The Open Citation Project
First Year Report to JISC

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

Version history
Version 1.0: submitted 31st August 2000
This version 1.1: 19th September 2000 (final)
Period covered in report (October 1999 - August 2000)
Contact: Steve Hitchcock, sh94r@ecs.soton.ac.uk

Summary

The Open Citation project, originally called 'Integrating And Navigating Eprint Archives Through Citation-linking', is well on target to realise its objectives during its first year. The main outputs during this period have been:

- A pilot demonstrator of reference linking covering the whole of the Los Alamos physics archives (arXiv)
- A programming interface for inter-archive reference linking
- Original data on usage of the Los Alamos physics eprint archives by authors and readers

In addition, two major developments since the proposal have strengthened the role of the project, and are already benefitting the longer-term prospects:

- The formation of the Open Archives initiative (OAi) to promote author self-archiving solutions
- The release of EPrints software to manage institutional and discipline-specific eprint archives, and which is compliant with Open Archives.

Members of the project teams at Southampton University and Cornell University are leading players in both developments, ensuring effective joint development will continue.

Glossary of distributed information and reference linking services

ArXiv. Collection of eprint archives based at Los Alamos covering physics, mathematics and computer science.
CogPrints. An eprint archive for cognitive sciences, modelled on the Los Alamos physics eprint archives and hosted at Southampton University.
Computing Research Repository (CoRR): an eprint archive of papers in all areas of computer science.
CookiePusher. A user interface that informs an SFX server of user context and preferences.
CrossRef. A commercial reference linking service for journal publishers.
Dienst. A protocol and server for distributed document libraries.
Digital Object Identifier (DOI). An identification system for intellectual property in the digital environment.
Distributed National Electronic Resource (DNER). A managed environment for those in the UK higher and further education community to access quality-assured information resources on the Internet.
Distributed Link Service (DLS). Software supporting link placement in third-party documents that are accessible anywhere on the Web.
Dublin Core. A metadata element set for labelling electronic resources.
Eprint archives. Classified and indexed storage and retrieval services for formal scholarly papers deposited by authors.
EPrints. Generalised software developed at Southampton University for managing eprint archives.
LinkBaton. A user interface to a link resolver that directs particular link types at user-specified resources.
Link resolvers. Two issues in resolving links to destination documents on the Web are: stability of the documents' location, and multiple versions. Link resolvers can help with one or both of these problems.
Networked Computer Science Technical Report Library (NCSTRL). A federated collection of TR report libraries maintained by different university computer science departments.
Open Archives initiative (OAi). Initially a forum to solve interoperability between author self-archiving solutions (eprint archives), now extended to support a wider range of digital resources of academic and scholarly interest.
OpenURL. A URL that transports metadata, or keys to access metadata, for a digital document or object for which the OpenURL is provided.
ResearchIndex. Software designed for 'autonomous citation indexing', in effect builds an ISI-like index of the online full-text scientific literature.
Santa Fe convention. Document specifying the technical requirements for implementing Open Archives.
SFX. A link resolver and server designed to give users of a given university or institutional library access to local resources and to networked subscription and non-subscription-based services.
SlinkS. The Scholarly Link Specification framework facilitates inter-publisher reference linking by providing a syntax and vocabulary for exchanging information.
SPIRES. Stanford Public Information REtrieval System. A collection of library databases covering high-energy physics, including journals and eprint archives.

For more information on these terms see the extended glossary in Appendix 3.

Contents of the report

1 Introduction

1.1 What's in a name? The project and the Open Archives initiative

Originally proposed as 'Integrating And Navigating Eprint Archives Through Citation-linking', the project's new title of the Open Citation Project more concisely reflects the activities of the project, placing it at the centre of important international developments yet remaining consistent with the proposal.

By adding links to freely accessible documents, references become 'open' (i.e. the referenced documents become accessible) to the user for further investigation. Access to such documents is provided by managed eprint archives, and with the advent of the Open Archives initiative (OAi) efforts are being made to make these archives 'open' to automated data services.

Characteristically, eprint archives contain papers archived by the authors themselves. As a result these low-cost automated archives are free to both authors and readers and are an excellent environment for linking citations. Users can be provided with a simple, one-click means of navigating from one document to another that it has cited, with no financial barriers to erode the effectiveness of the link.

Although users may be able to access eprints freely in these archives, automated data services could not, either because they were not permitted by the archive administrator, or because there was no interface or means to collect the desired data. Citation linking is one such automated data service. Large amounts of data have to be collected from the data sources, in this case the archives, to build the links. To tackle this problem the project proposal described the need to consider the requirements for interoperability between different services and archives distributed on the network.

Interoperability on this scale looks likely to be achieved through the OAi. The significance of this for citation linking is that such services will in principle be able to act independently on multiple archives wherever they are on the network. This is the scenario identified in the project proposal, so the project was ideally placed to both inform, and benefit from, the OAi.

Where the project made the connection between citations and links, the OAi promises to provide the environment in which the service can operate most effectively, and hence the project was appropriately re-titled the Open Citation (OpCit) project.

1.2 More archives, more users? The EPrints factor

At a recent OAi meeting delegates noted that two principal drivers for establishing Open Archives will be the availability of more archives and the launch of the first OAi services. Together the OpCit and EPrints projects are on the verge of delivering both.

The EPrints software has emerged from the CogPrints project (also funded by JISC and also led by Stevan Harnad). CogPrints, an eprint archive for cognitive sciences, was inspired by and modelled on Paul Ginsparg's physics eprint archives at Los

Alamos. Based at Southampton University, a mirror site for the physics archives, CogPrints was able to adopt much of the software that was used to run the archives but discovered that much of it had to be adapted to suit its users. Notably, the uncompromising user interface was difficult to use for non-physicists.

Recognising this problem was not specific to cognitive science, EPrints was set up to generalise the user interface for disciplines other than physics and to provide a simpler way of managing archives than the Los Alamos software offered.

As a result EPrints is generalised software for managing eprint archives. It addresses the need for simple set-up of archives, within institutions and other organisations, by departmental managers or librarians, information providers or publishers, i.e. by non-specialists in computing. It provides an interface for administrators, for authors to deposit papers, and for users to access papers. In short, EPrints supports properly maintained and reliable archives at a local level by groups that can assure the long-term integrity of an archive.

The first beta version of the EPrints software was made available in July and has already received outstanding commendation from beta test users including the California Digital Library.

EPrints is based on experience of running a real archive, and by fortune of timing is fully compatible with the requirements of the Open Archives initiative, so it is the most complete and usable service of its type. Being OAi-compliant means that any archive built using EPrints, wherever it is based, whatever size and whatever the subject of its content, can be viewed as an integral part of a larger community of archives.

EPrints is available from Southampton for free. The next step is to make it available as open source software both to enhance distribution and to spread the costs of development and maintenance.

The significance of co-developing software both to build Open Archives and support Open Archive services, in this case reference linking, has long been recognised within the OpCit project and will be fully exploited in OpCit's second year.

To see how EPrints is being used in the OpCit project, see section 2.4, Performance against proposal.

1.3 Overall objectives

Positioning itself within the framework being established by the OAi, the main research partners (2.3.3) identified the following key objectives for the OpCit project:

- Integrating and developing software solutions for reference linking in large-scale Open Archives
- Improving author and user interfaces for the archives
- Defining the operational semantics of documents (digital objects) to allow perfect linking

- Modelling interoperability between linking services and other digital library services
- Building interfaces to other linked information environments
- Promoting author self-archiving

While this develops and focuses the many objectives cited in the original proposal, it still contains most of those objectives, as shown in the performance review in section 2.4, and in the work plans commented on in section 2.3.5.

2 Activities and progress

2.1 Outputs

The main outputs during this period have been:

- A pilot demonstrator of reference linking covering the whole of the Los Alamos physics archives (arXiv)
- A programming interface for inter-archive reference linking
- Original data on usage of the Los Alamos physics eprint archives by authors and readers

2.2 Main activities

2.2.1 OpCit in context

Information environments organised by digital libraries are delivered via the Web but are distinguished by services that apply to contents deemed to be within them, determined not by physical location but by the nature and selection of those contents and the services that act on them. Examples of these environments might include the contents of libraries as in the Distributed National Electronic Resource (DNER), single-publisher collections such as the ACM Digital Library, larger collections of published journal papers accessed via DOI-based services, or distributed archives such as the Networked Computer Science Technical Report Library (NCSTRL). In essence, in these environments the Web is transformed from a document delivery service into a dynamic, computational framework.

The information environments being explored for reference linking by the OpCit project are outlined in Figure 1. Of immediate concern is the area bounded by the left-hand vertical arrows and the information environments denoted by 'Southampton' and 'Cornell'. The various tools and example information sources, which were beyond the scope of the project in year 1, are explained in Hitchcock *et al.* (2000). The next two sections describe how both partners are developing the computational framework for reference linking.

