Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-491

Towards an Open Infrastructure for Relating Scholarly Assets

Christopher MUNRO[a,1], Philip COUCH[a], Jon JOHNSON[b], John AINSWORTH[a], and Iain BUCHAN[a]
[a] *Health eResearch Centre, Division of Informatics, Imaging, and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, UK*
[b] *Centre for Longitudinal Studies, UCL Institute of Education, London, UK*

Abstract. Discovery of useful relationships between scholarly assets on the web is challenging, both in terms generating the right metadata around the assets, and in connecting all relevant digital entities in chain of provenance accessible to the whole community. This paper reports the development of a framework and tools enabling scholarly asset relationships to be expressed in a standard and open way, illustrated with use-cases of discovering new knowledge across cohort studies. The framework uses Research Objects for aggregation, distributed databases for storage, and distributed ledgers for provenance. Our proposal avoids management by a single central platform or organization, instead leveraging the use of existing resources and platforms across natural partnerships. Our proposed infrastructure will support a wide range of users from system administrators to researchers.

Keywords. Research informatics, distributed systems, publication archives, research objects, cohort studies, discovery networks, knowledge management.

1. Introduction

The tsunami of data generated, or leveraged by, social and biomedical sciences poses a significant challenge in knowledge discovery for both researchers and data investors (those collecting or enabling the collection of research-ready data).

For example, researchers may be interested in finding scholarly assets relevant to their work and describing how they have derived new knowledge from existing assets. At the same time, data investors may wish to track the outputs, such as published papers, based on their data – this is illustrated in a recent literature search by the Millennium Cohort Study wishing to track the outputs from its data [1].

The current models of knowledge discovery via structured metadata focus on large Data Archives and Research Data Portals. These centralized approaches involve the curation of metadata (often with extensive manual processes), primarily from retrospective documentation of surveys and other data collections. With the increasing availability of portals and platforms storing scholarly assets there is move to better reuse existing data to answer new questions [2]. To maximize the impact of this reuse asset

[1] Corresponding Author: Dr. Christopher Munro, Health eResearch Centre, Division of Informatics, Imaging, and Data Sciences, University of Manchester, Jean McFarlane Building, Oxford Road, Manchester, M13 9PL, UK. Email: chris.munro@manchester.ac.uk

producers need to specify the relationships between existing scholarly assets that asset consumers can navigate to discover the context of use.

To facilitate the discovery of new knowledge there needs to be open and shared specification of the aggregated assets, and an infrastructure to make these aggregations available. Such assets may be co-produced across multiple organizations/sources that do not necessarily share the source data, for example due to governance constraints. Akin to a car built of components from different factories.

The problem for discovery is two-fold: encouraging the creation of metadata around generated data resources and connecting metadata to its source through a chain of provenance at minimal additional cost. This discovery problem is like that of research using articles from serials and journals, which use exchange protocols such as OAI-PMH, and standards such as Z39.50. However, whereas the digital manuscript field benefits from having a small set of similar metadata models, in social and biomedical science data management there are many (and quite different) metadata models in use.

To address this problem we reviewed existing approaches, technologies, and standards, and we developed use-cases and specified requirements for an infrastructure to support discovery of relationships between scholarly assets.

2. Methods

2.1 Use Cases

We developed the use-cases below drawing on the experiences and needs of users of the HeRC e-Lab and CLOSER discovery platforms for collaborative research (across organizations and disciplines) using shared data sources.

The HeRC e-Lab is an online research collaboration platform which can combine and harmonize existing datasets, for example storing multiple birth cohort datasets, plus linked clinical data, as part of the STELAR project [3]**.** The CLOSER discovery platform [4] enables researchers to search and explore data from eight UK longitudinal studies.

We have considered a range of users from those driving research and consuming content from platforms, to the platform developers and data managers:

1. A researcher has obtained data from a Research Data Portal and generates some derived variables and wants to share what it is and how it was generated, and provide a citable link for publication.
2. A researcher is looking at a dataset available from a Research Data Portal and wants to know what other researchers have produced before deciding whether to explore a funding call.
3. A Principal Investigator wants to understand what data from their study has been used in published research and whether the data produced is being under-utilized.
4. A Platform owner would like to make their users' aggregations of assets searchable in a lightweight shared network, where there is no reliance on a single third party central platform being maintained and available.

