Open Science Portal Based on Knowledge Graph

Vasily Bunakov[0000-0003-3467-5690]

STFC UKRI, Harwell Campus, Didcot OX11 0QX, United Kingdom

Abstract. The work outlines an ongoing effort of building the Open Science Portal for STFC UKRI based on the knowledge graph assembled from various records of science. The graph is a result of interplay across the records that the organization maintains and the external quality records in reference databases supported elsewhere. The twofold role of persistent identifiers is illustrated as, on one hand, facilitators of building the knowledge graph and, on the other hand, as a specific means of the information enrichment within the graph once it is built. The business case and the implementation detail of the Portal is reported, as well as the projections for its further development and possible applications. The work is one of the outcomes of FREYA which is a Horizon 2020 project developing the persistent identifiers infrastructure and recommendations for Open Science.

Keywords: Open Science, knowledge graph, persistent identifiers, EU project

1 Introduction

The FREYA project [1] is developing the infrastructure for persistent identifiers (PIDs) and recommendations for their effective use as part of the European and global environment for Open Science. PIDs landscape is rich nowadays and includes identifiers for digital objects such as research publications or datasets, as well as for real-world entities such as people or organizations. The grand vision of FREYA is the "PID Graph" that creates relationships across objects and entities with PIDs and provides a basis for new services within research disciplines and across them.

The FREYA partners develop pilot applications that exploit parts of the PID Graph relevant to their respective research disciplines, also these applications in turn provide contributions to the PID Graph by opening up the information assets within the organizations using a common set of recommendations and where reasonable common technological solutions, too.

This work outlines a particular pilot application by STFC UKRI [2] in support of the Open Science agenda. First, the business case for the Open Science Portal is described, then the detail is given about data sources, technology used and examples of the Portal content, then plans for the further development of the Portal are discussed.

2 Business Case for STFC Open Science Portal

STFC (Science and Technology Facilities Council) is a part of UKRI (UK Research and Innovation) [3] which is the national funding agency investing in science and research in the UK. STFC is a funder of science and of the post-graduate education in the UK, and a funder for the research conducted by the UK scientists on large-scale scientific instruments abroad. STFC is also a research organization that operates or co-owns large-scale scientific instruments (facilities) and high-performance computing across a few UK locations, granting access to them to the UK and overseas visitor scientists who conduct their own experiments.

All the mentioned streams of STFC funding and funding-in-kind (facility time and computation time awarded) result in research artefacts such as journal papers and preprints, PhD theses, data and software. Tracking down these artefacts back to the instruments, organizations and people involved is important for the evaluation of STFC role in a number of research fields such as biomedicine, chemistry, materials science, engineering, particle physics, astronomy, also of its role in higher education.

Research artefacts that have been produced with the support of STFC funding or funding-in-kind are reflected in records of science, e.g. bibliographic information for journal papers, or records of data deposited in certain reference databases. These records of science and the artefacts behind them are handled by a variety of information systems within STFC, also some well-curated STFC-related records of science are managed by external providers as parts of their larger collections (examples being crystallography or biomedical databases or national services for PhD theses).

To fully account for STFC funding and funding-in-kind, there is a need to systematically collect and manage these records of science, including the discovery of connections across them. Apart from this objective of having a better accountability for public spending on science, there is an important aspect of knowledge preservation and knowledge discovery in the spirit of Open Science that encourages and supports reuse of research outcomes beyond the point of their origin. Open Science contributes to new research by other research organizations and individuals, raises the public awareness of science and its applications, also the organization itself can benefit from the more explicit and context-rich representation of its records of science with a single point of entry for their discovery.

Accountability and Open Science aspects are interrelated and can be supported by the same research information infrastructure that STFC are now building, with the first prototype of such infrastructure receiving support of FREYA project and having the focus on aspects that are specifically relevant to FREYA scope, i.e. persistent identifiers as a means of knowledge discovery and integration.

This new research information infrastructure is provisionally coined with the name of STFC Open Science Portal. This is going to be a publically available resource for the discovery of records of science that have been produced with the support of STFC funding or other flavours of sponsorship such as facility time awarded to visitor scientists. The records of science include information about research outcomes in any form (publications, data, etc.), records of STFC funding or other flavours of sponsorship,

organizational context of STFC-supported research, as well as research attribution to particular large-scale instruments (facilities and their beamlines).

