The One Repo: background, implementation and call for funding.

Michael P. Taylor[1,2], Sebastian Hammer[1]
[1]Index Data, Ruardean, UK; Copenhagen, Denmark; and Boston, USA. mike@indexdata.com.
[2]Department of Earth Sciences, University of Bristol, Bristol, UK. quinn@indexdata.com

May–June 2015

Abstract

As scholarly communication undergoes seismic changes, endless opportunities are opening up. In an open-access world, there is potential for Internet-enabled research on huge corpuses, discovering new correlations and making new connections. To facilitate these processes, we need platform that provides uniform access to the metadata and full text of all open-access articles, whether in repositories or open-access journals. The platform must provide data that is complete, up to date, high quality and open for every kind of re-use. The One Repo (http://onerepo.net/) is is that platform. It aims to make the entire open-access scholarly record available via a Web UI, embeddable widgets and various web-services, as well providing all of the metadata for direct download. It is built from battle-tested components that are in use in high-volume commercial systems. Numerous harvesting methods are used. The existing demonstrator presents a UI that integrates results from a small number of repositories and other sources. We seek funding to rapidly increase coverage. The One Repo has dramatic implications for scholarship, research and engineering across every field of human endeavour.

Table of Contents

The vision

> "To exploit the potential of the network all we need to do is get as much material online as fast as we can. We need to connect it up, to make it discoverable … at network scale the system is designed to *ensure* that resources get used in unexpected ways. At scale you can have serendipity by design, not by blind luck."
>
> — Cameron Neylon, *Network Enabled Research: Maximise scale and connectivity, minimise friction*.

Scientific publishing, and scholarly communication more broadly, are undergoing enormous changes. The traditional channels of commercial, peer-reviewed journals are seen as expensive, unwieldy, and inadequate as a means of assuring quality (however one defines this). As solutions, the Internet offers a confusion of different methods and models for information sharing, and research institutions are working hard to adapt.

While it is unclear exactly how the field will evolve, open access scholarly publishing will certainly play a central role. It is a natural evolution of the traditional scholarly communication model, but offers untold new opportunities for data mining, information re-use, and new discoveries (as well as enormous cost savings). Unlike paywalled publications, which hide knowledge in inaccessible silos, open access publishing invites and encourages rich connections between papers and projects.

To realize these benefits, it is essential that documents are easily findable. While there are numerous national and subject-specific repositories, and Google Scholar does a credible job of aggregating those that serve information in a well-structured fashion, none of these provide a comprehensive platform for innovative uses of research.

As a catalyst for all this potential progress, what's needed is a single source of data, an open aggregation of all the metadata about about all the scholarly articles from all the repositories – every record presented in the same simple, canonical format, and made freely available for everyone.

- It must be **complete**, so that searches can be directed against a single source and result in high-quality, exhaustive search results. It therefore cannot be limited to repositories which are well-behaved.
- It must be **up to date**, so that the most current papers can be found.
- It must have **high-quality data** in a uniform format, to facilitate analysis, re-use, and sophisticated post-processing. It must faithfully represent the metadata exposed by each vendor, even when vendors don't follow best practices for data formats.
- In the spirit of open access, it must be **open for any and all kinds of re-use**. Specifically, it must be offered on a liberal license for commercial and academic re-use, so that the entire ecosystem benefits.

We are building a repository with these four qualities, which can be the basis for any number of exciting projects and products. We call it The One Repo.

The problem

It was more than twenty years ago that Stevan Harnad published his "subversive proposal" that scholars should make their papers freely available on the Internet in institutional repositories (IRs). Thousands of these now exist, many of them implemented using well established software packages such as EPrints and DSpace.

The OpenDOAR directory of repositories contains metadata records describing not only IRs but also governmental and disciplinary repositories. At present, OpenDOAR lists some 2860 repositories. Assuming that perhaps another third to half as many IRs again exist but are not registered, the total number in the world is probably close to 4000 – and indeed this is about the number recorded in the alternative register ROAR.

Although these repositories in aggregate make an enormous amount of research freely available, the fragmentation of this knowledge across 4000 repositories using many different data formats makes it difficult to discover and awkward to use. In practice, IRs form an archipelago of isolated islands rather than a continent of discoverable knowledge. There have been previous attempts to solve this problem, such as the OARA project; but the present maturity of OA infrastructure means the time is now right.

