The Open Citation Project
Second Year Report to JISC

Reference Linking for Open Archives
http://opcit.eprints.org/

Version history
This version 1.0 submitted to JISC 31st August 2001
Period covered in report: from September 2000
First year report http://opcit.eprints.org/y1report/y1report-final.pdf
Contact: Steve Hitchcock, sh94r@ecs.soton.ac.uk

The Open Citation project is a collaboration between Southampton University, Cornell University and arXiv.org, funded by the Joint NSF – JISC International Digital Libraries Research Programme.

Summary

In its second year the Open Citation project has made significant progress on its primary deliverables and has initiated some important new developments.

The project has exceeded plan with some of its major deliverables, in particular, producing a citation-linked demonstrator, including citation analysis and ranking, for the whole of the physics arXiv. Progress has been made towards integration of this service with arXiv to a degree unanticipated in the previous report, which if successful would put the results of the project before a large user base on a long-lasting basis.

Attention to technical developments in Year 2 has resulted a small restructuring of previously reported work plans. In Year 3 the project will emphasise analysis, evaluation and dissemination. With its linking demonstrators, data mining results and user surveys, the project has amassed a rich resource that will be investigated more fully in the final year, giving important original insights on the impact of eprint archives in scholarly communication.

In terms of activities, the highlights of the year are:

- Development of a richer, more stable arXiv citation database
- Prototype citation-ranked search engine for arXiv
- Enhanced citation-linked arXiv demonstrator, includes ‘cited by’
- Progress towards integration with arXiv
- Demonstrator reference linking API in a presentation/rendering application
- Open sourcing of key software modules for reference linking
- New version releases of EPrints software and support tools
- Survey of users of eprint archives
- Proposals for extended metadata transport between OAi service providers
- A one-day seminar for our technical collaborators

Contents of the report

1 Introduction

In the first year report we highlighted the important relationship between the project and the then emerging Open Archives initiative (OAi) and the EPrints.org software effort. This year we highlight another key partnership, with the physics arXiv.

Led by Paul Ginsparg, arXiv has been a partner in OpCit from the outset. This gave the project access to full-text content and arXiv usage data, much of which had not been available or explored before. The project demonstrated and evaluated a reference-linked model of the whole archive during year 1. What was not clear was where this would lead next. Important new features were planned, possibly cross-linking with other archives. A remark from one of the principal technical support team at arXiv during the evaluation of the demonstrator indicated an unexpected opportunity: “I see no reason why we couldn't incorporate this in arXiv at an early stage”. There were qualifications, but this would put the approach promoted by the project before the largest possible user base on a long-lasting basis. Through the year we worked on numerous schemes that we hoped would enable integration of OpCit data with papers *served from* arXiv, but what has taken us closest to realizing this was not anticipated. This process is described in sections 2 and 4 of this report.

During the year the OAi has achieved wider recognition, especially among digital library projects, reflecting the broadening of its scope from eprint archives. In the UK the DNER has begun consulting about how it might support OAi. The OpCit project remains committed to OAi, and continues to support it actively: our Cornell partners now host OAi.

The process of clarifying the role of OpCit within the OAi framework led to a re-appraisal of the project’s data collection and management procedures, and indirectly to the current plan for integration with arXiv. Two new technical proposals supporting a stronger relationship between OAi archive maintainers and service providers are to be presented at an OAi workshop in Darmstadt during September 2001 [1].

Where OpCit-style reference linking is likely to play an important role for OAi in the future, Eprints.org software is a driver for OAi now. OpCit has supported EPrints financially during the current year, in which new version releases have appeared, notably to conform with version 1.0 of the OAi metadata harvesting protocol, a major revision to the earlier protocol (and subsequent revisions). The need to update the software rapidly and to prepare it for open-sourcing, the need to support a growing number of users, and other activities associated with the success of OAi, mean that OpCit support is insufficient. From October EPrints will be separately funded by JISC within the *OPSIS: OPen-Sourcing Institutional Self-Archiving* project. EPrints and OpCit together remain integral parts of the work at Southampton to enable academic institutions to build and maintain interoperable eprint archives with, for the user, the widest possible scope and the latest features and services. At Cornell, EPrints may replace Dienst, a software system and protocol for managing distributed digital libraries, on which the original OAi protocol was based. OpCit will add a reference linking module to EPrints to simplify the task of interlinking various archives.

2 Activities and progress

The main outputs during Y2 have been:

- A rich citation database based on arXiv physics archives
- Prototype citation-ranked search engine for arXiv
- Enhanced citation-linked demonstrator (v. 2.0), includes 'cited by'
- Demonstrator reference linking API in a presentation/rendering application
- Open sourcing of key software modules for reference processing
- New version releases of EPrints software and support tools

During Y1 the project produced a pilot reference linking demonstrator (v. 1.0) covering the whole of the arXiv physics archives, and defined a programming interface (an API) for inter-archive reference linking. In Y2 the reference linking demonstrator (v. 2.0) has been extended to present full citation analysis of the physics archives, and the API was implemented in a working example.

2.1 Citation analysis: building a citation database

The most important new feature of the linked arXiv demonstrator (v. 2.0) is the ability to discover what later papers in arXiv have cited a selected paper from arXiv. Linked references are useful and can save the user time, but the purpose of a reference is to direct a user to the cited source, which can be found, however laboriously, for any formally correct reference. In contrast, the user cannot derive a complete list of citations of a work unaided. Citations that lead a user forward in time map the development of an active area of research to its present by means of the most significant, most highly cited, papers. ISI has demonstrated the value of such services for many years. **OpCit is the first to successfully apply this approach to a large-scale eprint archive.**

At the heart of the service is a richer database of citations than was available in Y1, in which simple metadata – year, volume, page number – were used to identify a known paper in the archive. Where these data were recognised in a reference a link to the referenced paper was inserted.

