Cadernos BAD

Using persistent identifiers to track PhD outcomes

Frances Madden
The British Library, UK
frances.madden@bl.uk
Vasily Bunakov
Science and Technology Facilities Council, UK
vasily.bunakov@stfc.ac.uk

Abstract

The British Library and STFC are working together to connect metadata relating to PhD theses to use of research facilities provided by STFC. Through the creation of a graph which connects metadata across EThOS, the UK national e-theses repository, the STFC insitutional repository ePubs and the Diamond Light Source repository, links between the theses, authors and instiutions have been made. This pilot work has the potential to be extended to enhance the metadata stored within EThOS which is administered by The British Library. This work presents an exciting possibility for other institutions as this graph will present a method to enable the bi-directional updating of metadata held across different systems.

Keywords: Persistent Identifiers, DOI, research facilities, PhD supervisors

Introduction

National libraries and research funders have different but complementary interests in capturing quality metadata relating to e-theses. The British Library and Science Technology Facilities Council (STFC) have been working together to express and enrich thesis metadata through the use of persistent identifiers (PIDs).

The British Library manages EThOS (https://ethos.bl.uk/) the national repository for theses in the UK. It comprises over 500,000 records of theses in the UK dating back to 1758. Approximately half of these, 270,000, are available for full text download either directly through EThOS or through universities' institutional repositories. The database is populated through harvesting records from universities' institutional repositories.

STFC is a funder of research but also a provider of research facilities such as the ISIS Neutron and Muon source facility. It also funds the Diamond Light Source in conjunction with the Wellcome Trust. A large proportion of the work undertaken at these facilities is by PhD students, one estimate is that it is 70% of the experiments are undertaken by PhD students. STFC has an institutional repository, ePubs (https://epubs.stfc.ac.uk/) and the Diamond Light Source has its own repository Diamond DB (http://publications.diamond.ac.uk/) both of which contain publications which have been produced by STFC and Diamond staff or visitor scientists. These repositories include PhD and, occasionally, master theses by young researchers who conducted their experiments on Diamond (http://www.diamond.ac.uk/) and two other STFC facilities: ISIS Neutron and Muon Source (http://www.isis.stfc.ac.uk/) and Central Laser Facility (http://www.clf.stfc.ac.uk/).

FREYA (https://www.project-freya.eu/en) is a European Commission funded project which aims to extend and integrate PID infrastructure within Europe and globally. The partners involved are a combination PID service providers and disciplinary partners including STFC and the British Library. A major output of FREYA, the PID Graph, relates to the connection of different types of PIDs together to enable multi-step relationships to be discerned. Within the context of FREYA each partner is using the concept of the PID Graph to map their usage of PIDs and identify where PIDs can be added to entities to enable interoperability between systems and reduce double keying of information.

The British Library plans to enrich the metadata held within EThOS using PIDs when it is migrated to a new repository system which is in development. The enhanced metadata will be created and incorporated into the EThOS metadata at the point of migration. The repository project is now in its pilot phase, with the second phase, which includes the migration of EThOS, starting in January 2020. STFC have begun to enrich metadata within their own repository through analysing the connections between data in the ePubs repository, EThOS and data from ResearchFish, a UK research reporting tool. A smaller number or records have also been ingested from Oxford Research Archive (https://ora.ox.ac.uk/) to measure the potential of using university repositories for metadata enrichment.

Literature Review

Graphical databases are widely used by many technology companies but within the research sphere particularly, there are a number of initiatives using them. The possibilities presented by connecting different aspects of research information are being explored through a number of initiatives such as Scholix (http://www.scholix.org/), Research Graph (http://researchgraph.org/), OpenAIRE Research Graph (Manghi & Bardi, 2019) amongst others. The work of the Research Graph has been explored in Aryani et al. (2018) and the Scholix initiative and its implementation have been described in Cousijn et al. (2019).

