Submission for ECDL 2001 OAI Workshop

Enhancing OAI Metadata for Eprint Services: two proposals

Tim Brody, Zhuoan Jiao, Steve Hitchcock, Les Carr and Stevan Harnad
Open Citation Project, IAM Research Group, Department of Electronics and Computer Science, University of Southampton SO17 1BJ, UK
Contact email: sh94r@ecs.soton.ac.uk

The Open Archives Initiative has always maintained a distinction between data providers and service providers. This works at a functional level; some current projects are showing that it is less significant at an operational level. The Open Citation project harvests reference data from full-text eprint archives for reference linking, citation analysis and citation-ranked search. These are regarded as service provider functions where the end-user interacts directly with the services. The aim is to make these data available for export back to the full-text archives and to other service providers. Thus OpCit becomes a data provider too. The current Open Archives Protocol for Metadata Harvesting says nothing about full-text data harvesting for services such as these, nor about the export of processed data. This short paper outlines two proposals for progress on these issues.

1 Introduction

Emerging Open Archives services are blurring the distinction between data and service providers introduced in the original Santa Fe Convention framework documents describing the Open Archives Initiative (OAI). (Van de Sompel and Lagoze 2000) This is highlighted by the recently announced Kepler framework, a broker-based peer-to-peer network architecture, such as that used by Napster, which provides individual authors with software to set up personal archives, hosts a registration server and harvests data for subsequent dissemination. (Maly *et al*. 2001) This blurring between provider functions has significant practical implications for OAI.

The OAI Protocol for Metadata Harvesting (Van de Sompel and Lagoze 2001) mandates simple, Dublin Core based metadata to achieve interoperability between archives with low overhead for archive maintainers. This focus on simplicity appears to be vindicated by the apparent demise of NCSTRL, a forerunner of OAI as a collection of distributed archives. (Krichel and Warner 2001b) It suggests, however, the expectation of data transfer between services at a fairly low level of functionality unless OAI data and service providers can supplement the basic metadata for particular application areas, as allowed in the OAI protocol.

An example of enhanced metadata is the proposed Academic Metadata Format (AMF), a parallel metadata set that can be deployed with basic OAI metadata and which is designed to be used by the eprint archiving community, or more generally 'to advance scholarly communication over the Internet'. (Krichel and Warner 2001a)

This is a welcome development. Once an OAI service provider itself becomes a data provider, it is inevitable that data output in this case will not be simple document metadata as offered by the original document archives (which might be a better description than 'data provider'), otherwise, what service would the service provider be offering?

This paper considers the enhancement of OAI metadata from the perspective of the Open Citation (OpCit) project, which is demonstrating reference linking and citation analysis services by working with the largest OAI-compliant archive, the Los Alamos physics archives (arXiv). The approach described is intended to be generalised for other Open Archives, and is presented here to promote discussion and participation in the process.

2 Reference linking and citation analysis

To provide reference linking and citation analysis requires more data than is included in the OAI metadata. In principle, the full reference list is required (the full text is required for in-text, context reference linking, i.e. linking the occurrence of the citation within the text to the reference at the end of the text). OAI metadata does not provide mechanisms to expose and harvest full content. (Warner 2001)

An early OpCit reference linking demonstrator was described by Hitchcock *et al*. (2000). The reference database continues to be kept-up-to-date and has since been restructured to store richer data, improving accuracy and robustness. This has enabled the service to be extended, providing forward (in time) citation links as well as a Google-like search service that ranks results according to citations or hits.

The project has achieved this by working in partnership with the arXiv maintainers. To extend this approach to other archives, the project confronts two questions:

1. How can reference lists be extracted from Open Archives within the framework of the OAI?
2. How can processed data be exported back to the original archives so that it can be visible to the users of the archives (rather than as a standalone demonstrator)?

Since OAI data is intended to be interoperable, and OAI service providers are envisaged as cooperative, it is reasonable to assume that other services, search engines or Web portals for example, not simply the original archives, may also want to import the processed data. Thus a generalised interface to expose the data is desirable.