The development of the enhanced database has been described by Jiao [3t]. The main tasks involved in building the citation database were:

1. Extract reference lists from full-text papers
2. Discover metadata in references
3. Insert metadata entries in the citation database
4. Link references

The database is structured with the following tables:

- Reference table:
    - reference text (e.g. [1] Z. Bern, et al. Phys. Lett. B401:273)
    - metadata extracted from the references
    - reference ID (e.g. 1, 2 ... an assigned integer)
    - source paper ID (e.g. arXiv:astro-ph/0001001)
    - feature ID (e.g. v4:p20:y1999)
    - authors (lists all authors for a paper) and first author

- Publication table:
  - metadata from the archive data provider
  - feature ID
- Abstract table
  - source paper ID, article title, abstract
- Links table:
  - source paper ID, reference ID, target paper ID

This database has served the construction of a prototype citation-ranked search engine, which in turn has become the initial interface for the latest demonstrator, and produces the full citation analysis of a paper.

2.2 Citation-ranked search

Design of the prototype search engine, cite-baseSearch, was first undertaken as part of a final year undergraduate project. The student worked closely with the project researchers to obtain the necessary data, and also informed the ongoing development of the citation database.

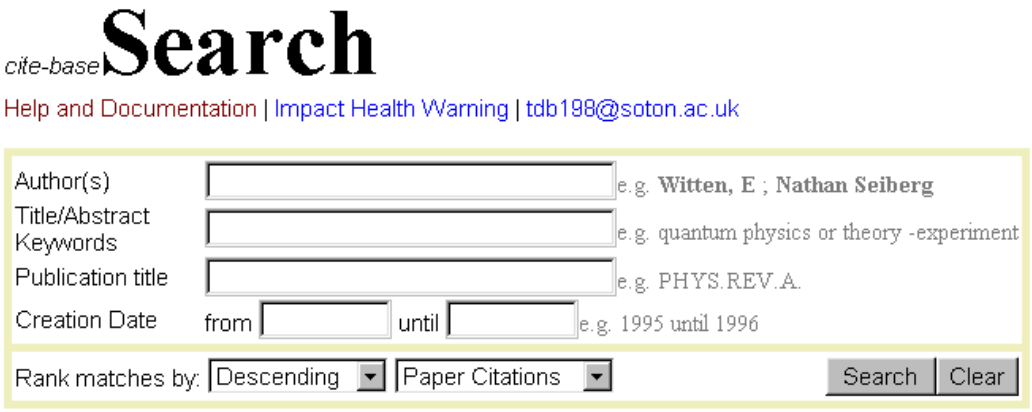

Figure 1. User interface for the cite-baseSearch citation-ranked search engine, now adapted for use in the OpCit linked arXiv demonstrator

This project provided the first example of a citation record for a selected paper. In this case the user selects the paper by typing information known about the paper in a conventional search interface (Figure 1). Any means of identifying a paper, such as a reference link, can serve as input to the search engine, so it became possible to use this output as the means of providing forward linking citation information as part of the OpCit demonstrator. Cite-base is no longer just a separate search service.

The impact of this project on OpCit has been significantly greater than this simple description suggests, as is explained below in section 2.4.

2.3 Reconsidering data collection: storage and presentation formats

The aim of making OpCit data available through arXiv had an effect on the data collection as well as the data dissemination process.

Most physics papers are deposited in arXiv in TeX formats. In October 2000 it became apparent that arXiv had begun rudimentary reference linking where the archive ID of a paper was given explicitly in a reference. Examination of various document formats available from arXiv revealed that the link data were added in the TeX version. To add OpCit's additional linking information – the project can link references to other arXiv papers without an ID – to that of arXiv we realised it would be necessary to add these data to the same source. Prior to that the project was working exclusively with downloaded PDF documents for both reference extraction and linking.

We started downloading papers in TeX as well as PDF, which presented new challenges in processing the reference data. Data extracted from 150k+ documents by different methods is unlikely to be identical. Analysis of the output from PDF documents suggested that reference lists could be extracted successfully from some 80% of documents in the physics archives, for TeX similarly (interpreting author-defined macros are a common problem with TeX documents, for example), but these were not necessarily from the same documents. Reference data are now extracted solely from TeX sources. By combining the two data sets marginally more references could be extracted, but this would be time-consuming and has not been attempted.

2.4 Data export

The project holds local copies of almost all papers from arXiv with reference links added. But the project has few users; arXiv has many more users. The problem was how to make the project data available through arXiv. We explored a number of options that seemed natural extensions of the work the project had done, but which were unsuccessful initially – some of the lessons from this are reported in section 4.

The real power of the OpCit approach lies in its citation database. Development of the citation-ranked search engine provided the clue to the way forward because it needed to use output from the database. The developer drafted a simple XML-based format to extract citation records from the database. The insight was to recognise that if a common format could be agreed, citation records could be exported using the OAi protocol. In this way any recognised OAi service, including arXiv, could potentially re-use OpCit data.

The feasibility of this approach becomes more apparent if the role of OpCit is considered schematically within the OAi framework of data and service providers, as shown in Figure 2.

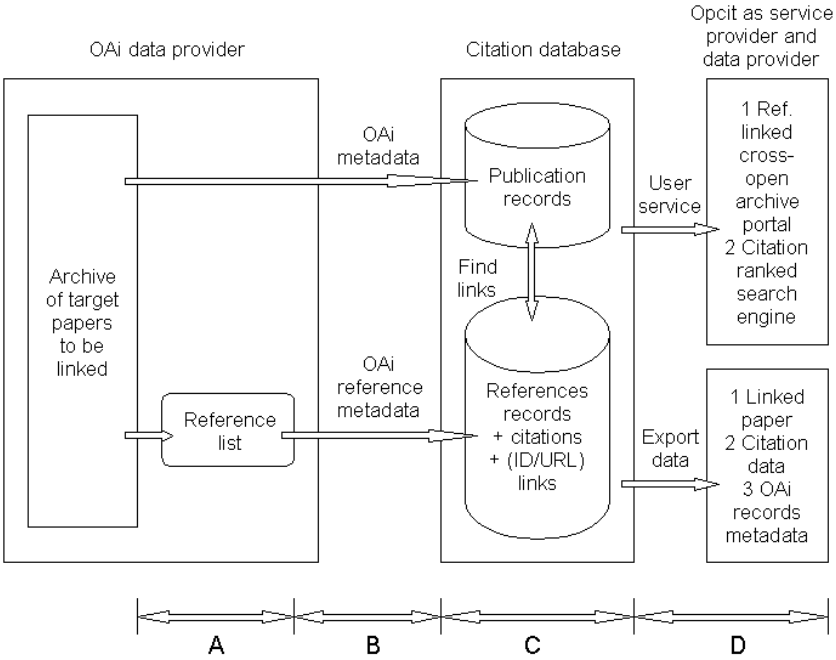

Figure 2. Schematic of proposed data input and output from the OpCit citation database within the OAi framework

