publications

MDPI

Article

Integration of a National E-Theses Online Service with Institutional Repositories

Vasily Bunakov [1,*] and Frances Madden [2]

[1] STFC UK Research and Innovation, Harwell Campus, Didcot OX11 0QX, UK
[2] The British Library, London NW1 2DB, UK; frances.madden@bl.uk
* Correspondence: vasily.bunakov@stfc.ac.uk

Received: 14 February 2020; Accepted: 30 March 2020; Published: 9 April 2020

Abstract: We present an information resource prototype that was developed by the FREYA project for the integration of a national e-thesis service and institutional repositories supported by a large national laboratory. The integration allows us to mutually enrich the metadata in the e-thesis service and institutional repositories with new entities and attributes, and can offer novel ways of reasoning over research outcomes that are supported by direct funding and funding-in-kind by large research facilities. The integrated information resource can be presented as a labeled-property graph for its exploration with a declarative query language and visualizations. We emphasize the role of persistent identifiers (PIDs), including for entities that are currently not necessarily or not consistently assigned PIDs.

Keywords: e-theses; institutional repository; research facility; persistent identifier; EU project; PID

1. Introduction

A national e-thesis service and repository supported by national laboratories present different parts of the repository spectrum. A strong commitment to Open Science allows these disparate sources of research information to be integrated and offers novel ways of measuring research outcomes. The open repositories movement is both a facilitator of this integration and a beneficiary, as the integrated resource presents opportunities for the development of new services that can be used by multiple parties including but not being limited to repositories that contributed to the integration.

EThOS [1] is the UK's national thesis service and provides an aggregated record of all doctoral theses awarded by UK universities dating back to 1787. Of the more than half a million titles listed in this database, approximately half (two hundred and seventy thousand) are available for free digital download, either from the EThOS database itself or via links to universities' institutional repositories.

As part of its contribution to the FREYA project [2], the British Library has worked with the Science and Technology Facilities Council (STFC) to link the EThOS metadata with that in STFC repositories. The ePubs institutional repository [3] and Diamond Light Source publications database [4] have been selected as the institution-specific resources for the integration of their selected records with EThOS. ePubs cover publications by STFC staff (including those produced jointly with visitor scientists on ISIS Neutron and Muon Facility and Central Laser Facility) and the Diamond database covers publications mostly produced by visitor scientists who perform experiments on the Diamond synchrotron co-owned by STFC and Wellcome Trust.

FREYA aims to build a graph which connects different entities in the research lifecycle through persistent identifiers. The project team has focused on addressing a range of "user stories" that are a specific format of user requirements identified by partners and stakeholders during a sprint conducted in summer 2018. This work aims to address the user story: "As a funder, I want to track down the outcomes and beneficiaries of PhD studentship awards that I granted" [5].

It is a stance of the FREYA project that any integration of information resources acquired from different repositories can be represented as a graph with nodes clearly identified by one or more persistent identifier (PID). Figure 1 outlines the research and research reporting environment that relate to the STFC-sponsored research and the underpinning metadata which is fed through to EThOS, taking the Diamond Light Source as a particular example of a large-scale facility. Some of the entities in this diagram are already typically assigned with PIDs, and some of the entities may require new PID types that are not yet fully adopted or developed. The level of the actual adoption of PID types in a particular research community is an important factor for making a decision about the PID types use in the integrated information resource. As an example, the level of adoption of persistent identifiers for persons is low in the repositories involved, so assigning PIDs to all Researcher or PhD Student entities would be laborious and could not be performed with the limited resources we had. Yet experiments on large-scale facilities quite often have persistent identifiers associated with them; this practice just has to be propagated from the facilities that already use it to other facilities within the same organization. In regards to the organizations involved, they are not too numerous, so it is feasible to assign persistent identifiers to nearly all of them. The level of adoption of PIDs for publications and data is already fairly good, so the integrated resource we are building can rely on reusing these PIDs rather than on assigning them as in the case of organizations.