A blog post also describes the concept of the PID graph as it is adopted in FREYA

(Fenner & Aryani, 2019). DataCite have also created a GraphQL API which allows for several APIs including the DataCite/Crossref Event Data Service, ORCID API, DataCite and Crossref APIs all be queried at once (Fenner, 2019).

Methodology

The methodology of the FREYA project as a whole has been to gather user stories to establish the information, which the community would find beneficial to have. A sprint in 2018 gathered over 70 user stories from across the FREYA partners, their communities, FREYA ambassadors and also through several user events such as the DI4R conference. These user stories all had a standard format: As a <role>, I would like to <capability> so that <benefit>. This enabled comparison across the user stories and to identify for which types of entities there was a desire for PIDs.

Of these user stories, there were two related to PhD theses. The first contributed by STFC, 'Tracking down PhD studentship outcomes, beneficiaries, co-funders and supporters'. There are numerous questions which can be answered through this:

- What organisations benefited from the PhD after it was completed, e.g. through employing the student
- How many studentships funded result in a completed thesis
- What artefacts can be identified which contributed to the thesis or its outcome

(https://github.com/datacite/freya/issues/69)

The British Library identified a user story, 'Linking people and research outputs to theses' which included a wish amongst potential PhD students to see an overview of the work of previous students' of a particular supervisor and what they went on to do in their careers. EThOS provides a useful platform to view these relationships as it can offer the opportunity to review supervision across different institutions. There is also a requirement from the British Library to be able to move seamlessly between the thesis and its related artefacts such as datasets and articles (https://github.com/datacite/freya/issues/35).

Between them these user stories identified a need for PIDs for the following entities:

- Theses
- People
- Organisations
- Datasets
- Articles
- Funders
- Grants

And that there needed to be capability to be able to link all of these entities together.

STFC made a pilot implementation of the integrated metadata from the various

repositories mentioned in the Introduction. The resultant integrated metadata is represented as a labelled-property graph in a neo4j database (https://neo4j.com/). Initial relations have been made between the nodes that represent repository records; these nodes have been further broken down to new types of nodes that represent Persons (Authors), Organizations (Universities) and Facilities. The University nodes have been marked up with organizational identifiers from Global Research Identifier Database (GRID, https://grid.ac/).

To demonstrate the potential for the expansion of the resultant graph onto domain-specific research data, a few nodes were created to represent chemical information in the Imperial College London Spiral repository (https://spiral.imperial.ac.uk/) and in the ChemSpider portal (http://www.chemspider.com/) which is supported by the Royal Society of Chemistry. These subject-specific extensions are a work in progress that requires effort beyond the boundaries of the FREYA project. The ultimate goal of this work would be to keep minting clearly identified chemical data records in university repositories and then allow PhD students to connect their theses to these records, rather than attempting the difficult task of harvesting chemical data from the theses full texts retrospectively.

Results

The pilot graph database incorporates metadata harvested from five disparate sources: EThOS, Researchfish, ePubs, Diamond DB and Oxford Research Archive, as well as identifiers for organisations from GRID. The graph contains over two thousand labelled nodes with over three thousand relations that allow various requests using the CYPHER query language (https://neo4j.com/docs/cypher-manual) and a simple visualization in a web browser.

