From Open Access Metadata to Open Access Content: Two Principles for Increased Visibility of Open Access Content

Petr Knoth
Knowledge Media institute
The Open University

An essential goal of the open access (OA) movement is free availability of research outputs on the Internet. One of the recommended ways to achieve this is through open access repositories (BOAI, 2002). Given the growing number of repositories and the significant proportion of research outputs already available as OA (Laakso & Bjork, 2012), it might come as a surprise that OA content is not necessarily easily discoverable on the Internet (Morrisson, 2012; Konkiel, 2012), more precisely, it is available, but often difficult to find. If OA content in repositories cannot be discovered, there is little incentive to make it available on the Internet in the first place. Therefore, not trying hard enough to increase the visibility of OA content would be a lost opportunity for achieving the main OA goals, including also the reuse potential of OA content. In this paper, we build on our experience in finding and aggregating open access content (not just metadata) from repositories, discussing the main issues and summarizing the lessons learned into two principles that, if adopted, will dramatically increase the discoverability of OA content on the Internet and will improve the possibilities of OA content reuse.

1. Repositories as large metadata silos

Institutional repositories[1] clearly serve a number of purposes, including collecting and curating digital outputs or providing statistics for measuring research excellence. The roles of repositories are well documented in the SPARC's position paper on institutional repositories (Crow, 2002). However, the primary goal of repositories is to open and disseminate research outputs to a worldwide audience. The SPARC's position paper specifically says:

"For the repository to provide access to the broader research community, ***users outside the university must be able to find and retrieve information from the repository****. Therefore, institutional repository systems must be able to support interoperability in order to provide access via multiple search engines and other discovery tools.* ***An institution*** *does not necessarily need to implement searching and indexing functionality to satisfy this demand: it* ***could simply maintain and expose metadata, allowing other services to harvest and search the content****. This simplicity lowers the barrier to repository operation for many institutions, as it only requires a file system to hold the content and the ability to create and share metadata with external systems."*

To fulfill their main purpose, it is crucial that repositories expose valid metadata in a standardised format, unambiguously linking metadata with content[2], allowing external systems to harvest both the metadata and the content. While this might seem trivial, we will show that locating and accessing content is currently a challenge. At the moment (even the most prominent) repositories are often seen by search engines and aggregation systems as large metadata silos rather than well organised content stores. Only a fraction of metadata records in institutional repositories are associated with content and many current repository solutions make content harvesting[3] an order of magnitude more complicated than metadata harvesting or even restrict it completely.

Yet, the primary goal of the OA movement is to provide open access to research outputs, i.e. content, not open access to metadata of research outputs. Open access to metadata is not disruptive, it only marginalises the benefits of open access, making little difference to the traditional publishing model.

2. OA metadata vs OA content

To better understand the metadata-content relationship, we created a sample set of 83 (mainly EPrints) repositories (with mostly pdf research outputs to which we restricted our study) and calculated the ratio of metadata to content. The statistics have been acquired from the CORE system, which aggregates metadata and content from many repositories and provides a range of services on top of this aggregation (Knoth & Zdrahal, 2012). We will use CORE data and build on the experience acquired in CORE also later in the paper. Altogether, the selected repositories contained 1,461,016 metadata records. For the experiment, we selected only those repositories about which we were confident the statistics would give us reasonably accurate results, as collecting these statistics for all repository systems is problematic due to content referencing and access issues that will be discussed in the next sections. Table 1 shows (a) the fraction of metadata records that provide a link to content, (b) the proportion of metadata with downloadable content and finally (c) the ratio of metadata records to content that is in addition machine-readable.

	metadata linked to content	content downloadable	content machine readable
Mean	54.1%	34.4%	27.6%
Median	39.5%	16.7%	13.0%
Standard deviation	39.2%	34.2%	31.0%

Table 1: Content to metadata ratio in OA repositories calculated on a sample of 83 repositories with 1,461,016 metadata records.

[1] In this paper, we use the term institutional repositories (which are the main interest of the Open Repositories conference), to refer also to subject-based repositories or archives or systems for depositing research outputs used by open access publishers. As a result, the conclusions of this paper and recommendations are equally valid for both the green (self-archiving) and gold (OA publishing) routes to OA.

