Data Discovery Paradigms Interest Group

Mingfang Wu

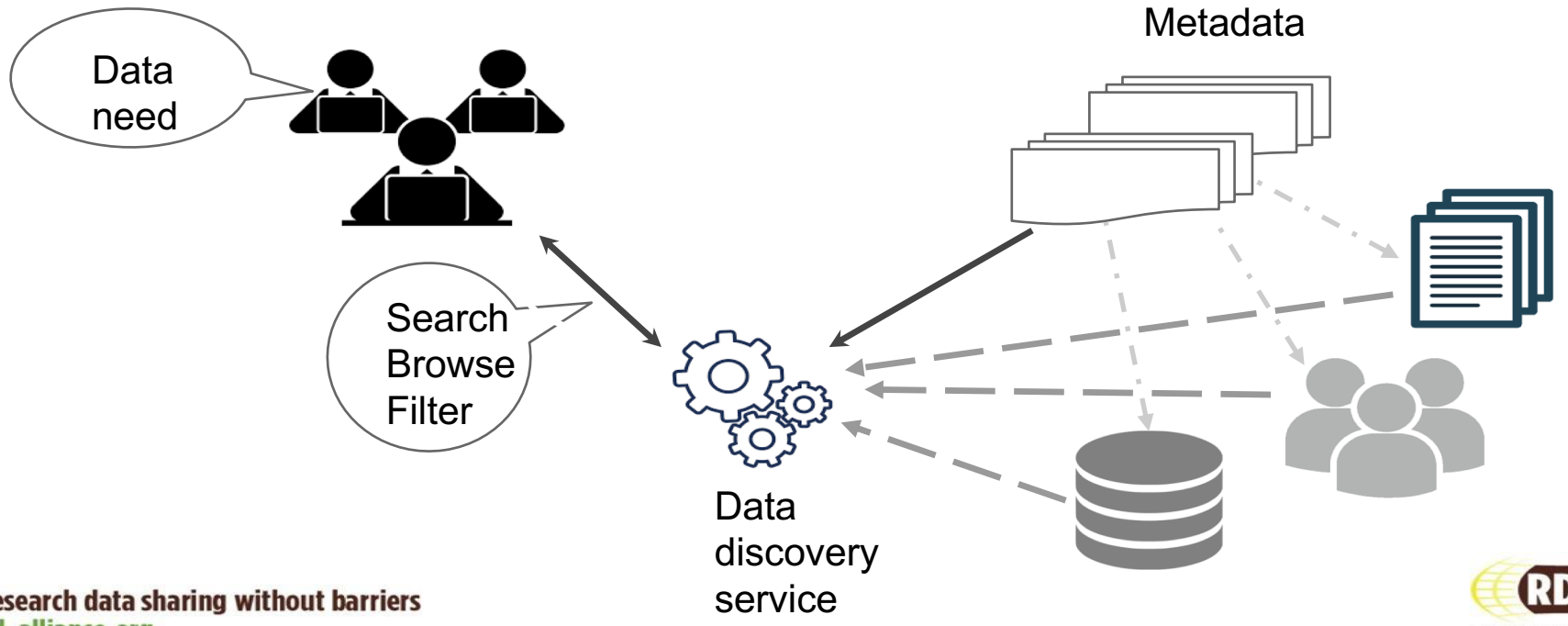
EOSC services collaborations and RDA

21 October 2019, Helsinki

research data sharing without barriers
rd-alliance.org

DDP Interest Group: Motivation

Helping to make research data **Findable** to support users in discovering data.



DDP Interest Group: Objective

- Provide a forum where representatives across the spectrum of stakeholders and roles can explore how to improve data discovery.

- Produce actionable recommendations for data producers, data repositories and data seekers.

Output I - Eleven quick tips for finding research data

Tip 1: Think about the data you need and why you need them.
Tip 2: Select the most appropriate resource.
Tip 3: Construct your query strategically.
Tip 4: Make the repository work for you.
Tip 5: Refine your search.
Tip 6: Assess data relevance and fitness-for-use.
Tip 7: Save your search and data- source details.
Tip 8: Look for data services, not just data.
Tip 9: Monitor the latest data.
Tip 10: Treat sensitive data responsibly.
Tip 11: Give back (cite and share data).

Best practices for data seeker

Can be used for learning and research skills training

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

Output 2 - User Requirements and Recommendations for Data Repositories

Nine requirements (from 79 use cases)

- Indication of data availability
- Connection of data with person/institution/paper/citations/grants
- Fully annotated data
- Filtering of data based on specific criteria on multiple fields at the same time
- Cross-referencing of data
- Visual analytics/inspections of data/thumbnail preview
- Sharing data in a collaborative environment
- Accompanying educational/training material
- Portal functionality similar to other established academic portals

Data repository operators can use the requirements for the following purposes:

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

Output 2 - User Requirements and Recommendations for Data Repositories

Recommendations:

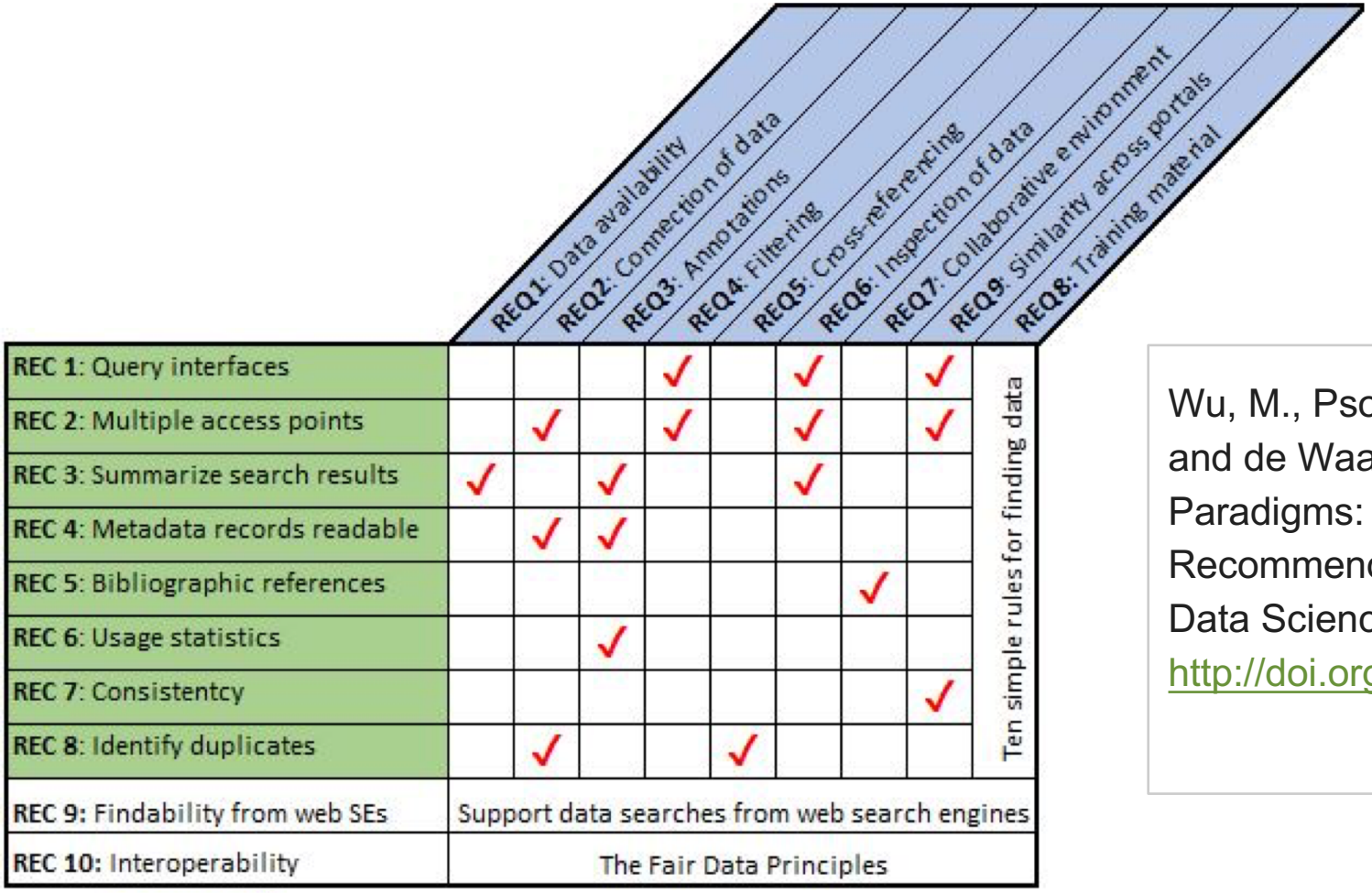
- Multiple query interfaces
- Multiple access points
- Assessable search result
- Readable and analysable metadata records
- Available bibliographic references
- Available data usage statistics
- Consistent interface
- Identifiable duplicats
- Findable from web search engines
- Interoperability with other repositories

Data repositories can take the ten recommendations:

- As guidelines when implementing a new repository
- As a checklist when conducting heuristic evaluation of an existing repository.

Data repositories can implement all or prioritise their implementation based on their user needs and available resources.

Output 2 - User Requirements and Recommendations for Data Repositories



Wu, M., Psomopoulos, F., Khalsa, S.J. and de Waard, A., 2019. Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. Data Science Journal, 18(1), DOI: http://doi.org/10.5334/dsj-2019-003

RDA RESEARCH DATA ALLIANCE

Output 3 - A survey of current practices in data search services

Goal:

- Choose appropriate technologies for search functionality
- Sharing experiences with relevancy ranking.

Khalsa SJ, Cotroneo P, Wu M (2018): A survey of current practices in data search services. DOI: http://doi.org/10.17632/7j43z6n22z.1

Survey highlights:

- Majority of participating repositories deployed either Lucene-based search systems or DB based SQL search.
- Most repositories deploy a system as it is (default settings), do not know which ranking model is deployed or can be modified, and do not apply any heuristics and other technologies to enhance search.
- Less than a quarter of repositories conducted evaluation of search quality, but none of them provide a performance measure.
- About half of the repositories have tried to boost their repository records to web search engines. The Sitemap is most used method followed by Schema.org.
- Repositories would like us to have recommendations on how to improve relevance ranking using a specific approach and evaluation standards.

DDP IG: ongoing task forces

- Metadata Enrichment
  - To describe and catalog various efforts to enrich research data metadata sets to satisfy several use cases.

- Data/Metadata granularity
  - To provide guidance for data managers and data providers that help them determine the best level of aggregation (LofA) to optimize user's discovery, access, interoperability and citability.

Thanks ...

Contact

fpsom@certh.gr
mingfang.wu@ardc.edu.au
sjsk@nsidc.org