Data Description Registry Interoperability WG

Interlinking Method and Specification of Cross-Platform Discovery

Created: September 2015
Last update: Jan 2016

Amir Aryani (orcid.org/0000-0002-4259-9774)
amir.aryani@ands.org.au / http://people.anu.edu.au/amir.aryani/

Purpose of this document

This document describes the outcome of the Data Description Registry Interoperability (DDRI)[1] working group and specification of the interoperability model implemented by the partners in this group. In addition, this document shows the testing of this model through an implementation called Research Data Switchboard, a collaborative project by the participants in this working group. The intended audience is the Research Data Alliance community.

Section 1. 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. The partners in this working group have addressed the problem of cross platform discovery by connecting datasets together on the basis of co-authorship or other collaboration models such as joint funding and grants.

This document provides a technical description of the interoperability model that is being used to create these connections. In addition, it demonstrates the application of this model by reporting on the number of connections created through bilateral information exchange between partners in this group.

Section 2. Method

This section describes the approach and connectivity methods that this working group has explored to discover how datasets can be linked across platforms.

[1] https://rd-alliance.org/group/data-description-registry-interoperability.html

Note: In the context of this section, a *node* is a dataset (*D*), publication (*P*), researcher (*R*) or grant (*G*), and a *relation* (*[:RelationLabel]*) is a connection between two nodes that has been identified in the metadata of a repository or registry.

Connecting DOI: DOI for publications and datasets has been linked across multiple registries where a node had the DOI as an identifier or citation element. The following is the most common connections that we have identified as useful through linking DOIs

- $D_1 \rightarrow P_1$
- $D_1 \rightarrow P_1 \rightarrow D_2$
- $D_1 \rightarrow D_2 \rightarrow P_1$

KnownAs relations: In order to identify and aggregate duplicated datasets and publication, we recommend connecting the nodes that have the same URL with the relationship type KnownAs. In this exercise DOIs are transformed to resolvable URLs and included in building KnownAs relationships. The following are three main relations that can be derived from this process:

- D_1 *-[:KnownAs]→* D_2
- P_1 *-[:KnownAs]→* P_2
- G_1 *-[:KnownAs]→* G_2

Co-Authorship: In order to connect datasets by co-authorship we recommend exploring two possible sources of information, ORCID and Google Search. ORCID enables link nodes to be linked where they are connected by an ORCID author identifier, and Google Search enables nodes to be linked where the title of the nodes has been presented in the content of the researcher page. The following relationships shows the most common connectivity between datasets where they are linked by a researcher.

- $D_1 \rightarrow R_1 \rightarrow D_2$
- $D_1 \rightarrow R_1 \rightarrow P_1 \rightarrow D_2$ *(also* $D_1 \rightarrow P_1 \rightarrow R_1 \rightarrow D_2$*)*
- $D_1 \rightarrow P_1 \rightarrow R_1 \rightarrow P_2 \rightarrow D_2$

Co-Funded Projects: One form of relation between datasets is through grants and co-funded research projects.

- $D_1 \rightarrow G_1 \rightarrow D_2$

In addition, grants can connect datasets by multiple degrees of separation via researcher (grant participants) and publications. Please note that these relations do not represent causality, instead they suggest semantic and contextual relationship. The following shows three examples of such relations:

- $D_1 \rightarrow P_1 \rightarrow G_1 \rightarrow D_2$
- $D_1 \rightarrow G_1 \rightarrow R_1 \rightarrow D_2$

- $D_1 \rightarrow P_1 \rightarrow G_1 \rightarrow R_1 \rightarrow D_2$

Section 3. Metadata Model

The input sources that are useful for our recommended approach contain heterogeneous metadata including DC, MODS/MET, MARC21, RIF-CS, etc. In order to make the nodes connectable we recommend applying a harmonisation process where a minimum metadata has been extracted from all sources into a simple and unified set of properties for each node. The following is the expected resulting list of properties for each node. Please note that:

- [M] properties must exists
- [O] Optional
- [I] properties should be indexed for quick search
- [L] properties must have lowercase value
- [C] properties have constant value
- N/A: not applicable
- all properties' names are lowercase