In the OpCit application the schematic in Figure 1 shows data input and output in the context of the OAI. The only parts of this schematic mandated by the OAI are the target paper archives and OAI metadata output. OpCit is responsible for the citation database and specified user services. The subjects of the questions above are the production of the reference list, and data export (stages A and D, respectively, in the schematic).

2.1 Extracting and parsing reference lists

Experience of large archives shows that document maintainers are unwilling to permit automated downloading of full texts. Alternatively, OpCit has developed software to identify and extract reference lists from papers, and this is available for free download and use by archive maintainers to create a separate collection of reference list documents, saving the main document server from possible overload.

This series of Perl modules is available from http://arabica.ecs.soton.ac.uk/code/doc/ReadMe.html, and includes:

- Markup_TeX.pm: inserts 'xxxOpCit' at the beginning of each reference in the TeX source file. This mark-up is used by 'Parser_DVI.pm to identify each reference;
- TeX2DVI.pm: converts TeX/LaTeX to a DVI file, then DVI to text by 'dvitype' (Unix command);
- Parse_DVI.pm: parses the text file created by 'dvitype' to produce a list of references;
- Citation.pm: parses each reference (citation) string to discover its metadata (authors, journal, volume, issue, etc.)

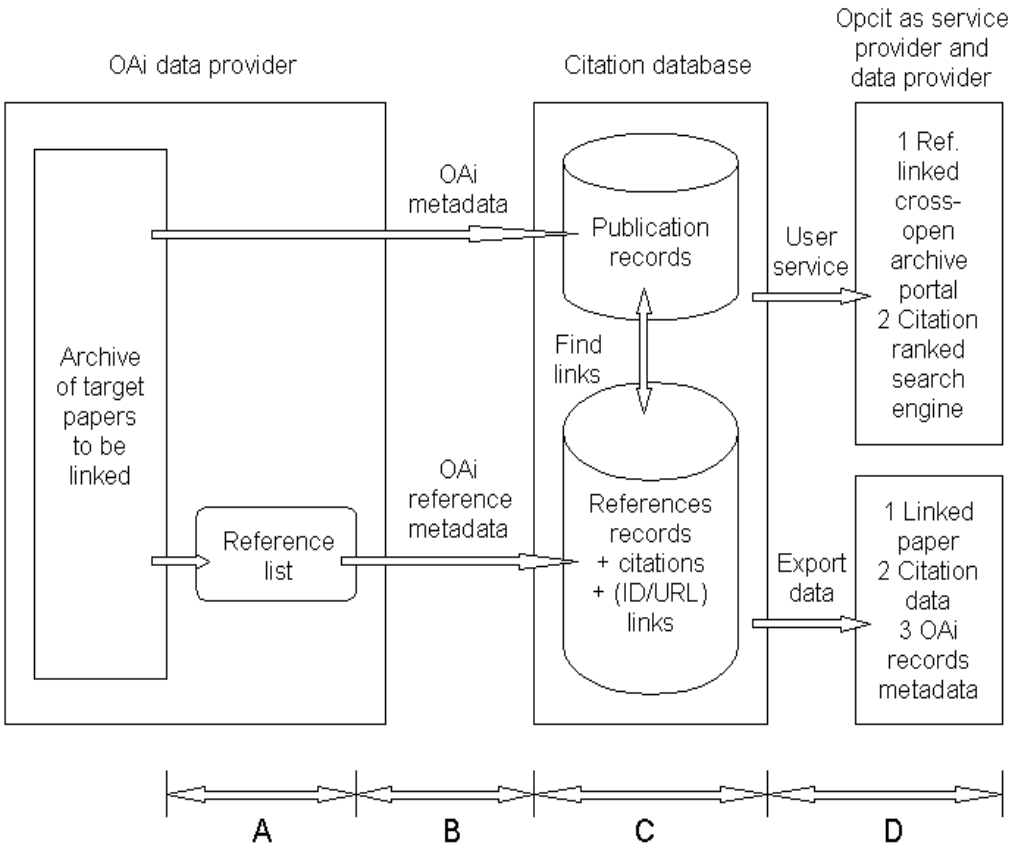

Figure 1. Proposed schematic of data input and output from the OpCit citation database

This software adds little manual overhead to maintenance beyond initial set-up, and imposes no requirements on authors. Initially these modules are optimised for arXiv because of the common TeX format found there, although other versions have been used with pdf and html. Perl scripts that make calls to the above modules can be developed locally.