[2] By the term *content*, we understand the object of the research output, as opposed to a metadata record describing the research output. In many circumstances, we will use the term content to refer to publication manuscripts (full-texts), however, content can also refer to research data or any other type of object described by a metadata record.

[3] In the context of this paper, content harvesting is synonymous to content aggregation. Please refer to (Knoth & Zdrahal, 2012) for more information about the value of OA aggregations.

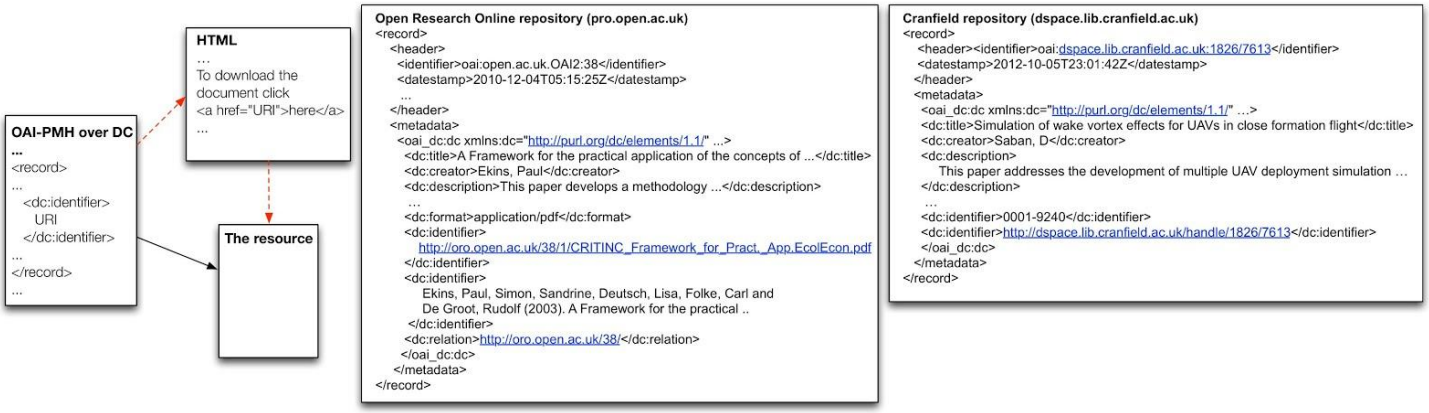

Figure 1: Different approaches to content identification. The Open Research Online repository (as most EPrints repositories) specifies as one of its dc:identifiers a link that resolves to the actual resource, while the Cranfield repository (as most DSpace repositories) identifies the resource by providing a link to a page from which the resource (if available) can be accessed. The dashed line shows the approach taken by the Cranfield repository and the full line the approach taken by the Open Research Online repository.

The results indicate that only 27.6% of research outputs in repositories are linked to content that can be downloaded by automatic means and analysed (e.g. indexed). In addition, due to substantial differences between repositories (as demonstrated by the standard deviation), the median repository will only provide machine readable content for 13% of its deposited resources. Furthermore, due to the methodology the repositories were selected in our experiment, it is likely that these statistics are in fact rather optimistic, as they do not include repositories that do not provide direct links to the content from their metadata. Such repositories form a significant (perhaps even the majority) group and we will focus on them in the next section. Quoting the (in)famous article (Salo, 2008), this yet again confirms that *"[The institutional repository] is like a roach motel. Data goes in, but it doesn't come out."*

Analysing all the reasons for this low availability of content would be out of scope for this paper, but among the main barriers certainly are legal and technical issues (Knoth & Zdrahal, 2012). Since we cannot easily resolve the legal aspects, we will primarily focus on the technical issues. The technical issues have a detrimental effect not only on finding and collecting OA content, but also on our ability to text-mine content and actually track and analyse the progress of the OA movement. By not being able to produce accurate and transparent statistics about how much OA content (not metadata) is available and track its growth, we might be loosing a strong argument for the adoption of OA.

3. Content referencing practises

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), as its name suggests, has been developed to support primarily harvesting of repository metadata. OAI-PMH supports representing metadata in multiple formats, but at a minimum repositories **must** be able to return records with metadata expressed in the Dublin Core format (OAI-PMH v2.0, 2008). Exposing metadata records using OAI-PMH, where records are encoded in the Dublin Core format, is the predominant solution supported by the majority of repositories today. If repositories want to allow other services to harvest and search their content (as required by the SPARC guidelines (Crow, 2002)), they must provide a link to the content as part of the exposed metadata.