The Portal is going to be a multi-purpose research information infrastructure that can be used by STFC and by external stakeholders. The Portal can contribute to knowledge preservation and knowledge discovery, raise visibility of STFC-sponsored research, demonstrate STFC adherence to the principles of Open Science, and support practical applications of these principles to research impact studies, professional engagement with other research organizations and funders, and to public engagement.

3 Data Sources and Technology for STFC Open Science Portal Implementation

The Portal ingests metadata from a few sources, leaving data, full-text publications and other research artefacts in their current respective locations, and represents the integrated metadata as the knowledge graph. The metadata is subject to a moderate level of harmonization when integrated, yet there is no intention to support higher levels of the metadata harmonization or unification. Records of science in the external quality sources are not ingested in the Portal but linked from it through the use of persistent identifiers or in some cases by other record matching techniques.

The sources where the Portal ingests metadata from:

- STFC publications repository [4]
- STFC data repository [5]
- Diamond Light Source bibliographic database [6]
- DataCite (records there that have been produced by STFC) [7]
- Unpaywall [8] to discover Open Access versions of publications
- (under consideration) Gateway to Research [9]
- (under consideration) Crossref COCI and other sources of citations [10]

The sources that the Portal links to:

- The Cambridge Structural Database [11]
- The British Library EThOS service [12]
- Europe PMC [13]
- Protein Data Bank in Europe [14]
- (under consideration) Zenodo [15]

For managing the integrated metadata, a community edition of the neo4j graph database [16] is used that is currently hosted on the developers' computers and in the STFC Intranet. The current size of the database is within tens of thousands of nodes and tens of thousands of relationships; this has a potential to grow to hundreds of thousands or a few millions of nodes, and hundreds of thousands or a few millions of relationships. The inflation of the database beyond hundreds of thousands or low millions of nodes and relationships is not expected at the moment.

For the records ingestion, OAI-PMH endpoints, bespoke APIs or bulk export features of the aforementioned sources have been used, which resulted in tabular or XML files. These have been further processed using XSLT transformations and Unix shell scripts, then uploaded in the graph database using standard neo4j tools.

Relationships across records are produced using elements of them associated with persistent identifiers where possible, otherwise fuzzy matching techniques are used, such as measuring distance between corresponding metadata elements, as was in the case of doctoral theses records [17]. The enrichment of scholar communication with persistent identifiers is a gradual process, and the broader PIDs proliferation should allow more efficient records matching in the future, saving the effort of building research knowledge graphs. This is a good example when best practices (of persistent identifiers minting and use) can augment and in certain cases replace technology (of fuzzy records matching).

Exploration and visualizations of the resulted graph and its parts (subgraphs) are made using queries in Cypher language [19] and standard Web components of neo4j. Further experiments with visualization are going to involve Apache ECharts [20] and JavaScript Cytoscape [21].

An example of a subgraph extracted from the integrated graph that is going to support the Open Science Portal is illustrated by Fig. 1. The metadata sources where these records are harvested from do not necessarily know about each other, but their integration in STFC Open Science Portal provides connections and allows cross-walks between any of the records of science involved.

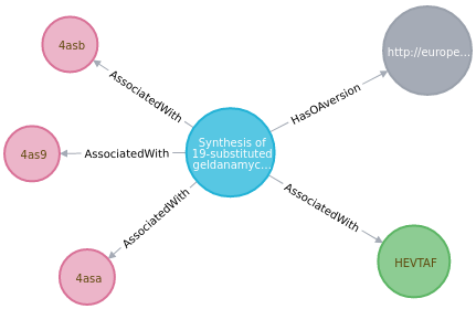

Fig. 1. Publication in Diamond bibliographic database (central node) connected to its Open Access counterpart in EuropePMC (top right), to Cambridge Structural Database record (bottom right) and to three Protein Data Bank records on the left.

The vision of the Portal is for it to become a single entry point for searching across all records of science that could be publications, dataset descriptions, grants etc. The full-text search across all these records of science are supported by cross-records indexes created in the graph database and powered by Apache Lucene [18]. At the moment, an out-of-the box indexing algorithms are used that are implemented by default in the database engine and that do not use the terminological structure of any research domain. It would be interesting to explore the efficiency of discipline-specific indexing, yet this requires a dedicated study which can be costly, too, especially if it involves

human evaluators. Also in the case of the STFC knowledge graph, there would be natural obstacles to the application of discipline-specific text indices, as STFC research is multi-discipline by its nature and thus the knowledge graph may incorporate records of science from many branches of science, as different as biology or physics or occasionally even archeology or history of art when samples are supplied by these disciplines for investigation on synchrotron or neutron sources.

