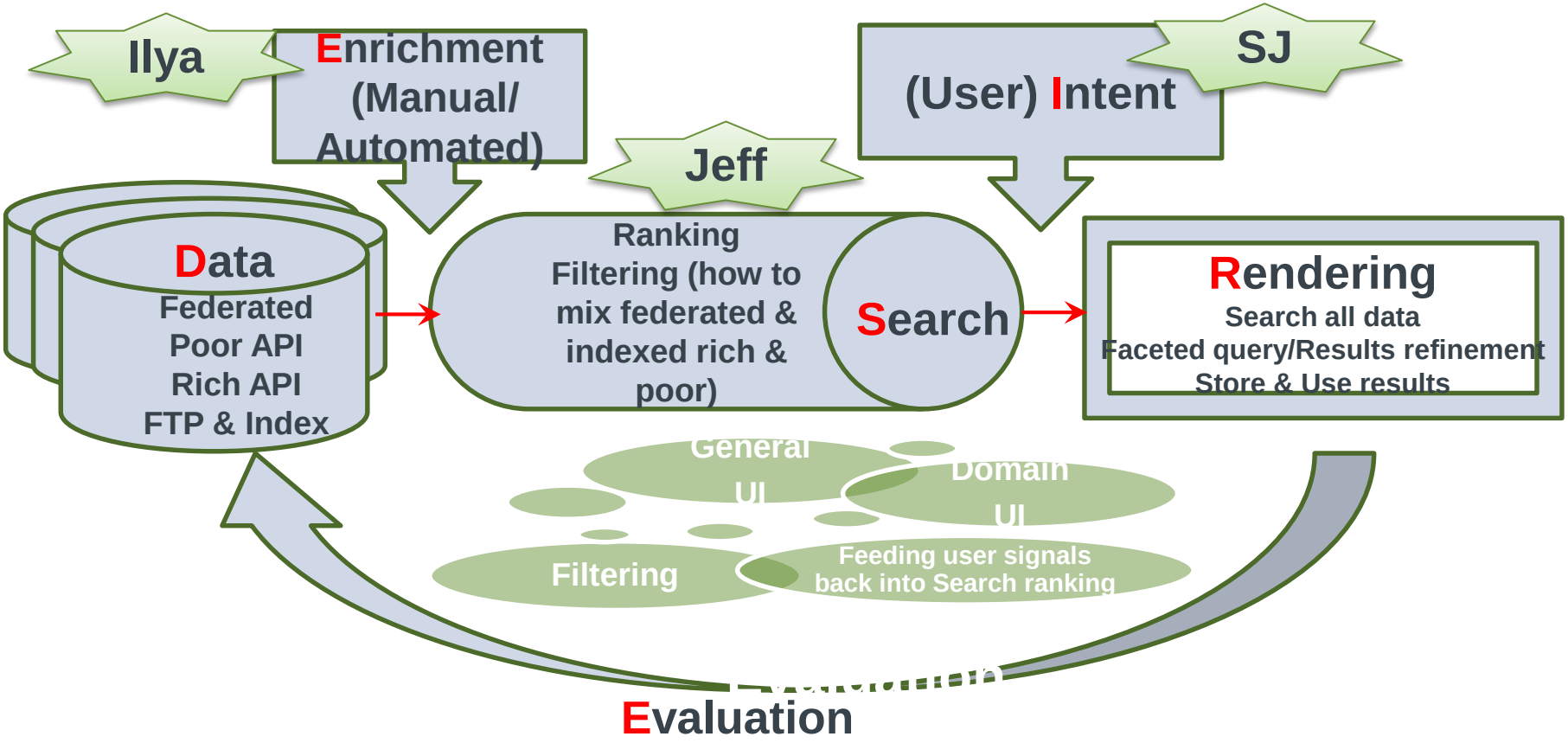
Relevancy Ranking - from a search



Relevancy Ranking - From a facet browsing



The DESIRE Model of Data Discovery (P8)



Initiatives on search problems

Text Document

Text REtrieval Conference (TREC) (1992 - date)

- OpenSearch
- Precision Medicine
- Complex Answer Retrieval
- Contextual Suggestion
- Federated Web Search
- Microblog
- Novelty Search
- Web Search Interactive Search

Image

CLEF:Information Access Evaluation meets Multilinguality, Multimodality,and Interaction (2010 - date)

- imageCLEF
- lifeCLEF
- eHealth
- Question Answering

Data

BioCADDIE 2016 Dataset Retrieval Challenge

Text Document in XML Markup

INEX: Initiative for the Evaluation of XML Retrieval (2002-2014)

- Social Book Search
- Linked Data
- Snippet Retrieval
- Relevance Feedback

Goals

- Provide a means or forum for data search community to share experiences on relevancy ranking.
- Help people choose appropriate technologies when implementing or improving search functionality at their repositories.
- Capture the aspirations, successes and challenges encountered from repository managers.