The trick is to recognise that while the data in stage D could be exported to another OAi service provider, if we could loop this two-dimensional figure it could be shown that the so-called 'service provider', i.e. the recipient of OAi data, in this case from OpCit, could be the original OAi data provider in A, in this case arXiv.

The format would need to include more metadata than is included in the basic OAi metadata set, but could still be compatible with the protocol for data transfer. Coincidentally, a member of the technical support team at arXiv was co-authoring such a format, the Academic Metadata Format (AMF), aimed at the eprint archiving community. A meeting with one of the co-authors led to revisions that would enable AMF to handle OpCit data - see http://amf.openlib.org/doc/ebisu.html

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. Further development of AMF will support identification systems for authors and institutions, beyond OAI's current document identifiers. Using unique identifiers for authors, for example, will allow users to find the output of a particular researcher. Current systems cannot differentiate two authors with the same name.

During Y3 a primary task will be to support AMF in and out of the OpCit citation database. Then we will discover if AMF serves the purpose of allowing other services to use data directly from the OpCit citation database, thereby adding a new twist to the relationship between OAi data providers and service providers.

2.5 OpCit ArXiv demonstrator now

The combination of the OpCit citation database and citation-ranked search engine has provided a flexible solution in terms of presentation, one example of which can be seen in the current demonstrator (v. 2.0), shown in Figure 3, which can be explored at http://arabica.ecs.soton.ac.uk/

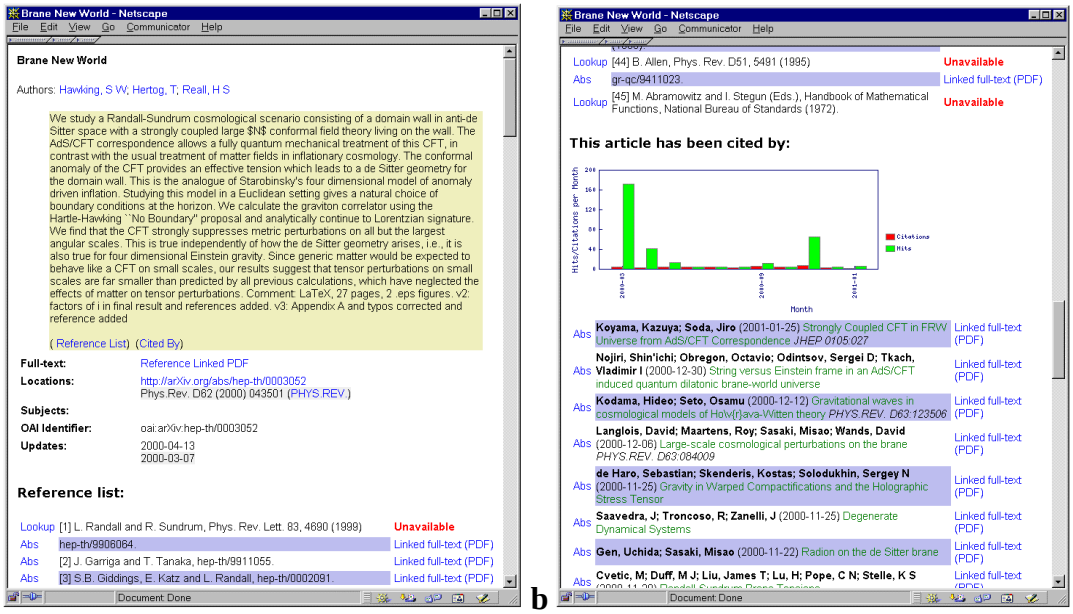

Figure 3. OpCit citation-linked version of arXiv v. 2.0, showing parts of the same record for a single paper: a, abstract and reference list; b, scrolling down the same page shows the list of papers citing the paper described by the record

This example provides continuing access to the reference linked PDF examples from Y1 combined with the new features created in Y2, among many other linked features. Figure 3a describes the version history of the paper and shows links to:

- author lookup in cite-base
- a reference-linked PDF version of the paper
- the arXiv page for the paper
- a journal lookup in cite-base
- similar records (abs) for the referenced papers
- linked PDFs for the referenced papers (some referenced papers are unavailable in arXiv)

Figure 3b shows graphically the popularity of the paper (based on limited user data from the arXiv Southampton mirror) and the incidence of citations over time. Links are presented to:

- similar records (abs) for the citing papers
- linked PDFs of the citing papers (all citing papers are available in arXiv)

This version of the demonstrator was announced to members of the project steering group on 20 August. It is immediately obvious that the interface needs further work, but we will await the results of experiments at arXiv to assess the presentational requirements of an updated version.

2.6 Demonstrator reference linking API in a presentation/rendering application

Complementary work at Cornell on developing reference linking has shown the benefit of working with data at a more abstract level. Where the Southampton work is migrating from a presentational to a representational form, Donna Bergmark at Cornell built a presentation application during Y2 while emphasising that the application programming interface developed in Y1 remained the critical element. It can now be seen how the methods defined by the Cornell OpCit team and the output of the Southampton experiments have converged. The design and formatting decisions taken by both teams allow scope for further integration during Y3.