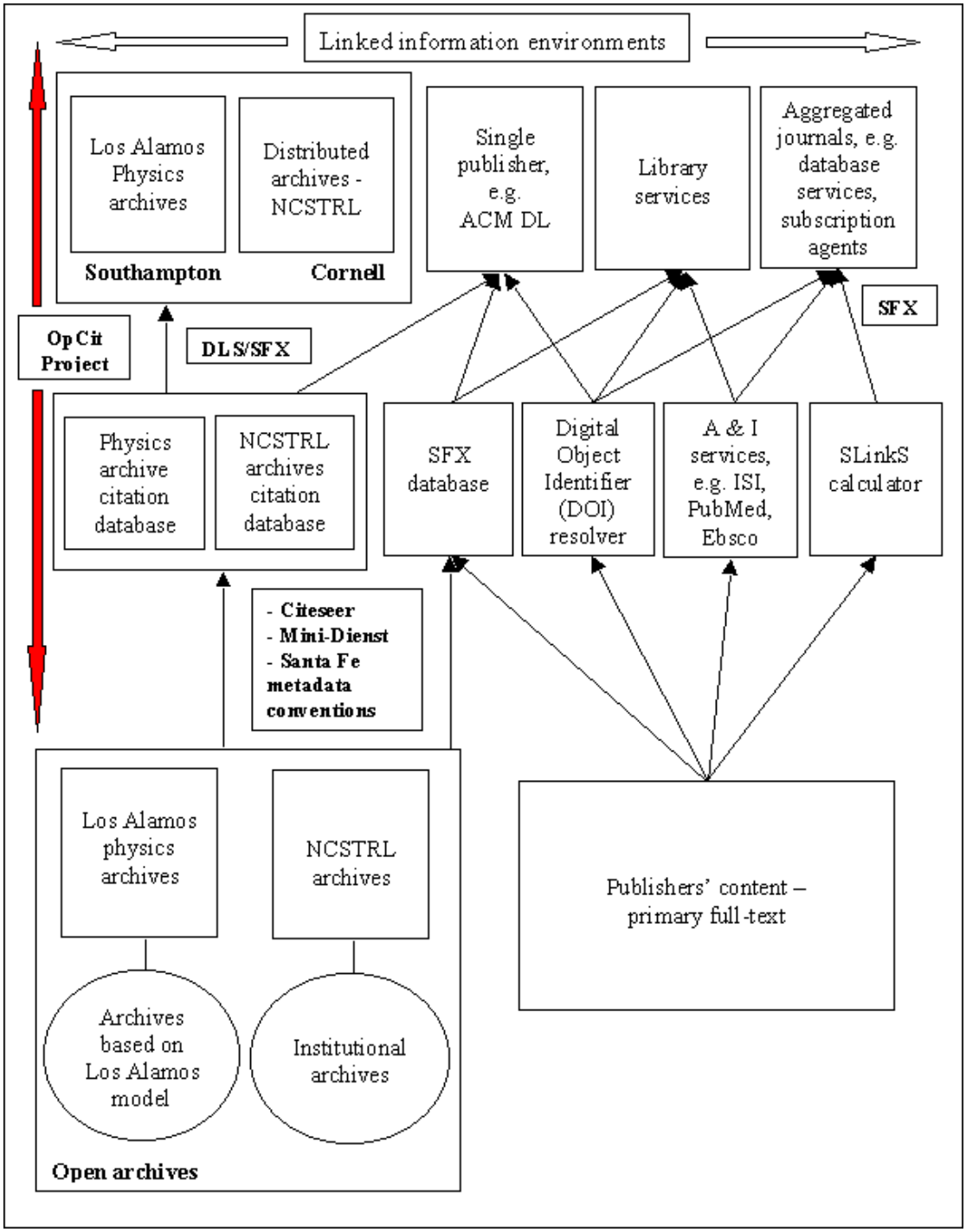

Figure 1. Information environments and link services: a range of possible scenarios and tools for OpCit and beyond

2.2.2 Reference-linked arXiv

The process of adding reference links dynamically to documents retrieved from an archive involves parsing the document during download to identify and read references automatically. These data are compared with a precompiled citation database, and links to the referenced works added where exact matches are found.

A similar method is used for OpCit to link papers in arXiv:

1. convert original documents to a format (e.g. plain text) for extracting references;
2. parse documents to identify and read references;
3. match reference metadata (e.g. volume, issue, year, page numbers) with metadatabase
4. present document to user with active links on matched references

A more sophisticated approach would build a richer citation database that could also be informed by stages 1 and 2, and this is one of the objectives of the work at Cornell described in the next section, and of continuing work at Southampton. A richer citation database will support further analysis of document usage, building on the work begun in section 2.2.4.

However the citation database works, there remain two key stages, of text recognition and link insertion. In a system where documents are archived by their authors this introduces certain constraints. Unedited texts will have more variability in presentation styles, especially in reference lists, and are more likely to contain errors. Physics papers in arXiv, ironically, have quite a consistent form for references, but by virtue of adopting a terse style. Presentation of linked texts must be based on a format acceptable to users of the archive rather than a format that suits linking technology.

The project's first demonstrator of linked physics (http://arabica.ecs.soton.ac.uk) returns the requested document, with links, in PDF (Figure 2), a format that can be derived from TeX, the most common submission format for the archives. This has proved to be an effective basis, with some qualifications, for further development (see the results of evaluations in section 4).

Future versions of the demonstrator will give users the option to link to third-party services. A feature of the interface in the current version directs the user to an intermediate page, the 'SFX' page, offering a choice: download the text from the archive, or look up some contextual information on the citation (see Figure 3). In future other options on the page might offer documents supported by an independent (non-Southampton) link service - a number are developing (see Figure 1) - which maintains some knowledge of user privileges and can offer alternative versions of the cited paper from other content providers, for example, library, journal aggregator or publisher services. Currently only the archive versions - abstract, original and link-enhanced (e.g. Figure 2) full texts are available.

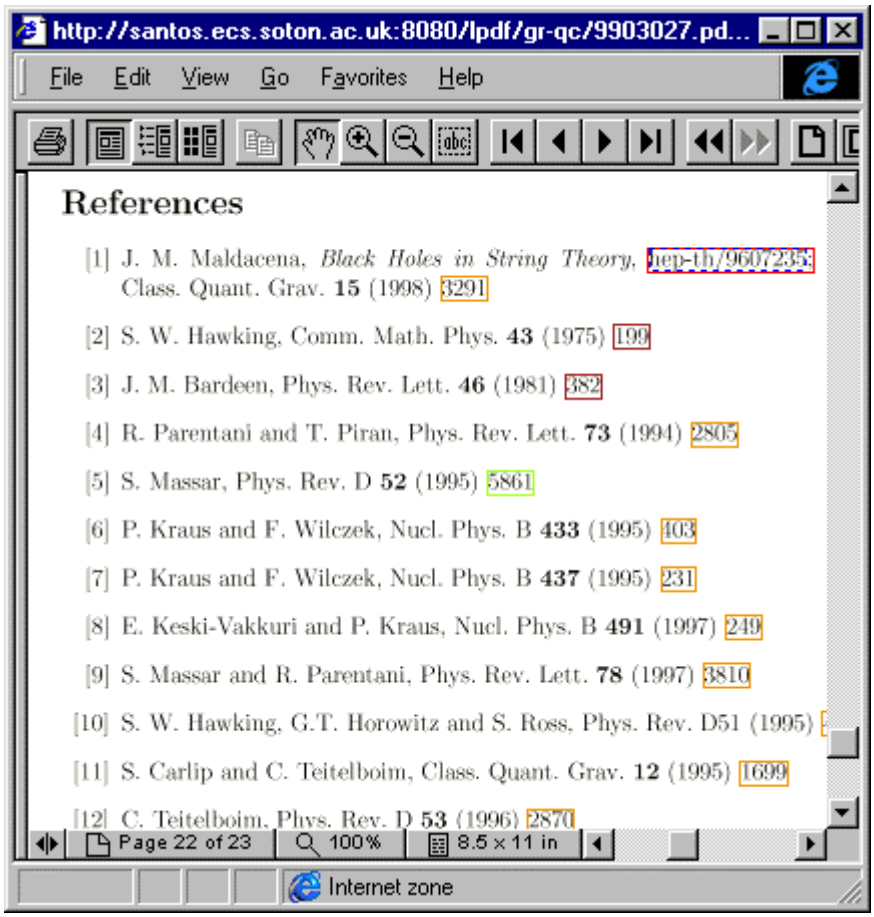

Figure 2. Original document from the physics archives presented in PDF enhanced with reference links (the active links are indicated by red and orange boxes; the different colours are a feature of the development version only)