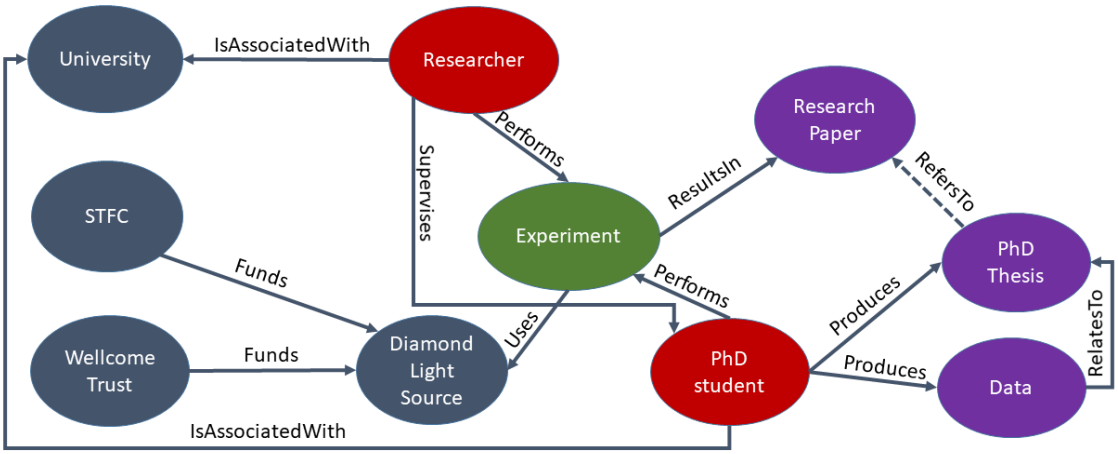

Figure 1. A schematic of PhD research performed on the Diamond Light Source [6], and the artefacts produced from it.

The information resource that properly reflects this operational environment with all its dependencies will be a commonly-built and commonly-used knowledge graph enriched with PIDs to enable consistent and easy matching and interoperability between entities. Building such an integrated information resource presents conceptual and technological challenges that are discussed in this paper, along with the benefits of the integrated resource use.

2. Metadata Sources and Methods of Their Integration

Apart from the aforementioned EThOS service and the two institutional repositories, we integrated records of science and reference information from a few other sources:

- University of Oxford Research Archive [7]
- Spiral repository of Imperial College London [8]
- ChemSpider service by the Royal Society of Chemistry [9]
- GRID research identifiers database [10]

A small number of organization-specific records managed by the Researchfish service [11], which is a popular CRIS (Current Research Information System) solution in the UK, were tried for the

integration, but discovering the true potential of CRIS systems for the integrated resource required a continuous access to CRIS records, as well as the development of reasonable techniques for matching them with other resources used. This was beyond the scope of the information integration exercise that we planned; also we did not intend to semantically align the resulted integrated resource with popular models such as the Common European Research Information Format (CERIF) [12], as the technology we used for the integrated resource exploration was based on the labelled-property graphs model [13] and on requests formulated opportunistically in a declarative query language, rather than relied on a specific metadata model. For future integrations though, if they involve substantial numbers of CRIS records, the use of CERIF model for matching CRIS records with other records can be a natural choice.

The purpose of the additional integrations that we performed was to evaluate the application of the same integration methods as for the main information sources, and further enrich the resultant knowledge graph with more persistent identifiers and additional metadata.

PIDs such as DOIs were used where possible for the records integration, and fuzzy matching techniques were used in other cases. For fuzzy matching, we selected theses within a limited timeframe of the thesis's publication date (plus or minus one year from the year recorded by EThOS) and measured the distance between the respective groups of thesis titles, which proved to be a reasonable technique for finding thesis counterparts across different information sources. Our approach to records matching differed from the one adopted by the Brazilian Digital Library of Thesis and Dissertations (BDTD) [14], who used both titles and author names for finding correspondences. The Levenshtein distance [15] was used as a measure of the titles proximity, with the threshold of making decision about "sameness" of the title consciously selected at 15 symbols after a few dozen record matching experiments, so that we could capture all true positive matches between title strings and did not have too many false positives; then the remaining false positive matches were filtered out by manual checks.

A community edition of the Neo4j graph database [16] was used for the records management and records matching using the APOC [17] implementation of the Levenshtein distance. The advantage of using the Neo4j graph database, which is a database that is well-optimized for managing and querying labeled-property-graphs, for the records integration was that it did not require the schema to be designed in advance, but rather allowed us to acquire and integrate records from different sources opportunistically and incrementally.