The following is an edited extract from the Cornell team's Annual Report to the NSF: Project Number IIS-9907892, June 30, 2001 (Full version at http://www.cs.cornell.edu/cdlrg/Reference%20Linking/AnnualReportYear2.ps)

In the first year, we 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 API that included four methods:

1. `getMyData()` - the digital object should emit standard metadata describing that object, i.e. title, authors, year of publication, etc.
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.

In the second year, we finished implementing 1 and 2, and implemented 4. Implementation of 3 is underway as a student Masters of Engineering project. The digital object was implemented as a Java class, called Surrogate. To instantiate a Surrogate for an online document, one simply calls the Surrogate constructor with the URL of the online document. We decided to concentrate initially on HTML documents; work on PDF documents is underway at U. of Southampton, and has just begun at Cornell on the ACM online library.

Having implemented the `getLinkedText` method, we were next interested in using this with an archive search engine. This required building a repository of Surrogates to correspond to items in the repository. Hence we added a fifth method, `save()`, which causes a Surrogate object to save its instance data as short XML files in the appropriate directory. Correspondingly, a special constructor was implemented that would re-instantiate a Surrogate object using the saved XML documents.

Thus with the current implementation one can analyze an entire repository, save the surrogates, and then have them resurrected one by one, "on demand ". In other words, the user wishing to retrieve a document from the archive could (if he or she chooses)

be shown the linked version by resurrecting the appropriate Surrogate and invoking the `getLinkedText` method on it.

A demonstration of the API being used by a browser application can be viewed at http://cs-tr.cs.cornell.edu/RefLinkingDemo
This demonstration fulfils our goal of using the link data obtained by the analysis of an online document in a presentation/rendering application. We used XSLT to transform the link data into JavaScript code segments suitable for retrieving linked full text.

The API has been distributed to our colleagues at the University of Southampton, and is available to other academic users – see section 2.7 below.

The following design decisions were taken and have been found to be good ones:

1. Do everything in XML.
2. Use Dublin Core for bibliographic data.
3. Use a URN (Uniform Resource Name) based on the work's bibliographic data.
4. Use <reflink> elements to contain a reference's bibliographic data. <reflink> can then be translated to any of a variety of actual links, such as plain URLs, XLinks, OpenURLs, and so on.
5. Send all debug messages to *syserr*. In this way, an application can simply display the XML output, without intervening error messages. (See the demo for an example.)

During Y2 of this work, we ran a number of experiments.

1. Intralink *D-Lib*. At the end of year one we had `getMyData()` and `getReferenceList()` working for a single (randomly selected) *D-Lib* paper. In Y2 we processed hundreds of *D-Lib* papers.
2. We experimented with various presentation methods and came up with the demo given above.
3. We looked at *D-Lib*'s well-formedness. Of 280 journal items, 60 were sufficiently ambiguous or ill-formed that they could not be converted into legal XML with any precision. Thus about 75% of *D-Lib* papers can be converted into XML and analyzed.
4. For the papers that can be analyzed, we extracted reference linking information at better than 80% accuracy.
5. Analyzing a *D-Lib* paper takes about 5 seconds on average, on a Sparc 10.

2.7 Open source OpCit software

2.7.1 Software for reference extraction

A major limitation of the OAi protocol is that it provides no means for archive data providers to make reference lists, which are vital for reference linking services such as OpCit, available to data harvesting services. This is because reference lists are part of the full texts and OAi does not mandate that papers are accessible to harvesters. OpCit is proposing an optional approach for archive maintainers in which the maintainers automatically produce separate copies of the reference lists for harvesting. To support

this proposal, the project is making software, as developed and used by the project to extract reference lists from original source documents, available for free download.

The project has also received requests from partners and others for help in processing reference data from papers, for which this software may also be useful

A series of Perl modules are available from
http://arabica.ecs.soton.ac.uk/code/doc/ref/readme.html

These include two modules for extracting reference strings from TeX source files, and two modules that are specific to references typically found in physics papers in arXiv

Reference extraction

- `Extract::Reference` inserts mark-up 'xxxOpCit' at the beginning of each reference in a TeX source file, converts TeX/LaTeX to a DVI file, then DVI to text by 'dvitype' (Unix command), and parses the text file to produce a list of references
- `Parser::Citation` extracts metadata (authors, journal, volume, issue, etc.) from a reference string

arXiv-specific modules

- `Parser::arXivCite` a special case of `Parser::Citation`, processes arXiv IDs (e.g. hep-th/0001001) found in references
- `Parser::arXivRef` processes multiple citations within a single reference entry, as often found in physics papers

This software adds little manual overhead to archive maintenance beyond initial set-up, and imposes no requirements on authors. Initially these modules are optimised for TeX files because this is the most common submission format to arXiv. Other versions of the modules have been used with PDF and HTML files and may be available in the future.

Some of the reported problems when using this software include:

- Cannot find the start of the reference section
- Use of "unexpected macros" in source code to represent the references
- References presented in inconsistent styles – papers in archives are not subedited
- One reference contains several citations (common in physics papers), making it difficult to identify where to place the link
- Difficult to separate metadata elements, e.g. Mizuno, A., Makishima, K., & **Tashiro, M. PASJ**, 51, 663

2.7.2 Software for the Reference Linking API

Also downloadable is the Cornell Reference Linking API in Java, which can be used to enhance the functionality of document retrieval engines. The Java implementation runs on Linux, Solaris, Windows2K and NT. A key part of the implementation code is based on software originally written at the University of Southampton, and subsequently enhanced at Cornell. All other packages used (e.g.Tidy, Xalan, etc.) are

open source or are available under general public license. To download this software go to http://www.cs.cornell.edu/cdlrg/Reference%20Linking/

The Java application program returns a requested paper to the browser overlaid with JavaScript code. Each reference in the paper is made into a live link that produces a user dialogue giving the complete reference string and allowing the full text, if available, for the reference to be retrieved.