Name	comment	Publication	Dataset	Grant	Researcher
key	A URL to the object or landing page without http://www. If an object has no URL then we do need to put it in this graph.	[M,I,L]	[M,I,L]	[M,I,L]	[M,I,L]
title	for Web:Researchers we use dc.title	[M,I]	[M,I]	[M,I]	[O,I]
authors	array of authors, or investigators (for grants)	[M]	[O]	[M]	N/A
initials		N/A	N/A	N/A	[O,I]
first_name		N/A	N/A	N/A	[O,I]
last_name		N/A	N/A	N/A	[O,I]
full_name	remove the title ("Dr. Mr. Mrs...") from the full name. DC.Title can be used to populate this field.	N/A	N/A	N/A	[M,I]
node_type	publication, dataset, grant,researcher, institution	[C,M,I,L]	[C,M,I,L]	[C,M,I,L]	[C,M,I,L]
publicaiton_type	paper, book	[C,M,I,L]	N/A	N/A	N/A
date	universal date and time. Publication date for datasets and papers and awarded (or start) date for grants.	[M,I]	[O,I]	[M,I][2]	N/A

[2]Publication_date for grants is the date that grant is awarded, if not available then grant start date should be used

node_source	Array of sources including Dryad, RDA, CrossRef, ORCID, Scopus, Web, CERN, figshare	[C,M,I,L]	[C,M,I,L]	[C,M,I,L]	[C,M,I,L]
nla	national library of australia identifier.	N/A	N/A	N/A	[O,I,L]
doi	remove 'doi' from the value	[O,I,L]	[O,I,L]	N/A	N/A
orcid	formatted like 0000-0002-1825-0097 and not like a URL	N/A	N/A	N/A	[O,I,L]
isni	some ORCID records might have isni number. format of isni is the same as ORCID	N/A	N/A	N/A	[O,I,L]
isbn	only for books	[O,I]	N/A	N/A	N/A
purl	remove Http://www.	N/A	N/A	[O,I,L][3]	N/A
grant_number		N/A	N/A	[M,I,L][4]	N/A
for_codes	Array of Field of Research codes. List available at http://www.abs.gov.au/ausstats/abs@.nsf/0/6BB427AB9696C225CA2574180004463E	[O,I,L]	[O,I,L]	[O,I,L]	[O,I,L][5]

Section 4. RD-Switchboard: Putting the Method into Practice

Research Data Switchboard is an open collaborative project that has served as a reference implementation for the recommended method and model described in sections 2 and 3. This project was initiated by the collaboration of the following international partners:

- ANDS (Australia)
- Dryad (US)
- CERN InspireHEP (Switzerland)
- figshare (US)
- da-ra and GESIS (Germany)
- Data Curation Unit (Greece)
- Data and Literature Interlinking Service (a Research Data Alliance initiative)
- OpenAIRE (European Infrastructure)

The system aggregates links between publications, datasets and research grants from national and international registries and utilises graph modeling technology to identify missing links

[3] For ARC and NHMRC it is mandatory; however, some international grants in future will not have any PURL
[4] In RDA XML (RIF-CS format), the element `Identifiers.identfier.arc (or nhmrc)` contains the grant number.
[5] FOR code for researchers identifies their research activity area.

between datasets. At the time of writing this paper, RD-Switchboard uses the Neo4j graph database and the Force Directed Graph Drawing Algorithm to visualise the links between datasets as demonstrated in Figure 1. Here, RD-Switchboard has identified the datasets co-authored by Australian researchers in Dryad and CERN data repositories, and linked them to datasets in the Research Data Australia repository.

