

# ***Metadata for Harvesting: the Open Archives Initiative, and how to find things on the Web.***

Philip Hunter and Marieke Guy.

## **Abstract**

The OAI Protocol for Metadata Harvesting offers the prospect of resource discovery tools far beyond what is currently available to users of the Web via standard search engines. This article illustrates how existing information about available resources can be repurposed fairly easily and cheaply using standard tools. However the publishing of this information throws up a number of practical and philosophical questions which have to be addressed by projects and institutions.

## **Keywords**

Open Archives Initiative, Metadata, Harvesting, Resource Discovery, ePrints, Dublin Core.

## **Introduction**

This paper is about the use of metadata harvesting, as a way of making information about institutional resources, in both electronic and in other formats, more widely available via a Web-based interface. This is a revised version of the paper presented to the Internet Librarian International conference in March 2002, since when we have had the advantage of discussions with Michael Nelson of the Open Archives Initiative (during a session on metadata at the first Open Archives Forum workshop in Pisa in May 2002), and also Herbert van de Soempel at the third workshop in Berlin in May 2003. [1]

Metadata harvesting is a method of making information available about resources, which has generated a lot of interest since the creation of the Open Archives Initiative in 1999. In the course of the paper we will fill in some of the background to the Open Archives Initiative, as well as the relevant aspects of the OAI Protocol for Metadata Harvesting. We will also be talking about and defining eprints, since these are the resources most projects wish to make available using the OAI Protocol for Metadata Harvesting.

A key focus of this paper is some aspects of the practical implementation of a metadata repository using existing metadata. Many institutions have existing metadata for their resources (essentially catalogue information, item descriptions, collection descriptions, etc., which can be used to identify and locate resources on the Internet) - this information about documents and digital objects can be (relatively) easily repurposed to create an OAI compliant metadata repository. UKOLN for example, has metadata for its eJournals (*Exploit Interactive*, *Cultivate Interactive*, and also for *Ariadne*); some of this has been used by us to explore the processes which are involved in the repurposing of metadata.

Beyond the technical issues in implementation however there are some important questions which have arisen, mostly revolving around Copyright and Intellectual Property Rights. These also have to be dealt with in the course of setting up a metadata repository on the Web, and these are briefly discussed here.

## **What the Open Archives Initiative is**

The genesis of the Open Archives Initiative can be traced back to the very early 1990s and the creation of the Los Alamos Pre-print Archive by Paul Ginsparg. This is still regarded as the premier example of how to run a pre-prints archive, and Paul Ginsparg is internationally recognised as one of the leaders in the area of new scholarly publishing architectures. ArXiv (as the archive is known) 'has evolved towards a global repository for non-peer reviewed research papers in a variety of physics research areas'. It also incorporates mathematics, non-linear sciences and computer science pre-prints. ArXiv moved to Cornell University in 2001. [2]

The Open Archives Initiative often use terms in very precise technical senses, and in a way which is sometimes perplexing to those outside the OAI community. In fact the precision of certain of the terms used within the OAI community helps to explain the architecture of open archives. 'Pre-print' is one of the core concepts handed on from the Los Alamos experience, signifying a non-peer reviewed document available in electronic format before formal peer-reviewed print publication. By extension, 'pre-print archive' is essentially an archive of non-peer reviewed papers in electronic format, made available before formal peer-reviewed publication. Definitions of what might be archived or made available with this technology have changed however, since the creation of the Los Alamos Archive in 1991. The OAI now talks about 'eprints', for example, a significantly different concept. Since there are other categories of document which might be archived and made available in a similar way, the terminology was expanded to accommodate these. Hence the document categories 'postprints' and 'eprints'. Postprints (as might be guessed) are documents for which harvestable metadata is available after the peer-review process. Collectively, they are all 'eprints': 'eprint' is defined in practice as the collective term for all other something-print items.