The Java application program does the following:

1 Imports one class from the Reference Linking API. In Java:
`import Linkable.API.Surrogate;`
2 Uses the URL for the paper to get a Surrogate, which determines reference linking information for that paper. In Java:

```
s = new
Surrogate("http://www.dlib.org/dlib/april98/04spink.html"
);
```

3 Uses the Surrogate to get linked text. In Java:
`byte[] linkedText = s.getLinkedText();`
The resulting byte array is an XML document suitable for further processing.
4 Runs an XSL stylesheet over `linkedText` to replace the <reflink> elements, which contain all the bibliographic information and urls of the reference, into JavaScript. This HTML document is the same as in item 3.
5 Outputs the HTML result of the XSL transformation to standard out, i.e. the browser screen.

A demonstrator of this process is at http://cs-tr.cs.cornell.edu/RefLinkingDemo/

2.8 A survey of users and non-users of eprint archives

Another component of OpCit initially undertaken as a final year undergraduate project was a survey of users of arXiv and of the CogPrints eprint archive, a three year old archive for cognitive scientists initially supported by the eLib programme. The work complements the mining of arXiv user data by the project ('mining the social life of an eprint archive') reported last year. That earlier work was based on evidence of what users actually do. This survey explored the users' own views and perceptions of what they *think* they do.

The initial findings suggest positive and negative implications for both ArXiv and CogPrints.

arXiv

- arXiv authors archive the earlier stages of their papers in the form of pre-prints, although many also go on to archive their refereed post-prints.
- arXiv authors may have lost sight of the comparative advantages of using archives, for example, that the archive offers a high level of visibility.
- arXiv users appear to overlook the causal role the refereed journal system is still playing in the research process. The responses indicate that many would

be happy with a journal-free landscape, even though most continue to submit to journals.
- arXiv authors are not held back by worries over copyright and plagiarism.

CogPrints

- CogPrints authors archive the later stages of their work in the form of refereed post-prints.
- CogPrints authors are aware of the importance of visibility through using both archiving and peer-reviewed publication.
- CogPrints authors cannot see the benefit and importance of archiving their pre-prints.
- CogPrints authors have concerns over copyright and plagiarism.

The preliminary results of the survey are at http://www.eprints.org/results/

The data produced by the survey are substantial and demand more detailed analysis, which the project is now undertaking, with wider dissemination in Y3.

2.9 EPrints update

Supported by OpCit, the first operational version (1.0) of the EPrints archive-creating software was released in December 2000. Development of EPrints was accelerated to keep up with the requirements of the completely revised OAi Metadata Harvesting Protocol v. 1.0, which was announced in January 2001. Compatible Open Archives code was included in EPrints v.1.1.

A replacement version of EPrints, with a completely new database, is at alpha-testing stage. EPrints 2 includes support for multiple archives on one server and international language support.

From October development and open sourcing of the core EPrints code will continue within the JISC Opsis project, although OpCit will build and test supplementary modules, e.g. for checking references in author submissions.

2.9.1 EPrints migration tool

To increase the number of EPrints-based archives, an EPrints 'migration' tool has been developed to upgrade existing archives and to rescue ailing ones by enabling them to become compliant with EPrints and therefore with OAi. The tool has been used to migrate the archives for both journals edited by Stevan Harnad, *Behavioral and Brain Sciences* and *Psycoloquy*. During the migration process the reference lists from papers in the archive were copied for inclusion in the EPrints record.

Another use of the tool may be to save the distributed institutional archives in computer science that have until now been managed within the large NCSTRL service, which is due to close at the end of September. Maintainers of the contributing NCSTRL archives could be invited to use the migration tool to upgrade to OAi.

3 Project management

3.1 Costs

Costs have been controlled and maintained within budget in most categories. An analysis will be submitted separately.

3.2 Staffing

There has been stability and continuity in staffing of the project. All permanent researchers who joined at the outset remain with project. There have been additional contributions from temporary contributors, including undergraduates.

In lieu of the 0.5 research assistant position in the original proposal, further temporary appointments have given the project flexibility to tackle specific tasks. In addition, as already noted, some costs of the EPrints.org developer will continue to be met by the project until Opsis funding begins.

3.3 People

All those involved in the people are listed at http://opcit.eprints.org/opcitpeople.shtml

The principal changes during Y2 were:

- At Southampton Chris Gutteridge replaced Rob Tansley as EPrints.org developer.
- The ArXiv team, including Paul Ginsparg and Simeon Warner, are moving from Los Alamos to Cornell University.
- Herbert Van de Sompel, who joined the Cornell team and contributed to project development, now becomes the first e-director of the British Library, where we hope to continue some joint developments.
- Students: two final year undergraduate students built projects in conjunction with OpCit:
    - Tim Brody (cite-baseSearch)
    - Catherine Hunt (survey of users of eprint archives)

3.4 Southampton-Cornell-arXiv Partnership

The partnership has continued to be productive. Principals and other senior members of the teams met at Open Archives meetings in Berlin and Geneva. Donna Bergmark from Cornell visited Southampton in July, giving talks on the API described above and on OpenURL, a method of linking currently being standardised by NISO, and attending the JISC-DNER Open Archives meeting in London. Collaboration on adopting and promoting the Academic Metadata Format (AMF), initially led by Simeon Warner at arXiv and Thomas Krichel, will we hope lead to greater visibility for this approach as well as supporting the project's implementation of a linked arXiv.

3.5 Technical seminar for OpCit researchers and partners

The project has made considerable progress towards implementing the Open Archives principle of data interoperability. An invitation-only technical seminar was hosted in Southampton in July 2001 to discuss these issues with technical developers from five partner projects. Most of the presentations from that meeting can be found at http://opcit.eprints.org/tech-seminar/

3.6 Steering group

Members of the project steering group are listed at
http://opcit.eprints.org/opcitpeople.shtml

The steering group has expanded to include new members with expertise in the areas covered by the project, all with a strong interest in collaborating with the project.