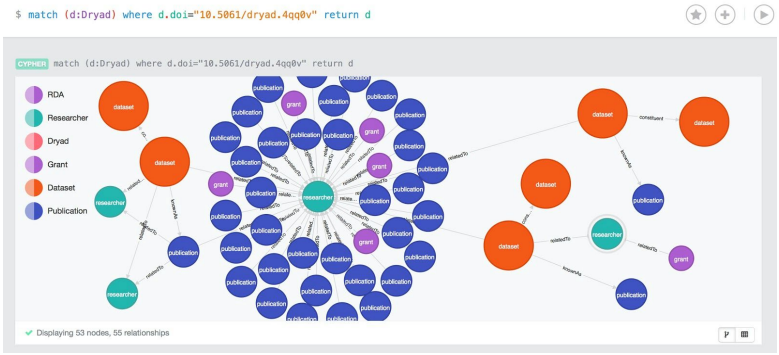

Figure 1: RD-Switchboard interface to the graph database using Neo4j

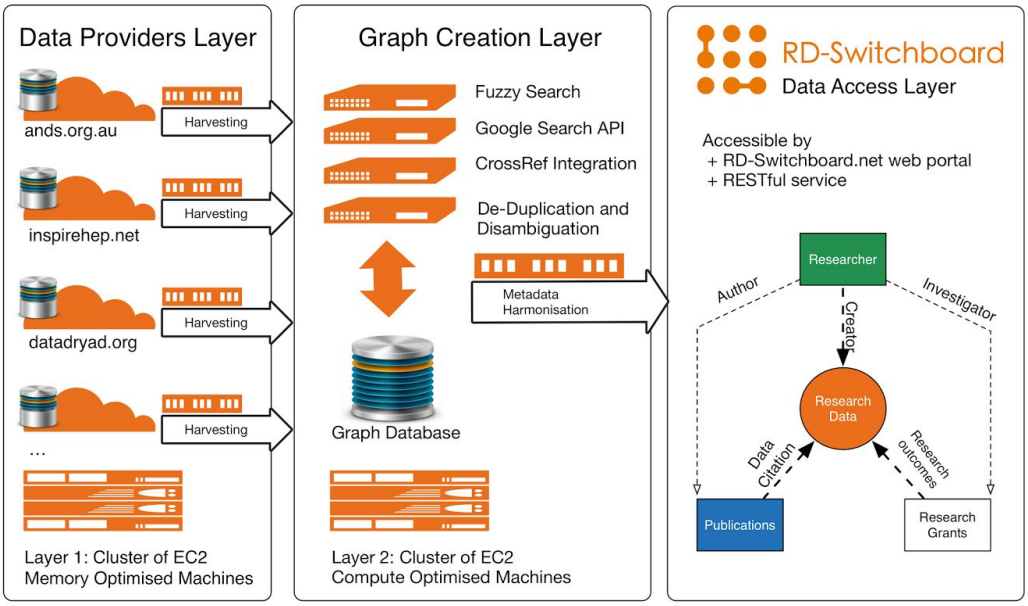

Figure 2: RD-Switchboard abstract architecture

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 Data Access Layer.

The source code for RD-Switchboard is accessible at http://github.com/rd-switchboard , and you can find futher updates on this project at http://rd-switchboard.org .

Section 5. Results of Connectivities between Nodes

This section shows the results of the connections identified by the RD-Switchboard platform. Please note that these results are valid at the time of writing this document (September 2015); however, they will change in future as the metadata input sources and the platform evolve.

The queries in the following table have been described in the Cypher language. More information available at http://neo4j.com/docs/stable/cypher-query-lang.html

Description	Cypher Query	Results
Connections between Dryad and ANDS	`match (n:dryad)-[*1..3]-(a:ands) return count(n)`	3,191
Connections between ANDS and CERN	`match (n:cern)-[*1..3]-(a:ands) return count(n)`	2,634
Connections between Dryad and ORCID researchers	`match (n:dryad)-[*1..3]-(o:orcid:researcher) return count(distinct(o.key))`	1,379
Connections between CERN and ORCID researchers	`match (n:cern)-[*1..3]-(o:orcid:researcher) return count(distinct(o.key))`	858
Connections between da-ra and DLI nodes	`match (d:dara)--(x:dli) return count(distinct(d.key))`	1,009
Connections between da-ra datasets and publications with DOI	`match (d:dara)--(x:crossref) return count(distinct(d.key))`	1,198

Section 6. Adoption

The outcome of the DDRI working group and the reference implementation (Research Data Switchboard) can be adopted using the materials available in GitHub repository (github.com/rd-switchboard). At the time of writing this section (Jan 2016) the code based has been adopted by three organisations:

- Australian National Data Service (ANDS) has adopted the code to create a service for discovery of connections between Australian datasets international scholarly works. This work is currently at the testing stage by domain experts.

- Australian National Computational Infrastructure (NCI) has adopted the proposed recommendation to create a graph of connections between datasets stored on NCI infrastructure. This work is currently in development phase.
- The University of Sydney has adopted RD-Switchboard to create a graph of connections between researchers and scholarly works including datasets and publications. This work is at the early stage of development.

Summary

In this document, we have described the interlinking recommended method and its reference implementation that was created by the collaboration between the partners in the Data Description Registry Interoperability WG. The reference implementation source code is accessible at http://github.com/rd-switchboard , and you can find futher updates on this project at http://rd-switchboard.org .