Data Discovery Paradigms Interest Group
September 19, 2017
RDA 10th Plenary Meeting, Montreal, QC
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:
▪ 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 engine 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, 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.
- Sep 17 (RDAP10): Present outputs & status from Task Forces, plan next steps.
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. Creating test collections for search evaluation and methods of evaluation
13. Upper-level ontologies for search
14. Common APIs (e.g. OpenSearch)
15. Collections and granules: build tool that enables guidance for data submitters on how data is organized
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. Cross over between domains: how to enable cross-walk between domains
21. “Return to the semantic”: schema has been populated by crowdsourcing rather than 1 researcher.
22. Implementing schema.org as it exists! How does it apply to science?
23. Identify list of prototyping tools, use by WG!
<table>
<thead>
<tr>
<th>Topic</th>
<th>Nr Points</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guidelines for making data findable</td>
<td>194</td>
<td>1</td>
</tr>
<tr>
<td>Use cases, prototyping tools and test collections</td>
<td>263</td>
<td>2</td>
</tr>
<tr>
<td>Metadata enrichment</td>
<td>232</td>
<td>2</td>
</tr>
<tr>
<td><strong>Relevancy ranking</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cataloging common API's</td>
<td>255</td>
<td>3</td>
</tr>
<tr>
<td>Data Citation practices and metrics</td>
<td>272</td>
<td>4</td>
</tr>
<tr>
<td>Granularity, domain-specific cross-domain issues</td>
<td>312</td>
<td>5</td>
</tr>
<tr>
<td>De-duplication of search results</td>
<td>293</td>
<td>5</td>
</tr>
<tr>
<td>Using upper-level ontologies</td>
<td>320</td>
<td>6</td>
</tr>
<tr>
<td>Search personalisation</td>
<td>348</td>
<td>7</td>
</tr>
</tbody>
</table>
1. Use Cases, Prototyping Tools and Test Collections:
   ▪ Identify a set of common data search use cases, leading to a set of requirements
   ▪ Meant to be useable by all data discovery services

2. Best Practices for Making Data Findable
   ▪ Three key actors: Data Provider, Data Seeker, Data Repositories

3. Relevancy Ranking:
   ▪ Choose appropriate technologies for search functionality
   ▪ Sharing experiences with relevancy ranking.

4. Metadata Enrichment:
   ▪ Map search improvements to metadata requirements
   ▪ Document the value of enriched metadata for improving search

4 Active Task Forces:
1. Goals of group and progress (= *this!*)
2. Overview of 4 active Task Forces:
   1. Use Cases, Prototyping Tools and Test Collections
   2. Best Practices for Making Data Findable
   3. Relevancy Ranking
   4. Metadata Enrichment
3. Discuss work of Task Forces:
   ▪ Shall we close off work on some of these Task Forces?
   ▪ What new task forces should we start?
4. Discuss overlap and synergies with other WG/IGs:
   ▪ Peter McQuilton (BioSharing/FAIR sharing IG)
   ▪ Collaborations with other Working/Interest Groups?
6. Next steps.
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 user requirements, from the use-cases.
There are several rich sources of use cases available
- different organizations
- using their own surveys or interviews
- in the context of improving their own data search services

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

We adapted these use cases into a single framework/schema:
- “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”
Heavy bias in the use-cases towards the “Researcher” role, i.e.:
  - Academics, Researchers, PhD/Master students

Distributed a survey using the same framework (i.e. “As a”, “I want”, etc), specifically targeting groups beyond the “Researcher” role:
  - Funders
  - Librarians

16 additional use-cases captured

82 use-cases in total
Manually identified the discrete roles in the use-cases
  e.g. “Researcher” comes in various forms: Professor, Principal Scientist, Early Career Researcher, Student(PhD/Master).