The project has remained in contact with members of the steering group on an individual basis. Attempts have been made to arrange formal meetings, but none would have attracted a quorum of members. One approach was to identify focal points, meetings at which members might be present. As before, the geographical spread of the group and diversity of interests made this difficult. Partly to counter this we have established a stronger UK component within the group. The plan is to hold a meeting in the UK at the beginning of Y3 to which all members will be invited.

New members of the steering group

- Ann Apps, MIMAS and DC Working Group on Bibliographic Citations
- Simon Buckingham Shum, The Open University (Knowledge Media Institute)
- Thomas Krichel, Long Island University (Palmer School of Library and Information Science), USA
- Charles Oppenheim, Loughborough University (Department of Information Science)

Leavers

- Ian Jones, left British Computer Society
- Cliff Morgan, stood down as Chair of DC Working Group on Bibliographic Citations

3.6 Work plans

Performance is assessed against the Y2 work plan, and against the original project proposal, in the next section. An amended work plan for Y3 is given in section 6.

3.7 Performance

3.7.1 Progress against Y2 plan

Objectives

- Extend reference linking and interoperability models to other Open Archives. *Continuing work with AMF and arXiv.*
- Model linkable objects in a repository architecture. *Implementation of the Cornell API during Y2.*
- Integrate other reference linking services in model applications. *An emerging infrastructure - based on progress with AMF, this year's revisions to the OAi protocol and the continuing standardisation of OpenURL - will boost ongoing OpCit efforts in this area in Y3.*
- Collaborate on the development of user interfaces for archives, with special emphasis on support for author submission and reference checking. *Modular components that can offer this functionality as part of EPrints have been written and await release of the fully tested EPrints v. 2.0.*
- Investigate the impact of linking on usage of the archives and on citation patterns. *Requires that linking and citation services become a regular part of archive users' experience. Will become possible if OpCit data can be integrated with arXiv.*
- Support maintainers of Open Archives with information, tools and services for reference linking. *Open source releases of software for reference processing and reference linking API.*
- Build a gateway to Open Archives and eprint archives. *Registered Open Archives are listed on the OAi Web site. Need to reassess how useful such a gateway would be in terms of the growth of OAi and the number of archives that remain outside OAi. This will be part of the dissemination work in Y3.*
- Knowledge-based citation services using ontologies, information contexts; also visualisation and clustering methodologies. *This work will focus on new measures of citation impact to be developed in Y3.*

JISC/NSF milestones and deliverables

- Release of v. 2.0 linking implementation across arXiv physics archives. *Done.*
- Report on preliminary evaluation. *To follow. V.2.0 OpCit demo released on 20 August. Some limited evaluation possible; more extensive evaluation when arXiv integration is complete.*
- Extension of linking to distributed NCSTRL archives. *NCSTRL to close in Sept. 2001. This will be dependent on migration of NCSTRL archives to OAi?*

3.7.2 Progress against original proposal

1. **Redesigning and universalizing the author deposit procedure, interface and infrastructure**. Continued collaboration with EPrints at Southampton and Cornell. Cornell will replace Dienst with EPrints and plans to add a reference linking service to EPrints, built on the API to simplify the task of interlinking various archives.
2. **Redesigning the user interface, its capabilities and its infrastructure**. Implemented in Southampton v. 2.0 demonstrator, although caution that this is likely to be amended in a full arXiv version.
3. **Extracting citation data from all papers in the archive**. arXiv papers downloaded to Soton in TeX formats (for reference extraction) and in PDF (for reference link presentation). Database schema designed to support 'forward' reference linking implemented. Software tools developed for reference extraction (some modules specialised for the physics archives) now available for download by other projects.
4. **Generating hypertext links for all citations in the archive.** Implemented in v. 1.0 demonstrator, links to all cited papers held in the arXiv. Citation database upgraded and included in v. 2.0.
5. **Automatic addition of hypertext links for all papers in the archive.** Implemented in v. 1.0 demonstrator and now incorporated in v. 2.0 for the 90+ per cent of papers for which a PDF copy can be generated.
6. **Optimizing the deposit procedures and formats of (1**). Further work to be based on EPrints and the Cornell interface.
7. **Upgrading the citation-navigating capabilities of (2).** *The initial user interface for the OpCit v. 2.0 demonstrator is to be refined and improved. The integration of OpCit data in arXiv will be a test of the ability of the project to influence the design of the interface for citation navigation in other services.*
8. **Bibliometric analysis of citation and usage.** There has been further analysis and reporting during Y2 [3, 6, 3t-8t] of the data collected in Y1. More extensive reporting is planned for Y3. Analysis to be combined with the findings of the survey of authors and users of eprint archives.
9. **Develop a family of generic tools (establishing a set of standards for low-level interoperability, i.e. communicating meta-data and meta-information within the current archive network).** Developed by OAi; generalized to EPrints; now being extended to AMF.
10. **Application and further development of Open Journal software.** Software modules developed in Y2 are now openly available. These are specialised tools designed for reading and linking physics-style references in TeX/LaTeX documents. Also downloadable is the Cornell Reference Linking API in Java, which uses some of the code developed in the earlier Open Journal project. Tools for working with PDF documents, which have potentially wider application, are still not released due to licensing restrictions.

4 Learning from experience

4.1 Undergraduate projects

Other JISC-NSF projects have reported the benefits of designing PhD student projects that would inform the JISC-NSF work. OpCit may be unusual in working with undergraduates for specific project tasks, such as the data mining performed by two summer students and reported in Y1. This year two final year undergraduate projects at Southampton have performed research that has subsequently been integrated into the work of OpCit, notably the citation-ranked search engine, and the survey of users of eprint archives, both described above. Conceived and designed to inform OpCit, in both cases the students were supervised separately by an academic tutor (also the OpCit project director at Southampton, Professor Stevan Harnad). There was close liaison with research staff on implementation, particularly on the search engine, although the project was careful not to interfere with the primary objective for the students, which was to produce works that would count towards their final degrees.

