From Data Search to Data Showcasing: The Role of Semantic Technologies in a New Service

Peter Cotroneo[1], Wouter Haak[1], Gabriel Oscares[1], Eleonora Presani[1], Abhinav Rohatgi[1], and Paul Groth[1]

[1] Elsevier
[2] Radarweg 29. Amsterdam 1043 NX
p.cotroneo,w.haak,g.oscares,e.presani,a.rohatgi,p.groth@elsevier.com

Abstract. Universities and other research institutions increasingly want to present and showcase the datasets their researchers produce. In this work, we describe how Elsevier has leveraged semantic technologies in the form of knowledge graphs and cross-organizational metadata in order to create a dataset showcasing service.

Introduction The scientific community is increasingly concerned with the availability of datasets in order to improve the scientific process.[3] This has led to a large ecosystem of research data repositories where researchers can make their datasets available including data specific repositories (e.g. the Protein Databank), general data repositories (e.g. Zenodo), institutional data repositories (e.g. University of Melbourne data repository), to field specific repositories (e.g. ICPSR). Institutions interested in understanding their output in terms of data thus face a challenging proposition: they need to be able to discover datasets across all these repositories and provide a navigable index. This navigable index is what's termed a showcase. While they can, and in some cases do, require researchers to register their datasets with a central portal, this is a time consuming process especially in institutions with thousands of researchers who work in multiple disciplines. In this work, we describe the use of two kinds of semantic technologies: 1) cross-organizational metadata and 2) a knowledge graph of research to build a data showcasing service. Importantly, the launch of this service has been a critical component of Elsevier's Research Data Management value proposition.

System Description A prerequisite to the the creation of the showcasing service was the ability to index existing data from multiple repositories. Elsevier had already developed a research data search engine[4] with the primary use case being to support researchers in their search for data.[5] The search engine provides deep

[3] Mesirov, J.P.: Accessible reproducible research. Science **327**(5964), 415–416 (2010). https://doi.org/10.1126/science.1179653
[4] https://data.mendeley.com/datasets
[5] de Waard, A.: Research data management at Elsevier: Supporting networks of data and workflows. Information Services & Use **36**(1-2), 4955 (Sep 2016). https://doi.org/10.3233/ISU-160805

indexing (i.e. both metadata and data) of over 30 data repositories. A key part of the search engine is that it normalizes the metadata provided by the multiple data repositories to a common schema. This, for example, ensures that we index the author name, institution, license, links to publications, etc in a common way. Using this information, we can essentially look-up all datasets with an associated institution. However, this is not as trivial as it first seems. First, we face the issue that many of the repositories do not provide an institution name at all for their datasets. Second, the institution name is provided as free-text. This is where semantic technologies come into play.

Cross-Organizational Metadata: To obtain institutional names associated with datasets, we use the notion that research data is often associated with scholarly articles. Over the last several years, the Scholix initiative[6] of the Research Data Alliance and the World Data System was formed to allow organizations to exchange metadata about the links between datasets and literature. It is supported by over 15 organizations including Elsevier, DataCite, OpenAire, Crossref, ANDS, and EBI for example. Concretely, a common schema[7] was designed to express these links. Data repositories then register these links with one of the Scholix hubs; for example with DataCite or CrossRef. The OpenAire Scholexplorer Service harvests and aggregates these links and exposes these using a web API. Thus, there is a shared semantics about what is contained in these links and how to address them. Elsevier's system then uses the Scholexplorer service to enrich its knowledge graph of research to have links between articles and datasets.

A Knowledge Graph of Research: Scopus is a database of the world's scientific literature. It forms a knowledge graph connecting 69 million unique article entities, with 70,000 institutional entities and 12 million author entities stored using standard search engine technologies. Our showcasing service uses this knowledge graph to first identify all the articles associated with an institution. It then filters out all articles that contain a link to a dataset and that is also contained in the data search index. Thus, using data search we can generate a searchable showcase page limited to datasets published by the institutions. Furthermore, we can use the knowledge graph to disambiguate free text institution names when available in the data search index.

Conclusion In this work, we briefly described our approach to creating a new service that would be difficult to build without two semantic approaches: knowledge graphs and cross organizational shared metadata. It was not just enough to provide disambiguated entities, it was necessary to be able to have links between entities - being able to jump from institution to article to dataset. Likewise, a shared schema across providers is crucial in being able to deliver this service at scale across multiple independent and heterogenous data repositories. Overall, we have seen that the time to market has been reduced using these approaches.

[6] http://www.scholix.org/

[7] Burton, A., Fenner, M., Haak, W., Manghi, P.: Scholix Metadata Schema for Exchange of Scholarly Communication Links (Nov 2017). https://doi.org/10.5281/zenodo.1120265