Progress

- Identify issues on search ranking from within a repository and evaluation methods.
- Identify current practices in relevancy ranking for data search through a survey questionnaire (draft). (Please provide your feedback by 24th April.)
- Explore possible testbeds to address data search challenges, some possibilities may include:
  - Elsevier can provide AWS EC2 instances for a relevancy test bed. The Elsevier team could probably clone the machines that they used during the recent bioCADDIE Challenge.
  - ANDS can provide a corpus of metadata from the Research Data Australia repository.

Future Activities

- Conduct the survey, analyse and share survey result.
- Identify potential collaborative projects from the survey.
- Prioritise and coordinate activities from the survey, for example, compare common ranking models.

Aspiration

- Build test collections with real world data search tasks for data search community to work on.

Contributors

Peter Cotroneo (Elsevier)
Beth Huffer (LLC)
Siri Jodha Khalsa (U. Colo.)
Jens Klump (CSIRO)
Dawei Lin (DAIT, NIAID/NIH)
Anita de Waard (Elsevier)
Mingfang Wu (ANDS)

We welcome more participants to the task force!

All materials from the Relevance Ranking Task Force are available from this Wiki page at the RDA site.

Data Discovery IG – Task Force E

Use Cases, Prototyping Tools and Test Collections

research data sharing without barriers

rd-alliance.org

Goals and Aims of the Task Force

Primary goal

identify the key requirements evident across data discovery use-cases from various scientific fields and domains

Why?

- improve the provided data discovery services
- understand the continuously evolving methods of data discovery employed by the end-users

Particular Objectives:

1. Identify the questions / aspects necessary to capture use-cases / user scenarios
2. Perform a survey aiming for a wide audience, across disciplines / domains.
3. Organize the information gathered in order to identify common aspects / categories / clusters
4. Extract user-profiles, and therefore user requirements, from the use-cases.

Capturing use cases

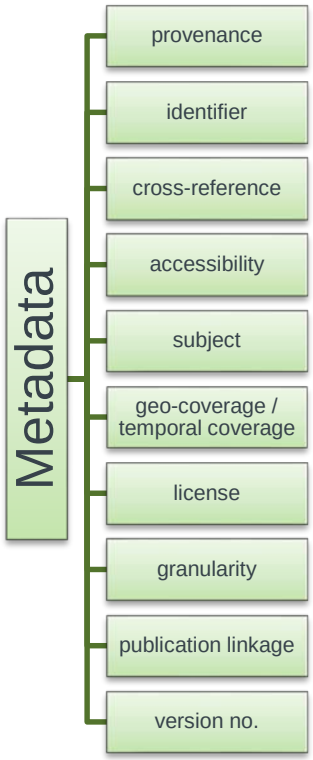
- Overall gathered >60 use-cases

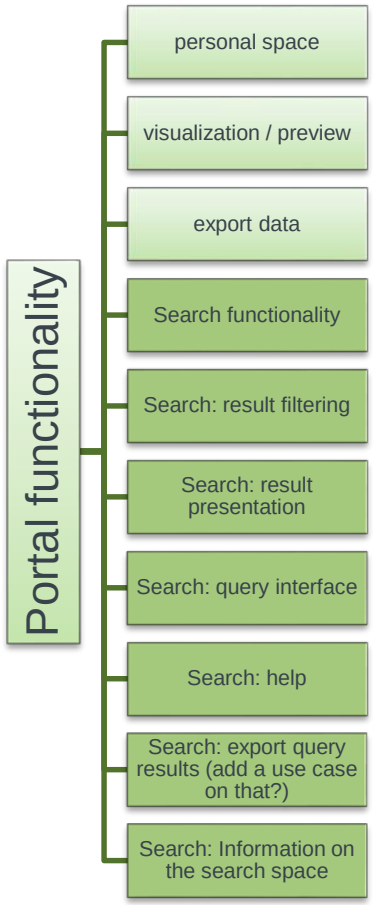
- Structure
  - "*As a*" (i.e. role)
  - "*Theme*" (i.e. scientific domain/discipline)
  - "*I want*" (i.e. requirement, missing feature, supported function)
  - *"So that*" (i.e. the user need that is addressed)
  - "*Comments*"

