Final Report to the NSF:

The Open Citations Project: Integrating and Navigating Eprint Archives through Citation Linking

Project Number IIS-9907892

Project Background and Motivation

For decades there have been frequent and intense discussions about the real and potential effects that open access to research papers may have on the scholarly community and associated communities (government and industrial laboratories, educators, etc.). While some of these discussions precede the Internet, certainly the rapid emergence of the Web has amplified them. The ability of the Web to immediately broadcast research results combined with the increased rapidity of results in many fields have added new pressure on publishers, scholarly societies, libraries, and other information intermediaries to development new solutions.

The recognition of the Web as a technical foundation for new publishing solutions has led to a number of changes in scholarly publishing of the last decade. A class of new alternatives collectively known as *Eprint Archives* has emerged that offer scholars the opportunity to *self archive* research papers without the mediation of traditional publishers. Some well-known instances of these eprint archives include the physics arXiv[1] developed originally at Los Alamos National Laboratory and run now at Cornell University, the CogPrints archive[2] for cognitive sciences run out of University of Southampton, the Cryptology ePrint Archive[3] run out of UCSD, and the more journal-oriented BioMed Central[4]. A few of these, most notably arXiv, have led to structural changes in scholarly exchange in their target communities.

These new publishing solutions raise important questions in a number of areas. What are the economic models for these archives? How do these new publishing vehicles interact with traditional models for academic promotion and tenure? How does peer review as a vehicle for quality assurance transfer to this new domain? What are the cultural barriers within specific communities that interfere or promote these changes?

The Open Citations Project[5], referred to as OpCit for the remainder of this report, has over the past three years provided an international focus for research in eprint publishing solutions. While its main concentration has been technical, the project has moved

[1] http://www.arxiv.org
[2] http://cogprints.ecs.soton.ac.uk
[3] http://wprint.iacr.org
[4] http://www.biomedcentral.com
[5] http://www.opcit.eprints.org

forward with the belief that solutions to a number of core technical issues provide the context for examining these non-technical contextual issues. In addition, the project has been examining and developing metrics for the effectiveness of open archiving by extending traditional citation analysis techniques. We are confident that as a result of the OpCit Project, a broadly-based campaign for raising awareness of open access, embracing the Open Archives Initiative (OAI) and the Budapest Open Access Initiative, can now be complemented by software for building open-access eprint archives, GNU EPrints (also known as eprints.org software), and a citation-ranked search engine for open archives, Citebase. Together these tools enable authors to provide open access to their papers, and to measure impact by citation measures as well as usage measures.

Summary of Project Activities

The principal partners in the Open Citation Project were Southampton University's IAM Group, the Digital Library Research Group at Cornell University, and arXiv, at the outset of the project based at Los Alamos and now hosted at Cornell. The project has consistently followed a management model with Southampton taking the lead, as the primary recipient of funding, with the Cornell group undertaking research in support of the goals established by Southampton. This report combines results from both groups, reflecting the cooperative nature of the work throughout the project.

The method used by the project at Southampton has been to build tools to measure and analyze citations from the 200,000+ papers stored by the arXiv physics archives, the largest eprint archive of its type. These data have been complemented, experimentally, with data on how the archives are used, e.g. which papers are viewed most. Collectively the citation and usage data are stored in Citebase, a citation database which provides a user interface for search and discovery, and a machine interface for analysis of this rich data source by other services.

The major activity at Cornell in connection with the Southampton effort has been development of the Open Archives Initiative Protocol for Metadata Harvesting[6] (OAI-PMH), which has been funded by a number of NSF grants and outside funding. The OAI-PMH is now internationally recognized as a fundamental building block for information federation. With the emergence of OAI-PMH and the consequent emphasis on institutional archives, it was evident there would be a need for large numbers of local, institution-based archives smaller than arXiv, but which would need to operate on similar principles — low cost, largely automated deposit, offering indexing and dissemination of author-archived content. Software used to build CogPrints, a cognitive sciences archive modelled on arXiv, was rewritten to make it OAI-compliant, and then to make it generic. This became the basis of GNU EPrints[7], which was further developed as part of OpCit to generalize the author and management interfaces for open-access archives.