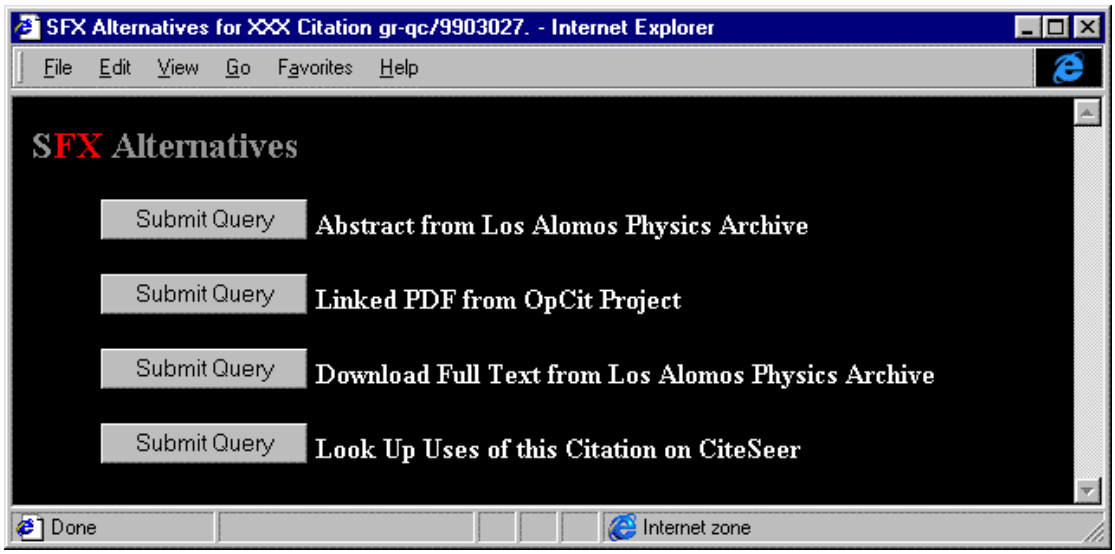

Figure 3. Activating the link for reference [5] in Figure 2 optionally returns an 'SFX' router page offering the user a choice of sources

2.2.3 An API for reference linking

While researchers at Southampton undertook the linking of the Los Alamos physics archives using tools first developed for the Open Journal project, Cornell took a more general view of the reference linking problem as it applied to online scholarly material, to determine a framework for linking the online literature and to construct tools that would encourage the linking of electronic publications.

An application programming interface (API) for reference linking was developed. The motivation for this work was to answer 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 principal methods:

- getMyData() - the digital object should emit standard metadata describing that object, i.e. title, authors, year of publication, etc.
- getReferenceList() - the digital object should say what its list of references is (this is the fixed number of references contained in the online document).
- getCitationList() - the object can say what other works the object knows have cited it. (This list grows as more items are analysed.)
- 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.

The API is realized by a Surrogate class, with one Surrogate object per archive item (an item is a document held in an online archive). The methods of the Surrogate object reflect the reference linking API just described. A Java interface was written that specified the method signatures for the class; finally a "null implementation" was developed which added fields to the class and null implementations for the interface methods. The null implementation was working by the end of 1999.

A complete implementation is underway. The output of each of the four methods given above is an XML file that can be used for further analysis, or for rendering a linked document, or for many other purposes. Currently the first two methods are almost fully coded, using HTML input from *D-Lib* magazine as a test bed. Tools from other XML projects, and from the Distributed Link Service at Southampton, are being used. Cornell’s software can directly call Southampton citation analysis code, to analyse the reference section of an online document. Some tools from the ResearchIndex project will be used to handle Postscript documents, and other Southampton tools to handle PDF.

We need to finish implementing the reference linking API in Java, and move the Surrogate objects into a persistent database (such as MySQL or Cornell’s FEDORA repository architecture for digital objects). Another goal is to fold the reference linking API into the widely-used Dienst architecture and protocol. Not only would this provide persistent storage for the Surrogate objects, it will simplify the task of interlinking various archives, e.g. within the OAi framework where Dienst is the chosen protocol for communication with archives.

2.2.4 Citation analysis: mining the social life of an eprint archive

Citation-linking offers further benefits over and above natural navigability for users; it also offers new ways of analysing and understanding how the literature is used. Author-end citation analysis, based on explicit citations in published papers, has for many years been used to reveal the lineages and dynamics in the growth of the offline research knowledge. Online usage patterns, based on Web logs and download frequency, show what readers are actually doing. Citation analysis and usage patterns, or reader-end citation analysis, can be correlated in a way that raw "hit" rate (download frequency) alone cannot do, and will also enable the impact of linked citations to be tracked.

Reader-end citation analysis is entirely new informetric territory, because citation searching could only be done offline until now, so there was no automatic way to analyse how readers actually go about it. The OpCit project is able to investigate, for the first time, the wealth of data generated by users of the physics eprint archives at the Los Alamos site and the Southampton mirror site. A large number of patterns and graphs have been produced, and can be found linked from the project's Research Web page (http://opcit.eprints.org/opcitresearch.shtml).

For example, it has been discovered that more highly cited papers show higher and more sustained download frequency, i.e. that posting a paper in the archives can increase its impact factor, as shown in Figure 4.

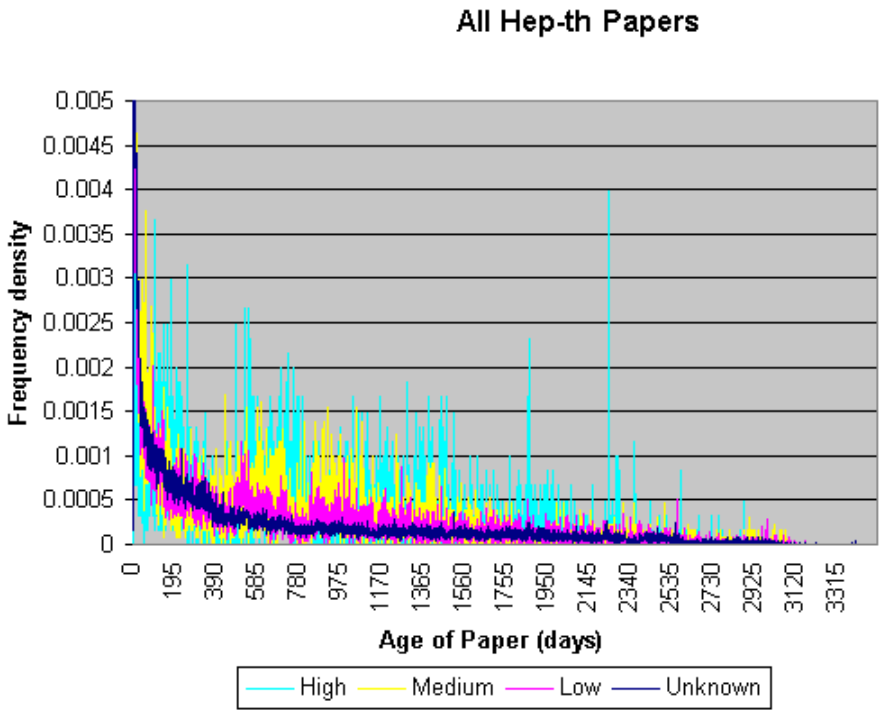

Figure 4. Download frequency for high-, medium- and low-impact papers in the high-energy physics – theoretical (hep-th) subarchive

Analysis to identify further significant features is ongoing and will be reported soon.

Such data will also be used to provide feedback for optimizing the features of the EPrints interface, to monitor and document open archiving and citation navigating practises, and to chart the course of both knowledge creation and use as this new medium evolves its own niche in scholarly and scientific research practise – the "scholarly skywriting" continuum – to reveal previously hidden embryological stages of learned inquiry and interaction.

A related question, which may ultimately only be answered through services such as developed by the OpCit project, is: do links have an impact on citation? This requires *usage on a large scale and over a reasonable period* of the fully-linked archives, and is only really feasible if the OpCit links become an integral part of the Los Alamos archives and a natural feature for users. Evaluating the demonstrator, a response from Los Alamos indicates this may be in prospect (see section 4.2).

A citation-linked online literature makes new forms of usage and impact analysis possible that will not only enable us to better understand, predict and direct developments in this new medium, but it will permit much finer-grained monitoring and analysis of the online evolution of our digitized knowledge.

2.3 Project management

2.3.1 Costs

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

2.3.2 Staffing

Southampton staffing for the project has been as per plan with one significant exception. We were not able to recruit a suitable candidate for the 0.5 RA position. Instead we have funded two second year undergraduate students to cover some of the project work – an investigation of author and reader usage of the Los Alamos physics eprint archives – for an extended period during the main summer vacation. This has been especially productive, providing the project with a large volume of original data to be analysed for patterns of usage and a more detailed understanding of the social life of a large eprint archive (section 2.2.4).

In addition, some of the work planned for this RA – notably an author deposit interface for eprint archives – has emerged from the EPrints project.

Overall, this staffing issue has not held back the work of the project. Indeed, it forced a re-appraisal of work plans, enabling recruitment to be targetted for particular objectives.

For year 2 we will reconsider the role of the 0.5 RA should a suitable candidate emerge. Other possibilities are that we might more explicitly cross-fund the continuing development of EPrints as there is a natural convergence with Opcit, and we may consider further short-term appointments for specific tasks.