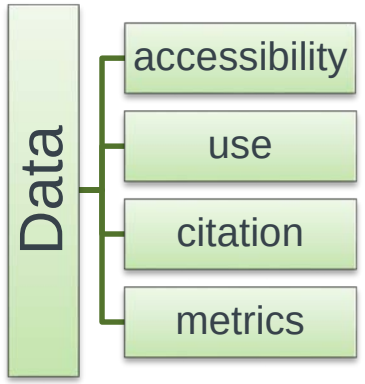
- Major Sources
  - ✓ UK Research Data Discovery Service use cases
  - ✓ User stories as purposed for the agile methodology
  - ✓ Falling Water User Interview Responses
  - ✓ BioCADDIE
  - ✓ Spatial Data on the Web

Categories

- Parsing through the use-cases, identified the following major categories







Collected Requirements – Part 1/2

User Type	Requirement	Comment	Use Cases
Researcher / PhD Student	**Indication of Data availability**	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 potentail data embargo. Ideally should have an evident big button for "Download"	4, 9, 14, 24, 28
Funder / Researcher / PhD Student	**Connection of Dataset with person / institution / paper / citations**	This allows for ranking of datasets, personalization of information displayed as well as accountability. Also this information can be used for grant application as well as for comparative studies (datasets across papers). Finally, allow for the upload of manuscript for direct connection.	19, 20, 21, 22, 23, 25, 27, 33, 42, 47, 51, 61
Librarian	**Educational / training material**	Help researchers manage and discover data in a methodical and seamless manner	62,
Researcher / PhD Student / Master Student	**Filtering of Datasets based on specific criteria on multiple fields at the same time (such a release date, geo coverage, date range, specific events). Ideally filtering based on gradual refinement of query. Also support of custom tags.**	Support targeted studies (e.g. find global temperature records for volcanic eruptions in the last century; find articles on bronze age in Britain)	3, 6, 8, 10, 16, 18, 24, 29, 32, 36, 39, 40, 52, 56, 64, 65, 66

Collected Requirements – Part 2/2

User Type	Requirement	Comment	Use Cases
Researcher / PhD Student / Master Student	**Full annotation on dataset (including granularity, origin, licencing, provenance, and method of production, times downloaded).**	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 eg 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.	11, 17, 30, 31, 37, 38, 41, 46, 48, 49, 53, 54, 58, 59
Researcher / PhD Student	**Cross-referencing across datasets (same or different repositories). This can be provided as a single entry point for reference data (i.e. a meta-repository)**	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	1, 2, 54,
Researcher / PhD Student	**Sharing data (either whole dataset or particular records) in a collaborative environment**	Need to rerun at the last minute to check nothing has been published since last study/search.	5, 7, 35,
Researcher	**Provide visual analytics / inspection of data / thumbnail preview**	Decide if this data set is right for a research purpose	13, 15, 43, 44

Future Steps

- Some tasks are completed but:

- ❑ Transform requirements to recommendations
- ❑ Identify the best fitted use-cases to generate scenarios for testing against a testbed

- ❑ Explore prototyping tools for data discovery
- ❑ Investigate additional user roles within the context of the involved use-cases

Data Discovery IG – Best Practices

research data sharing without barriers

rd-alliance.org

Goals of the Task Force

Explore current practices of making data finable and recommend best practices to the data community

Three key perspectives:

1) Data producers/data contributors
   - https://docs.google.com/document/d/1V-1z8jFlEoaxeNgX6NsSslEYvkIqIzdYQyYqEGmSyms/edit
2) Data repositories/registries
   - https://docs.google.com/document/d/1qCI0ZPVknibPGJGP3OI1xn_lvtqfwlM6tFFL1aw0DBA/edit
3) Data seekers/consumers
   - https://docs.google.com/document/d/130e2k5RJDUOk14T8VwPD6qMLmRmdRoTzKbF_QW6-8Ks/edit#heading=h.terqcqv1fg7

Data Producers/Contributors

Bill Michener, Natalia Atkins, Amy Nurnberger, Mingfang Wu

- Data should be **preserved and accessible** through a certified or community-recognized data repository.
- **Persistent identifiers**, such as DOIs, should be assigned so that the user can discover and acquire specific data products.
- Data should be understandable via **comprehensive metadata** that adhere to community-accepted metadata standards.
- Data and metadata should be **checked for quality** via appropriate quality assurance and quality control methods that are described in the metadata.
- Data discoverability can be facilitated through the addition of **descriptive keywords and text** (e.g., abstract, project description) to appropriate metadata elements.
- Data and metadata **file names** should be clearly, descriptively and uniquely labeled.
- Provide clear guidance on how to **cite and acknowledge** the data contributor(s).
- Provide **ORCiDs** for the data contributors so that they can be disambiguated and be properly acknowledged.
- **Offer guidance on data usage**—i.e., how the data have been used and, if possible, could reasonably be used in the future.
- Ascribe a **data sharing and usage license** to the data—providing potential users with a simple and standardized way to understand how your data cab be shared and used based on conditions that you choose.