Google does not solve the problem

The *de facto* discovery tool for most purposes is the Google search engine; and, for scholarly research, its cousin Google Scholar. These are helpful, but far from complete solutions. While Google Scholar indexes articles from many different sources, including commercial databases and some repositories, it does not solve the repository problem for several reasons:

- It is not focussed on repositories, and has no mandate to focus on them.
- Its coverage is patchy and haphazard.
- There is no clear statement of what sources are and are not covered.
- There is no accountability to a board or the wider public.
- There is no API, and screen-scraping is prohibited.

This last point is crucial for practical purposes. The *only* thing that can be done with Google Scholar is to read its results from a screen. They cannot be automatically queried, aggregated, analysed, harvested, mined, backed up or otherwise used.

Perhaps worst of all, Google has no commitment to Scholar. It provides the service at present, but it could be withdrawn at any time. (Google has a history of doing this, for example recently closing down Google Wave, Google Reader and Google Code.) The scholarly community simply cannot rely on an opaque, closed, unaccountable and long-term unreliable service.

National repositories do not solve the problem

Some valuable initiatives already exist to gather repository content within individual countries: for example, JISC's CORE (COnnecting REpositories) in the UK, and HAL in France. As significant as these are, however, they reduce the number of places to be searched from 4000 repositories to (potentially) 200 countries. What is needed is a *single* point of access for the whole

world – one which provides not just a discovery tool but a shared platform for further work.

National solutions have on occasion been mothballed at the end of experimental periods: for example, the ARROW project of Australia closed in December 2008, and the DARE project in the Netherlands ended in 2006 (although Narcis fulfils some of the same role). Such demises arguably indicate that national solutions are not broad enough, and therefore do not offer enough value, to sustain themselves for the long term.

The solution

We offer The One Repo (http://onerepo.net) as a solution to these challenges. This is a system, already existing in proof-of-concept form, to gather all the content of all the world's repositories and open-access journals into a single database, in a uniform format – freely accessible to all as harvestable data, by a set of web APIs, as embeddable widgets, and as a Web UI.

The One Repo is not a research project, but is built on battle-tested components that are in use in high-volume commercial systems. It has been proven robust, efficient and scalable.

Architecture and access

The One Repo consists of the following cleanly separated layers, providing access to the database at different levels:

- The canonicalised union database itself, available for harvest via OAI-PMH and other mechanisms.
- A standards-compliant (SRU) search-and-retrieve web-service that can be run against the union database, any of its constituents, or any defined subset (e.g. a national subset).
- A set of widgets which present results from the union database. These can be embedded in other web pages to include One Repo content in, for example, web portals, library catalogue systems or individual IRs.
- A web user-interface, open to all without registration, built from widgets.
- A linked open data representation, which allows navigation by software, and possibly crowd-sourced submissions of corrections, additional information about authors or documents, etc.

This commitment to open access at every level of the stack is crucial to ensure that data, having been acquired, centralised and canonicalised at some cost, cannot be engulfed by a new silo.

Methods of harvesting

The database is built and maintained by harvesting metadata records from existing repositories and publishers. Harvesting works by any of these methods:

- Metadata transfer using the OAI-PMH protocol
- Bulk download of records in any XML format
- Bulk download of records in any MARC-based format

- Any XML-based harvesting API
- Any web-based UI can be scraped when no better solution is available. This is supported by a connector framework, based on Mozilla's Web engine, by which it is possible to provide a JSON-over-XML web-service for any site.

Crucially, data may be harvested in any metadata format (or none, in the case of web-scraping), and is normalised into a single uniform schema. All databases are treated essentially equally within the One Repo.

Whenever possible, comment streams, reviews and discussions associated with documents will be harvested and associated with the relevant entry in the union database.

Content policy

The policy of The One Repo is to accept all tadata deposited in the included repositories and all open-access papers provided by publishers, including the following:

- Metadata for gold open-access articles, with links to the full text.
- Metadata for green open-access archived papers, with links to the full text.
- Metadata records describing papers that are *not* available. These are important for at least three reasons. First, in some cases, they describe papers that will become freely available after the expiry of an embargo period; second, such metadata records provide a means of discovering the author and requesting a copy directly – a process that may be facilitated by an “ask author for a copy” button; and third, records of papers that should be available (but are not) are important data for tracking compliance of open-access policies.
- Links to associated data-sets, such as specimen photos, matrices for phylogenetic analysis, databases of observations and survey results. (We do not plan to replicate the data objects themselves.)
- Comments, reviews, and discussions that are clearly associated with a harvested article.