Categorize across two dimensions
  1. The implicit data issues stemming from each use case,
  2. The actor/audience that should take responsibility to address a data issue

The data issue tags resulted in 24 vocabularies
  Tags classified into three major categories: Metadata, Portal Functionality and Data, with tags as sub-categories.
Analyzing Use-cases (2/2)

Metadata (24)
- provenance (5)
- geo-coverage / temporal coverage (3)
- accessibility (2)
- publication linkage (2)
- identifier (1)
- cross-reference (1)
- subject (1)
- license (1)
- granularity (1)
- version no. (1)

Portal functionality (43)
- personal space (2)
- visualization / preview (2)
- export data (1)
- Search functionality (20)
- Search: result presentation (7)
- Search: result filtering (4)
- Search: query interface (3)
- Search: help (1)
- Search: export query results (1)
- Search: Information on the search space (1)

Data (6)
- accessibility (3)
- use (1)
- citation (1)
- metrics (1)
Extracting Requirements

- Capture the user perspective in the data discovery process
- Grouped all use cases based on the context of the “I want” field
  - i.e. the specific data discovery need that is not being currently met
- After extracting requirements, circulated second survey for ranking
  - 31 responses, ranking each requirement individually in the scale of 1 to 5
Ranked requirements

- **REQ1**: Indication of data availability
- **REQ2**: Connection of data with person / institution / paper / citations / grants
- **REQ3**: Fully annotated data
- **REQ4**: Filtering of data based on multiple fields at the same time
- **REQ5**: Cross-referencing of data
- **REQ6**: Visual analytics / inspection of data / thumbnail preview
- **REQ7**: Sharing data in a collaborative environment
- **REQ8**: Accompanying educational / training material
- **REQ9**: Portal functionality similar to that of other established academic portals
Final Task Force Outputs

✓ Use-cases in a Google Spreadsheet, formatted for further analysis
✓ Document outlining the work done and the key outcomes
✓ White paper / Article connecting Use-Cases with Recommendation for Repositories (in progress)
Data Discovery Paradigms IG

Best Practices Task Force

William Michener, Mingfang Wu

research data sharing without barriers
rd-alliance.org
Goals of the Task Force

Primary goal

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

To whom?

- Data Provider: is a type of agent responsible for the creation and/or dissemination, accessibility of data to a consumer
- Data Repository: provides a service for human and machine to make data discoverable/searchable through collection(s) of metadata
- Data Seeker: searches for data to satisfy a need for data
Where are recommendations from?

Data search requirements:
- Use cases & requirements
- User search behaviors
- The FAIR data principles

Current practices:
- Scan existing data repositories
- W3C recommended Data on the Web Best Practices
- Use experience or usability principles
- Literature
Ten Recommendations

- REC 1: Provide a range of query interfaces to accommodate various data search behaviors. *(REQ 4, REQ 6, REQ 9)*
- REC 2: Provide multiple access points to find data (e.g. search, subject browse, faceted browse/filtering). *(REQ 2, REQ 4, REQ 6, REQ 9)*
- REC 3: Make it easier for researchers to judge relevance, accessibility and reusability of a data collection. *(REQ 1, REQ 3, REQ 6)*
- REC 4: Make individual metadata records readable and analysable. *(REQ 2, REQ 3)*
- REC 5: Be able to output bibliographic references. *(REQ 7)*
- REC 6: Provide feedback about data usage statistics. *(REQ 3)*
- REC 7: Be consistent with other repositories. *(REQ 9)*
- REC 8: Identify and aggregate records that describe the same data object. *(REQ 2, REQ 5)*
- REC 9: Make records easily indexed and searchable by major web search engines *(Make data searchable to web search engines)*
- REC 10: Follow API search standards and community adopted vocabularies. *(The FAIR data principles - interoperability)*
Mapping between the REQ, the REC and the Ten Rules

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>REC 1</strong>: Query interfaces</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 2</strong>: Multiple access points</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 3</strong>: Summarize search results</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 4</strong>: Metadata records readable</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 5</strong>: Bibliographic references</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 6</strong>: Usage statistics</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 7</strong>: Consistency</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 8</strong>: Identify duplicates</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>REC 10</strong>: Interoperability</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Final Task Force Outputs

- White paper / Article connecting Use-Cases/Requirements with Recommendations for Data Repositories *(in progress)*

- Ten simple rules for finding research data *(close to finish)*
Data Discovery Paradigms IG