The focus of the Open Science Portal for the foreseeable future is going to remain the metadata, yet data visualizations are going to be incorporated opportunistically where effort to produce them is moderate. The Fig. 2 illustrates a rich context around a record in the INSDB (Inelastic Neutron Scattering Database) [22] which is published as a part of the earlier mentioned STFC data repository [5]. The orange circle on the top right with the "RecordedAs" relation to it represents a record in the STFC ePubs repository that is currently unaware of the most of the information context in this graph (apart from the corresponding journal paper). Some of the entities in this graph, e.g. research instrument or organizations involved, do not currently bear persistent identifiers but have a good potential for acquiring them through the further refinement of the graph. The Fig. 3 contains visualization of the data file associated with the same INSDB record. Visualization (which is interactive when accessed in a Web browser) is made using Plotly JavaScript library [23]. URLs on the top refer to the data record and to the paper that reports on the data, so that this visualization gives a good idea of both what the data is and where further detail of it can be found through the resolution of persistent identifiers that the paper and the data URLs are based on.

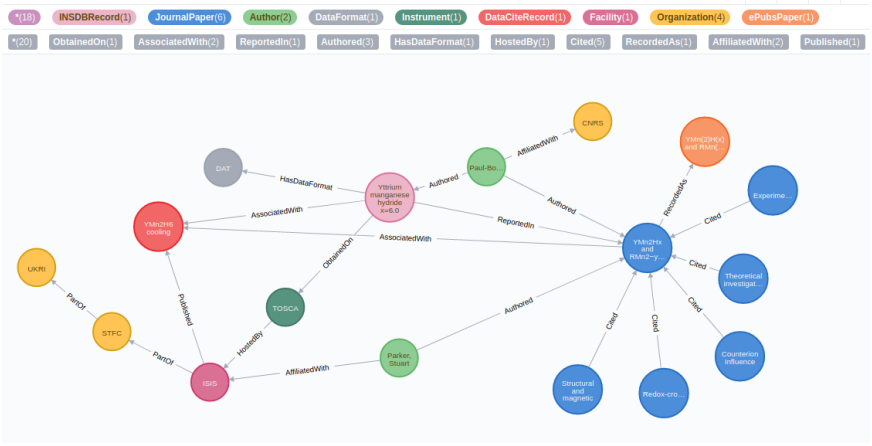

Fig. 2. Rich research information context around INSDB (Inelastic Neutron Scattering Database) record that involves a description of the original experiment in DataCite, people and organizations involved, and resulted research paper cited by other papers.

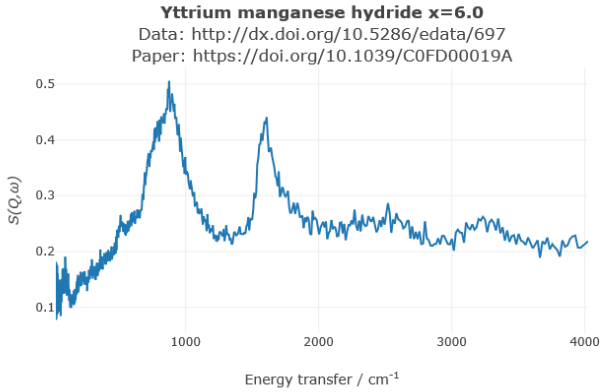

Fig. 3. Visualization of a dataset from INSDB (Inelastic Neutron Scattering Database) that is associated with the INSDB record included in the graph on Fig. 2.

4 Further Development of the Open Science Portal

The development of the Open Science Portal has been focused so far on the data sources integration in the back-end graph database, with illustrative visualizations to see what usage scenarios are principally possible. The ultimate goal is to make the Portal a fully functional Web application with features of a full-text search across a variety of entities (publications, dataset and software descriptions, grants information) and with contextual visualization of metadata (subgraphs). The target audience for the graphical user interface are going to be STFC staff, visitor scientists, other funders and policy makers.

Apart from the graphical user interface, an API based on the GraphQL technology [24] is going to be implemented. Implementing a simple GraphQL endpoint to a graph database is straightforward and can be realized with standard plugins. A more sophisticated approach may be required though owing to the diverse nature of the metadata sources integrated; this might be based on measuring the popularity of certain attributes of the graph database nodes and relations [25]. The open API should allow third party developers to build their own applications around the STFC records of science.

The existing graph will benefit from disambiguation of certain entities in it, and assigning them with persistent identifiers that are not immediately available as they were not present in the source records that the graph is composed of. Organization names are the prime candidate for such disambiguation; GRID.AC [26] or ROR [27] services can be sources of persistent identifiers for organizations, also Crossref Funder Registry [28] for organizations that are funders of science.