The Open Archives Initiative has two main points of focus: the speeding up of scholarly communication, and a general opening up of access to communities interested in these resources. An important convention held in Santa Fe in 1999 established the OAI: this meeting was originally called the 'Universal Preprint Service meeting', and was initiated by Paul Ginsparg, Rick Luce and Herbert Van de Soempel. At the meeting they defined the goal of the Open Archives Initiative as:

*to contribute in a concrete manner to the transformation of scholarly communication. The proposed vehicle for this transformation is the definition of technical and supporting organizational aspects of an open scholarly publication framework on which both free and commercial layers can be established.*

The convention established a ‘combination of organizational principles and technical specifications to facilitate a minimal but potentially highly functional level of interoperability among scholarly eprint archives’[3]

## **Rising costs of Scholarly Communication**

Stevan Harnad of Southampton University has argued very persuasively (in a paper called ‘For whom the Gate Tolls’) that the current scholarly communication model actively succeeds in defeating its original idea. In that, in return for a paper being accepted for a scholarly journal, the scholar signs away his copyright control of his work, so that the publisher has sole publishing rights to the material. The costs of publishing in the traditional paper format are now so great that libraries are charged enormous subscription fees for access to the journal, which means that fewer and fewer can afford these subscriptions. And ultimately, the community which the scholar wishes to address in many cases no longer has access. In extreme cases this could mean that the author himself can no longer go into his departmental library, and see his own work on the shelves. [4]

The SPARC publishing initiative was set up in 1998 specifically to explore new publishing models for the scholarly community as alternatives to the traditional publishing model in serials. [5] The Open Archives Initiative from the following year is a further level of response to the perceived crisis in scholarly publishing, and builds on the experience of pre-print and eprint archives worldwide. The OAI argues that:

- 1 The explosive growth of the Internet has given scholars almost universal access to a communication medium that facilitates immediate sharing of results.*
- 2. The rapidity of advances in most scholarly fields has made the slow turn-around of the traditional publishing model an impediment to collegial sharing.*
- 3. The full transfer of rights from author to publisher often acts as an impediment to the scholarly author whose main concern is the widest dissemination of results.*
- 4. The current implementation of peer-review – an essential feature of scholarly communication – is too rigid and sometimes acts to suppress new ideas, favor articles from prestigious institutions, and cause undue publication delays.*
- 5. The imbalance between skyrocketing subscription prices and shrinking, or, at best, stable library budgets is creating an economic crisis for research libraries.*

The OAI also argue that: ‘e-print archives exemplify a more equitable and efficient model for disseminating research results. An important challenge is to increase the impact of the e-print archives by layering on top of them services – such as peer review – deemed essential to scholarly communication. This is the focus of the Open Archives Initiative’. [6] When this paper was first given at the ILI conference at Olympia in London, we demonstrated to the audience how papers submitted to this site were available worldwide on the *day* of submission. This is a very valuable prize for the academic community. Given the sizeable price rises indicated by academic serials publishers in October 2002, the argument for widespread takeup of the eprints idea by the academic community gains strength with every day that passes.

## Technical aspects of the OAI approach to improving Scholarly Communication.

We are now going to look at the technical side of what the Open Archives Initiative is about. It should be understood that the OAI has laid down a *minimal* set of what is required for interoperability. It also should be understood that the OAI Protocol is principally about the exchange of metadata. Though it is by its origins motivated by the need to find electronic resources, the protocol specifies virtually nothing about this side of scholarly communication. It is also not about the specification of particular metadata formats, though it expects as a minimum something like Dublin Core metadata. The Santa Fe recommendations on interoperability were restricted to interoperability at the level of Metadata Harvesting. For this they described a set of simple metadata elements, to enable ‘coarse granularity document discovery among archives; the agreement to use a common syntax, XML to represent and transport both the Open Archives Metadata Set (OAMS) and archive specific metadata sets; and thirdly, the definition of a common protocol [the Open Archives Dienst Subset] to enable extraction of OAMS and archive-specific metadata from participating archives’. [7]

The Santa Fe Convention presents a technical and organization framework which is designed to facilitate the discovery of content stored in distributed eprint archives. Because the technical recommendations have been implemented by a number of institutions, it is now possible to access data from eprint archives through end-user services. At the moment, mainly via harvesting services which provide Web interfaces to the aggregated metadata exposed by data providers.