Of most significance, EPrints builds archives that comply with the OAI Protocol for Metadata Harvesting (PMH). This means that any content deposited within an EPrints-based archive will become visible to users of independent OAI services, such as Citebase,

[6] http://www.openarchives.org
[7] http://www.eprints.org

immediately enhancing the chances of discovery. Authors depositing papers in an EPrints archive are not required to have any knowledge of OAI metadata, as it is generated automatically.

Connecting papers in open archives and a citation database is a method for automatically extracting metadata and reference lists from the papers. There are many different applications for reference linking. The Cornell team considered the question "what would be the ideal behavior of a digital object that supported reference linking (both incoming and outgoing)"? Answering this question led to an application programming interface (API) for reference linking.

All three components have been tested, evaluated and demonstrated to be useful by third-party users, and will continue to be developed and integrated within new projects and products beyond the lifetime of the OpCit project.

Details on Project Activities

CiteBase

Citebase is a citation-ranked search and impact discovery service that measures citations of scholarly research papers that are available on the Web in the larger open access, OAI disciplinary archives - currently arXiv, CogPrints and BioMed Central. Citebase harvests OAI metadata records for papers in these archives, automatically extracting the references from each paper. The association between document records and references is the basis for a classical citation database.

The primary means for users of accessing this database is the Citebase Web interface[8]. The user can classify the search query terms (typical of an advanced search interface) based on metadata in the harvested record (title, author, publication, date). In separate interfaces, users can search by archive identifier or by citation. What differentiates Citebase is that it also allows users to select the criterion for ranking results by Citebase processed data (citation impact, author impact) or based on terms in the records identified by the search, e.g. date. It is also possible to rank results by the number of 'hits', a measure of the number of downloads and therefore a rough measure of the usage of a paper. This is an experimental feature to analyze the quantitative and the temporal relationship between hit (i.e. usage) and citation data, as measures of impact. Hits are currently based on limited data from download frequencies at the UK arXiv mirror at Southampton only.

The combination of data from an OAI record for a selected paper with the references *from* and citations *to* that paper is also the basis of the Citebase record for the paper. A record can be opened from a search results list. The record contains bibliographic metadata and an abstract for the paper, from the OAI record. This is supplemented with four characteristic services from Citebase:

[8] http://citebase.eprints.org/

- Graph of this Article's Citation/Hit History
- All Articles Cited by this Article (Reference List)
- Top 5 Articles Citing this Article (option to view All Articles Citing this Article)
- Top 5 Articles Co-cited with this Article (option to view All Articles Co-Cited with this Article)

Another option presented to users from a results list is to open a PDF version of the full paper. This option is also available from the record page for the paper. This version of the paper is enhanced with linked references to other papers identified to be within arXiv, and is produced by OpCit. An earlier evaluation[9] found that arXiv papers are the most appropriate place for reference links because users overwhelmingly use arXiv for accessing full texts of papers, and references contained within papers are used to discover new works.

Prior to a more recent evaluation[10], Citebase had records for 230,000 papers, indexing 5.6 million references. By discipline, approximately 200,000 of these papers are classified within arXiv physics archives.

GNU EPrints

EPrints is software for building open-access archives aimed at institutions and special-interest communities, and is now used by nearly 60 archives.