2.3.3 People

Research is coordinated at two main sites, at Southampton University and at Cornell University, with support for content from the principals of the Los Alamos eprint archives. The following people work for the project at these sites:

IAM (Intelligence, Agents, Multimedia) Research Group, Southampton University

- Principal Investigator, Professor Stevan Harnad
- Chair, Professor Wendy Hall
- Technical Director, Les Carr
- Project Manager, Steve Hitchcock

Researchers

- Zhuoan Jiao
- Rob Tansley (EPrints)
- Tim Brody (June - Sept. 2000)
- Ian Hickman (June - Sept. 2000)

Department of Computer Science, Cornell University

- PI, Carl Lagoze
- Researcher, Donna Bergmark

arXiv.org e-Print archives, Los Alamos National Laboratory

- Principal, Paul Ginsparg
- Simeon Warner

2.3.4 Partnership with Cornell

Prior to the project launch Carl Lagoze visited Southampton where he gave an open seminar on the work of the Cornell Digital Library Research Group.

To launch the collaboration Cornell hosted an informal visit in Ithaca for the Southampton researchers, in early November 1999. During this meeting we formulated our first year group work plan and also exchanged experiences with various tools and interfaces. This meeting involved Carl Lagoze and Donna Bergmark and others (from Cornell), Steve Hitchcock and Zhuoan Jiao (of Southampton). Steve Lawrence of NEC and Eric Hellman of Openly Informatics, both on the project steering committee, visited Ithaca at the same time to talk about their reference linking work (Eric also visited Southampton soon after to discuss software partnerships). A return visit was made by Bergmark to Southampton during May 2000 to exchange software and ideas.

2.3.5 Steering group.

The steering group provides comment and feedback on the technical design and implementation within the framework not just of this project but also of the wider interests of those seeking to enhance the process of scholarly communication and

reference linking. In this sense it is not a management group, nor does it direct the project.

The steering group is also a very strong driver for working with other projects (see section 6). This is illustrated in the project's initial evaluation process (section 4.2).

The project organised an informal dinner for members of the steering group attending the second meeting of the OAi and the ACM Digital Libraries conference, both held in San Antonio, TX, in June.

The San Antonio meeting highlighted the difficulty of gathering the full steering group. To fulfil the JISC requirement to hold formal meetings of the steering group, the project has considered convening a UK (or European) subgroup, but the composition of the group and the international context of the project suggest this may be of limited value

The project continues to monitor opportunities to hold a full steering group meeting where possible.

Members of the current steering group are listed below:

- Bill Arms and Joe Halpern, Association for Computing Machinery (ACM)
- Mark Doyle, American Physical Society
- Michael Friedman, Highwire Press, USA
- Eric Hellman, Openly Informatics, Inc., USA
- Ian Jones, British Computer Society (BCS)
- Steve Lawrence and Lee Giles, NEC Research Institute, Princeton, USA
- Cliff Morgan, Dublin Core Working Group on Bibliographic Citations
- John Ober, California Digital Library, University ePub Initiative
- Heath O'Connell, Stanford Linear Accelerator Center (SLAC) Library
- Norman Paskin, International DOI Foundation
- Ed Pentz, CrossRef, Publishers International Linking Association, Inc. (PILA)
- Paul Taylor, Queen Mary and Westfield College, London (Department of Computer Science)
- Herbert Van de Sompel, University of Ghent, Belgium (Automation Department of the Central Library)

2.3.6 Work plans

Where the early project work plans set out the tasks identified from the original proposal, the success of the project in Year 1 has allowed us to extend the scope of the work in Year 3 to include new research. This has yet to be fully elaborated as it will depend on the results of other work now in progress at Southampton and Cornell Universities, but it presents an opportunity to bring forward new ideas in a real application environment.

Otherwise the three-year work plans give full details of the project's activities and deliverables. Summary versions for each year were submitted previously. A unified work plan for all three years, with links to all work produced by the project so far, can be found at http://opcit.eprints.org/plans/workplans.html.

2.4 Performance against proposal

2.4.1 Components of citation linking project

What follows are the stated objectives from the original proposal with quoted extracts for explanation and notes on progress. The target of the proposal is the Los Alamos physics eprint archives (arXiv).

1. Redesigning and universalizing the author deposit procedure, interface and infrastructure.

"whatever it takes to make all deposits interoperable specifically for citation extraction and linking will also help to make them interoperable in other respects"

- Deposit procedure: EPrints has developed an interface for author deposit. The software is being beta tested.
- Interoperability and metadata: This work is being led for the OAi by our partners at Cornell and is specified in the OAi's Santa Fe agreement. There is already significant integration with EPrints and with the programming interface for reference linking being developed at Cornell for the project. Continues in Y2.
- Reference checking: a new feature to be tested in Y2 by integrating reference databases with the EPrints interface.

2. Redesigning the user interface, its capabilities and its infrastructure.

"navigation of the entire archive can continue via citation-links, with no need to launch another top-down search. Algorithmic content classification ... will also be incorporated."

- Navigation (linking): implemented in current demonstrator. In Y2 this will be set up as an OAi service that can be invoked either on any Open Archive database, or on any EPrints-based archive.
- Enhanced interface - integrated services: to be developed in Y2. We plan to work on this with the California Digital Library, among others.
- Classified content: Y3

3. Extracting citation data from all papers in the archive.

"convert (papers) into a citation-linkable form ... designing the automatic tools that will actually be linking them"

- Linkable formats: arXiv papers converted to PDF.
- Linking tools: tools developed to read and link PDF documents, specialised for the physics archives.
- Data gathering tools: built as part of Cornell's API.
- Reference databases: simple metadatabase built for 'backward' (in time) reference linking. More sophisticated database schemas to be designed to support 'forward' reference linking and to collect data from Southampton text recognition tools and Cornell API tools.

4. Generating hypertext links for all citations in the archive.

Implemented in current demonstrator.

5. Automatic addition of hypertext links for all papers in the archive.

Implemented in current demonstrator.

6. Optimizing the deposit procedures and formats of (1).

To be continued in Y2, based on EPrints and the Cornell interface.

7. Upgrading the citation-navigating capabilities of (2).

The reference linking capability applied to arXiv in Y1 will be generalised to other Open Archives in Y2.

8. Bibliometric analysis of citation and usage.

> "will help us to better understand, predict and direct developments in this new medium - Author-end citation patterns; Reader-end citation-based navigation patterns"

Original data generated on the patterns of use of arXiv by authors and readers. Analysis and reporting to continue in Y2.

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 ... and (with) other resources"

Developed for the OAi; generalized to EPrints.

10. Application and further development of Open Journal software.

This toolset has been specialised for the requirements of reading and linking physics-style references. An XML ‘deciter’ enables references extracted from documents to be formatted for other applications and other document formats. Continues in Y2, when it is anticipated that some components will be co-developed for applications by research and commercial partners.

2.4.2 Further enhancements

a. Other kinds of links.

> "keywords, author-names, glossaries/indices"

Aim to adapt citation navigation approach to add author/keyword navigation to arXiv. EPrints is more transparently set up to collect relevant data from author input. As a test it is planned to load a subset of documents from arXiv into EPrints.

b. Revision/update linking.

> "automatic forward and backward linking between versions"

To be investigated.

c. Commentary links.

Depends on practices by authors of archived papers. Can be implemented by EPrints, which supports this explicitly.

d. Journal links.

> "links to the version of a paper in the journal's own official online archive; links to cited papers in the journal's online archive that do not appear in (arXiv)"

The project is working with partners to apply interfaces to online journals services, such as Openly Inc.'s LinkBaton and ExLibris' SFX. Other partners include content providers such as the American Physical Society, Highwire Press and the publishers' collaborative association for journal linking, CrossRef. It should be noted that others are developing these interfaces, and the project does not intend managing journal content directly, but linking to third-party content. Examples in Y2.

e. Peer review.

> "Referees can be directed to (the) citation-enhanced draft; allow referee reports to be linked just as commentaries are"

Adapting EPrints to do this. Continues in Y2.

f. Links to proprietary databases.

> "(to) journal archives, archives of scanned contents of journal back issues, electronic books, and secondary publisher databases but strategic questions about charges"

As for **d**, content will be managed elsewhere, but linking through OpenURL (currently being considered for standardisation by NISO) to library-based services may be possible in Y2.

g. Links to other public archives.

Superseded by OAi and EPrints.

h. Links to authors' home server archives.

c.f. **c**. Depends on practices by authors of archived papers. Implemented in EPrints.

3 Learning from experience

Perhaps the most prominent lesson from year 1 concerns the different mandates of the respective funding bodies in this international collaboration, and the practical effects on projects. Where JISC invests in development with an emphasis on supporting those projects that might produce and support continuing services within the academic community, NSF has a stronger focus on research.

Broadly this can support a highly effective R&D continuum: it has informed and invigorated the collaboration between Cornell and Southampton in this particular project. The differences emerge in reporting and project management procedures. The most obvious manifestation was JISC's insistence on a consortium agreement between project partners, although there have been other, minor issues.