In summary, the OAI has specified

- A protocol for the exchange of metadata (a new version of the protocol replaced the first version in the summer of 2002).
- That XML should be the syntax for representing and transporting the metadata
- That metadata should be exposed to end-user services
- That metadata should be harvested to facilitate the discovery of content stored in distributed eprint archives

### More definitions

Making your metadata available to third parties is always spoken of as ‘exposing’ metadata. This means that you have placed your metadata, wrapped in the appropriate XML, in a place which can be accessed by a third party. In terms of the jargon, this means that you are a ‘Data Provider’. Data providers ‘expose’ metadata. This does not necessarily mean that you (as a Data provider) maintain an eprints archive. It means only that you are supplying metadata for resources available *somewhere*. In practice however most data providers will also have a full-text archive available whose location can be indicated in the metadata records. But in OAI jargon, the institution which exposes the metadata is a data provider. Similarly, putting metadata on the Web doesn’t make your institution into a data provider, strictly speaking, even if the metadata is being used to make a locally browsable eprints service available. To

conform to the OAI definition of a Data provider you have to be registered with a third party who can harvest your exposed metadata, and provide some kind of user services. You have to register, in order that the Service providers know that you are there and that you are exposing metadata for them to harvest.

The idea of the Service Provider is that these should provide and develop third-party value-added services. The most common service available so far is the aggregation of metadata records into a single searchable archive. This means that (effectively) the contents of a large number of metadata repositories can be cross-searched. For this to be possible obviously the aggregated metadata has to be visible to users on the Web via some kind of interface which supports queries. The OAI makes no specifications about the nature of these interfaces, and the facilities available will vary a great deal from service to service. Though most at the moment offer a choice of simple or advanced query interfaces, not dissimilar to those available with library OPACS.

As we have seen, the OAI has defined a number of concepts extremely closely. Others the OAI has scarcely defined at all. This is because the focus of the OAI is on the Protocol for Metadata Harvesting. It is worth repeating that what exposed metadata describes, and what services might be provided, are not major concerns of the OAI. This is partly a conscious gambit by the OAI to avoid restricting the possible uses of the open archives idea. They are developing an enabling technology, not defining what may be done with it.

Given that there are projects out there exploring the use of the protocol to describe multimedia objects, there is some unclarity as to just how far 'eprints' should be used as an umbrella term for all of these. There is the question of electronic documents with multimedia components or learning modules embedded in their structure – are these eprints, or are they something else? So far the consensus seems to be that an 'eprint' is sufficiently defined as a 'document-like object'. Even if we aren't sure what it is, or what this might turn out to be some way down the line of OAI development.

Continuing our exploration of OAI definitions of terms, their published documents used to speak of metadata about eprints being held in repositories (the word 'archive' was not used in this context). The eprints themselves are described as being held in 'archives'. It is worth saying that 'archive' in this sense does not carry with it the baggage which the term carries outside the OAI and eprints context. Being in an archive does not imply that the eprint has been appraised in anyway. It does not mean that it has been selected by those in charge of running the archive. It also does not mean that it will be preserved at that location as an object of value. Individual archives will have their own policies which might be found by looking at the archive's Web site, but these are likely to differ according to their own priorities for digital preservation.

In March 2002 the terms 'repository' and 'archive' were used within the general OAI community in a fairly strict sense. You could have an eprints archive for example, but (as we have pointed out) the metadata for this archive was held in a repository. This would mean that to speak of a 'metadata archive' within an OAI context would have been understood to be incorrect. Part of the reason for this distinction was to keep apart the ideas of harvestable data, and the full-text document objects within the

architecture of the OAI. Since then it has become clear that the use of 'archive' within the OAI is not restricted to document objects – the term also can be used to refer to collections of harvestable metadata. It remains important of course to be clear about what is being referred to, but otherwise it is permissible to speak of both collections of metadata and collections of document-like objects as 'archives'. For the purposes of clarity in this paper, we will continue to speak of metadata in terms of repositories.

## **Some Issues in implementation**

One of the issues which arises out of the ease with which existing metadata can be repurposed is whether or not an institution actually *wants* to commit to giving access to objects they hold because they have exposed metadata for these objects, and the information has become available to the public via Service Providers. The answer is not always going to be an automatic yes, since there are resource implications to be considered.

In practice those who are setting up metadata repositories about ePrint Archives are providing, where possible, HTTP hyperlinks to full text documents contained in the archives. In the case of UKOLN's eJournals, our metadata records point directly to the original pages of the ejournals. However this also is not part of the original OAI specification. You don't have to do it in this way at all. You might simply point to a site where further information might be found, or even just give a physical location for the object which the metadata describes.