Ten Simple Rules for Finding Research Data

K. Gregory, S.J. Khalsa, W. Michener, A. de Waard, M. Wu

research data sharing without barriers

rd-alliance.org
Best Practices for Making Data Findable TF created teams to develop
- Best Practices for Data Providers
- Best Practices for Data Repositories
- Best Practices for Data Seekers

The latter produced *Ten Simple Rules for Finding Research Data*
- Authored by Kathleen Gregory, Siri Jodha Khalsa, Bill Michener, Fotis Psomopoulos, Anita de Waard, and Mingfang Wu
- Intended for submission to PLOS
1. Think about the data you need and why you need them.
2. Select the most appropriate resource.
3. Construct your query.
4. Make the repository work for you.
5. Refine your search.
6. Assess data relevance and fitness-for-use.
7. Save your search and data source details.
8. Look for data services, not just data.
9. Monitor the latest published data.
10. Give back.
For Details

- Visit our demo at the coffee breaks
  - Wednesday, 20 September
  - Will feature outputs from each DDP Task Force

- Documents linked from our P10 page
  - https://www.rd-alliance.org/ig-data-discovery-paradigms-rda-10th-plenary-meeting
Help data repositories choose appropriate technologies when implementing or improving search functionality at their repositories.

Capture the aspirations, successes and challenges encountered from repository managers.

Provide a means or forum for sharing experiences with relevancy ranking.

Aspiration: Build test collections with real world data and search tasks for data search community to work on.
Activities Discussed at P9

- Conduct the survey, analyse and share survey result.

- Survey goals
  - Identify potential collaborative projects from the Survey
  - Prioritise and coordinate activities from the survey, for example, compare common ranking models.
  - Serve as a benchmark to be looked back on in future to assess how much and in what ways data search has improved.
1. What are characteristics of each repositories? (5)
2. What are system configurations (e.g., ranking model, index methods, query methods)? (7)
3. What are evaluation methods and benchmark? (10)
4. What methods have been used to boost searchability to web search engines (e.g., Google, Bing)? (2)
5. What other technologies or system configurations have been employed? (5)
6. Wish list for future activities for the RDA relevance task force (2)
Geographical Distribution

Total 94
Survey result highlights ...
Data repositories have objects other than data

- Datasets: 91
- Software: 24
- Published Papers: 21
- Methods: 16
- Thesis: 13
- Others: 23

Total: 94
Data repositories use common search systems

Total 94

Ranking models from those who chose the systems: Lucene, Solr and ElasticSearch

Total 51
Open source and available skills are top reasons for choosing a search system
Majority didn’t conduct any kind of evaluations

- 11 Created test collection
- 11 Informal evaluation
- 6 Log analysis

No one provided any performance measure

Yet, nearly half of them believe their search engine will satisfy most users
Respondents’ preference of future activities (n=82)

Our aspiration of creating TREC like test collection doesn’t resonate with respondents well.

Recommendations and solutions are preferred!
Future Activities (P10-P11)

- Finish survey analysis, write up a report/paper.
  - The survey is still open for next 4 weeks (by 20th Oct.). Your participation is more than welcome!
  - https://www.surveymonkey.com/r/RDA_relevancy_ranking

- Publish survey data
- Start new activities from the survey
114 people responded to the survey!

Links to the final survey instrument and summary of per survey questions are available from the Relevance Ranking Task Force Wiki page at the RDA site.
Data Discovery Paradigms IG

 Metadata Enrichment Task Force

Beth Huffer, Ilya Zaslavsky

research data sharing without barriers
rd-alliance.org
Background and Scope

▪ Formed in April 2017

▪ Objective:
  ▪ To describe and catalog various efforts to enrich research data metadata sets to satisfy several use cases
Deliverables

1. A catalogue of automated metadata enrichment tools, together with information about what type of metadata they are able to produce, and the use cases for such metadata;

2. A brief report on how metadata enrichment correlates (or doesn't) with other aspects of data discovery.
Planned Activities

▪ Review DDPIG survey results regarding metadata enrichment;

▪ With an initial focus on automated metadata enrichment tools and services, identify and document:
  ▪ The specific method(s) being employed by each tool or service;
  ▪ The types of metadata (e.g. methods, tools, location, provenance) being produced by each;
  ▪ The use cases (e.g., improving search, enabling faceted browsing, facilitating data integration) those metadata are being used for;
Cross-reference survey responses about metadata enrichment efforts with other responses to look for possible correlations. For example, are repositories that perform metadata enrichment more or less likely to:

- Analyze query logs?
- Measure search engine performance?
- Tune relevancy rankings using internal resources?

