An Open Access Triptych: IsItOpenAccess, the Open Citations Corpus and IDFind

Richard Jones, Cottage Labs, richard@cottagelabs.com
Mark MacGillivray, Cottage Labs, mark@cottagelabs.com
David Shotton, University of Oxford, david.shotton@zoo.ox.ac.uk
Cameron Neylon, PLOS, cneylon@plos.org
Emanuil Tolev, Cottage Labs, emanuil@cottagelabs.com

Abstract: This proposal presents three new services which, although independent, are natural companions and which have great significance for the repository community. They are: IsItOpenAccess- a service which takes a set of identifiers and provides a report on the open access status and licensing conditions of each; The Open Citations Corpus - an open access database of bibliographic records and the citation links between them; IDFind - a proof-of-concept service for identifying identifiers. We discuss the relationships between these services (and others in use in the sector) and how they can benefit the repository community.

This proposal presents three new services which, although independent, are natural companions and which have great significance for the repository community. They are:

IsItOpenAccess (IIOA) [1] - a service which takes a set of identifiers and provides a report on the open access status and licensing conditions of each. Currently it supports only DOIs and PMIDs, and only a handful of publisher web-pages provide us with good licensing information. It is driven by a desire by funders and institutions to verify that published papers are compliant with funder Open Access mandates, but it has potential uses beyond this for institutions, repositories, and those who wish to use literature to create new works. The data is stored as an extension to BibJSON [2], and freely available, although it does not store bibliographic data alongside the license and identifier data. IIOA has been commissioned by PLOS and is currently being hosted by Cottage Labs.

The Open Citations Corpus (OCC) [3] - an open access database of bibliographic records and the citation links between them. Initially populated from the reference lists of all open access articles in PubMed Central [4], this database is currently being expanded with references from arXiv [5] preprints and (soon) from subscription-access as well as open-access journals via CrossRef [6]. The service will synchronise regularly with the source data sets, continually checking for new citations between articles and enriching the database as it goes, and is designed such that new source data sets can be added easily. The result is a large linked database in BibJSON, that can also be exported as RDF, with many valuable identifiers being used to identify the bibliographic records. The OCC is being run by the University of Oxford with Jisc [7] funding.

IDFind [8] - a proof-of-concept service for identifying identifiers. There is a proliferation of identifiers in use just in our sector: ISSNs, ISBNs, DOIs, PMIDs, ORCIDs, UUIDs; can we always tell what kind of identifier we are looking at? This may not be too difficult for humans (although still not always straightforward), but it is very hard for machines; so if we want to automatically enhance metadata by creating canonical forms of identifiers, we may not be able to do so easily. IDFind addresses this issue by providing an API which will try to identify your identifiers and give you back information about them. IDFind has been developed by Cottage Labs without any external funding to-date.

Although each service clearly occupies its own feature space, they are highly complementary and overlap with each other at the edges. They each rely heavily on identifiers and the ability to understand and resolve them. Both OCC and IIOA use BibJSON as a common format.

OCC and IIOA could utilise IDFind for identification of identifier types and potentially canonical forms of those identifiers and forms which can be dereferenced. For example, the DOI "10.1371/journal.pone.0035089" might have the canonical form "info:doi:10.1371/journal.pone.0035089" but the form which can be dereferenced would be "http://dx.doi.org/10.1371/journal.pone.0035089"

Meanwhile OCC could share bibliographic information (including other potentially useful identifiers) with IIOA and IIOA could share licence information with OCC, creating a more complete information set in within corpus. Furthermore, the OCC holdings could be routinely run through IIOA, pre-empting user needs to find the licence information, by looking it up in advance and caching it. *Figure 1* shows the cycle of potential relationships.