It will be most productive to think of the Open Science Portal as a new piece of the research information infrastructure that the organization itself and other parties can use for multiple purposes. Some of the possible applications of the Portal are mentioned in the "Business case" section, yet it is in the nature of every infrastructure to find its uses that are beyond the initial thinking of the stakeholders needs. So when developing the

Open Science Portal, a good attention should be given to non-functional requirements such as maintainability and extendibility of the underpinning knowledge graph.

Acknowledgements

The work is supported by FREYA project funded by the European Commission under the Horizon 2020 programme (Grant Agreement number 777523). The views expressed are those of the author and not necessarily of the project or the funder.

The author thanks Natalie Johnson (Cambridge Crystallography Data Centre) for matching the Cambridge Structural Database records with those in STFC repositories.

The author thanks Johanna McEntyre and Christine Ferguson (European Bioinformatics Institute) for their advice on using Europe PMC and Protein Data Bank APIs.

The author thanks Stewart Parker (STFC ISIS Neutron and Muon Source) for the example of the Inelastic Neutron Scattering database record connected to the record of the original experiment on ISIS facility.

The author thanks his colleagues in FREYA project for their feedback on the prototype demonstrations.

References

1. FREYA project, https://www.project-freya.eu/, last accessed 2020/09/08.
2. Science and Technology Facilities Council, https://stfc.ukri.org/, last accessed 2020/05/31.
3. UK Research and Innovation, https://www.ukri.org/, last accessed 2020/05/31.
4. ePubs: STFC publications repository, https://epubs.stfc.ac.uk/, last accessed 2020/05/31.
5. eData: STFC "Long Tail" data repository, https://edata.stfc.ac.uk/, last accessed 2020/05/31.
6. Diamond Light Source bibliographic database, https://publications.diamond.ac.uk/publications/searchpublicationsquick, last accessed 2020/05/31.
7. DataCite search, https://search.datacite.org/, last accessed 2020/04/29.
8. Unpaywall: An open library of scholarly articles, https://unpaywall.org/, last accessed 2020/04/29.
9. Gateway to Research: UKRI gateway to publicly funded research and innovation https://gtr.ukri.org/, last accessed 2020/04/29.
10. COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations, http://opencitations.net/index/coci, last accessed 2020/04/29.
11. The Cambridge Structural Database, https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/, last accessed 2020/04/29.
12. The British Library EThOS service, https://ethos.bl.uk/, last accessed 2020/04/29.
13. Europe PMC portal, https://europepmc.org/, last accessed 2020/04/29.
14. Protein Data Bank in Europe, https://www.ebi.ac.uk/pdbe/, last accessed 2020/04/29.
15. Zenodo repository, https://zenodo.org/, last accessed 2020/04/29.
16. neo4j graph database, https://neo4j.com/, last accessed 2020/04/29.
17. Bunakov, V., Madden, F.: Integration of a National E-Theses Online Service with Institutional Repositories. Publications 8(2), 20 (2020). doi: 10.3390/publications8020020
18. Apache Lucene, https://lucene.apache.org/, last accessed 2020/04/29.
19. Cypher Query Language, https://neo4j.com/developer/cypher-query-language/, last accessed 2020/04/29.

20. Apache ECharts data visualization framework, https://echarts.apache.org/, last accessed 2020/09/08.
21. Cytoscape JavaScript library, https://js.cytoscape.org/, last accessed 2020/04/29.
22. Inelastic Neutron Scattering Database, https://www.isis.stfc.ac.uk/Pages/INS-database.aspx, last accessed 2020/04/29.
23. Plotly JavaScript Open Source Graphing Library, https://plotly.com/javascript/, last accessed 2020/04/29.
24. GraphQL query language, https://graphql.org/, last accessed 2020/04/29.
25. Bunakov, V. Metadata Integration with Labeled-Property Graphs. In: Garoufallou, E., Fallucchi, F., William De Luca, E. (eds.) Metadata and Semantic Research. Communications in Computer and Information Science, vol. 1057, pp. 441-448. Springer International Publishing, Cham (2019).
26. GRID: Global Research Identifier Database, https://grid.ac/, last accessed 2020/04/29.
27. ROR: Research Organization Registry, https://ror.org/, last accessed 2020/04/29.
28. Crossref Funder Registry, https://www.crossref.org/services/funder-registry/, last accessed 2020/04/29.