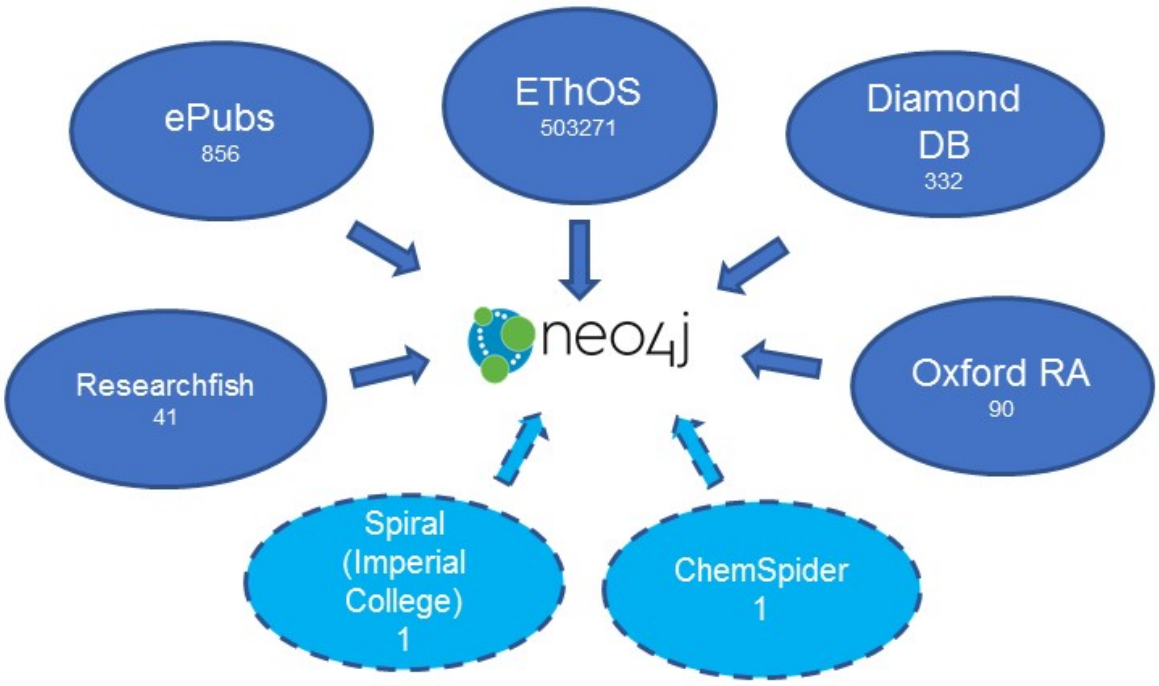

Figure 1: Various data sources have been used to populate the graph database

Dark blue refers to automated ingestion, light blue refers to manual ingestion to illustrate specific cases. The numbers indicate the numbers of records ingested.

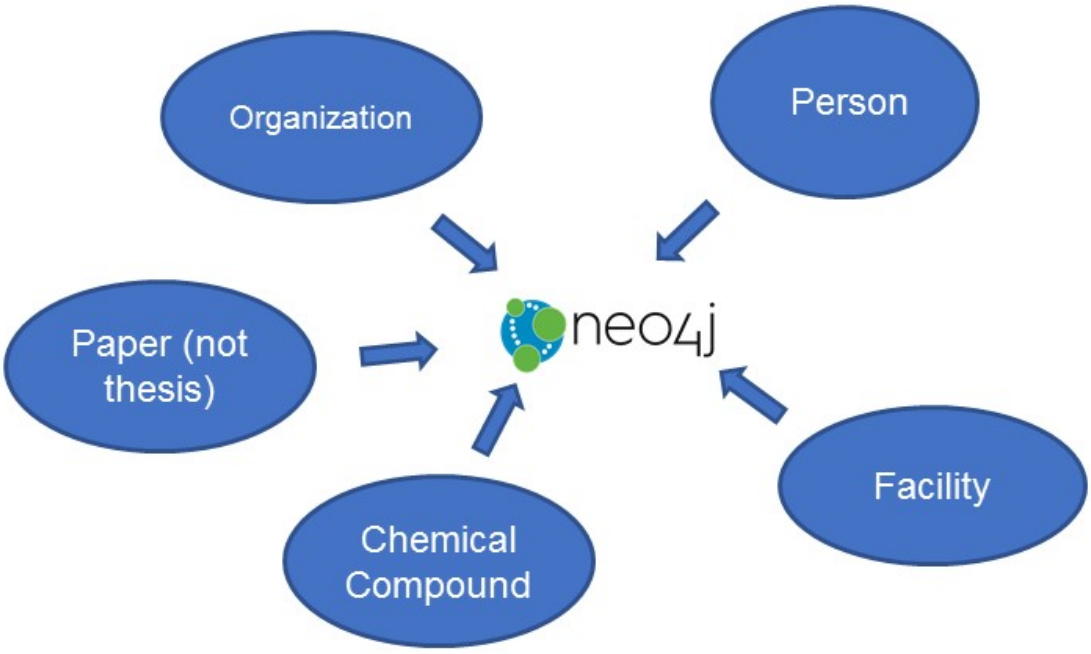

Figure 2: Inputs into database

More node types of relating to the sources named in the figure were created in the graph, with PIDs assigned where they were available.