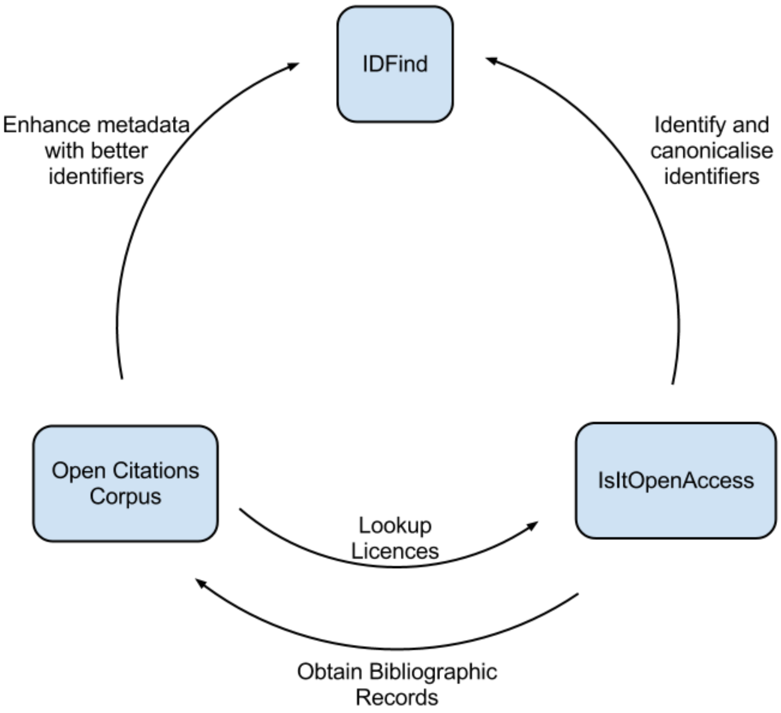
IDFind
Enhance metadata with better identifiers
Identify and canonicalise identifiers
Open Citations Corpus
IsItOpenAccess
Lookup Licences
Obtain Bibliographic Records

Figure 1: An Open Access Triptych

We should be clear at this point that these relationships do not yet exist. IDFind is a prototype service, while OCC and IIOA are relatively new, and still have long lists of features that need to be implemented. Nonetheless, it is a short enough step to join them up in this way and the benefits of this kind of service collaboration cannot be understated. we should also remember that there are other services in the wild that could be incorporated to further extend the power of this network, such as the University of Oxford's Journal Article Identifier Resolver [9] (which finds alternative identifiers for journal articles), or ImpactStory [10] (which calculates altmetrics for online resources).

So how is this relevant to repositories? Each of these services, individually, clearly has some value to the repository community and to repository holdings.

Metadata held by repositories could be enhanced from OCC, or identifier records cleaned up through IDFind. Repositories could obtain citation information from the OCC and use the information to connect full-text holdings together.

Meanwhile, from a strategy perspective, IIOA provides a very powerful mechanism whereby institutions can obtain two key pieces of information:

1. Is my institutional Open Access mandate being complied with?
2. Are the materials held as Open Access in my institutional repository there legitimately?

Knowing the answers to these questions will aid in the enforcement of any institutional OA mandates, and may be used to go on and answer a third important question:

3. What proportion of Open Access articles coming from my institution are also being archived in my institutional repository?

By combining these three services (and potentially more) we can improve the quality, quantity and reliability of information available for institutions to answer such questions.

But there are also ways that repositories can contribute to the information held by these services. Repositories could supply additional metadata to enhance the OCC, or could provide links to full-text versions of Open Access articles. The full-texts could be provided to OCC to mine for citation information; this is already happening with full-texts held in arXiv, for example.

This presentation will begin with a brief introduction to each of the services: IsItOpenAccess, the Open Citations Corpus and IDFind. We will cover their motivations, functional behaviour, and value to the community. We will go on to explain how they fit together, and other services in the sector which could also fit into a network of cooperating systems for the benefit of the community. We will then discuss in detail the relationship such a network of systems would have with repositories, and how repositories are *part* of that network, not just consumers from it.

We will conclude by looking at any gaps in the services and a roadmap for future developments and collaborations.

[1] IsItOpenAccess beta: http://iioa.cottagelabs.com
[2] BibJSON: http://bibjson.org
[3] The Open Citations Corpus: http://opencitations.net/
[4] PubMed Central: http://www.ncbi.nlm.nih.gov/pmc/
[5] arXiv: http://arxiv.org
[6] CrossRef: http://crossref.org
[7] Jisc: http://www.jisc.ac.uk
[8] IDFind (test server, may not always be responsive): http://test.cottagelabs.com/idfind
[9] Journal Article Identifier Resolver http://www.miidi.org/sameas/
[10] ImpactStory: http://impactstory.org/