2.2 Citation database

The parsed reference data extracted using these scripts are stored in the classic citation database structure shown in Figure 1 along with the conventionally harvested OAI metadata. Comparing the reference records with the publication records for an archive enables links to point at those referenced documents held in the archive. Also, for each publication record it is possible to use the reference records to determine if it has been cited. This can be displayed in a number of ways, as shown in Figure 2.

Within the project this database, called cite-base, is also used to serve the cite-baseSearch engine (http://cite-base.ecs.soton.ac.uk/cgi-bin/search) that ranks results according to citations or hits, as selected by the user.

2.3 Data export

While the project is capable of managing data capture, database maintenance and the user interface within the current framework, a more intriguing prospect is exporting data, principally back to the archive maintainers so that the services demonstrated above can be made available to users of the archive, but also to other OAI service providers.

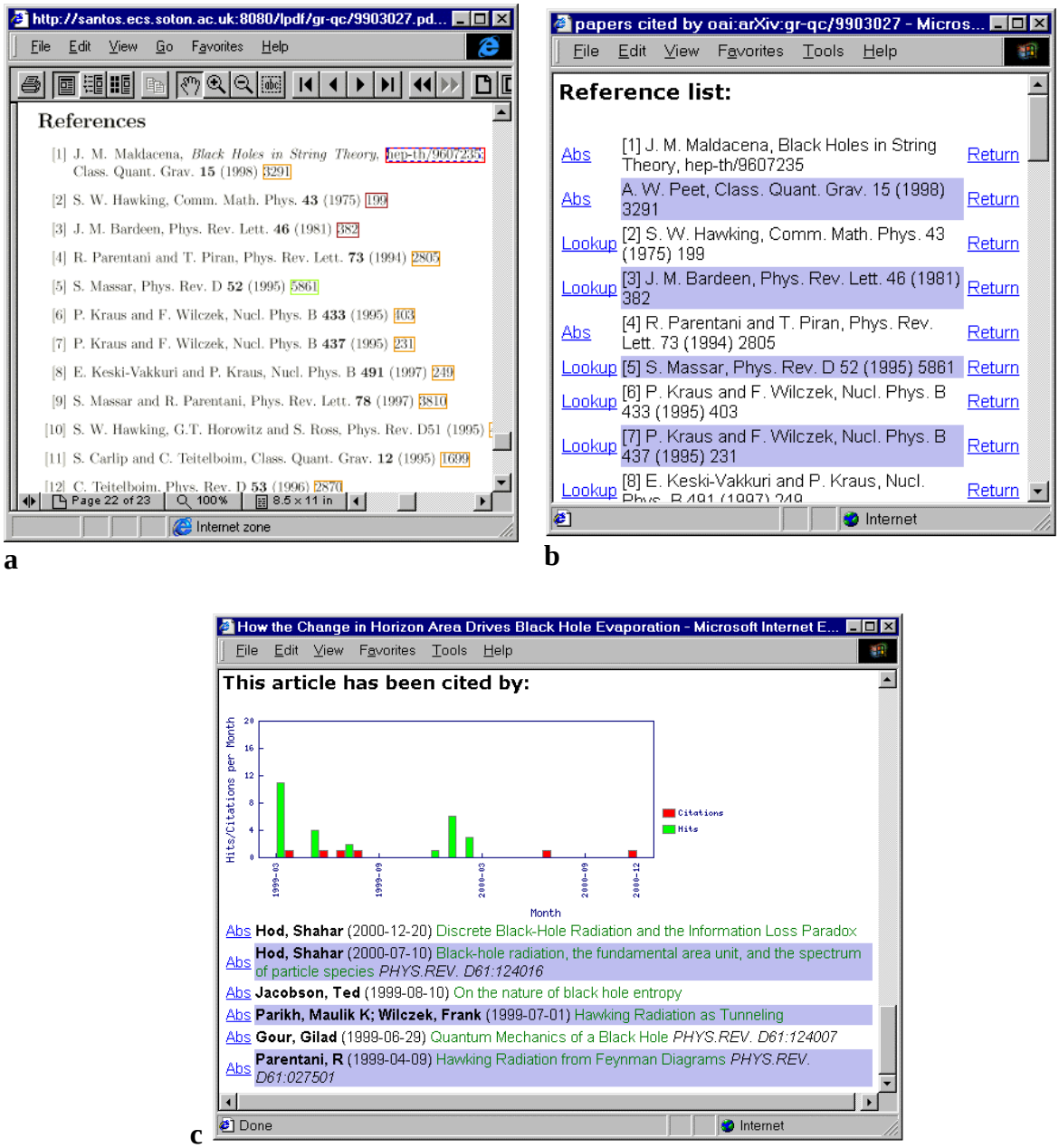

Figure 2. Reference and citation lists for a paper in the physics archives: **a**, reference links added to the original pdf; **b**, the same reference list extracted from cite-base with direct links (Abs) to texts and Lookup links to search for e.g. other works by the cited authors where the cited document is not available; **c**, citations *to* the paper from which the reference list was taken.

OpCit's immediate requirement is for a metadata format to express the reference lists (and hence the referenced articles), along with fields such as title, abstract, etc. This can be expressed using Dublin Core metadata, specifically the *relation* attribute. An in-house format, "opcit_dc" (Figure 3), augmenting Dublin Core with a *citation* attribute, was developed to store structured data for references (title, author, year of publication). *Citation* allows the reference and linked identifier to be expressed (with hindsight this is probably possible using Dublin Core qualifiers, without having to modify the schema).