Unfortunately, as we will show now, content harvesting from repositories is problematic precisely due to inconsistent or ambiguous approaches to linking metadata records with the content the metadata records describe. As demonstrated in Figure 1, some repositories, such as Open Research Online (and typically EPrints), provide as one of the dc:identifiers a direct link that resolves to the identified resource. On the other hand, other repositories, such as the Cranfield repository (and typically DSpace), identify a resource by providing a link to an html page (*splash-page*) on which one can find the download link. To better understand how this inconsistency in approach emerged, we will now investigate the appropriate standard.

The OAI-PMH specification states on this topic that:

*"The nature of a resource identifier is outside the scope of the OAI-PMH. To facilitate access to the resource associated with harvested metadata, repositories **should** use an element in metadata records to establish a linkage between the record (and the identifier of its item) and the identifier (URL, URN, DOI, etc.) of the associated resource. The mandatory Dublin Core format provides the identifier element that **should** be used for this purpose."*

The main issue boils down to the semantics of the *identifier of the associated resource*. While it might seem this implies the identifier should resolve to the identified content, in fact, the specification only states it should establish a linkage of the record with the identifier of the associated record (whatever this identifier might be). Coming back to Figure 1 and the example of the Cranfield repository, it seems that the identification of an html page describing a resource, instead of the resource, does not violate the OAI-PMH specification. In fact, the OAI-PMH specification provides examples, such as *<dc:identifier>http://arXiv.org/abs/cs/0112017</dc:identifier>*, which also identifies only a page from which the resource can be downloaded. On the other hand, the specification quite openly states the nature of the identifier is outside of its scope. So overall, it seems standards are fairly weak about something which is, in fact, crucial to the OA mission.

Regardless of which content referencing approach conforms to the standard, the current situation is problematic for content harvesting systems due to a number of reasons including that it is not clear whether (a) the dc:identifier provides a link to another identifier or the actual content and (b) the harvesting system is sufficiently informed about the item it is looking for (e.g. the format of the document, the number of items, the availability of the item, etc.). As a result, a system performing content

harvesting might need to implement repository specific content wrappers or use some intelligent techniques to crawl, locate and identify the correct links to content on the page identifying the content (which was designed to be read by people not by computers). At the same time, content harvesting systems can create unnecessary load on the repositories when they are required to crawl their websites instead of just requesting the relevant data.

4. A pragmatic approach to content referencing

As the OAI-PMH standard allows other data representations to be used in addition to the required Dublin Core description, solutions using richer formats, such as MPEG-21 DIDL or METS, have been proposed in (Van de Sompel et al.) already in 2004. However, the adoption of these solutions has been unfortunately low. Though there is evidence some solutions exist, in fact, out of more than 280 repositories harvested by CORE, none of them uses MPEG21-DIDL or METS. There is at the moment an ongoing work in developing another protocol called ResourceSync (ResourceSync draft, 2013) that might also help in this situation, however, even if released today (and successful) it will likely take at least several years until it will be adopted. Yet, there is a real need for a solution today, as the inability to perform content harvesting is already clearly inhibiting the benefits of open access.

From the perspective of a harvesting system, it is vital that repositories unambiguously identify resources. Accurate identification of content is, and should be understood as one of the primary responsibilities of an institutional repository and, as a result, attention should be paid to ensure content harvesters are capable of finding the links from metadata to content. This is something that should be checked and monitored on an ongoing basis as the inability to harvest content compromises a critical functionality of the repository. However, we believe that to achieve this, it is important that there are available tools that help repository managers to check and monitor the validity of the repository metadata including the metadata-content linkage. Such tools can also suggest metadata format changes when necessary and help repository managers to keep metadata cleaner, thereby promote interoperability. The CORE's Repository Analytics[4] are a step in that direction, helping repository managers to track harvestability of their content.

Therefore, the solution we propose in the next section is pragmatic rather than exciting. It focuses on generating maximum benefit to the sector for a minimum investment. The solution deliberately sticks to the use of current standards to minimise adoption time. It aims at respecting differences across systems and offers backwards compatibility. Finally, it emphasizes the need for easy to use compliance mechanisms to assist repository managers in ensuring systems interoperability. To enable content harvesters in the OA domain to locate content, repositories should be required to adhere to the following principle.

