Research Graph: Building a Distributed Graph of Scholarly Works using Research Data Switchboard

Amir Aryani, Australian National University, amir.aryani@anu.edu.au, orcid.org/0000-0002-4259-9774
Jingbo Wang, National Computational Infrastructure, Jingbo.Wang@anu.edu.au,
orcid.org/0000-0002-3594-1893

Session Type

Presentation

Abstract

In this paper, we discuss an open collaborative project called Research Graph derived from the outcome of the Research Data Alliance (RDA) working group on Data Description Registry Interoperability. This project addresses the problem of connecting scholarly works across heterogeneous systems. The RDA working group recommendation provided a solution for connecting publications and research data (data in research) across multiple open access repositories using co-authorship model and jointly funded research projects. Research Graph adopts and extends this work by creating a distributed graph that connects open access repositories to close research management systems traditionally locked behind the firewall. In addition, the distributed graph addresses the challenge of scalability and enables individual universities and repositories to hold a small and manageable graph and synthesis this graph with trusted partner organisations.

Conference Themes

- Supporting Open Scholarship, Open Data, and Open Science
- Repositories of high volume and/or complex data and collections
- Managing Research Data, Software, and Workflows
- Integrating with the Wider Web and External Systems

Keywords

Linked open data, research discovery, collaboration network, research graph

Audience

Librarians, research administrators, technologists, decision makers, project leaders, evaluators, government agency representatives, and strategists should find the presentation informative regarding existing data resources and goals of Research Graph, and the resulting capability to create and use large quantities of connected metadata regarding scholarly outputs for research discovery.

Background

Driven by the rapid development of data storage technology and the increasing demand for open science, the number of research repositories is growing fast and researchers more than ever have access to a range of research information systems and data infrastructures such as open access publication repositories, discipline-specific data repositories, and national (regional) level infrastructures. The problem is that these infrastructures are often operating in silos; that is, there is no easy way to make connections between their research results, especially unpublished datasets, and external research work. Even their publications will be isolated items if there is no cross-reference in the literature.

In this presentation, we report on the progress of an open collaborative project called Research Graph[1]. This project addresses the problem of cross-platform discovery of research data , publications and even grant information by operating set of services that connect these records across multiple research information systems. This work is derived by leveraging the Research Data Switchboard [1] software developed as part of the Research Data Alliance working group on Data Description Registry Interoperability[2]. The group had participants from Australian National Data Service (ANDS), Dryad (US), CERN InspireHEP (Switzerland), figshare (UK), da|ra and GESIS (Germany), Data Curation Unit (Greece), OpenAIRE (European Infrastructure), ORCID, and DataCite. The participants in the group have provided substantial metadata records including publications, datasets, researcher information and grant records that are currently available in a form of graph database hosted on the AWS cloud.

The main objective is to connect research publications and 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. Research Graph extends the RDA working group outcome by creating a distributed graph that connects open access repositories to closed research management systems. These research management systems hold valuable information about connections between grants, publications and researchers. However, traditionally these systems are operating behind the firewall and disconnected from open access repositories.

Content

In this presentation, we discuss how to create a graph from the research management systems with the minimum technical requirement, and how to link this graph to a larger Research Graph database that contains funding information, and collections of research data and open access repositories . This approach works for research management systems with no support for RDF and Ontology; however, we are extending the Research Graph model to support VIVO [2] ontology and as such the Research Graph APIs will support exchanging information in VIVO RDF.

Note regarding this presentation: given the limited time of the talk, we focus on the overall architecture and workflows for connecting researchers, publications, research datasets and grants using Research Graph components,. The technical discussion will be limited to abstract function of the components so there would be no need for software engineering expertise to benefit from this talk. However, the materials and extended technical discussions will be available online to all conference attendees.

Registry Objects: Research Graph has five registry objects in its metamodel: `Researcher`, `Publication`, `Dataset`, `Grant`, and `Relation`. Figure 1 illustrates these objects. Further information about these objects is available at researchgraph.org/schema.

Please note that in this model `Relations` are defined as a separate registry object that connects two URIs. However, for simplicity we abbreviate:

Object$_1$-[:relationLabel]→ Object$_2$

to Object$_1$ → Object$_2$

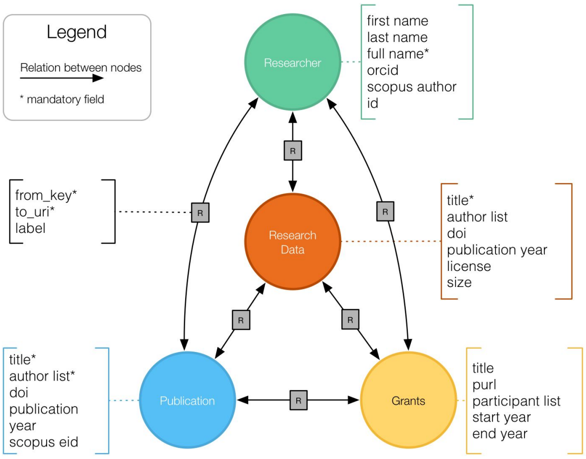


Figure 1: Research Graph Metamodel

[1] www.researchgraph.org

[2] https://rd-alliance.org/groups/data-description-registry-interoperability.html

Multiple degrees of separation: two registry objects can be linked indirectly via other objects. E.g. a publication (`P`) and a grant (`G`) can be linked in two degrees of separations via a researcher (R) who is the author of the paper and investigator of the grant: `P → R → G`

In Research Graph, we use the concept of multiple degrees of separation to identify how a publication or a research datasets is connected to other related works. This concept also used in the synthesis function (explained later in the paper) to connect the registry objects across platforms.