Data Repositories/Registries

Mingfang Wu, Siri Jodha Khalsa, Martin Fenner

- Provide a **range of query interfaces** (e.g., simple search, advanced search, map search) to accommodate various data search behaviors.
- Provide data seekers **multiple access points** to find data (e.g. search, subject browse, faceted browse/filtering).
- Make it easier for researchers to **judge relevance** of a data collection, e.g. by highlighting query terms, describing data accessibility, providing data previews, showing data coverage
- Make records easily **indexed and searchable** by major web search engines.
- Be able to link to (and display) information/resources related to **data creation and usage** (e.g. paper publication, provenance information such as models, software, instruments, projects, researchers, derived data, data derived from, creators etc.) if possible.
- Be able to **output a bibliographic reference in popular formats** (e.g. Evernote, Bibtex, etc.) so that a researcher can save this information for later reference or share it with colleagues.
- Provide feedback about **data usage statistics** (e.g. metadata viewed, data viewed, data downloaded, data cited if possible, etc.)
- Follow search **API standards and community adopted vocabulary** to enable interoperability with other search services.
- Make data **searchable from different platforms**, e.g. computer, iPad or smartphone.

Data Seekers/Consumers

Siri Jodha Khalsa, Mingfang Wu, Bill Michener, Anita de Waard

- **Resource selection**: Search with a domain-specific portal or aggregator before using a general-purpose search engine (Google, Bing, etc.).
- **Query construction**: Check “search help” page from data portal for how to construct advanced queries. The more a search context is described, the more likely relevant data will be retrieved and ranked high.
- **Refine search**: Use facets and query recommendations to broaden or narrow down a search.
- **Track new and evolving data**: Make use of search APIs to monitor latest published data in a field or data of dynamic nature.
- **Acknowledge the producer**: Give back to the data producer/contributor by citing data properly.

Next Steps

- Review and revise the three sets of practices
- Merge into a paper that can be submitted to an appropriate journal

Agenda Today:

16:00- 16:10: Introduction of the group's goals and progress (= *this!)*

16:10 -16:55: Overview of each of the 3 active Task Forces:

- Relevancy Ranking Task Force
- Use Cases, Prototyping Tools and Test Collections Task Force
- Best Practices for Making Data Findable Task Force

17:00 Discuss new Task Forces and Coordination :

- Volunteers to lead other top-5: Common API's, Metadata Enrichment?
- What other topics to start?
- Do we want to turn Task Forces into Working Groups?
- Collaborations with other Working/Interest Groups?

Ranking of Topics From Survey:

Topic	Nr Points	Rank
Guidelines for making data findable	194	1
Use cases, prototyping tools and test collections	263	2
Metadata enrichment	232	2
Relevancy ranking	255	3
Cataloging common API's	255	3
Data Citation practices and metrics	272	4
Granularity, domain-specific cross-domain issues	312	5
De-duplication of search results	293	5
Using upper-level ontologies	320	6
Search personalisation	348	7

Long List of Topics at RDAP8:

1. Deduplication and cross-repository issues
2. Identifiers and how they help in search
3. Data citation: how do we access/use?
4. *Relevancy ranking for structured data?*
5. Enrichment tools for faceting and ranking
6. Domain-specific vs. generic issues: interfaces and enrichment
7. Different discovery platforms for Open Search, science-focused OS profile?
8. Metadata standards to enhance data discovery, e.g. schema.org and such
9. Models and methods of personalization
10. *Identify core elements of Findability*
11. Automated integration of records; granularity and findability
12. *Common APIs (e.g. OpenSearch)*
13. Upper-level ontologies for search
14. *Creating test collections for search evaluation and methods of evaluation*
15. Collections and granules: build tool that enables guidance for data submitters on how data is organized
16. *Guidelines for making your data findable! Best practices based on experiences.*
17. *Identify collections of use cases for users: e.g. browsing vs search*
18. Measures of data quality: and impact of findability
19. Define series of reference datasets – can be used to do these metrics
20. Identify list of prototyping tools, use by WG!
21. Cross over between domains: how to enable cross-walk between domains
22. “Return to the semantic”: schema has been populated by crowdsourcing rather than 1 researcher.
23. Implementing schema.org as it exists! How does it apply to science?