If NSF requires a less prescriptive approach to reporting, clearly this approach must work across a broad range of its programmes. Equally, JISC recognises from experience that the process of introducing services requires regular communication with a wider number of participants. The prospect of continuing development and introducing new services demands new sources of finance, possibly involving commercialisation, and raises issues of ownership and rights, especially where technology development is concerned. These were among JISC's motivations for requiring a consortium agreement.

It is inevitable that if the documentary requirements of the funding agency of one project partner are not the same as those of another, there is less incentive for the two partners to cooperate on these matters.

Fortunately the partners in the Opcit project have been able to understand and resolve these differences. The project has worked hard to produce clear and common objectives, so there is a shared commitment to the plans, timescales and deliverables.

If what one partner may perceive as a secondary requirement introduces a significant cost, however, the scope for cooperation reduces. Without even exploring the likely contents of a consortium agreement, the project realised that a substantive agreement would incur costs that did not appear to be justifiable unless desired by all project partners and required by both funders. Although stressing the role of the agreement for certain projects, subsequently JISC made the agreement optional.

Although different, NSF and JISC have created a vibrant new R&D environment for the promotion of digital libraries. The willingness of projects, partners and funding bodies to learn and adapt within this framework suggests the programme will be stronger should it be extended in future.

4 Evaluation

Evaluation by the project is ongoing. A three-year plan of evaluation is included in Appendix 1.

During year 1 the project has performed preliminary evaluations of the linked archive demonstrator with two interested groups:

1. arXiv authors
2. the project steering group

Each group offers a different perspective, one an intimate knowledge of the content of the archives, the other expertise in the technology and publishing framework. A summary of each evaluation follows, with more complete results in Appendices 2.1 and 2.2 respectively.

4.1 Evaluation by arXiv authors

Should the project's demonstrator of citation linking for the physics archives be adopted as a real service, then users of that service would be the intended beneficiaries. This questionnaire-based evaluation, managed via email, was intended to get preliminary feedback from those users. It was not meant to be a detailed evaluation, which would not be appropriate for the first (v 1.0) demonstrator. Instead, it was designed to obtain simple responses that would confirm (or not) the validity of the approach and to prompt comments that would guide the implementation of subsequent versions.

Circulation of the questionnaire was restricted to less than 100 users, at the request of our colleagues at Los Alamos. In an attempt to maximise the response rate, authors of the papers in the demonstrator with most links were selected to receive the questionnaire.

The response suggests a well established community of users with well focussed needs. Details of the results of the questionnaire are in Appendix 2.1

4.1.1 Main findings

- The demonstrated approach to linking is endorsed with some qualifications.
- All aspects of the interface need more user-focussed development.
- The use of PDF as the linking format continues to present some users with problems, e.g. speed of download, viewer not supported, lack of familiarity with the format.
- Two most-desired enhancements to the demonstrator are to:
    1. make it an integral part of arXiv
    2. links to online journals.
- Two results indicate that arXiv papers are the most appropriate place for reference links:

1. These users overwhelmingly use arXiv for accessing full texts of papers and as their principal resource discovery tool.
2. Citations are used (predictably) to discover new works to greater or lesser degree (not clarified in these results).

4.2 Evaluation by OpCit steering committee

Although presented in questionnaire form as above, this is less a survey, more a prompt for comments. Further, its real focus is not the current demonstrator, although some welcome and constructive comments on it were received, but future versions. The demonstrator on which this evaluation was based is less inclusive of the work of other partners than future versions will be, but it will be the basis of later versions that will be more inclusive.

The quality of responses received was high, but the overall number of responses was disappointing, and the lack of inclusiveness would have been a factor.

All members of the steering group received the questionnaire by email, with individual notes to direct their comments. Detailed results from the questionnaire are in Appendix 2.2.

4.2.1 Highlighted comments

- "Everytime I stop by the demo, I'm excited about this technology."
- "Overall a great start. This project should be able to evolve quite rapidly into a great overlay over arXiv.org and be a valuable portal for researchers. Citeseer integration will be terrific."
- "If you can provide a simple PDF->PDF+OpCit converter that puts in SFX-type links then I see no reason why we couldn't incorporate this in arXiv at an early stage"
- "I'm on a campaign to get ACM to convert directly from Latex to PDF (rather than going through SGML); this may play a role in that process."

4.2.2 Main findings

- Support more links, e.g. to online journals, using alternative resolver services, e.g. LinkBaton, SFX, DOI, to provide these.
- Quantify link reliability.
- Improve the user interface:
    1. Provide user option to see links on (html?) wrapper pages as well as on PDF full-texts.
    2. Intra-text links (e.g. from citations to references) recommended.
    3. The 'SFX' interface needs to be reconsidered and updated - needs to be 'smarter' and offer more working options (taking care not to infringe the interests of ExLibris', which is commercialising SFX).
    4. PDF slow; bitmapped fonts and scanned images should be avoided.
    5. The browsing interface for locating papers could be improved.
    6. Link colours (boxes) can be confusing.

- Define and package linking software components for non-Southampton applications, e.g.
    1 to add PDF+OpCit links option to arXiv
    2 to support publisher linking services
    3 to improve pre-publication production systems

5 Future developments

Future OpCit developments are intended to support further releases of the linked demonstrator, each iteration adding new content, new services and new features, from Southampton, Cornell and elsewhere, eventually in a fully distributed environment. The programme of planned releases is shown in Table 1.

Version	Archives	Added features	Release period
V2.0	arXiv physics archives	Forward linking, links to online journals services and other digital library services	2Q2
V2.5	NCSTRL, CoRR, WWW conference series (ACM DL)	Inter-link distributed archives	4Q2
V3.0	All arXiv (inc. maths) and v2.5 (other Open Archives as appropriate)	Knowledge linking	3Q3

Table 1. Future releases of the OpCit demonstrator

6 Contacts with other projects

A remarkable convergence of technologies and services to build interdisciplinary and institutional archives with links to other reference sources such as online journals and aggregated services is in prospect. The project plans to work with a range of partners to develop the following components in this framework:

- User interface: EPrints, California Digital Library
- Interoperability with other archives: OAi
- Links to published resources and online journals: SFX (OpenURL), LinkBaton (Openly Informatics), DOI/CrossRef, American Physical Society (APS), Highwire Press
- Indexing and bibliometrics: ResearchIndex (NEC Princeton), Stanford Linear Accelerator Center (SLAC)

We intend to extend this list of collaborators as the project releases more tools and resources.

7 Project publications and presentations

Publications

Harnad, S. and Carr, L.
Integrating, Navigating and Analyzing Eprint Archives Through Open Citation Linking (the OpCit Project)
Current Science, 2000 (special issue honour of Eugene Garfield) (in press)
http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.citation.htm

Les Carr, Steve Hitchcock, Wendy Hall and Stevan Harnad
A usage based analysis of CoRR. A commentary on: "CoRR: a Computing Research Repository" by Joseph Y. Halpern
ACM Journal of Computer Documentation, May 2000
http://opcit.eprints.org/comment/JCD-commentary.html

Steve Hitchcock, Les Carr, Zhuoan Jiao, Donna Bergmark, Wendy Hall, Carl Lagoze and Stevan Harnad
Developing services for open eprint archives: globalisation, integration and the impact of links
Proceedings of the fifth ACM Conference on Digital Libraries, San Antonio, Texas, June 2000.
http://opcit.eprints.org/dl00/dl00.html

Presentations

Stevan Harnad is giving a series of talks at various information forums on
"How and Why to Free the Give-Away Refereed Research Literature Online Through Open Archiving"
Full list of meetings with dates at
http://listserver.sigmaxi.org/scripts/wa.exe?A2=ind00&L=september98-forum&F=l&S=&P=27435

Steve Hitchcock, *et al*.
Extending Reference Linking for Open Archives
DLI 2 All-Projects Meeting, Stratford-upon-Avon, June 2000
http://opcit.eprints.org/talks/stratford/title.html

Steve Hitchcock, *et al*.
Developing services for open eprint archives
Fifth ACM Conference on Digital Libraries, San Antonio, Texas, June 2000
http://opcit.eprints.org/talks/dl00/title.html

Stevan Harnad
The journal is dead: long live the journal!
Keynote presentation at *Death of the journal: will PubMed Central kill the journal in Health and Life Sciences? ASLIB Biosciences Group Annual Conference*, London, May 2000

Donna Bergmark
An Architecture for Reference Linking
Internal presentations, Cornell University, April, May 2000
http://www.cs.cornell.edu/cdlrg/Reference%20Linking/An%20Architecture%20for%20Reference%20Linking_files/v3_document.htm

Donna Bergmark
Current Projects in Reference Linking
Internal presentation, Cornell University, December 1999
http://www.cs.cornell.edu/Prism/private/MeetingSlides/REFLINK/BergmarkTalkForProjection/

Donna Bergmark and Steve Hitchcock
The Open Citation Project
Digital Libraries Initiative All-Projects Meeting, Cornell University, October 1999
http://www.cs.cornell.edu/bergmark/DLICornell/title.htm

Appendix 1. Three year evaluation plan

A1.1 Introduction

Through collaboration between Southampton and Cornell Universities, and by building on the outputs of earlier projects, the Open Citation project is able to combine research into generic linking structures and interoperability requirements with early-stage implementations and evaluation from year 1. These preliminary, qualitative evaluations will be limited in scope and distribution, the purpose being to inform more complete and more sophisticated versions to follow in years 2 and 3. Usability tests and fuller evaluations will be performed on those versions.