Data objects deposited with third-party services such as GenBank, FigShare or Morphbank are currently considered out of scope.

Where we are now

The One Repo exists as a demonstrator, presenting much of the UI functionality at http://onerepo.net/.

The entire software stack is in place, including the scalable, redundant back-end, and the harvesting infrastructure. Access is available to the web services at various levels as noted above, although comprehensive documentation is not yet available.

What is missing at the moment is primarily content. At present, about twenty repositories and journal sites are harvested. Nevertheless, the demonstrator is already a useful tool for research – but much less valuable than it will become as its coverage increases.

Where next

The underlying software of The One Repo scales well to very large numbers of data sources, but supporting each new data source requires an investment of time. The present set has been chosen to encompass a range of geographical locations, subject areas, repository frameworks, etc. As resources permit, our development goals for The One Repo are as follows:

- Support for **many more data sources.** Ultimately, we aim to include the content of every institutional repository in the world, plus appropriate subject and government repositories, and open-access journals. In the push towards achieving full coverage, use can be made of existing aggregation efforts such as the UK and French national repositories mentioned above.
- Support for **more metadata fields**, and better mapping to achieve uniform semantics across disparate data sources. (We already support the RIOXX profile for those databases that provide these fields.)
- Creating **developer documentation**, describing the various web service APIs that provide access to the harvested data and live searching, and how to incorporate One Repo widgets.
- Support for **richer searching** – for example, the ability to limit results to records for which full-text is available, or for which peer-review has been carried out (where available metadata makes this possible).
- Provision of the harvested data as **linked open data** using RDF and Turtle.
- Facilities for **compliance-checking** against institutional and funder open-access policies. At the lowest level, this will require APIs capable of searching for many documents at once – for example by means of a list of DOIs. At a higher level, specific compliance-checking for individual policies can be implemented with understanding of appropriate requirements.
- Support for **community participation** in efforts that are outside of the scope of the initial project. An example of this might include subject term or author name normalization projects or other efforts to enhance the quality of the stored metadata.

Call for funding

In order to expand the coverage of The One Repo, we seek charitable funding. This project is for the world. Investment funding would not be appropriate, as the project does not seek to make a profit: in order to remove all barriers to participation, we do not plan to charge any participants. We are looking for funders with a global vision.

Costs come under three main headings.

1. Adding support for new repositories. Costs vary widely depending on how a given repository is implemented, what metadata profiles it supports, what harvesting methods must be used, etc.
2. Development and documentation. Little of this is required, as the majority of the One Repo software is already in place, and documentation exists in other forms that can readily be adapted. This activity includes supporting the emerging community around the service with

an attractive interface and simple, powerful tools for re-using the data and developing new use cases.

3. Ongoing operation. Again, costs vary based on the number are size of repositories, frequency of re-harvesting, etc., but our technology platform allows us to scale the One Repo with the growing open-access community at a predictable cost. This covers hosting, hardware, network bandwidth, etc., but most importantly maintenance of harvester configuration as harvested systems change, and enhancement as they improve to include more and better metadata.

Conclusion

The present system of scholarly publishing in journals is widely recognised as expensive, slow and inefficient. The system of institutional repositories, intended in part to solve this problem, has so far proven an ineffective solution due to inconsistencies and the difficulties of discovery. The One Repo represents an opportunity to solve this problem once and for all, using a technical solution based on proven, scalable technology, with dramatic implications for scholarship, research and engineering across every field of human endeavour.

Appendix: document version history

v2.1 (30 June 2015). Add some background on precedent.

v2.0 (24 June 2015). New document structure based on more clearly articulated value proposition. Every section rewritten and much obsolete material removed. Various sections re-ordered. As of this version, the document is co-authored by Sebastian Hammer.

v1.3 (19 June 2015). Remove financial estimates to an appendix in a separate document.

v1.2 (14 May 2015). Minor wording changes. Add some links, adopted from the blogged version of the early parts of this whitepaper.

v1.1 (13 May 2015). Fix some minor typos and redundancies. Add version history.

v1.0 (not numbered, 12 May 2015). First released version.