Data Discovery: Past and Way Ahead

Jonathan Yu[1], Simon J D Cox[1], Joel Benn[2], Adrian Burton[3], Mingfang Wu[4]

[1]CSIRO Land and Water, Melbourne, Australia, jonathan.yu@csiro.au, simon.cox@csiro.au
[2]Australian Research Data Commons, Canberra, Australia, joel.benn@ardc.edu.au
[3]Australian Research Data Commons, Canberra, Australia, adrian.burton@ardc.edu.au
[4]Australian Research Data Commons, Melbourne, Australia, mingfang.wu@ardc.edu.au

INSTRUCTIONS

A decade has passed since the research community started the movement of sharing, publishing and citing data, in response to the world-wide call for open research. Since then, there have been more than thousands of data repositories established across the globe with the aim of making data more discoverable. Research Data Australia (RDA)[1] is a national research data registry and portal managed by Australian Research Data Commons (ARDC) under NCRIS funding. RDA syndicates metadata from hundreds of data repositories from Australian institutions, research organisations and government organisations.

In recent years, new platforms for data discovery and global repositories have been established, e.g. Google Dataset Search[2], Zenodo, GBIF, SeaDataCloud, eBird, OneGeology. These platforms provide alternative venues for publishing and discovering data, and typically focus on specific services and communities. Research data providers across Australia may have to consider how to syndicate their metadata to have better data discovery opportunity, that may lead to use different syndication protocols and maintain multiple crosswalks, and yet to see what benefits are to their institutions. There are un-answered questions such as if there are needs that aren't able to meet by existing and emerging platforms.

The BoF will provide an opportunity to bring data providers and data repository operators together to discuss their current use of RDA data catalogue and data registry, current challenges in data discovery, and provide perspectives on opportunities and options for how Research Data Australia fits in future data publishing and discovery ecosystem.

The format of the BoF will include a presentation of survey results from current RDA data providers, with a structured discussion on themes such as:

- How and why research organisations publish data, pain points
- Metadata profiles - maintaining multiple profiles (RIF-CS[3], DCAT[4], schema.org[5]) and syndication protocols (OAI-PMH, schema.org crawl)
- Future directions in metadata syndication
- Joined up discovery: software, models, services, instruments, grants/projects, organisations, controlled-vocabularies?

Through structured discussion, we aim to achieve shared understanding of the research data provider community around ongoing data discovery needs and priorities that will inform the next phase of ARDC's development of the RDA service.

NB Participants of this BoF are highly recommended to attend the oral presentation "Future Directions in Data Discovery" (just before this session) for more context.

1 https://researchdata.ands.org.au/
2 https://toolbox.google.com/datasetsearch
3 https://documentation.ands.org.au/display/DOC/About+RIF-CS
4 https://www.w3.org/TR/vocab-dcat/
5 https://schema.org/