Could this work for other projects? The arrangement has benefitted the project enormously. As conceived neither project was critical to OpCit, but both have exceeded expectations in terms of results and their impact on the project. The students were of high calibre. One was already known to the project through earlier summer work, noted above. That student (Tim Brody) is expected to rejoin the group as a postgraduate researcher in the new academic year, and to continue working with the project in some way. OpCit project management and student supervision were strictly separated. If this combination of factors can be emulated, there is no reason why other projects could not benefit similarly.

4.2 Database not documents: control of the interface

It is a feature of digital libraries that they tend to be document-centric because this is the way users work. Even if the activities supporting services in a digital library involve data processing, the way they are perceived is often in terms of the output of the service, which typically means helping users to find information that is usually contained in documents. For example, the output of a search engine is a list of documents; the output of the first OpCit demonstrator (v. 1.0) was a collection of documents with added reference links.

In addition, documents, and the context in which they are presented, denote the origins and ownership of the work, of author and producer. Control of the interface is key for many services. It’s not surprising therefore that this document-centric thinking infects the work of developers as well as users. In a more open environment – in terms of data interoperability, as well as rights management – it becomes less important for every contributing service (‘service providers’ in OAi terms) to preserve its own document-based output. It’s a lesson that OpCit has learned during Y2.

With its demonstrators OpCit has produced a service aimed at users, and it would like to attract as many users as possible. ArXiv users are the obvious target since the OpCit service is an enhanced version - in the view of OpCit, a view that needs to be examined - of the service they already use.

As indicated above, arXiv was willing to help, but in what form could OpCit data appear in arXiv? Documents seemed the natural method of interchange. Documents are what OpCit harvests from arXiv, due to the limitations of the OAi metadata. There were a number of options for putting OpCit in arXiv:

- ArXiv publicises OpCit
- OpCit becomes a new mirror in the arXiv network

On reflection, one is not integral to arXiv and the other is not practical. Other approaches were:

- ArXiv links to OpCit-linked versions of papers held on OpCit's server.
- OpCit gives arXiv its software tools for processing references so that arXiv can create an OpCit-like service.

In effect the first approach would be an additional option in the user's choice of downloadable arXiv formats. This was not immediately dismissed and various interfaces were created to examine how this might work at a practical level. None were satisfactory, perhaps due to the growing realisation that this approach would not be robust enough to serve in the current arXiv network. In addition, documents require an interface that has to be defined by the document producer, in this case by OpCit, which was unlikely to be acceptable to arXiv.

OpCit has effectively shown the feasibility of a linked archive, and in the second approach arXiv would rebuild it independently. This overcomes the problem of building a robust service, but adds significant overhead to arXiv's activities. Nevertheless, it was arXiv's initial suggestion. Licensing problems with OpCit software that included components from Adobe's PDF library prevented us from pursuing this at the time of the suggestion.

The prospect of data and documents processed by OpCit appearing as an integral part of arXiv, even on a limited basis, was not progressing as fast as was hoped.

The event that changed perceptions of the problem was the need to share OpCit data with the cite-baseSearch project. Cite-base needed to augment the metadata it had imported from arXiv, using the OAi protocol, with link data from OpCit. Although cite-base was a local project at Southampton the augmented metadata could be exported to cite-base using OAi. Since the OAi protocol is not specific to Southampton, it could be seen that here was a method that could be used to make OpCit data available to arXiv. The prevailing wisdom was that, in OAi terms, OpCit acted as a service provider harvesting data from arXiv, the data provider. In the new scheme OpCit becomes the data provider.

OAi metadata elements are not rich enough to describe reference lists so en extended format had to be created. A solution was emerging, and has been consolidated by working with arXiv developers and others to build the Academic Metadata Format (AMF), a Dublin Core based format. It still has to be proven. If we had persevered with the project's first principles for reference linking - data not documents - we may have avoided this diversion. Ultimately we are all driven by the impact we can create with the user, and data driven services are no different.

5 Evaluation

The OpCit arXiv demonstrator v. 2.0 will become the focus of detailed evaluation early in Y3. If the integration of OpCit services in arXiv is successful, evaluation will be able to proceed on a number of fronts. It may be possible to invite all arXiv users to participate in a formal evaluation - what do users *think* of the service. In addition, the project will explore arXiv user data, as before, but seeking to identify any changes in the patterns of usage possibly caused by the new linking services - what do users *do* with the service. We will proceed with a smaller target group of evaluators if arXiv integration is unlikely to be possible before the end of 2001.

Interpretation of the evaluation results will be informed by the ongoing analysis of the eprints user survey. In this way the project will be able to report the first comprehensive picture of the motivations and activities of authors and users of eprint archives, and their reactions to major new services.

6 Future developments and work plan

The OpCit project reached a major milestone with the release of its v. 2.0 linked arXiv demonstrator. Much of Y2 was devoted to achieving this, as this report has revealed. Other parts of the work, the student projects for example, have contributed in excess of plans. Some other Y2 tasks still need to be tackled.

In particular, in Y3 the project will emphasise analysis, evaluation and dissemination. With its linking demonstrators, data mining results and user surveys, the project has amassed a rich resource that demands to be investigated more fully.

In addition the project will undertake to widen the scope of its interoperability activities, promoting OAi, EPrints and OpenURL

These activities are outlined below with a plan of work for Y3 shown in Table 1.