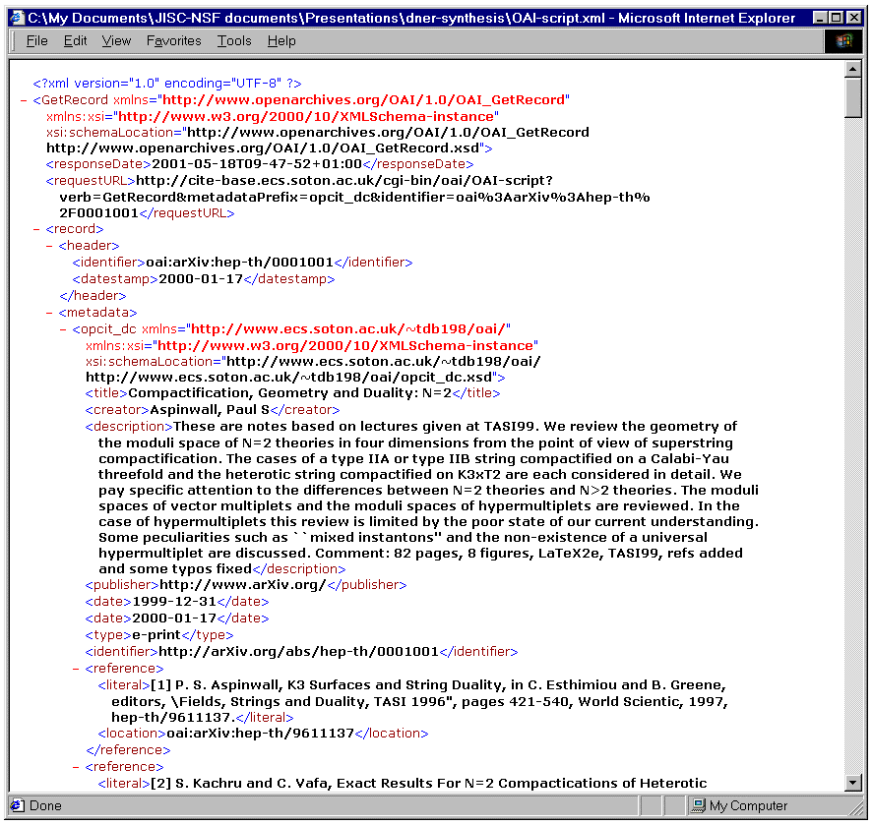

Figure 3. Example citebase record in "opcit_dc" format

The development of the AMF offers the chance to extend this approach for OpCit and OAI archives that handle research papers. AMF is a relational model for data, e.g. two documents are related by a reference. These relations can be expressed in either direction, e.g. AMF can express all the papers by an author, or all the authors of a paper. AMF will work best when the "noun" objects (texts, people, organisations) can be uniquely identified, which will require new identification systems. Current metadata is quite "weak", so the ability to be able to compare the impact of authors, say, is difficult within a large community.

It is planned to use AMF to transfer reference data between OpCit and arXiv. Another Open Archive, RePEc (Krichel and Warner 2001b), which is a database of papers on economics, plans to use AMF to implement the OAI protocol, and such a rich information resource could be used by others to implement, e.g. academic Web portals.

3 Conclusion

The Kepler framework is allied with a federated OAI search service called Arc. Data harvesting by Arc revealed that not all archives strictly follow the OAI protocol, and although the OAI validates registered data providers for protocol compliance and conformance with XML, this verification is not complete. (Liu *et al*. 2001) There are two possible responses: at the level of the OAI protocol, or work-around solutions by service providers. It seems clear the OAI wants to boost content-based archives by offering as few barriers as possible to data providers, so the Arc developers seem to anticipate the latter response.

In the case of the Open Citation project and arXiv, service and data providers together recognise the need to supplement the basic OAI metadata to improve functionality and performance.

Two proposals for the OAI community to consider emerge from this brief description of the reference linking and citation analysis work of the OpCit project:

- Richer formats are required to supplement the basic OAI metadata and to expose data for transfer between service providers and archive maintainers. The Academic Metadata Format appears to be a good foundation for this, and others are encouraged to participate in the development and review of AMF so that it might be adopted with some consensus.
- Reference lists need to be available for automated download from archives. Modular software aimed at archive maintainers, which automates the extraction and mark-up of references from submitted papers, is available for free download from the OpCit project.

References

Hitchcock, Steve, *et al*. (2000) Developing Services for Open Eprint Archives: Globalisation, Integration and the Impact of Links. *Proceedings of the Fifth ACM Conference on Digital Libraries* (ACM: New York), 143-151, June. Version available at http://opcit.eprints.org/dl00/dl00.html

Krichel, Thomas and Warner, Simeon M. (2001a) A Metadata Framework to Support Scholarly Communication. Submitted to the *Dublin Core Conference*, Japan, October. Draft available at http://openlib.org/home/krichel/kanda.html

Krichel, Thomas and Warner, Simeon M. (2001b) Disintermediation of Academic Publishing through the Internet: an Intermediate Report from the Front Line. *ICCC/IFIP Conference on Electronic Publishing*, Canterbury, UK, July. http://openlib.org/home/krichel/sants.html

Liu, Xiaoming, *et al*. (2001) Arc - An OAI Service Provider for Digital Library Federation. *D-Lib Magazine*, Vol. 7, No. 4, April. http://www.dlib.org/dlib/april01/liu/04liu.html

Maly, Kurt, Zubair, Mohammad and Liu, Xiaoming (2001) Kepler - An OAI Data/Service Provider for the Individual Access. *D-Lib Magazine*, Vol. 7, No. 4, April. http://www.dlib.org/dlib/april01/maly/04maly.html

Van de Sompel, Herbert and Lagoze, Carl (eds) (2001) The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 1.1, 2 July. http://www.openarchives.org/OAI/openarchivesprotocol.htm

Van de Sompel, Herbert and Lagoze, Carl (2000) The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, Vol. 6, No. 2, February. http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Warner, Simeon (2001) Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial. *HEP Libraries Webzine*, No. 4, June. http://library.cern.ch/HEPLW/4/papers/3/