A1.2 What we're evaluating

The project is developing and demonstrating reference linking services for eprint archives, notably those archives that conform to the Open Archives initiative (OAi). The principal resource is the Los Alamos physics archive, the largest and pre-eminent archive of its type, with over 100,000 documents. Initially a locally-stored version of that archive is being maintained as though it were a mirror but with all documents stored in a particular delivery format (pdf). This **first demonstrator (v1.0) links every reference that points to another document that can be found in the archive to that document**. This version is built and is currently being investigated.

Linking a reference to the document it cites is generally referred to as 'backward linking', as it points backward in time to an earlier document. Linking forward in time, to discover later documents to have cited a given paper, is also desirable (although in the faster-moving environment of eprint archives the distinction between forward and backward, before and after, can become blurred). Forward linking demands a more powerful database of papers and their references than has been built to date, but **we plan a new version (v2.0) of the demonstrator with forward links and a better user interface in year 2**.

References can **also link to journal articles and other digital library services**, not just to eprints. Developing link services of this type will depend on relationships with external partners and will not be implemented by the project alone, which is focussing on free-to-access repositories of academic papers. We have a number of such partners on our steering committee and hope to test links to their services in v2.0.

Additionally, at this stage we want to **test an author deposit interface for archives, including a reference checking** feature. It's not clear what the target archive will be but the primary interface will be based on that used by the EPrints software. It must be a large archive, ideally one that has already adopted EPrints.

It is planned to **extend these linking services to include other Open Archives in other disciplines (v2.5)**, notably those that support the OAi interoperability requirements. This is important because distributed archives need to enable services, such as linking services, to collect relevant data. The target area for this demonstration is likely to be computer science archives, since these are large and

distributed (though some are disorganised). The largest computer science archives – the Networked Computer Science Technical Library (NCSTRL), and the Computing Research Repository (CoRR), will become OAi compatible, and there are various supplementary resources (including journal and conference publications) that the project has access to.

The blurring of forward and backward links leads us to consider the archives not in a time dimension but as a knowledge space. Addition of 'knowledge-based' linking services- e.g. using ontological reasoning to represent a sophisticated conceptual model of document terms and their relationships - will be the principal feature of **the final project demonstrator (v3.0), which will attempt to inter-link the earlier version releases (v2.0 and v2.5),** and other Open Archives where possible**.**

A1.3 The evaluation plan

The plan is to evaluate the demonstrators produced by the project. There is currently no plan to evaluate the operation and implementation of the project itself.

The evaluation plan will have formative, i.e. feeding back into the development of the project, and summative, i.e. outputs of the evaluation, elements.

A1.3.1 Focus

The nature of the project defines the following areas for evaluation:

- presentation of links (principally links on references; also knowledge-based links)
- effective integration of services (including online journals, other digital library services, etc.)
- the user interface (reference checking services)
- effective integration of resources (Open Archives, different disciplines)

A1.3.2 Methodology

Feedback will be generated from these contributing groups:

- specialist test user groups for each release selected from user and author lists for the linked archives – emphasis on presentation of links and integration of resources
- project steering group, includes archive administrators, developers, database providers, librarians, publishers, representatives of standards organisations and learned societies – emphasis on integration of services
- special-interest groups, e.g. members of the Open Archives initiative, developers of reference linking services – especially for later releases
- local (Southampton) user group for pre-release usability testing of interfaces

These are geographically broad groups, so questionnaires and electronic communication will of necessity be the main means of inviting feedback.

A1.3.3 Requirements

The evaluation will require the following activities

- Formal questionnaires (specialist and general users)
- Group meetings (steering group – may be geographically restricted)
- Individual meetings (local users)

A1.3.4 Summary timetable

Version	Archives	Added features	Release period	Evaluation
V1.0	arXiv physics archives	Backward (in time) linking	2Q1	4Q1
V2.0	arXiv physics archives	Forward linking, links to online journals services and other digital library services	2Q2	3Q2
V2.5	NCSTRL, CoRR, WWW conference series (ACM DL)	Inter-link distributed archives	4Q2	1Q3
V3.0	All arXiv (inc. maths) and v2.5 (other Open Archives as appropriate)	Knowledge linking	3Q3	4Q3

A1.3.5 Deliverables of the evaluation

- Reports on each release to be produced for the steering group
- Dissemination within JISC
- Papers and presentations

Appendix 2. Evaluation reports from questionnaires on OpCit demo v. 1.0

Appendix 2.1 Summarised responses from arXiv authors

Totals

No sent out 94; bounced 6; total out 88
No. returned 19
% return 21.6
Est. error margin 22.5%
Period of evaluation: 27 July to 13 August
Date of this report: 22 August

Evaluation

(Note. The selected target group of users included three subsets. The results from these subsets are indicated primarily for further analysis. The sum of the results from these subsets, shown in bold, is the main interest here.)

1 How useful did you find the reference linking in the demo as it is now? (select ONE by placing an X in the brackets)

(1.1) [1+6+3 **10**] Very
(1.2) [4+1 **5**] Moderately
(1.3) [2+1 **3**] Hardly
(1.4) [ ] Not

2 How useful would you find it if a larger proportion of the reference links reached their targets? (select ONE)
(2.1) [1+7+3 **11**] Very
(2.2) [3+1 **4**] Moderately
(2.3) [2+1 **3**] Hardly
(2.4) [ ] Not

3 What feature would you most like to see added to the service? (rank from 1 highest)
(3.1) [2 4+49+9 **62 points**] Integration with the underlying (physics) archives
(3.2) [4 2+27+7 **36 pts**] Links to databases of secondary content
(3.3) [1 5+38+12 **55 pts**] Links to online journals
(3.4) [3 3+27+5 **35 pts**] Links to other archives
(3.5) [5 1+17+1 **19 pts**] Links to services/holdings in your institutional library
Other (3.6) [
- Links to old but famous articles (there is KEK scanned archive, but it is not sufficient)
]

4 What features would you like to see modified? (select ONE OR MORE by placing an X in the brackets)
(4.1) [3+1 **4**] Link presentation
(4.2) [**5**] Link reliability
(4.3) [1+5+1 **7**] More links
(4.4) [4+2 **6**] User interface, e.g. better support for browsing
Other (4.5) [
- In the hypertex code I wrote I also made links from the paper body to the reference list and equations, figs, etc. These seem to disappear in your linked examples. As a general comment, most browsers don't allow to get back the hyperlink to the original location - for example, if I click on a link to an equation and check that equation, in most cases I like to go straight back to the place where I was, by clicking back. Most browsers don't support this (at least as far as I know) and this severely limits the functionality (the one I use is the Nextstep dvi previewer which is the only one I know that works satisfyingly).
- make links at the text instead in the references
]

5 Which source do you use most to access the full texts of papers? (select ONE)
(5.1) [1+10+5 **16**] arXiv.org
(5.2) [**1**] Electronic journals
(5.3) [ ] Print journals
Other (5.4) [(please specify)]

6 Which service do you use most to discover new papers? (select ONE)
(6.1) [1+10+4 **15**] arXiv.org
(6.2) [1+1 **2**] SLAC/SPIRES
(6.3) [**1**] Web of Science
Other (6.4) [(please specify)]

7 How often do you discover new works through citations? (select ONE)
(7.1) [2+2 **4**] Always
(7.2) [6+3 **9**] Frequently
(7.3) [1+4 **5**] Occasionally
(7.4) [ ] Not at all

8 Comments/suggestions
(8.1) [
- An opinion expressed by many people is that the service providing the number of citations could be improved. In particular, papers that are referred in the replaced versions are not being counted.
- I cannot view any of the links, I'm asked to supply a ``supported web browser``, which I apparently do not have.
- I saw the links but I couldn't get them to work. But maybe this is just because I don't know how to do web linking properly in PDF.
- Plz never include bitmapped fonts in the pdf files, as they are not very readable.
- The idea to use colours to discriminate between a linked paper, a non-archive paper, etc., at first sight seems nice (though I would prefer an ensuring green for a linked one and an alarming red for one out of reach). However, I realize that many of my colleagues have to use screens which offer only black and white - no colours.

