eResearch Australasia 2015

Research Data Switchboard: Finding Connections to Your Data

Amir Aryani, Adrian Burton, Andrew Treloar
Australian National Data Service, Canberra, Australia
{amir.aryani, adrian.burton, andrew.treloar}@ands.org.au

INTRODUCTION

Driven by the rapid development of data storage technology, the number of research data repositories is growing fast and researchers more than ever have access to a range of data repositories including university data storage, discipline specific repositories and national (regional) level data infrastructures. The problem is that these infrastructures are often operating in silos; that is, they cannot connect their datasets to the related research or datasets in other platforms.

In eResearch Australasia 2014, ANDS presented the Research Link initiative[1] as a coordinated effort for establishing and maintaining the connectivity between research data and elements in the research ecosystem including publications, grants and researchers. One of the outcomes of this initiative has been a new infrastructure called Research Data Switchboard (RD-Switchboard.org) a collaborative project by ANDS, and number of other international partners in a working group of the Research Data Alliance[2]. The current members of that working group are: Australian National Data Service (ANDS), Data Archiving and Networked Services (DANS - Netherlands), CERN (European Organization for Nuclear Research), DataCite, da|ra (German Initiative), Data Curation Unit (Part of Athena Research Centre, Greece), DataPASS, Dryad[3] and VIVO Cornell (from United States), and Thomson Reuters Data Citation Index. This group was formed as a partnership between registries and data infrastructures who have identified the need for connecting their infrastructure to the global network of scholarly works and find other datasets, paper and grants that can be associated to their research datasets.

RD-SWITCHBOARD

RD-Switchboard addresses the problem of cross platform discovery by operating online services that connect datasets across multiple registries. The best metaphor for these services is the SEE ALSO section in online bookstores, where customers are invited to look at other products by the same author, related topics or similar publishers.

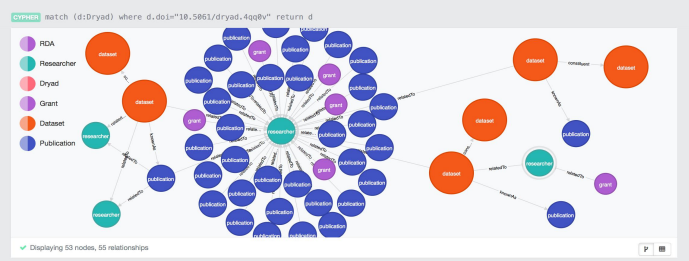

Figure 1: RD-Switchboard interface to the graph database using Neo4j

The main objective is to connect datasets together on the basis of co-authorship or other collaboration models such as joint funding and grants. The system aggregates links between publications, datasets and research grants from national and international registries and utilises graph-modelling technology to identify missing links between datasets. Figure 1 demonstrates how RD-Switchboard uses the Neo4j graph database and the Force Directed Graph Drawing Algorithm to

[1] Adrian Burton, Amir Aryani, "Research Linking Initiative: Toward Interoperability of Research Data". eResearch Australasia, Melbourne, Oct. 2014

[2] The Data Description Registry Interoperability (DDRI) working group: www.rd-alliance.org/group/data-description-registry-interoperability.html

[3] www.datadryad.org

visualise the links between datasets. Here, RD-Switchboard has identified the datasets co-authored by Australian researchers in Dryad and linked them to datasets in the Research Data Australia repository.

In addition, RD-Switchboard uses Google API, Fuzzy search algorithm and graph clustering technology to disambiguate authors and link them to their datasets.

Example of such connections: Where *D* is a Dataset, *A* is an Author, *P* is Publication and *G* is a Grant

- D_1->P->D_2 Two datasets are linked through a paper, i.e. data citation
- D_1->A_1->P->A_2->D_2 Two datasets are linked because their authors collaborated in other works
- D_1->A_1->G->A_2->D_2 Two datasets are linked because their authors collaborated in a grant

Note: For the simplicity of these examples, links are represented according to the direction of graph crawling algorithm, but are actually bidirectional.