Building a local Research Graph using Research Data Switchboard: Creating a research graph database from research information system or an open access repository is a simple and straightforward process. To accomplish this goal we use Research Data Switchboard inference components[3]. As illustrated in Figure 2, the process involved harvesting, inference and metadata harmonisation. As part of the presentation, we will provide examples of these steps using NCI[4] records and discuss how these steps can operate as a pipeline. This demonstration will be particularly beneficial to repository managers, technologist and data infrastructure providers.

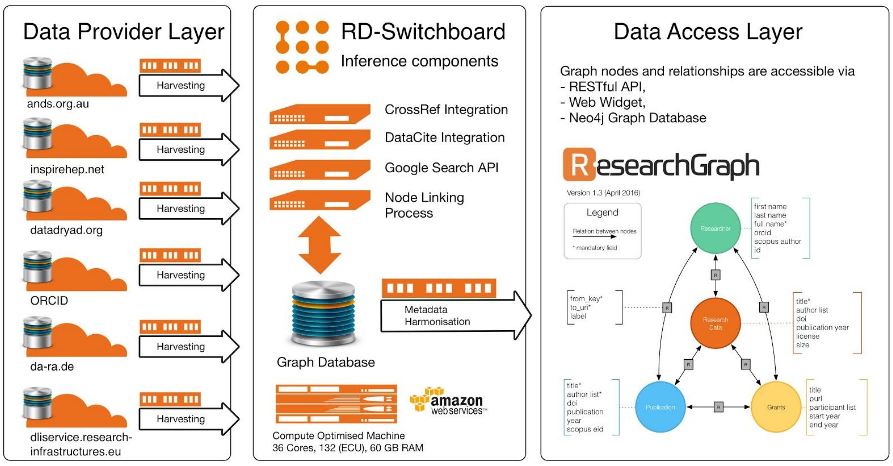
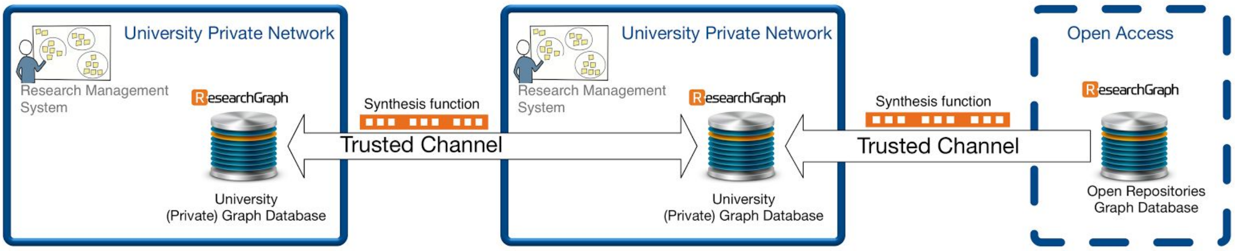


Figure 2: Research Data Switchboard workflow includes harvesting metadata, operating the inference components and harmonising the metadata records into the Research Graph model.

The outcome of this process is a graph database in Neo4j[5] format. We use Neo4j technology to store the connections between registry objects.

Synthesis function

The next step is to connect a local graph database to external systems. In the early stage of this project, we had a complex large scale graph by harvesting records from many repositories. However, this approach is not scalable, and it also excludes closed research information systems that traditionally operate behind the firewall. To address these issues, we have implemented synthesis function that links the graphs across private and public (open and closed) systems based on the trusted partnerships (Figure 3).



University Private Network
Research Management System
ResearchGraph
University (Private) Graph Database
Synthesis function
Trusted Channel
University Private Network
Research Management System
ResearchGraph
University (Private) Graph Database
Synthesis function
Trusted Channel
Open Access
ResearchGraph
Open Repositories Graph Database

[3] https://github.com/rd-switchboard/Inference

[4] http://nci.org.au

[5] https://neo4j.com

Figure 3: Using Trusted channels to synthesis the graph across multiple boundaries.

The result enables two partner institutions to connect their graph and only transfer nodes that are linked by multiple degrees of separation. At the time of writing this paper, the synthesis function operates with three degrees of separation to connect private repositories such as University Sydney graph database to external informations such as CERN inspireHEP, figshare, and Dryad DSPACE repository. In this presentation, we will demonstrate the example of these synthesised graphs, and we use graph visualisation and metrics to compare the connectedness of the registry objects before and after running the synthesis function.

Conclusion

In summary, this presentation demonstrates the potential interoperability between open access repositories and research information management systems. We will demonstrate a new set of capabilities that can facilitate connecting publications, grants, research datasets and researchers across platforms. These new capabilities make research outcomes more connected and discoverable, leading to identifying new research collaboration opportunities.

Acknowledgment

The authors would like to thank and acknowledge the significant contributions of Dr. Adrian Burton from Australian National Data Service (ANDS), Nathaniel Lewis (University of Sydney), Dr. Marta Poblet (RMIT University), Brigitte Hausstein, Dr. Peter Klas, Dr. Benjamin Zapilko (GESIS), Dr. Anusuriya Devaraju (CSIRO), and the members of the RDA DDRI WG.

References

1. Amir Aryani, Data Description Registry Interoperability WG: Interlinking Method and Specification of Cross-Platform Discovery, Research Data Alliance, doi:10.15497/RDA00003

2. Börner, K., Conlon, M., Corson-Rikert, J., Ding, Y. (eds.) *VIVO: A Semantic Approach to Scholarly Networking and Discovery,* Morgan-Claypool, 2012. p. 1-175.