Wouldn't it therefore be better to attach some symbol or letter - like "L" for linked, "O" for out of range, etc. - to each citation?
- Good luck.
- The only problem I had was configuring my acroreader to use the netscape browser. I had to find the path to our browser...
- The Acrobat PDF reader on all our local machines at our institute is helplessly slow. For this simple reason, all the wonderful services you suggest to use are simply totally inconvenient for us; it's MUCH faster to find papers "by hand" on the net. It's a pity, because the idea is good.
- I think it is a good time to separate astro-ph into two or more categories
]

Appendix 2.2 Summarised responses from the OpCit steering group

Totals

17 mails out; 13 sites
7 responses
Period of evaluation: 19 July to 4 August
Date of this report: 21 August

Evaluation

(Mark selected answers with X where appropriate; you are welcome to elaborate)

How effective is this as a demonstration of reference linking?
Very xxx **3**
Moderately xxxx **4**
Hardly 0
Not 0

- Moderately. I don't like getting linked to PDF documents; I have to invest a minute of download time to decide if I'm interested. I'd also like to go to the journal, not to a preprint, when I have the option.
- Moderately. The problem is, when I tried one of the papers (with SFX on) and clicked on a red box, it said "A web browser has not been specified". Going through the procedure didn't really help. Also, there were no links in the text itself (orange boxes?) only red boxes at the end, which made this a less effective demonstration.
- I tend to say "very" when my comments re the SFX explanation are considered. Also, I feel that ScreenCams could provide some additional value. But, from my own experience with that, I feel that it would be great to have something that would stream.

What features need to be improved?
o **Better search**
- I don't expect to be able to use a search facility when I'm linking directly from a citation to an article. This is a different scale of things - e.g. it's like trying to be an Ingenta or ISI.

- Better search would be nice
- I didn't see a search
- Do you mean searching for links within a document? I'm confused -- I didn't think there was any searching going on for users.
- It is somewhat awkward to break an arXiv id into year, month, and article number, it should be a single text entry box.
  If you stick with the current scheme, the article number box should add leading zeros as needed. If you go with a single box, the Los Alamos way of allowing current year or current month to be inferred should be supported (i.e, entering a '1' should correspond (today) to 0007001, entering a 6001 should be 0006001, etc.

o **Link reliability**
- So far, in the demo, it's really good. You know from looking at a link if it will work or not. However, in future versions, I'd rather only see the links that actually work, i.e., before making the link, make sure that you can link to the target doc.
- Journal links don't seem to work well
- Didn't test this too extensively. However, picking a paper at random (OK, it is one of mine), hep-th/9201076 all links have incorrect data (just a year for preprints, no journal name for journal articles) - so links don't work. Colors also seemed incorrect - for the two preprints cited, they should be orange and not red since the arXiv ID wasn't there.
- Link reliability: quite an issue, isn't it? Actually a very interesting research matter on its own.

o **Link presentation**
- presentation is great. You might want to think about making the red boxes into blue boxes, since blue is the default link color for most browsers, but that's nitpicking.
- a link button with a graphic would be more transparent to use
- The orange and red boxes are acceptable, although my guess is that highlight the link in red or orange rather than putting a box around it would look better.
- I would really like to see the links available either within the PDF OR separately as part of a wrapper page. Following link trails is inconvenient if you have to download a full PDF every time rather than going from wrapper page to wrapper page. Being able to extract links from a PDF into some kind of a tagged XML format (including reference type and tagging the components of the citation) would a very valuable tool.
  For the demo, having coded boxes for reasons things don't link is fine. Presumably in a live version, you would just not display anything for this references.
- I am colour-blind -- like about 50% of the men on this planet -- and this colour indication is really problematic for me. One could also think of indicating these things with figures. Have you seen the way the CNRI people show handle links on their own pages? Quite inspiring.

o **More links**
- in our HTML articles we make internal article links. Links from a figure mention ("See Fig. 1") to the figure, from a footnote mention to the footnote, from a reference mention to the reference. You could try to make those intra-document links, but let me tell you they're a lot harder than just reference links. I think that reference links are probably the number one bang-for-the-buck link you can put in scientific articles.
- More links in the text would have been nice
- link from main text to refs would be good
- Links to SPIRES and ADS would be great.
- addition of alternative resolvers i.e. other servers that provide service links (see OpenURL) addresses the "more links" issue.

o **User interface**
- fine for me, but you should probably do some usability testing, if you haven't already.
- I'm not convinced that I found the SFX on/off feature very useful - it just seemed to get in the way, but this probably says more about my lack of appreciation of the "appropriate copy" problem.
- I had problems with the user interface.
- The SFX-type page should display detail of the object it is trying to find and be a little more smart about what buttons are offered.
- Should be a way to choose which arXiv mirror you go to. Links shouldn't go from PDF to PDF. Much better to link to a wrapper page (link LANL abstract page) that gives users more information (abstract, PDF size, etc.) before final download decision. SFX page should give more information about the two papers being linked - at least display the arXiv id's (but full reference metadata would be even better). SFX base URL should be settable to point to other resolvers. OpenURL spec should probably be adopted. SFX page could use a lot of prettification. Adding a discovered arXiv id to reference might be a nice feature.
- I feel that your own SFX menu-screen could use some cosmetics. On http://arabica.ecs.soton.ac.uk/demo.html , I suggest to put the header for the explanation of what SFX-style is also on top of the page, not in the table. Plus, it would help to say -- in the explanation on SFX-style -- that "on" means the user will get a list of extended services (including full-text links), while "off" means that only links to full-text are provided.
  In the SFX explanation, the sentence re "SFX in principle, not SFX as such" is very confusing. Why not say that the provision of extended services via an intermediate menu-screen fits into the SFX framework, but that in the demo you don't actually use an ExLibris SFX server.

Other (specify)
- Your highlighted demo PDF files contain bitmapped fonts rather than Type 1 fonts. This should be fixed. PDFs of my paper I downloaded were Type 1 though.
- Many of the PDFs are very poor quality (produced from scanned images rather than Adobe Capture or directly from the origination program, I guess)

What features would you like to see added ? (rank from 1 highest)
o Inclusion of other linking services **8 points**
o Links to databases of secondary content **8 pts**
o Links to online journals **24 pts**
o Links to other archives **11 pts**
o Links to services/holdings in your institutional library **11 pts**
Other (specify)

- All these should be handed off via virtual links such as LinkBaton.
- I don't think you have to provide all of this yourself. Just open up your environment via OpenURL, so that others can demonstrate their creativity in creating linking services.
- Links to online journals are our #1 concern.
In terms of the demo, I don't know what databases you might link to, but in our biomedical articles, we have been asked to link genetic sequences to GenBank and some article references to PubMed. These are both nice features, and will be required at some point, but as I said above, the reference linking is most important for now.
The main thing that we would need added to the software before we could use it would be much better linking to online journals. That includes:
A. recognizing the references to each journal, in spite of authors who seem to be trying to make it difficult by using weird abbreviations for things.
B. algorithms for creating article links into each online journal. Our journals all work the same, but many other journals don't. Since you can't really take the time to test each link before you place it, you'll need a list of online journals that you can link to and the method to turn volume, issue, and page into a URL for each one.
C. possibly link status icons, based on journal info. For example: a free journal would have a little icon next to links to it, notifying people that they don't need a subscription to view the link. Subscriber-access only journals might have a different icon (or no icon, since this is pretty much the default :( ), and journals with pay-per-view might want to advertise that in links. We've done some work on A., B., and C., at least for our own journals

How can the work of the project feed into your work?

- We would like to run all of our PDFs through the software to generate reference links to all of our other journals. Secondly, we'd love to link to other publishers' journals and to PubMed. Additionally, we are looking into Manuscript Submission and Review systems right now, where an author would submit a PDF of his paper and the journal staff would review it online. If we could link the references in that (totally randomly formatted) paper, it will save reviewers tons of time in checking those references.
- Its most direct relevance is a) to my work on DC-CITE, and b) to CrossRef.
- If you can provide a simple PDF->PDF+OpCit converter that puts in SFX-type links then I see no reason why we couldn't incorporate this in arXiv at an early stage (we give two PDF options: PDF and PDF+OpCit links). If an SFX-type server is used then there would presumably be no need to for us to store any resource databases. We are currently testing OpenURL linking from our abstract pages (http://sfx1.exlibris-usa.com/openurl/openurl.html). It would be good to be able to specify a BASE-URL (taken from cookie) which gives the user's chosen server as

input to such a converter. However, I imagine this creates a whole bunch of issues about whether you want to use the type of ids/parameters specified by OpenURL. Our PDF usually has links already: either within the document, to other arXiv documents, or to URLs. I would like to see the internal links preserved as mentioned below.
- The free-form reference parsing and dynamic adding of references to PDF would be great for handling manuscripts within the editorial process (pre-production). Even in production it would be helpful to add links to our PDFs (though there is an argument against this - adding links is a dynamic process and is viewed as an added-value service - making them part of the PDF rather than a wrapper page means 1) a user looking at a saved PDF may not be getting the most up-to-date links and 2) they are able to retain the added-value service after their subscription lapses. SFX style linking addresses both of these problems though (a user would have to go back to the publisher gateway and this could be subscriber controlled). This is not necessarily the dominant view around here - having links in the PDF is certainly convenient, but I thought I would mention it.