The database was populated by, first, the creation of nodes for each record ingested and "sameness" relationships between the nodes. Then, certain attributes of these initial nodes were converted into new nodes linked with new relationships. Table 1 contains descriptions of the relationships initially created in the database. These relationships in part reflect the operational model on Figure 1 but could be further expanded with new ones, e.g., to reflect on the supervisor-to-PhD relationship or on references from a PhD thesis to published research papers.

Table 1. Relationships created in the graph database and their explanation.

Relations Created	Relations Meaning
AwardedDegreeTo	Connects a university and student awarded with the degree
Authored	Connects a student and a thesis that she authored
SameThesisAs	Connects different manifestations of the same PhD thesis
ExperimentedOn	Connects a student and a research facility she experimented on
Sponsored	Connects a student and a funder who sponsored her

For the exploration of the relationships created, we used an out-of-the-box graphical user interface of Neo4j that allows queries in Cypher language [18] and can display query results in the tabular or graphical formats. Overall, the graph database proved to be a productive tool for metadata integration and visualization.

3. Features of the Resultant Graph

The graph resulting from the records integration allows visual reasoning over research conducted by PhD students on large research facilities. The subgraph in Figure 2 presents research conducted by Imperial College London PhD students on two facilities: Diamond Light Source [6] and ISIS neutron and muon source [19].

Figure 2. The subgraph with Imperial College London PhD students who experimented on two research facilities. Imperial College London is indicated by the central node, the research facilities are represented by the nodes on the left and on the right and the students by the smaller nodes in between. The two nodes at the bottom illustrate two students who used both facilities in their research.

The grouping of researchers experimenting on the different facilities is evident, with two researchers at the bottom of the diagram who conducted experiments on both facilities. The university, Imperial College London, is indicated by the central node and the facilities are indicated by the larger nodes on the left and on the right.

As an add-on to the graph, we also looked in the possible connections between the thesis records in the British Library and STFC collections with the National Compound Collection [20] that is now a part of the ChemSpider service [9]. This collection contains numeric and structural data extracted from a few hundred representative PhD theses supplied by chemistry departments of UK universities. The integration of the National Compound Collection records in the knowledge graph allows connections between publications and data. On the one hand, it enriches the existing theses records, and on the other, it increases the provenance of data records in ChemSpider and gives them valuable context supported in some cases by the availability of the full text of the thesis.

Most of the National Compound Collection records were successfully integrated in the graph and matched with the EThOS records, yet no matches were found between the National Compound Collection and the STFC repositories' records. This effect can be attributed to the limited size of the National Compound Collection as PhD students are a prominent category of visitor scientists at the STFC facilities, over 50% of visitors on some facilities in some years, and STFC sponsors PhD research through funding studentships [21]. However, the results of the ChemSpider integration can be seen in Figure 3, and they contribute to the vision of a common graph as a research information infrastructure with various uses, not limited to the initially envisaged case of matching records between institutional repositories and EThOS. In particular, matching EThOS records to ChemSpider and other data services such as university data repositories may bring richer information context to the records of science and raise the visibility of both EThOS and connected resources.

Figure 3. A subgraph demonstrating the ChemSpider integration. It shows the University of Bristol (central node), PhD students where some of them experimented on two research facilities (larger nodes on the left and on the right) and the thesis nodes connected to students with the "Authored" relationship. The nodes in the bottom left of the diagram indicate a thesis from ChemSpider with "same" relationship to the EThOS thesis and with the chemical compound node on the far left.

The natural limitation of manual data extraction from theses is exemplified by the National Compound Collection and suggests that better data management practices could be introduced in the UK universities' research departments. The capture and recording of research data could happen during the actual PhD research, not after it, removing the severe cost implications and the issues of data quality by extracting from full-text files. This consideration has led to productive discussions, with at least one university now introducing new practices for PhD research data curation.

Data assets are going to be captured from the very start of a PhD research assignment with persistent identifiers (DOIs) assigned to these assets. This means data will be ready for citation in the PhD thesis itself and then can be included in reference management software automating linking between PhD thesis records and data records, with all the aforementioned mutual benefits for the thesis repositories and data repositories.