Principle 1: Content referencing. *Open repositories should always establish a link from the metadata record to the item the metadata record describes using a dereferencable identifier pointing to the version held in the repository. The dereferencable identifier should be provided in the appropriate metadata element in the used metadata format (i.e. dc:identifier in the case of Dublin Core).*

The implications of the principle. Repositories can use different metadata standards to describe resources offered using OAI-PMH. For example, repositories exposing complex objects might want to describe them in MPEG21-DIDL or METS. However, at minimum they must offer Dublin Core metadata, where each Dublin Core record uses at least one of its dc:identifiers to identify the described object (content in the repository) using a dereferencable identifier, i.e. the identifier must resolve to the object it identifies (not into another identifier). In the special case of Dublin Core, which can be used to describe simple objects only (Van de Sompel, 2004), if more dereferencable URIs are present (which is syntactically correct), we suggest to use the first dereferencable identifier as the identifier of the described object. The identifier should resolve to the version of the object held in the repository, which is important for being able to acquire repository content statistics.

The application of this principle of a dereferencable identifier is easily applicable in the domain of open access publications, as free availability of research outputs on the public internet directly follows from the definition of OA (BOAI,2012), thus every OA content can be described using a dereferencable identifier. Content not described in this way should not be treated as OA content. This approach might not be possible in some other domains, where copyright law might prevent direct dereferencing of the content. As a result, it would not be appropriate to handle the content referencing issue by modifying the Dublin Core or other standards, as adoption of this principle in the open access domain seems much more appropriate. The adoption of this principle would also lead to greater transparency of the OA content, allowing systems to provide more realistic statistics of Open Access repository content, avoiding various more or less anecdotal situations, such as the one demonstrated in Figure 2, where a repository contains 23,880 items not referencing any real content. As mentioned earlier, repositories can and should be supported by validation tools to check and monitor compliance with this principle as this compliance test can be automated.

5. The accessibility of repository content to machines

Correct content referencing is an essential requirement for content harvesting, but it does not on its own guarantee the harvestability of the content. One of the ways how content harvesting can be restricted by the content provider, is by using the Robots Exclusion Standard. This is a convention to prevent programs, such as crawlers, from accessing all or part of a website which is otherwise publicly viewable. It can also be used to specify a minimum delay with which a robot can repeatedly access the repository website.

To give an example of implementation of these restrictions in repositories, let us demonstrate them on Arxiv.org (Figure 2). According to the specification, this repository restricts access to all its pdf content to machines through the Arxiv.org website. The exception is Googlebot with unrestricted access, other selected search engines can access the content, but only at the rate of one document per 20s. At this rate, it will take more than 6 months to transfer the content of the whole repository. Such approaches clearly violate the principles of open access. To be fair, Arxiv.org outputs are available in the Amazon storage and can be downloaded by anybody providing that they pay the fee for the bandwidth. However, this, in fact, violates not only the

[4] http://core.kmi.open.ac.uk/repository_analytics

principles of OA, but also the openness of Internet. It does not matter what the cost is, what matters is that it creates a dangerous precedent where the data provider charges a fee for offering services to the one receiving the services.

DSpace @ Cambridge (www.dspace.cam.ac.uk)

```
<record>
   <header>
     <identifier>oai:www.dspace.cam.ac.uk:1810/220941</identifier>
     <datestamp>2009-10-23T14:49:26Z</datestamp>
     <setSpec>hdl_1810_219204</setSpec>
   </header>
   <metadata>
     <oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" … >
       <dc:title>Dark Item</dc:title>
       <dc:creator>Dark Item</dc:creator>
       <dc:description>Dark Item</dc:description>
       <dc:publisher>Dark Item</dc:publisher>
       <dc:date>Dark Item</dc:date>
       <dc:date>Dark Item</dc:date>
       <dc:date>Dark Item</dc:date>
       <dc:type>Dark Item</dc:type>
       <dc:identifier>Dark Item</dc:identifier>
       <dc:identifier>Dark Item</dc:identifier>
       <dc:rights>Dark Item</dc:rights>
     </oai_dc:dc>
   </metadata>
</record>
```