In its current incarnation, the name GNU EPrints reflects that it is open source and freely available under the GNU General Public License and conforms to the strict GNU guidelines for free software. The last major release of EPrints, version 2.0, appeared in February 2002, although it has been updated (now on version 2.2.1) to conform with the latest OAI-PMH (also version 2) announced in June. Features of EPrints version 2 include:

- Internationalized metadata stored as Unicode
- Support for multiple archives on one server
- An improved user interface

EPrints is extending its focus on institutional research papers. It is now configurable for adoption as a journal-archive, e.g. *Behavioral and Brain Sciences* and *Psycoloquy*, by new open access journals or established journals converting to open access, and will include the facility to manage peer review and peer commentary.

Reference Linking API

The API automatically extracts metadata and reference lists from papers using four principal methods:

[9] http://opcit.eprints.org/evaluation/v10/v10evaluation.html
[10] http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report.html

1. getMyData() - the digital object should emit standard metadata describing that object, i.e., title, authors, year of publication, etc. in Dublin Core format.
2. getReferenceList() - the digital object should say what its list of references is (this is the fixed number of references contained in the online document).
3. getCitationList() - the object can say what other works the object knows have cited it. (This list grows as more and more items are analyzed.)
4. getLinkedText() - returns the original content of the digital object but with link information added to it so that each reference can be used to go directly to an online copy of the referenced work, if an online copy is available.

Each component produced by these methods can be seen in a typical Citebase record, but this approach is generalizable to other reference linking applications.

A few Java classes were defined to support reference linking in an object oriented way. These methods can be invoked on the surrogate, a special class in the API that encapsulates data regarding a particular online digital object. To use the API, a new surrogate is instantiated, passing it the URL of the online digital object for which information is to be gathered.

The bulk of the analysis within the API program is done by the surrogate constructor. This call downloads the online work, turns it into XHTML, parses the XHTML, and extracts information, such as citations and references. The next call on the API invokes the method that returns the references in the form of an XML document, which is then converted to a string and printed.

It is anticipated that repositories will at some point contain reference linking data, so the API was later extended to support persistent storage of surrogates. Once a surrogate is instantiated, it can be saved to a repository, if desired. Thus one could build a repository of surrogates, which could later be re-instantiated and have the basic API methods invoked on them.

The API was used to build several applications against online journals (*D-Lib Magazine*, *Journal of Electronic Publishing*, ACM Digital Library). With five methods (the original four, plus save) the API was found to be sufficiently usable. The main limitation of the software is that not all HTML pages are equally easy to analyze, e.g. some HTML is badly written and cannot be converted into XHTML and, therefore, cannot be parsed. This is likely to remain a problem on the Web for some time.

Project Legacy

All three components described above are usable and will continue to be so beyond the conclusion of OpCit. What is available, the means of access, and plans for maintenance of services, are noted below:

- Citebase is now up-to-date and indexes arXiv fully. Citebase can be searched by users at http://citebase.eprints.org/. A machine interface for data sharing with other services is operational, and Citebase is listed as an OAI 2.0-conforming data

provider (http://www.openarchives.org/Register/BrowseSites.pl). Researchers at Old Dominion University have harvested Citebase data as part of their Archon federated digital library on physics (http://archon.cs.odu.edu/), as has the OAI search engine OAIster (http://oaister.umdl.umich.edu/o/oaister/viewcolls.html#c), and arXiv is a possible (re)harvester of Citebase data too. The citation database will continue to be updated and expanded in terms of coverage. Both interfaces to Citebase will continue to be developed and maintained.

- GNU EPrints is available as open source software and is downloadable from http://software.eprints.org/. Machine requirements for running GNU EPrints are other open source components including Linux, Apache Web server, Perl and a MySQL database. GNU EPrints will continue to be developed and maintained.
- The Reference linking API was written in Java and is downloadable from the OpCit project site at Cornell http://www.cs.cornell.edu/cdlrg/Reference%20Linking/. There is no further work planned on the API.