As we mentioned earlier, there are projects out there looking at the full range of resources for which metadata might be made available, including multimedia objects, etc. However there is already in every university a core set of documents which usefully might be used to populate an eprints archive. These include postprints (as defined earlier), as well as the kind of preprint literature pioneered by the Los Alamos archive. Other kinds of literature might find their way into a university eprints archive. Many documents produced by universities are not regarded as 'published' items, though they circulate widely. These include reports on almost any subject relevant to a university, and the documentation produced in the course of the work of a department. In this way an institution might build up an archive reflecting its research interests, which might be browsable worldwide, if the university so chooses. Not all of this material might be of interest on a worldwide basis; however, putting these materials in an archive which is referenced by metadata records would make them a great deal easier to locate in the future than they are now.

On the other hand, doctoral theses might benefit from being placed in an eprints archive, and having metadata for their contents exposed by a Data provider. The tradition in other parts of Europe, particularly Germany and the Scandinavian countries, has been (for hundreds of years) to publish these theses in book form, as accepted by the university authorities. In the UK there is no such tradition, and access to these in an eprints archive would be a low cost way of making them available.

Collection level description (as opposed to item level description, which involves the creation of many metadata records) is a low-cost way of making information about resources available. Making collection level details of these resources available in the

form of a metadata description which is then exposed to harvesting and aggregation services, creates a low-cost way of making entry-level data about collections available to researchers worldwide. [8]

## **Service Provision in practice**

The OAI specifies nothing about the functioning of Service Providers beyond the technical framework which provides the interoperability underpinning the harvesting of metadata. Currently existing Service providers provide metadata aggregation, and provide search interfaces. It is possible for them to provide other value-added services. One of the reasons why the OAI Protocol for Metadata Harvesting is formulated as a limited set of prescriptions is to leave open as much space as possible for the development of services on top of the basic protocol. Rather than presuming that the Initiative knows exactly how services based on the protocol will develop, they are assuming the opposite (no-one imagined the current state of the Web in 1991, and it would be foolish to dare to define the nature of the Web in 2012 on the basis of the last ten years). One of the most successful Service Providers at the moment is ARC - by March 2002 there were more than 65 repositories registered with ARC: by the first week of November there were 113 repositories available via this service. [9]

ARC has a search Facility – the *Exploit Interactive* and *Cultivate Interactive* metadata records are regularly harvested by ARC. If this repository is searched, say, for author ‘Philip Hunter’, the results page shows all the hits for that particular author. The page also shows the various metadata repositories which returned relevant records (in this case UKOLN e-journals). The results contain links to the metadata records for the query results (i.e., for the articles the user is searching for), and the metadata records link to the full-text in either the relevant eprints archive, or (in the case of the UKOLN e-journals, the original magazine articles. The metadata records may be detailed or basic; the records may point to locations on the Web; to paper based publications in no particular place; to general library classifications; or to actual library locations.

## **Repurposing existing Metadata for a harvestable Repository**

This section of our paper looks in brief at a practical implementation of metadata repurposing – in this case the repurposing of metadata created for two of UKOLN’s electronic journals – *Exploit Interactive* and *Cultivate Interactive*. First we give a short overview of the conversion processes and techniques which might be used; of how to export existing metadata to a database; and of the use of standard report creation techniques to wrap the metadata in XML.

The metadata was extracted from the journals by using a search facility. Each article in both magazines has its own metadata record: these records contain a number of standard fields which comprise a subset of the Dublin Core fields, but which are not marked up in XML. The fields exported were: Title, Creator, Description, Date, Type, Format, Identifier, Source, Language. The records were converted into a CSV format (comma separated values), which can easily be imported by many databases and

spreadsheets (Access, Excel, etc). In this particular case the field values were imported into an Access database.

With the metadata safely contained within the database, we then used the standard Access query and report facilities to wrap the metadata fields with the XML tags. This technique will be familiar to many who have produced automated reports based on data kept within a database (or even within flat files which contain information in a table format). Instead of wrapping information extracted from a database by query with a standard text, the field values were wrapped with the appropriate XML tags.