Arxiv.org (http://arxiv.org/robots.txt)

```
# robots.txt for http://arxiv.org/ …
# Indiscriminate automated downloads from
#   this site are not permitted
# See also: http://arxiv.org/RobotsBeware.html
# $Date: 2012/04/27 15:58:32 $
User-agent: *
...
Disallow: /pdf/
Disallow: /html/
...
User-agent: Googlebot
...
Allow: /pdf
Allow: /html
...
User-agent: Yahoo! Slurp
...
User-agent: msnbot
Crawl-delay: 20
...
Allow: /pdf
Allow: /html
...
```

Figure 2: On the left, an example of a repository item not referring to any real content (there are 23,880 instances of the "Dark Item" in this repository). On the right, an example of restriction access to machines as implemented in Arxiv.org. While Google has unrestricted access, it will take other selected search engines over 6 months to retrieve the content. Other than selected systems cannot access the content at all.

To be consistent with the definition and goals of open access (BOAI, 2002), we suggest repositories should conform to the following principle.

Principle 2: Content accessibility to machines. *Open repositories must provide universal access to machines with the same level of access as humans have. It is the role of open repositories to allow machines harvest the entire content of the repository in a reasonable time to enable harvesting systems to acquire and maintain up-to-date information about the repository content.*

The implications of the principle. Repositories should ensure their content could be access by programs, for example to collect data for text-mining or indexing. This access must be universal, i.e. it should not discriminate or provide an unfair advantage to any particular system (with the exception of abusive behavior). Programs, such as content harvesters, must be allowed to access the repository at such frequency that allows them to keep information about the repository current. However, some repositories might contain both OA materials as well as non-OA materials. In such cases, it is acceptable and, in fact, beneficial to clearly restrict access to the non-OA materials. If a repository does not comply with this principle, it should not be considered an Open Repository.

6. Conclusions

We have shown that the proportion of open access content that can be collected by harvesting systems from repositories is fairly low in comparison to the metadata. As this inhibits the benefits of OA, we tried to identify the main technical barriers to the full transition from open access metadata to open access content. Consequently, we formulated two principles that can be easily and timely adopted by repositories to resolve this issue. These principles, fully consistent with the OA mission, require open repositories to provide dereferencable identifiers for open access content and allow fair access to robots wishing to process the repository content. To put these principles into practise, we emphasize the importance of compliance validation and monitoring tools assisting repository managers in improving their systems.

References

[BOAI, 2002] Budapest Open Access Initiative. (2002) http://www.opensocietyfoundations.org/openaccess/boai-10-recommendations

[Crow, 2002] Crow, R. (2002). The case for institutional repositories: a SPARC position paper. *ARL Bimonthly Report 223*.

[Knoth & Zdrahal, 2012] Knoth, P. and Zdrahal, Z. (2012) CORE: Three Access Levels to Underpin Open Access, *D-Lib Magazine*, 18, 11/12, Corporation for National Research Initiatives, http://dx.doi.org/10.1045/november2012-knoth

[Konkiel, 2012] Konkiel, S. (2012) Are Institutional Repositories Doing Their Job? https://blogs.libraries.iub.edu/scholcomm/2012/09/11/are-institutional-repositories-doing-their-job/

[Laakso & Bjork, 2012] Laakso, M., & Björk, B. C. (2012). Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC Medicine*, *10*(1), 124.

[Morrison, 2012] Morrison, Louise (2012) 5 reasons why I can't find Open Access publications. http://mmitscotland.wordpress.com/2012/08/06/5-reasons-why-i-cant-find-open-access-publications-2/

[OAI-PMH v2.0, 2008] The Open Archives Initiative Protocol for Metadata Harvesting Version 2.0 (OAI-PMH), Impementation Guidelines (2008). http://www.openarchives.org/OAI/openarchivesprotocol.html

[ResourceSync draft, 2013] ResourceSync protocol draft. 2013 http://www.niso.org/workrooms/resourcesync/

[Salo, 2008] Salo, D. (2008). Innkeeper at the roach motel. *Library Trends*, *57*(2), 98-123.

[Van de Sompel et al, 2004] Van de Sompel, H., Nelson, M. L., Lagoze, C., & Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-lib magazine*, *10*(12), 1082-9873.