One other significant legacy of the project is planned work within the OAI framework on ePrint specialized profiles. Experience with the OAI indicates that is a core with sufficient utility and extensibility to be useful in a variety of contexts. As a result of experience within OpCit the OAI is planning to proceed with leveraging the extensibility capabilities of the core protocol to provided eprint-specific functionality:

- Develop a metadata for exposing references via the OAI-PMH.
- Develop a metadata standard for exposing author information via the OAI-PMH.
- Develop a standard profile for exposing eprint repository information via the OAI-PMH Identify verb.
- Work with the JISC-funded RoMEO project[11] to develop simple standards for exposing rights metadata from eprint repositories.

Selected Project Publications

C. Gutteridge (2002) "GNU EPrints 2 Overview". Author eprint, Dept. of Electronics and Computer Science, Southampton University, October, and in *Proceedings 11th Panhellenic Academic Libraries Conference,* Larissa, Greece, November http://eprints.ecs.soton.ac.uk/archive/00006840/

D. Bergmark and C. Lagoze. An architecture for automatic reference linking. In *ECDL, September 2001*, Sept. 2001. <http://www.cs.cornell.edu/bergmark/www10.pdf>.

D. Bergmark, W. Arms, and C. Lagoze. An architecture for reference linking. Technical report, Cornell University, Digital Library Research Group, Oct. 2000. TR 2000-1820.

D. Bergmark. Automatic extraction of reference linking information from online documents. Technical Report TR 2000-1821, Cornell Computer Science Department, Nov. 2000.

[11] http://www.lboro.ac.uk/departments/ls/disresearch/romeo/

D. Bergmark. Link accessibility in electronic journal articles. Technical Report TR 20001793,Cornell Computer Science Department, March 2000. http://www.cs.cornell.edu/bergmark/LinkAnlaysis.ps

D. Bergmark., Phempoonpanich, P. and Shumin Zhao, S. (2001) ”Scraping the ACM Digital Library”. *SIGIR Forum*, Vol. 35 No. 2, Fall http://www.acm.org/sigir/forum/F2001/bergmarkFinal.pdf

S. Hitchcock, L. Carr, Z. Jiao, D. Bergmark, W. Hall, C. Lagoze, and S. Harnad. Developing services for open eprint archives: globalisation, integration and the impact of links. In *5th ACM Conference on Digital Libraries, San Antonio, June 2-7, 2000.*

S. Hitchcock, L. Carr, Z. Jiao, D. Bergmark, W. Hall, C. Lagoze, and S. Harnad. Developing services for open eprint archives: Globalisation, integration and the impact of links. In *5th ACM Conference on Digital Libraries, San Antonio, Texas, June 2 - June 7, 2000, also titled ACM Proceedings of Digital Libraries, 2000 (DL2000),* San Antonio, Texas, 2000.

S. Hitchcock., Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., Lagoze, C. and Harnad, S. (2002a) “Open Citation Linking: The Way Forward”. *D-Lib Magazine*, Vol. 8, No. 10, October http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html

Steve Hitchcock, Les Carr, Wendy Hall, Stephen Harris, S. Probets, D. Evans, and D. Brailsford. Linking electronic journals: Lessons from the Open Journal project. *D-Lib Magazine.* December 1998.

S. Hitchcock., Woukeu, W., Brody, T., Carr, L., Hall, W. and Harnad, S. (2002b) “Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service”. Evaluation report, IAM Dept., University of Southampton http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report.html

T. Brody., Carr, L and Harnad, S. (2002) “Evidence of Hypertext in the Scholarly Archive”. *Proceedings of HT'02, the 13th ACM Conference on Hypertext*, University of Maryland, June 2002 http://opcit.eprints.org/ht02-short/archiveht-ht02.pdf

OpCit Principals

Cornell University

- Donna Bergmark
- Carl Lagoze (Principal Investigator)

University of Southampton

- Tim Brody

- Les Carr
- Christopher Gutteridge
- Wendy Hall
- Stevan Harnad (Principal Investigator
- Steve Hitchcock (Project Manager)
- Zhuoan Jiao
- Robert Tansley