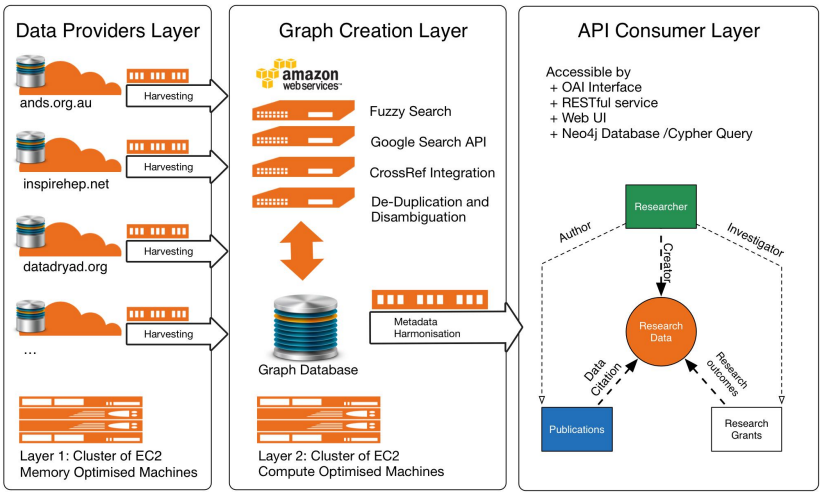

Figure 2: RD-Switchboard abstract architecture

The abstract architecture level of the RD-Switchboard is presented in the Figure 2 where a Data Provider layer enables data repositories to import metadata records into the platform, a Graph Creation layer aggregates information and uses Google API and other services to identify missing connections, and the outcome is an accessible API Consumer Layer.

INTEGRATION OPPORTUNITY

There are two main areas where the RD-Switchboard infrastructure will be applied. Firstly universities and research institutions will be able to add metadata about their research (grants and papers) to RD-Switchboard and leverage this system to discover connected datasets to their research in external repositories. Secondly data providers and repositories will be able to use RD-Switchboard API to discover connections to their datasets beyond their infrastructure; for example, Dryad can find grants and other datasets in Research Data Australia related to dryad records by co-authorship. For more information about the progress of this project and collaboration opportunities visit www.rd-switchboard.org.

SUMMARY

RD-Switchboard is a new collaborative project by ANDS, CERN (European Organization for Nuclear Research), Dryad and number of other international partners in the Research Data Alliance. This platform enables finding connections between research datasets across national and international data registries, and this infrastructure can be leveraged by Australian research institutions and universities to find their datasets in the international repositories.

ABOUT THE AUTHORS

Dr Amir Aryani is the co-chair of the Data Description Registry Interoperability WG in Research Data Alliance and the project lead for the Research Data Switchboard. He is working in the capacity of a project manager for Australian National University (ANDS), and part of this role is to manage ANDS interoperability projects with international partners. He has completed his PhD in the field of software evolution at the school of computer science, RMIT university, and he has peer-reviewed publications in fields of Software Engineering, Software Evolution and eResearch.

Dr Adrian Burton is Director of Services at the Australian National Data Service (ANDS). In this capacity he has a keen interest in national services that enable data publication, data discovery and data citation as well as the human support services that build the capability of researchers and research organisations to take advantage of data infrastructure. Adrian has provided strategic input into several national infrastructure initiatives, including Towards an Australian Research Data Commons, The National eResearch Architecture Taskforce, and the Australian Research Data Infrastructure Committee. Adrian is active in building national policy frameworks to unlock the value in the research data outputs of publicly funded research.

Dr Andrew Treloar is Director of Technology at the Australian National Data Service (ANDS). In 2008 he led the project to establish ANDS. He is currently co-chair of the Research Data Alliance (http://rd-alliance.org/) Technical Advisory Board and was a Visiting Fellow at the Data Archive and Network Services organisation in the Netherlands in 2013/14 (http://dans.knaw.nl/). His research interests include data management and scholarly communication. He never seems to be able to make enough time for practising his 'cello, or reading, but does try to prioritise talking to his chickens and working in his vegetable garden and orchard. Further details at http://andrew.treloar.net/ or follow him on Twitter as @atreloar.