The resulting document was then saved as a text format file using a standard word-processing application – in this case Microsoft Word. Saving the document as a text file minimises the file accumulating any word-processor specific characteristics which might corrupt the file and make it unreadable by a third party harvester.

What you then have is a single file containing all the metadata records you have chosen to extract from the database, wrapped up in XML tags. This is what is required by a third party harvester. However there may be some minor formatting problems which need attention, such as the replacing of apostrophes with `<&apos;>` and ampersands with `<&amp;>` etc., and possibly the removal of some undesirable spacings. If the extension which marks the file as a text only document is change to the XML file extension, the file can be viewed in an XML compliant browser such as Internet Explorer (versions 5.5 and above). Harvesters can deal with multiple documents, so you might want to split your records up into individual files, one per record, since there might be local reasons for preferring the data to be made available in this way. This can be done using very simple PERL scripting – essentially the script sections each sequence of xml tags and spits these out (as they say) as individual files.

Clearly a broad range of skills is necessary to carry out these procedures, but none of them are particularly complex, and most of these skills are available within the kind of institutions which might want to make existing metadata available for harvesting. These skills were of course available in UKOLN; in addition we had some related experience of supplying metadata for another service within the Cultivate project.

## **Further issues in implementation**

How easy is it to install the software? Version 1.1 of the widely used ePrints software, available from [eprints.org](http://eprints.org), was often reported to be tricky to install. It ran on Unix boxes under Solaris, and on PCs running Linux; It did not like to be installed on machines running several processes. It also requires (according to many systems administrators who were asked to install it) a broad skill set (Perl; Apache; MySQL). Version 1.1 has now been replaced with version 2.0. However a number of institutions have not upgraded their software, possibly on account of the struggle involved to install version 1.1. [10]

One of the reasons why the idea of open archives took root so quickly is that it appears to open the way to a new level of information equity across the world, since there are equal publishing opportunities for institutions worldwide who have access to

the Web. Exclusion from publication because of publisher preconceptions about institutions will be minimised; and there will be opportunities for the expression of new ideas; The Budapest Open Access Initiative was instituted on the basis of this perception. [11] This will be a major element in the take up both of the OAI and related open archives ideas in parts of the world which are currently not well represented in academic journals.

IPR issues are perhaps the most important to consider, and so far universities have not given much attention to the implications. For instance, are the preprints in your archive legally author-owned preprints, or are they owned by a publisher? This is a tricky question, which, perhaps – in the short term at least – might best be regulated by contract between the university and the academic, and between the academic and the publisher. This is because the laws of copyright around the world were drafted long before the advent of the Web, and have yet to catch up.

A number of universities now make a legal claim to the ownership of the intellectual property produced within the institution, though they do not enforce this claim for the most part, or even at all. This means it is no longer certain whether or not material submitted to your archive is owned by the author: it might actually be owned by the institution. In the long-term this might be a major issue, if publication via a university Web site is seen by academics as the surrender of ownership in a piece of work, as is the case generally in publication via traditional paper journals.

It is also worth considering whether or not you want every resource (for which you have metadata) to be available to users. I forget who authored the following example which I heard at a conference a few years ago, but it is a good one. Suppose you run a museum which holds George Formby's ukelele in the basement, in a controlled environment to protect it: the information that you have the object is in your catalogue, and the catalogue is available in your museum. Your catalogue is also available in many central and university libraries in the UK, and in a few university libraries around the world. The University of Austin, Texas has a copy. Each year, your museum receives three requests to view the ukelele.

Then your catalogue information is converted into harvestable metadata. Your metadata is harvested by several Service providers and aggregated with other related metadata collections. Now anyone searching on 'ukelele' using the facility provided, will find the object. Anyone searching on 'George Formby' will also find the object. Maybe someone links directly from their George Formby fan page to the metadata record for the object, and to the associated photographs of the object. Suddenly instead of the information being seen by a handful of inquirers, the metadata is now before the eyes of thousands of interested people. And all of them know where the ukelele is kept. So instead of three requests per year to examine the ukelele, your museum now receives fifteen hundred requests per year. Very quickly you have (at the least) a conservation issue on your hands.