2.2 Requirements

We identified the following requirements for a system to serve the use-cases above:

- Store assets in distributed infrastructures and search them in a unified way.

- Search and create aggregations of existing assets (e.g. publications, datasets), and specify relationships between them (via a web-based interface and programmatically for external platforms).
- Identify and authorize users, for example using OpenID, or ORCIC ID.
- Harvest assets' records and aggregations from existing platforms.
- Assist/guide users in creating 'profiles' for aggregations to ensure sufficient context is recorded and that the aggregations have enough metadata to be usefully searched.

3. Results

Cross platform/format search has been driven largely by serials, journals and libraries using the centralized approaches described below. The move toward common standards has been very helpful in this area, e.g. the CORE platform [5] which harvests multiple repositories via the OAI-PMH standard protocol, and journals using Z39.50.

Federated search: one platform facilitates search over all resources/service providers. The records from the service providers are stored at the data archive sites. The records' schema needs to be the same to easily add a new provider, unless a *mediator* is written to allow new record formats to be ingested, with additional cost. As data archives manage the assets the search interface is always up to date and there are fewer issues with synchronization. Search performance can be poor as the process involves waiting for the query results from each service provider.

Cross archive search via harvesting: data are first gathered from sources and then stored locally in the search platform, resulting in improved performance but with inherent synchronization issues. This is the approach of large search engines, but requires a common data format e.g. OAI-PMH. There is further complication if a source goes down and the record still exists in the search platform.

To enable discovery of content across platforms, we investigated a variety of possible approaches and technologies:

Aggregation*:* a key requirement for the infrastructure we propose is a method to aggregate resources that can be identified (e.g. dereferenced by a URI). The OAI-ORE standard for reuse and exchange defines Resource Maps [6] and these can also be made identifiable to allow discovery [7]. The Research Objects standard for aggregating resources [8] builds on OAI-ORE and incorporates formalized annotations, capture of provenance metadata, as well as minimum requirements for metadata, and the use of checklists [9] to assess the quality of aggregations.

Blockchain: there is emerging interest in utilizing Distributed Ledgers to benefit from the decentralized capabilities and provenance tracking properties of blockchain. The original implementation, bitcoin, is purely currency based but variants such as Multichain or Ethereum allow assets to be stored and blockchain additions validated. Blockchain has several weaknesses for this discovery role, e.g. it is not designed for large document storage, e.g. records of 1MB for Datacoin. The cost of adding to the block chain scales with the size of the block chain, this is challenging for an infrastructure designed to be used over a long period. Only the transaction information (but not the contents of the block chain) is searchable. Once records are added they are immutable which could problems in governance rich areas such as data access.

Distributed databases: distributed databases support the decentralized nature of our proposed infrastructure. Elastic search (which incorporates a search engine), and,

Couchbase are examples document stores and Cassandra is a columnar database. Graph variants include Titan, Neo4j and OrientDB with querying via the Gremlin graph querying language. Compared to blockchain these perform better for searching (in blockchain only the transaction history is searchable), and share the decentralized nature and redundancy but lack the in-built provenance capturing abilities. They also more naturally support file storage, for example Couchbase can store 'blob's of JSON up to 20MB. They also support the distributed nature (adding extra nodes), for example with Cassandra new nodes can be added to the cluster using a certificate approach.

3.1 Proposed Approach

Figure 1 contains an illustration of our approach. We propose an infrastructure that uses an aggregation format such as Research Objects (ROs), to be stored in a network of distributed databases, which supports the addition of new organizations. Identification and authorization could be handled via open ID, or ORCID ID [10] and usernames and passwords to cater for users unable to access the other methods.

The infrastructure incorporates a web interface for users (e.g. researchers, data investors) to add and search aggregations. The interface could be authenticated to, via Open ID or ORCID. This enables use of the system by those with low resources, whereas platform hosts may interact via software.

Both e-Lab and CLOSER are nodes in this example, and would host an instance of the distributed database, with the capability to push both records of content via OAI-PMH, and ROs.

The Figure illustrates the example of a dataset, contained in both the CLOSER portal and the HeRC e-Lab. There are ROs in the e-Lab aggregating existing data, scripts and results, and ROs in the CLOSER platform contain the questions, studies, sweeps and instruments.