How would you like to see your work feed into the project?

- we might be able to set up a service that would link PDFs for other people.
- I would very much like to see CoRR hooked into this.
- Support for either DOI linking or pragmatic citation linking (APS link manager) for instance would be a valuable addition. APS could work to supply accurate metadata for improved matching of Phys. Rev. citations
- wouldn't it be nice to let users choose an SFX resolver. I can at least think of 3: your system, our SFX server in Ghent and a demo SFX server from Ex Libris. This approach would clearly illustrate the place of the SFX framework in reference linking. And it would somehow make the mentioning of Ghent and SFX as partners in OpCit more real.

  Implementing this possibility only requires the implementation of the CookiePusher and the OpenURL (see http://www.sfxit.com/OpenURL). In most environments taht we have dealt with so far, the work involved is marginal. Initially -- to make things simple -- it could even start by only providing OpenURLs for arXiv identifiers. As you can see from the specs, the OpenURL is extremely simple in this case.

Comments/suggestions
- For most tex papers our current system uses hypertex to provide links within papers and to other archive articles. Clearly your approach is more general and will deal with the increasing number of non-tex submissions. I'd like to see your system produce the intra-paper links from references to the references section (though I imagine some people would prefer the reference to go directly to the referred-to paper - actually less useful in my opinion since in must cases one just wants to know which paper/worker is being referred to). I see no reason why you should not create links to journals and archives for references that cite both. Obviously, linking between archives is relatively straightforward (though I notice that references split by a line-break are not found, eg [132] in hep-th/0001001). I think the real test is identifying references to journals and perhaps even books.

- 1. did you know arXiv now IS OpenURL-aware? The Ghent SFX-server already provides quite some services for incoming OpenURL's carrying arXiv identifiers. We fetch metadata back from arXiv using the OAi Dienst Subset.
2. I am quite interested to understand more about the technique you use to actually insert links in the references: are the links being re-written dynamically at every download, or do you have a link being inserted once (in an batch process) that points at a front-end that connects to something like a handle system for the real resolution of the link (the DOI-style approach)? The more I work on these matters, the more I feel the latter is a very interesting path. With this respect you may be interested in our recent DOI-related experiment documented at http://sfxserv.rug.ac.be:8888/public/xref/ , and to be published as soon as I have time.

Appendix 3. Extended glossary of distributed information and reference linking services

ArXiv. The pre-eminent collection of eprint archives based at Los Alamos covering physics, mathematics and computer science. Launched in 1991 covering just high-energy physics, the archives contain over 130,000 deposited papers and are mirrored internationally at over a dozen sites.
CogPrints. An eprint archive for cognitive sciences, modelled on the Los Alamos physics eprint archives and hosted at Southampton University.
Computing Research Repository (CoRR): an eprint archive of papers in all areas of computer science. CoRR is one of the Los Alamos arXiv eprint archives, and is also one of the federated NCSTRL libraries.
CookiePusher. A user interface that informs an SFX server of user context and preferences.
CrossRef. A commercial reference linking service for journal publishers. Uses a Digital Object Identifier (DOI)-based link resolver.
Dienst. A protocol and server for distributed document libraries. At its most basic Dienst is a protocol for communicating information, e.g. metadata, about scholarly works. A subset of Dienst is the protocol for harvesting metadata from archives complying with the technical requirements of the Open Archives initiative (OAi Dienst subset). Information is carried in the form of Web protocol (http) requests. Its largest application is the Networked Computer Science Technical Report Library (NCSTRL).
Digital Object Identifier (DOI). An identification system for intellectual property in the digital environment. Developed by the International DOI Foundation on behalf of the publishing industry, its goals are to provide a framework for managing intellectual content, link customers with publishers, facilitate electronic commerce, and enable automated copyright management. CrossRef is an agency assigning DOIs to publishers for use its reference linking service.
Distributed National Electronic Resource (DNER). A managed environment for those in the UK higher and further education community to access quality-assured information resources on the Internet. Resources include scholarly journals, monographs, textbooks, abstracts, manuscripts, maps, music scores, still images, geospatial images and other kinds of vector and numeric data, as well as moving picture and sound collections.
Distributed Link Service (DLS). Software supporting link placement in third-party documents that are accessible anywhere on the Web. Like link resolvers, the DLS uses database lookup to determine possible link destinations, but doesn't need to produce an intermediate page to present links to the user. Various implementations and software components are available for different applications. Used in OpCit to support reference linking, the DLS reads PDF documents to identify references, formats the extracted reference data and compares with a pre-built database to discover locations for the referenced documents, where possible returning the results as links in the original full-text document. This process runs at the moment a document is requested by a user.
Dublin Core. A metadata element set for labelling electronic resources. The elements represent a broad, interdisciplinary consensus about the core set of elements that are likely to be widely useful to support resource discovery on the Web.

Eprint archives. Classified and indexed storage and retrieval services for formal scholarly papers deposited by authors. The oldest and largest eprint archives cover physics and are based at Los Alamos (arXiv). Most eprint archives are subject-based and are stored on a main server and mirrored – an exact replica of the content is regularly copied – at other international sites. Some single-discipline services register content from many servers at different institutions, such as the Networked Computer Science Technical Report Library. It is anticipated that single-institution archives will develop with broad access coordinated by Open Archive services and EPrints.
EPrints. Generalised software developed at Southampton University for managing eprint archives. It addresses the need for simple set-up of archives, within institutions and other organisations, and provides an interface for administrators, for authors to deposit papers, and for users to access papers.
GenBank. An annotated database of all publicly available DNA sequences, maintained by the US National Institutes of Health.
LinkBaton. A user interface to a link resolver that directs particular link types at user-specified resources. Like SFX, LinkBaton promotes personalization and localization in link services. Developed by Openly Informatics, Inc., it directs links to universal services such as online booksellers or stock ticker services. It is planned to offer journal reference links via chosen services.
Link resolvers. Two issues in resolving links to destination documents on the Web are: stability of the documents' location, and multiple versions. Link resolvers can help with one or both of these problems. Link resolvers can also control access. Instead of containing a direct URL, a Web link can send a request to a resolver, which might return a document or offer a selection of documents. Examples of resolvers for scholarly reference linking include CrossRef, LinkBaton and SFX.
Networked Computer Science Technical Report Library (NCSTRL). A federated collection of TR report libraries maintained by different university computer science departments. Like a specialised eprint archive, the reports are freely accessible to readers via the Web. Coordinating services – central indexing and communications between servers - are managed using Dienst.
Open Archives initiative (OAi). Initially a forum to solve interoperability between author self-archiving solutions (eprint archives), now extended to support a wider range of digital resources of academic and scholarly interest. Besides eprints and electronic texts, such resources include science and social science data sets, visual materials, archival collections, geographic information system data, sound and music, and video. Interoperability hinges on a fundamental distinction between the archive-functions, which include data-collection and maintenance, and end-user functions. In this approach, there are data providers and service providers. Data providers (such as individual eprint archives) support a simple harvesting protocol and provide extracts of metadata in a common, minimal-level format in response to requests from service providers. Service providers use extracted metadata to build higher level, user-oriented services, such as catalogues and portals to materials distributed across multiple eprint sites. The approach and its protocols were documented in the "Santa Fe convention".
OpenURL. A URL that transports metadata, or keys to access metadata, for a digital document or object for which the OpenURL is provided. A compliant resolver can read the OpenURL. Services that plan to generate and output OpenURLs include arXiv, CrossRef, ISI and the Open Citation project. There have been preliminary discussions about developing OpenURL as an ANSI international standard.

PubMed. Online search service that provides access to Medline, an abstracting and indexing service covering biomedical journals, and other related databases from the US National Library of Medicine.
ResearchIndex. Software designed for 'autonomous citation indexing', in effect builds an ISI-like index of the online full-text scientific literature. Developed at NEC Princeton, it has been used to search and index computer science papers on the Web.
Santa Fe convention. Document specifying the technical requirements for implementing Open Archives.
SFX. A link resolver and server designed to provide users of a given university or institutional library with access to local resources and to networked subscription and non-subscription-based services. Promotes the concept of 'localization' in reference linking. By receiving information on who or where a user is, the user's 'context', it is argued that the resolver can offer the most appropriate resources. In particular, the service can identify users that are authorised to access commercial services, e.g. journals licensed to a given site. Developed at Ghent University in Belgium, SFX is being commercialised by ExLibris.
SlinkS. The Scholarly Link Specification framework facilitates inter-publisher reference linking by providing a syntax and vocabulary for exchanging information. Services such as LinkBaton are used this framework.
SPIRES. Stanford Public Information REtrieval System A collection of library databases covering high-energy physics including journals and eprint archives. Hosted by the library of the Stanford Linear Accelerator Center (SLAC).

This report coordinated by Steve Hitchcock, sh94r@ecs.soton.ac.uk
19th September 2000