4. Applications of the Resultant Graph

The integrated knowledge graph can be considered a common research information infrastructure that can be used by the British Library and STFC, also potentially by other parties such as the universities where the theses originate, or by data services providers such as the Royal Society of Chemistry with the aforementioned ChemSpider service. Potential uses and applications of such a common infrastructure include:

- Gap analysis for repositories' coverage
- Disambiguation between entities
- Records enrichment
- Common machine interfaces (APIs) exploited by contributors into the common graph, as well as by third parties
- Records connection across repositories

4.1. Gap Analysis for Repositories' Coverage

With respect to the gap analysis use case, the records matching exercise performed highlighted over four hundred records in the EThOS repository with empty "Sponsor" metadata where STFC in fact sponsored the PhD research via monetary funding (postgraduate studentships) or via grants-in-kind

by offering research time for experiments on large research facilities. This can be attributed to the fact that only thirty thousand of the half a million EThOS records contain funder information due to university repositories not capturing this information or it not being harvested into EThOS. Furthermore, over a hundred cases were discovered where the "Sponsor" metadata element is not empty in EThOS but STFC or Diamond Light Source were not mentioned as research sponsors, despite evidence of it discovered through the EThOS records matching with STFC and Diamond repositories. Altogether, about six hundred cases were identified where STFC and Diamond profiles as research sponsors could be raised by the proper attribution to them in the EThOS records.

4.2. Entity Disambiguation

For the entity disambiguation use case, it was noted that STFC can be designated by a few different names and abbreviations in EThOS as it is a free text field; the records matching exercise performed allows harmonization of the organization naming across EThOS records and thus again improves the attribution of PhD sponsorship. The use of organizational identifiers such as GRID or the recently launched Research Organization Registery (ROR) [22] could create relationships between variant names of the same organization. The integration of the authority control for PhD research host organizations and PhD research sponsors would be a valuable addition to the existing information systems; it would allow the PIDs assignment during the records creation.

4.3. Records Enrichment

With respect to the records enrichment use case, we mixed up the author names coming in different formats from different sources, so that the Author attribute of the graph node contains multiple variants of the same name. This improves findability of theses when performing full-text search using an author name.

4.4. Developing Machine Interfaces

For the machine interfaces use case, a straightforward solution could be implemented using a standard plugin that creates a GraphQL [23] interface for the Neo4j graph database. However, owing to the metadata coming from disparate sources so that not all nodes of similar types bear the same metadata, we may not want to expose all of the metadata available in the graph through the same interface. The more sophisticated approach was considered in [24] and can be applied to the graph database described here. Another aspect of the work from the FREYA project has resulted in DataCite building a GraphQL API which allows for several APIs to be queried simultaneously including that of DataCite, ORCID and ROR and will soon include the full Crossref API too.

4.5. Connecting Records Across Repositories

For the use case connecting records across repositories, it seems wise not to implement it in an ad-hoc manner but base it on a reasonable machine interface to the graph when it is ready.

4.6. Other Use Cases

There may be other uses of the knowledge graph built by the parties that contributed metadata into it, as well as by the third parties who did not contribute to the graph but may benefit from its content and machine interfaces. The proper stance towards this knowledge graph would be considering it a new multi-purpose research information management infrastructure. It is in the nature of any infrastructure that it is used not only for the cases initially devised when it was constructed, but by other uses and applications that were not or simply could not be considered initially owing to the changing technological and cultural circumstances.

5. Future Plans

Further plans for the created knowledge graph include its enrichment with PIDs, specifically for more organizations and for large-scale instruments in facilities [25], also building machine and user interfaces to enable easier access and further utility of the graph. EThOS will move to a new platform in 2020/21 and the migration will offer an opportunity to augment the metadata with this additional information.

STFC considers the inclusion of references in the institutional repositories (theses records in them) to the EThOS records; such references can be a part of the STFC Open Science Portal prototype that will include references to the quality external repositories beyond the organization walls.

The above-mentioned effort of a UK university to modernize its practices around capturing data assets associated with PhD theses warrant its sharing with other universities, for which a dedicated project focused on the best practices of managing PhD research artefacts in universities is likely to be required.

Another interesting question on its own could be what governance model would be appropriate for the knowledge graph as a common research information management infrastructure, to ensure its productive expansion and its long-term sustainability. These considerations can, again, contribute to the definition of a collaborative project with a focus on best practices supported by a reasonable technology.