- Evaluation, analysis, dissemination of data mining, user survey, OpCit demonstrators and other OpCit results
- Continuing efforts to integrate OpCit with arXiv: develop and promote AMF
- Add OpenURL services
    - links to OpCit linked demonstrator; work with OpenURL resolver services; build an advanced OpenURL generator to turn references in PDF/TeX/LaTeX/HTML papers to OpenURL requests when viewed
- Advanced citation analysis – new measures of impact
- Implement and test EPrints components for reference checking
- Evaluate OpCit and Opsis project software
- Migrate non-OAi institutional archives in computer science

Table 1. Plan of work for OpCit Y3

	Oct - Dec'01	Jan-Mar'02	Apr-Jun'02	Jul-Sep'02
Evaluation, analysis, dissemination of OpCit results	■	■	■	■
Integrate OpCit with arXiv	■			
Add OpenURL services	■	■		
Advanced citation analysis	■	■	■	
EPrints components for reference checking	■	■		
Evaluation of OpCit and Opsis project software			■	
Migrate non-OAi archives (e.g. NCSTRL)	■			

7 Contacts with other projects

- From the JISC-NSF programme, useful contacts were established with the Harmony and Cross Domain Resource Discovery projects. All three projects were subsequently identified to join DNER Z projects cluster. OpCit is happy to contribute to this cluster.
- The project continues to promote AMF with a view to other groups joining its development.
- Close collaboration with EPrints.org will be maintained. A bid to extend EPrints from the archiving of pre- and post-refereeing research to the archiving and sharing of research data has been submitted to the EU-DataGrid Project through the UK e-Science programme.

8 Project publications and presentations

Full list of project publications at http://opcit.eprints.org/opcitpapers.shtml

Papers

[1] Tim Brody, Zhuoan Jiao, Steve Hitchcock, Les Carr and Stevan Harnad
Enhancing OAI Metadata for Eprint Services: two proposals
Accepted for the *Experimental OAI-based Digital Library Systems Workshop*, Darmstadt, September 2001, to be held in conjunction with the *5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*
http://opcit.eprints.org/ecdl-oai/oai-ecdl01.html

[2] Donna Bergmark and Carl Lagoze
An Architecture for Automatic Reference Linking
Cornell University Technical Report, TR2001-1842, 2001, to be presented at the *5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Darmstadt, September 2001
http://www.cs.cornell.edu/cdlrg/Reference%20Linking/tr1842.ps

[3] Stevan Harnad, Les Carr and Tim Brody
How and Why To Free All Refereed Research From Access- and Impact-Barriers Online, Now
High Energy Physics Libraries Webzine, Issue 4, June 2001
http://library.cern.ch/HEPLW/4/papers/1/

[4] Donna Bergmark
Automatic Extraction of Reference Linking Information from Online Documents
Cornell University Technical Report, TR 2000-1821, November 2000
http://www.cs.cornell.edu/cdlrg/Reference%20Linking/extraction.pdf

[5] Donna Bergmark, William Arms and Carl Lagoze
An Architecture for Reference Linking
Cornell University Technical Report, TR 2000-1820, October 2000
http://www.cs.cornell.edu/bergmark/ReferenceLinkingArchitecture.ps

[6] Stevan Harnad and Leslie Carr
Integrating, navigating and analyzing eprint archives through Open Citation Linking (the OpCit Project)
Current Science Online, Vol. 79 No. 5, 10th September, 2000 (special issue in honour of Eugene Garfield)
http://tejas.serc.iisc.ernet.in/~currsci/sep102000/629.pdf

Viewpoints

[1v] Stevan Harnad
Why I think research access, impact and assessment are linked
Times Higher Education Supplement, Vol. 1487, 18 May 2001, p. 16
Extended version at http://www.cogsci.soton.ac.uk/~harnad/Tp/thes1.html

[2v] Stevan Harnad
The Self-Archiving Initiative
Nature, Vol. 410, 1024-1025, 2001
Nature Web Debates version, 26 April 2001
http://www.nature.com/nature/debates/e-access/index.html

Presentations at conferences, meetings and workshops

[1t] Various presentations from the project, and from invited partners and colleagues
OpCit tech seminar: Interoperable data for scholarly communication, Southampton, July 2001. Hosted by the OpCit project
http://opcit.eprints.org/tech-seminar/

[2t] Stevan Harnad
For Whom the Gate Tolls?
http://www.cogsci.soton.ac.uk/~harnad/Tp/resolution.htm
Invited presentations based on this paper have been given at:

- *UKOLN/DNER Open Archives Meeting: Developing an agenda for institutional e-print archives*, London, July 2001
- *NSF Digital Libraries Initiative 2/IMLS Principal Investigators Meeting*, Roanoke, VA, June 2001
- *Change and Continuity in Scholarly Communications*, Royal Netherlands Academy of Arts and Sciences, Amsterdam, June 2001
- and elsewhere http://cogsci.soton.ac.uk/~harnad/talks.htm

[3t] Steve Hitchcock and Zhuoan Jiao
Enhancing the Dynamics of Large-Scale E-print Archives
JISC/DNER Synthesis Meeting for International Digital Libraries Research Projects, Bath, UK, May 2001
http://opcit.eprints.org/talks/dner-synthesis/title.html

[4t] Steve Hitchcock
Citations and Linking in Large-Scale E-print Archives
"What's E-Publishing ever going to do for us?" *Animal Health Information Specialists Annual Conference*, London, April 2001
http://opcit.eprints.org/talks/ahis/title.html

[5t] Les Carr
Distributed Eprints Archives and Scientometrics
http://documents.cern.ch/archive/electronic/other/agenda/a01193/a01193s3t5/transparencies/carr.pdf
AND
[6t] Stevan Harnad
Peer Review in the On-line Era
http://www.cogsci.soton.ac.uk/~harnad/Tp/Peer-Review/index.htm
Both at *Workshop on The Open Archives Initiative (OAI) and Peer Review Journals in Europe*, Geneva, March 2001

[7t] Les Carr
You Can Get There From Here!
OAi Open Meeting held to mark the European public release of the specifications of the OAi interoperability architecture, Berlin, February 2001
http://www.ecs.soton.ac.uk/~lac/GettingThere.html

[8t] Tim Brody and Stevan Harnad
Generalizing the Self-Archiving Principle Across Disciplines Via Interoperable Distributed Eprint Archiving: What France Can Do Now?
Coordination meeting for arXiv mirrors, Lyon, France, October 2000
http://www.cogsci.soton.ac.uk/~harnad/Tp/Tim/sld001.htm