Visualising Research Graph using Neo4j and Gephi

Amir Aryani^[1], Jingbo Wang˜, Hao Zhang^, Andy Xiang^, Zhaolian Zhou^, Kun Wang^

^Australian National University, ˜National Computational Infrastructure (NCI)

Session Type

- Developer Track

Abstract

The goal of this presentation is to provide an insight into the potential interoperability between open scholarship systems. We demonstrate how to export the publication metadata from DSPACE repository and link this information to ORCIDs (Researchers), Funding Records (grants) and research data (data in research) using the Research Graph model and open source software. Furthermore, we demonstrate how to transform this information to Neo4j graph database that enables us to run queries such as finding related publications to a grant with multiple degrees of separation. Finally, we will use the Gephi visualisation tool to plot the large graph and identify the clusters of research activities.

Conference Themes

Select the conference theme(s) your proposal best addresses (remove the others):

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

Keywords

Research Graph, Neo4j, Gephi, Collaboration Network, Visualisation

Background

In this presentation, we demonstrate how to visualise the network of connections between publications, researchers, grants and research datasets using graph visualisation tool called Gephi. In addition, we discuss how the information from open access repositories can be linked to international identifier services, funding information and research data infrastructures.

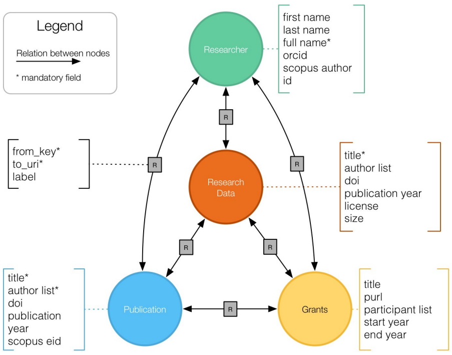

Figure 1: Research Graph Metamodel

We use Research Data Switchboard (rd-switchboard.org) software components and the Research Graph metamodel. Although this is a technical presentation, we limit the code walkthrough to the main dataflow between various components; hence, making the talk more accessible to a larger audience.

Research Graph[2] has been derived from the work of the Research Data Alliance working group[3] on connecting research data across multiple repositories. The group had participants from Australian National Data Service (ANDS), CERN InspireHEP (Switzerland), figshare (UK), da|ra and

[1] Amir Aryani is the corresponding author for this work: amir.aryani@anu.edu.au

[2] http://researchgraph.org

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

GESIS (Germany), DataCite, and a number of other international, regional and discipline-specific research data infrastructures . 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 based on the Research Graph metamodel hosted on the AWS cloud. In this presentation, we will use this database to connect the DSPACE records to other repositories and identify the collaboration networks across these platforms.

We use Neo4j graph database to hold the connections and process the links between publications, researchers, datasets and grants. Furthermore, we export this information to the Gephi graph visualiser to highlight the connections across the domain, national and international repositories.

Presentation Content

The process consists of three main steps: (1) exporting data from DSPACE to a Neo4j graph database (2) linking the graph database with ORCID, funding information and other open access repositories using the Research Graph synthesis function. (3) exporting the graph to Gephi graph visualizer and highlight the clusters. This process has been illustrated in Figure 2.

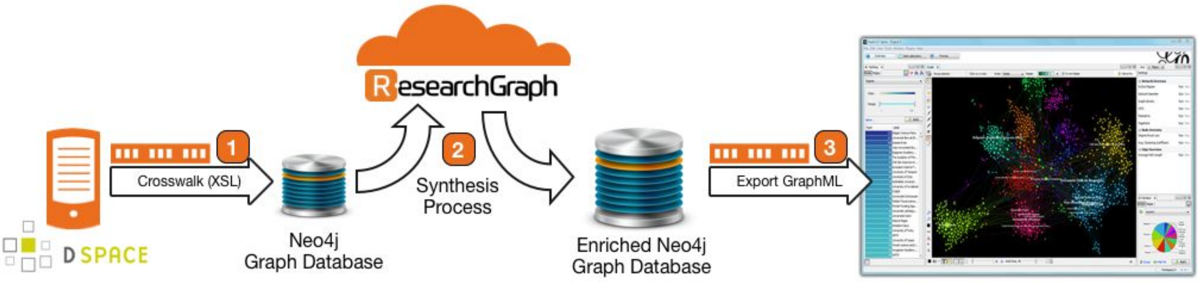

Figure 2: The process of transforming DSPACE records into a collaboration network and visualising the results in the Gephi graph visualizer.

In addition, we demonstrate how to use the Cypher query language for the following queries:

- Finding a datasets or publication by DOI, finding researchers by ORCID and searching for grants and research projects by PURL
- Discover highly connected registry objects such as high impact publications
- Find the shortest path between two registry objects such as the set of connections between two researchers or two research grants

Example of these queries are:

```
match (n:researcher) where n.orcid='0000-0002-7875-2902' return n
```

In the presentation, we will use DSPACE Cambridge open access repository[4] as the input data, and the outputs will be a Neo4j graph database and a Gephi project file. This information will be accessible to the conference participants for further exploration.

Additional References:

- Research Data Switchboard code: https://github.com/rd-switchboard
- Research Graph schema: https://github.com/researchgraph/schema
- Neo4j Cypher refcard: https://neo4j.com/docs/cypher-refcard/current
- Learn how to use Gephi: https://gephi.org/users

[4] https://www.repository.cam.ac.uk