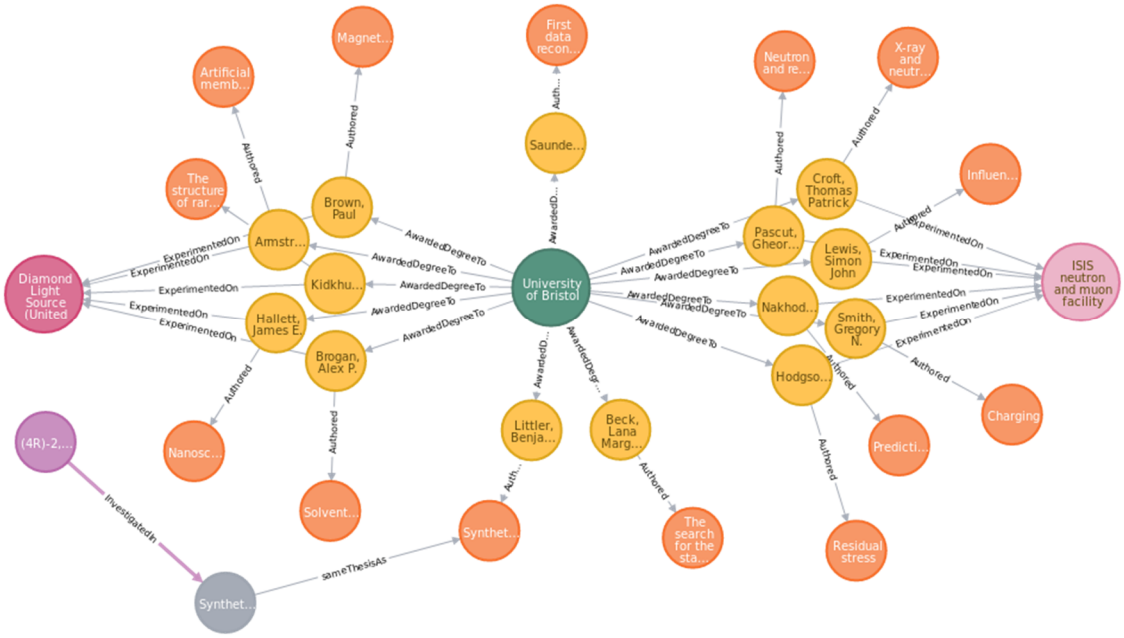

Figure 3: Example of subgraph that represents University of Bristol PhD researchers, facilities they used and their research outcomes

The green node represents the University of Bristol, the yellow nodes are authors, the orange nodes are PhD theses and the pink nodes are the Diamond and ISIS research facilities. The grey and purple nodes represent a sample from the ChemSpider database, for a PhD thesis record and a chemical compound data record respectively.

Discussion

As a research funder and a provider of research facilities, STFC can now gain a much

greater oversight of the effectiveness of the research funding it provides and a better means to track the outcome of the research and relate it to other publications from the author to enhance research discovery and impact activities.

This initial work presents an exciting opportunity for The British Library to enhance the metadata in EThOS. The current EThOS platform, a highly customised ePrints repository cannot support one to many relationships, e.g. it is not possible to capture PhD supervisor's ORCID IDs because generally there is more than one supervisor and it is not possible to relate the ID to the supervisor. The graph also offers many possibilities of validation and metadata enhancement and the comparison with other repositories is promising to improve the metadata referred to within EThOS.

