Future Directions in Data Discovery

Mingfang Wu[1], Joel Benn[2], Adrian Burton[3], Simon Cox[4]

[1]Australian Research Data Commons, Melbourne, Australia, mingfang.wu@ardc.edu.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]CSIRO Land and Water, Melbourne, Australia, simon.cox@csiro.au

INTRODUCTION

The first letter "F" in the FAIR data principle [1] stands for data findability, the principle recommends "F4(meta)data are registered or indexed in a searchable resource.". 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 [2], and by June 2019, the registry had more than 2000 repositories[1]. A similar trend has also been observed in the Research Data Australia (RDA)[2] - a national research data registry and repository managed by Australian Research Data Commons (ARDC). The existence of these data repositories has greatly improved data discoverability.

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, especially when required data are from several repositories; also, for data repositories that harvest metadata from a number of sources with different metadata schemas. This presentation outlines current and future trends on data discoverability and how data reositories can take advantage of these trends and meet data discovery challenges.

DATA DISCOVERY ON HORIZON

Dataset Search via Web search engines

Web search engines such as Google (http://www.google.com) have been developing dataset search tools. In September 2018, Google announced its Dataset Search tool (https://toolbox.google.com/datasetsearch). Although this is a Beta version, its index covers metadata from a number of large US (e.g NASA) and European (e.g. OpenAIRE) repositories as well as RDA. Our RDA log indicates the dataset search tool has attracted increasing traffic to RDA since the tool was launched, this is not surprising given that the standard Google web search tool has almost become a de-facto information discovery tool for researchers. It is very likely researchers/data seekers are picking up the Google dataset search tool and using it for data discovery in future.

To have metadata records indexed and be searchable by the Google dataset search tool, a data repository needs to: 1) Embed valid structured data markups into dataset metadata landing pages. The structured data markups are for describing properties of a dataset with vocabulary from schema.org (or W3C's Data Catalog Vocabulary), the markups are represented in JSON-LD (preferred or in RDFa). Inserting structured data markups into metadata landing page (in html) will help text analysis parser to process a landing page, extract data properties and provide richer search options and search result presentation, thus improve user's data discovery experience. 2) Include URLs of landing pages into the repository's sitemap and add the sitemap to the site's search console so Google crawler can find and reach landing pages. In this way, the Google dataset search tool enables data repositories make their data discoverable via a single web platform for broader data discovery.

[1] http://www.re3data.org/, accessed on 5th of June, 2019
[2] https://researchdata.ands.org.au/

Applying common set of vocabulary (either in schema.org or DCAT) to describe research resources enables improved metadata interoperability across data repositories, increases re-use and sharing of metadata. Schema.org provides a core, minimalistic vocabulary for describing the kind of entities that most common web applications need. However, schema.org expects and has enabled domains of practice to extend this core by design [3]. Like other domains of practice, the research data community has their own needs for extending this core to describe research data and its relationships to other resources. These extensions include specific research data types and the properties they possess, domain relevant and type specific to persistent identifiers, etc. There are some communities that are addressing these issues and have planned extensions to the core of schema.org to meet their own community needs, for example, bioschemas.org[3] for life science, and science-on-schema.org[4] for earth and environmental science.

Community response

According to a recent survey[5] that was carried out by the Research Data Alliance Data Discovery Paradigms IG[6], more data repositories are implementing structured markups metadata in landing pages, using defined metadata schema, or else with planned extensions of schema.org in one way or another.

RDA, as a national research data portal that aggregates metadata from about 102 academic institutions and government agencies, has also implemented schema.org markups&sitemap to make metadata beholding indexed and searchable by Web search engines. Since 2015, we have marked the 130k metadata landing pages in Schema.org for their discoverability through Web data search tools. There are two advantages in doing this: 1) As a national data catalogue, the syndication of metadata landing pages to a Web data search tool would not require the same syndication activity from each of the contributing 100 disciplinary/institutional repositories (many of whom currently are technically unable to do this). By having relevant, consistent information in a centralised Australian-wide catalogue there is a greater chance that all the data will be indexed, particularly as smaller sites are often less visible: greater economies of scale and lower maintenance costs are also achieved. This benefit has already been achieved: all datasets registered with Research Data Australia are indexed and findable in the Google Dataset Search beta product. 2) We are taking advantage of the Web architecture to make data more discoverable, particularly as a log analysis of our catalogue's activities indicates that about 90% of the traffic are from Web search engines.

New opportunities

In addition to providing the Schema.org&sitemap capacity over the national Research Data Australia aggregation, Australian Research Data Commons (ARDC) is also supporting repositories to build up their own so that they can participate directly in this new paradigm of web search engine enabled dataset discovery. As more data repositories make their data discoverable via the web architecture with common vocabulary, there brings a potentially new method for metadata/content syndication among repositories, enables federated search across repositories of a specific domain or related domains relevant to a research need. For example, most metadata aggregators (e.g. RDA) currently apply the OAI-PMH method to harvest metadata from data providers; if a data provider and an aggregator use two different metadata schemas, there requires a mapping between two schemas, number of mappings increase if a data provider feeds metadata to multiple aggregators and vice versa. If all data repositories (either data providers or aggregators) implement Schema.org&sitemap, resources spent on developing, tracking and maintaining mappings between any pair of schemas will be greatly reduced. An aggregator such as RDA can send crawler to harvest metadata from those data repositories' sitemap as web search engines' crawler does and extract metadata from metadata landing pages accordingly. New data discovery applications can be built such as federated search across resources of a specific domain, or related domains relevant to a research need; applications can support a spectrum of data search needs from free text search to SPARQL queries. It also makes data recommendations across data repositories easier and possible, for example, we are exploring this new metadata syndication method and recommendations among portals RDA, NRDP (National Research Data Portal) from Korea (as maintained by the Korea Institute of Science and Technology Information), the

3 https://bioschemas.org/
4 https://github.com/ESIPFed/science-on-schema.org/
5 http://bit.ly/2JZxXjK
6 https://rd-alliance.org/groups/data-discovery-paradigms-ig

EarthCube Project 418[7] data providers including Biological and Chemical Oceanography Data Management Office (BCO-DMO, https://www.bco-dmo.org/), Continental Scientific Drilling Coordination Office (CSDCO, https://csdco.umn.edu/), Interdisciplinary Earth Data Alliance (IEDA, https://www.iedadata.org/) among others.

REFERENCES

1. 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
2. Pampel H, 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; p. 287—288. Available from: http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1369620.
3. Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: Evolution of Structured Data on the Web. Communications of the ACM, 59(2), 44–51. doi:10.1145/2857274.2857276

[7] https://www.earthcube.org/p418