Author Contributions: V.B. conceived the idea of the graph connecting institutional repositories and EThOS, and developed a database. F.M. supplied data and requirements. Both authors participated in the planning of the study and in the writing of this article. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by funding from the Horizon 2020 FREYA project, Grant Agreement number 777523. The views expressed are the views of the authors and not necessarily of the project or the funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. EThOS E-Theses Service. Available online: https://ethos.bl.uk/ (accessed on 6 April 2020).
2. FREYA: Connected Open Identifiers for Discovery, Access and Use of Research Resources. Available online: https://www.project-freya.eu/ (accessed on 6 April 2020).
3. STFC UKRI Institutional Repository. Available online: https://epubs.stfc.ac.uk/ (accessed on 6 April 2020).
4. Diamond Light Source Publications Database. Available online: http://publications.diamond.ac.uk/pubman/searchpublicationsquick (accessed on 6 April 2020).
5. Tracking Down PhD Studentship Outcomes, Beneficiaries, Co-Funders and Supporters. Available online: https://www.pidforum.org/t/tracking-down-phd-studentship-outcomes-beneficiaries-co-funders-and-supporters/99 (accessed on 6 April 2020).
6. Diamond Light Source. Available online: https://www.diamond.ac.uk/ (accessed on 6 April 2020).
7. University of Oxford Research Archive. Available online: https://ora.ox.ac.uk/ (accessed on 6 April 2020).
8. Spiral Repository of the Imperial College London. Available online: https://spiral.imperial.ac.uk/ (accessed on 6 April 2020).
9. ChemSpider Service by the Royal Society of Chemistry. Available online: http://www.chemspider.com/ (accessed on 6 April 2020).
10. GRID Research Identifiers Database. Available online: https://grid.ac/ (accessed on 6 April 2020).
11. Researchfish Service. Available online: https://researchfish.com/ (accessed on 6 April 2020).
12. CERIF (The Common European Research Information Format). Available online: https://www.eurocris.org/cerif/main-features-cerif (accessed on 6 April 2020).
13. Labeled-Property Graph. Available online: https://en.wikipedia.org/wiki/Graph_database#Labeled-property_graph (accessed on 6 April 2020).

14. Lautaro, J.M. Improving LA Referencia Metadata by Linking Research Profiles to Repositories: The Case of the Brazilian Digital Library of Thesis and Dissertations (BDTD) and the Lattes CV Platform. Available online: https://www.conftool.net/or2019/index.php/Paper-P3D-435Matas_a.pptx?page=downloadPaper&filename=Paper-P3D-435Matas_a.pptx&form_id=435&form_version=final (accessed on 6 April 2020).
15. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
16. Neo4j Graph Database. Available online: https://neo4j.com/ (accessed on 6 April 2020).
17. Neo4j APOC Library. Available online: https://neo4j.com/developer/neo4j-apoc/ (accessed on 6 April 2020).
18. Cypher Query Language. Available online: https://neo4j.com/developer/cypher-query-language/ (accessed on 6 April 2020).
19. ISIS Neutron and Muon Source. Available online: https://www.isis.stfc.ac.uk/ (accessed on 6 April 2020).
20. Andrews, D.M.; Broad, L.M.; Edwards, P.J.; Fox, D.N.; Gallagher, T.; Garland, S.L.; Sweeney, J.B. The creation and characterisation of a National Compound Collection: The Royal Society of Chemistry pilot. *Chem. Sci.* **2016**, *7*, 3869–3878. [CrossRef] [PubMed]
21. STFC-Funded PhD Students. Available online: https://stfc.ukri.org/funding/studentships/studentship-terms-conditions-guidance/statistics-and-questionnaires/stfc-funded-phd-students/ (accessed on 6 April 2020).
22. ROR. Available online: https://ror.org/ (accessed on 6 April 2020).
23. GraphQL Query Language. Available online: https://graphql.org/ (accessed on 6 April 2020).
24. Bunakov, V. Metadata Integration with Labeled-Property Graphs. In *Metadata and Semantic Research*; Garoufallou, E., Fallucchi, F., Eds.; Springer: Cham, The Netherlands, 2019.
25. Bunakov, V. Metadata for Large-Scale Research Instruments. In *Metadata and Semantic Research*; Garoufallou, E., Sartori, F., Eds.; Springer: Cham, The Netherlands, 2018.