Submit follow-up questions to survey respondents, if indicated.
From the Relevancy Ranking Survey:

- **Q28: If you use any technologies to enrich metadata, please list them below**
  - 16 responses to this open-ended question; most refer to manual curation with or without specialized editors of markup tools
  - Just 3-4 refer to in-house custom scripts or specialized metadata enrichment pipelines to impute metadata
  - It is more likely to have automated metadata evaluation than automated enrichment

- Next: explore the mentioned systems; follow up with respondents, via phone interviews or an additional survey
Questions?

▪ Contact

Beth Huffer
beth@lingualogica.net

Ilya Zaslavsky
zaslavsk@sdsc.edu
1. Goals of group and progress (= this!)
2. Overview of 4 active Task Forces:
   1. Use Cases, Prototyping Tools and Test Collections
   2. Best Practices for Making Data Findable
   3. Relevancy Ranking
   4. Metadata Enrichment
3. Discuss work of Task Forces:
   ▪ Shall we close off work on some of these Task Forces?
   ▪ What new task forces should we start?
4. Discuss overlap and synergies with other WG/IGs:
   ▪ Peter McQuilton (BioSharing/FAIR sharing IG)
   ▪ Collaborations with other Working/Interest Groups?
5. Next steps.

Agenda Today:
1. Consider the work of some Task Forces complete?  
   ▪ Use Cases; Best Practices
2. Next steps for:  
   ▪ Relevancy Ranking TF; Metadata TF?
3. Continue with List from P8?  
   ▪ Cataloging common API's  
   ▪ Granularity, domain-specific cross-domain issues  
   ▪ De-duplication of search results  
   ▪ Using upper-level ontologies  
   ▪ Search personalisation
1. Proposals for future DDPIG work coming out of relevancy ranking survey
   1. Create and maintain an environment in which community members can implement and test search algorithms and provide technical support to each other.
   2. Facilitate creation of a corpus or several corpora that would be made available to the community to facilitate benchmark testing of data search systems
   3. Develop evaluation standards and/or evaluate existing standards for data discovery.
   4. Develop detailed recommendations on how to improve relevance rankings using a specific approach that the current group recommends.

2. Other ideas for new Task Forces can we start?
Suggestions from RR Survey

1. Detailed recommendations on how to improve relevance rankings using a specific approaches.
2. New data discovery topics, like including primary data into search, using of visualizations to represent results, new concepts of discovery.
3. Facilitate improved relationships with journal publishers
4. Ranking in linguistic corpus search, e.g., in terms of maximally different linguistic contexts for hits
5. Intelligent search
6. Clarity on the degrees of relevancy and the means to define this
7. The need to fund software development and maintenance for repositories developed with research funds
8. Evaluation of search engine rankings - comparison with peers.
1. Discuss overlap and synergies with other WG/IGs:
   - Peter McQuilton (BioSharing/FAIR sharing IG) [5 min]
   - Collaborations with other Working/Interest Groups? Suggestions from P9:
     - For the use-cases and data-citations. Scholix is an active WG
     - Data repositories WG
     - Repository Platforms for Research Data IG
     - Metadata IG. Metadata completeness -> Dataset quality. Making sure that Metadata is complete. Reusability and requirements for making the most use of a dataset.
     - National Data Services IG
     - Datacitation WG
     - Libraries for Research Data IG
     - Data repository WG
     - Long tail data IG
     - Library/librarians data IG
Data Discovery Paradigms IG

Action items and next steps

research data sharing without barriers
rd-alliance.org
1. Review of Actions coming out of this meeting
   ▪ Action 1 (responsible person)
   ▪ Action 2 (responsible person)

2. Next Steps