There are many potential obstacles to take-up: communities have their own reasons, some valid, some not, for avoiding involvement with eprints. Medics don't want non-peer-reviewed materials to be publicly available (at all); librarians are concerned that the development of eprints might result in budget reductions (a very real concern, even if libraries would at last be able to provide access to all relevant research for

their users). There is the question of whether or not eprints archives might be promoted as locally browsable resources, rather than as part of a global resource. If the archive is local, academics may not feel the submission of their work to be worth the time involved. On the other hand, there may be those who would not want documents to be available on a world-wide basis.

Another community which might be difficult to persuade to contribute to an institutional archive is of course the physicists, since Physics (and other sciences) already have eprint archives. From their point of view, submitting material to an institutional archive would be pointless duplication (this opens the question of whether archives should be created on an institutional basis, or on a subject basis). Another question which has to be addressed by institutions is whether or not items will be appraised before metadata is produced for harvesting. We know of one eprints project which in its early days had more documents in its appraisal buffer than were listed publicly, because they hadn't decided how to appraise the items.

## Conclusion

The OAI Protocol for Metadata Harvesting offers the prospect of resource discovery far beyond what is currently available to users of the Web via standard search engines, none of which (currently) make any substantial use of metadata. The technology is simple and robust, and take up is well under way. We have illustrated how existing metadata can be repurposed fairly easily and cheaply using standard tools. However, we have also illustrated that the technological side of the process is probably the least problematic: eprint archives and the exposure of metadata throws up a number of practical and philosophical questions which have to be addressed by projects and institutions.

## References:

1. Open Archives Forum -Workshop in Pisa, May 2002) – <http://www.oaforum.org/otherfiles/pisa-nelson.ppt> [Microsoft Powerpoint file]; and Ariadne [Hunter, Philip] “Open Archives Forum – First Workshop: Creating a European Forum on Open Archives” *Ariadne* Issue 32 – <http://www.ariadne.ac.uk/issue32/open-archives-forum/> [June/July 2002]
2. ArXiv Preprints archive at Cornell: <http://arxiv.org/>
3. Santa Fe Convention: [http://www.openarchives.org/sfc/sfc\\_entry.htm](http://www.openarchives.org/sfc/sfc_entry.htm)
4. Harnad, Stevan: “For Whom the Gate Tolls: How and Why to Free the Referred Research Literature Online Through Author/Institution Self-Archiving”: [November 2001]. – <http://www.ukoln.ac.uk/events/open-archives/presentations/harnad.pdf> [PDF file]
5. SPARC: The Scholarly Publishing and Academic Resources Coalition – <http://www.arl.org/sparc/home/>
6. Santa Fe Convention: [http://www.openarchives.org/sfc/sfc\\_entry.htm](http://www.openarchives.org/sfc/sfc_entry.htm)
7. Santa Fe Convention: [http://www.openarchives.org/sfc/sfc\\_entry.htm](http://www.openarchives.org/sfc/sfc_entry.htm)
8. Collection Description Focus (UKOLN): A Select Bibliography on Collection-Level Description (and related issues) – <http://www.ukoln.ac.uk/cd-focus/bibliography/>

9. ARC – A Cross Archive Search Service (Old Dominion University Digital Library Research Group – <http://arc.cs.odu.edu/>)
10. Eprints.org - <http://www.eprints.org/> Version 2.2. of the Eprints software was released 31/10/02. See: <http://software.eprints.org/newfeatures.php>
11. Chan, Leslie and Kirsop, Barbara, “Open Archiving Opportunities for developing countries: towards equitable distribution of global knowledge”: *Ariadne* Issue 30, [January 2002]. - <http://www.ariadne.ac.uk/issue30/oai-chan/>

**Philip Hunter** is currently leading the Open Archives Forum - an EU-funded project to promote Open Archives ideas in Europe. He was editor (and occasional scribe for) [\*Ariadne\*](#) for five and a half years (from October 1997 to March 2003) as part of UKOLN's Information & Communications and Web-support teams. He is now part of the UKOLN Research Team.

**Marieke Guy** is part of UKOLN's QA Focus team, which ensures that projects funded as components of the JISC Information Environment comply with standards and recommendations and make use of appropriate best practices. She has also co-ordinated technical support and advice services for the NOF national digitisation programme. Marieke was editor of *Cultivate Interactive* magazine until December 2001.