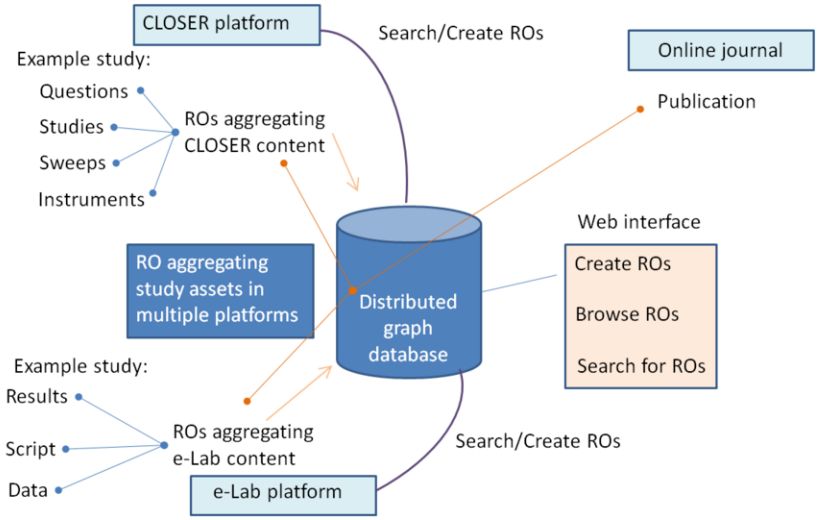

Figure 1: Outline of proposed asset relationship discovery framework, illustrated with e-Lab and CLOSER platforms and study data, using a currency of (extended) Research Objects (ROs).

4. Discussion

We have studied theoretical and practical solutions for discovering new knowledge from the relationships between digital scholarly assets – reviewing existing technologies and standards; exploring use-cases; specifying requirements; proposing an open technical and operational framework that can leverage existing platforms quickly.

Two key strengths of the proposed framework are: 1) it does not need to be maintained by a single organization – operations are distributed and thus resilient; and 2) it builds on existing standards, platforms, and assets.

The framework allows users (e.g. data investors, or researchers building on existing work) to aggregate related assets from different locations or platforms, enabling the discovery of relationships that were previously fragmented across the distributed assets. We propose using ROs to specify their aggregations and publish them in ways that are easy to find, share and reuse. The exchange of ROs provides a mechanism for discovery between different platforms – this can be a partial exchange where encapsulated source data may not be shared (e.g. resources are within a private portal, that requires additional access authorization) but their metadata are sufficient for discovery of new knowledge (e.g. contact details to request information from the portal). This highlights the need to specify profiles for 'types' of ROs, although this is currently missing from the RO specification, minimum information checklists [8] could be developed to facilitate different profiles.

The use of a distributed database for ROs increases availability by avoiding reliance on a single platform, and this approach enables trusted networks to be developed by incorporating extra nodes.

We are currently linking the e-Lab and CLOSER platforms using the framework presented here as part of UK Research Councils' move toward more collaborative research using shared digital assets.

References

[1] **D, Kneale, et al., et al.** *Piloting and producing a map of Millennium Cohort Study Data usage: Where are data underutilised and where is granularity lost?* London : EPPI-Centre, UCL Institute of Education, 2016.

[2] *Increasing value and reducing waste: addressing inaccessible research.* **Chan, Dr An-Wen, et al.,** 2014, The Lancet, pp. 257-266.

[3] *The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together.* **Custovic, A, et al.,.** 2015, Thorax, pp. 799-801.

[4] *CLOSER Discovery.* [Online] https://discovery.closer.ac.uk/.

[5] CORE system. [Online] https://core.ac.uk/.

[6] Discovery of resource maps. [Online] http://www.openarchives.org/ore/1.0/discovery.

[7] Resource maps. [Online] http://www.openarchives.org/ore/1.0/primer.

[8] *Why Linked Data is Not Enough for Scientists.* **Bechhofer, Sean, et al.,** Future Generation Computer Systems, 2013, Vol. 29.

[9] *MIM: A Minimum Information Model vocabulary and framework for Scientific Linked Data.* **Gamble, Matthew, et al.,** 2012 IEEE 8th International Conference on E-Science (e-Science).

[10] Introduction to the ORCID Public API. *ORCID.* [Online] https://members.orcid.org/api/introduction-orcid-public-api.