This initial work has also assisted with the identification of areas for improvement of metadata within the EThOS database. For example, there were 454 instances where the funder field in the EThOS metadata was blank but STFC had records of funding the thesis. It was also possible to match several variations of the Funder name against the Funder field in order to provide a more comprehensive picture of funding information. Science and Technology and Facilities Council was represented in several different ways:

- Science and Technology Facilities Council
- Science and Technology Facilities Council (STFC)
- Science & Technology Facilities Council
- STFC
- Science and Technology Facilities Council (Great Britain) (STFC)

It can reasonably be assumed that similar variations in funder names can be seen across other funders. Also because it is not possible to duplicate the field it is also likely that there would be instances where two funders are included in a single field and these would need to be rationalised. It is hoped to incorporate funder IDs as well as grant IDs within the metadata of EThOS as part of a planned replatforming process due to be take place in 2020/21.

By incorporating more identifiers in the EThOS metadata it will be possible to standardise the metadata with less effort. Some fields within EThOS are already standardised such as 'Current Institution' which refers to the institution which holds the thesis. This is a controlled list, but it is hoped to include this as an API call against an organisational identifier in the future. We will also attempt to standardise and include PIDs with some fields which are currently free text such as the Awarding Institution field. This field relates to the institution which awarded the thesis, generally this will correspond with Current Institution field but where the institution name has changed or if there is another awarding body such as Council for National Academic Awards (CNAA), this field will be different. We would hope to ensure that ISNI records associated with the institution include the former and alternate names of the institution.

It is also hoped to be able to support new types of grant identifiers, DOIs, as they emerge and are developed within the EThOS records (Brown, 2019).

The extension of the graph with information from domain-specific repositories, such as chemical information, has been a smaller part of the project so far, but it has a clear potential for growth if supported by the improved practices of research data curation in research-intensive universities. It will also augment the discoverability of the thesis if the backward linking can take place.

In addition, this process has identified areas in which the institutional repositories have better quality metadata than EThOS. For example, there were some theses that were found within the facility repositories as being supported by the facility; however they had different funder information in the Funder field in EThOS. The plan is that this graph will be expanded with data from other institutional repositories and funder databases which would allow for the enhancement of the metadata and the creation of stronger connections linked using PIDs requiring less data validation. By being able to harvest more metadata from institutional repositories and format it consistently, more connections within the graph can be created.

This work presents an exciting possibility for other institutions as this graph will present a method to enable the bi-directional updating of metadata held across different systems. All outputs from this work will be made available openly to support other institutions who are keen to undertake similar enhancement exercises.

Conclusion

This work presents an exciting first step on the road to enhanced and interconnected metadata. The creation of graphs such as this, particularly where they are based on persistent identifiers presents an opportunity to understand the research landscape, particularly relating to theses, more thoroughly. Through continuing this work beyond the life of the FREYA project, as EThOS is migrated to a new repository, the utility of EThOS as a service will increase.

References

Aryani, A., Poblet, M., Unsworth, K., Wang, J., Evans, B., Devaraju, A., Hausstein, B., Klas, C.P., Zapilko, B., Kaplun, S., (2018) A Research Graph dataset for connecting research data repositories using RD-Switchboard. *Scientific Data.* 5. 180099. https://doi.org/10.1038/sdata.2018.99

Brown, J. 2019. Funders and infrastructure: let's get building. Crossref Blog. https://www.crossref.org/blog/funders-and-infrastructure-lets-get-building/

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E. and Simons, N., 2019. Bringing Citations and Usage Metrics Together to Make Data Count. Data Science Journal, 18(1), p.9.

http://doi.org/10.5334/dsj-2019-009

Fenner, M., & Aryani, A., (2019) *Introducing the PID Graph*. FREYA Blog. https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

Fenner, M., (2019) The GraphQL API is open for (pre-release) business. DataCite Blog. https://doi.org/10.5438/qab1-n315

Manghi, P., & Bardi, A. (2019, March). The OpenAIRE Research Graph - Opportunities and challenges for science. Zenodo. http://doi.org/10.5281/zenodo.2600275