Integrating, navigating, and analysing open Eprint archives through open citation linking (the OpCit project)

Stevan Harnad* and Leslie Carr

Intelligence/Agents/Multimedia Research Group, Electronics and Computer Science Department, Southampton University, Highfield, Southampton, United Kingdom SO17 1BJ

The Los Alamos Eprint Archive (LANL) is a public repository for a growing proportion of the current research literature in physics. The Open Citation-linking Project (OpCit) is making this resource still more powerful and useful for its current physicist users by connecting each paper to each paper it cites; this can be extended to all the rest of the disciplines in other open archives designed to be interoperable through compliance with the Santa Fe Convention. A citation-linked online digital corpus also allows powerful new forms of online informetric analysis that go far beyond static citation analysis, measuring researchers' usage of all phases of the literature, from pre-refereeing preprint to post-refereeing postprint, from download to citation, yielding an embryology of learned inquiry.

0.0 Gene Garfield's scientometrics before the era of online open eprint archiving

Bibliographic citation is the mother of all hyperlinks. If Gene Garfield[1] had come of age in the online era he would still have been focused on monitoring, searching, accessing, linking, and analysing the research literature, but he would have gone about it in a radically different and far more powerful way. Indeed, in many respects, Gene's ideas and efforts were before their time: they wanted a PostGutenberg digital corpus all along[2].

Current Contents helped enfranchise scientists from (what was then called) the third (and second) world, as well as many from the less prosperous institutions of the first world. By bringing them at least the weekly contents pages of the vast pre-digital journal literature to which their institutions could not afford to subscribe (and for which their research activities scarcely afforded the time for the legwork of shelf-browsing in any case), Gene made it possible to let their fingers do the walking. Reprint requests could be mailed to the authors of the papers that non-first-world scientists needed to read (or, at the better-heeled first-world institutions, secretaries or students could be dispatched to the journal shelves to photocopy them).

But turnaround times were still slow and uncertain: The requested reprints were a long time in coming, if they came at all, in the non-first world; and even in the first, a lot of time and resources were wasted retrieving reprints for which a glance at their abstracts or full texts, once they were in hand, immediately revealed that it had been a false alarm. And even the relevant 'hits' had to be committed to growing, groaning reprint shelves in labs, which eventually created storage, navigation and retrieval problems of their own.

Nevertheless, on balance, the increased access to the literature that this system of monitoring and requesting provided was undeniably a benefit to research and researchers, and increased both productivity and impact.

And Gene's other major contribution, *Science Citation Index*, made it possible to monitor and measure that productivity and impact. Gene Garfield did not invent the 'publish or perish' metric of productivity, but he certainly fine-tuned it, with the citation-ratios of papers, authors and journals helping to supply promotion/tenure committees (as well as library serials-selection committees and sometimes even research funding committees) with a greater variety of beans to count – supplementing the peer-review system itself, which, if I were not a vegetarian, I would call the real meat of research assessment[3].

Nor was citation analysis merely an evaluative metric: It was also a way of charting the present, past and future course of research, sorting out the pedigrees of ideas and findings, and in general doing what might be called the quantitative 'embryology' of knowledge.

All that, despite the obstacles that both enterprises faced in the papyrocentric Gutenberg era. Journal contents pages had to be gathered, photo-copied, cut/pasted, collated, printed and mailed around the world by the Institute for Scientific Information (ISI) every week; citation lists had to be laboriously retyped, compiled, analysed, printed and again mailed around the world in vast, heavy, cumulative compendia. What a different world it would have been if all those data – journal contents pages and refer-

*For correspondence. (e-mail: harnad@soton.ac.uk) ence lists – had been digital to begin with! Nothing to photocopy, key-in, scan or cut/paste; and, if we threw in the Net along with the bytes, nothing to print or mail either! And as long as we are digi-dreaming, why not throw in the full texts of all those articles too? Then it is not only ISI that no longer needs to bother with digitizing or mailing, but researchers too can burn all their remaining reprint request cards – and free their shelves of offprints!

If the entire journal corpus had been digital and online, far more intelligent and customized automatic alerting services could have been devised than the mere scanning of weekly contents pages; and instead of passively waiting to be alerted, researchers could actively search and navigate the entire journal literature – not only through subject-, keyword-, and even full-text-searching, but also through citation-searching of a completely interlinked corpus. Searchers could even set thresholds for the impact levels of the papers, authors and journals to which they wished to restrict their search. For the earlier embryological stages of papers, hit-rates (down-load frequencies) for the preprint could supplement citation-rates for the reprint. And this rich, dynamic and growing embryonic corpus would have been the database for Gene's pioneering bibliometric analyses, with online user-based measures such as citation surfing, downloading, and hit-immediacy to complement the offline author-based measures such as publishing, citing and citation-immediacy.

This digi-dream is now becoming a reality, thanks to the open archives initiative (http://www.openarchives.org), interoperable open archiving software (http://www.eprints.org) and the Open Citation (OpCit) Linking Project (http://opcit.eprints.org).

1.0 The target

It is easy to state what would be the ideal online resource for scholars and scientists: All research papers in all fields, systematically interconnected, effortlessly accessible and rationally navigable from any researcher's desk worldwide[4,5].

To implement this immediately, the entire preprint and reprint literature would need to be available online in a usable, unified form. It is not, yet. But in physics a sufficiently large and representative subset of it is already available (see http://xxx.lanl.gov/cgi-bin/show_monthly_submissions).

So the work on that subset is now underway, and successful results based on it will not only generalize to the rest of the literature, once it is all online, but they will help to draw it all online more quickly.

2.0 The LANL Archive

The subset in question is the Los Alamos National Laboratory (LANL) Eprint Archive (http://xxx.lanl.gov)[6,7] which already contains 130,000 papers and is growing at an annual rate of 25,000 papers, with over 50,000 users daily, and *15 mirror sites around the world*. LANL also contains the *Computing Research Repository* (CoRR), which can be accessed directly through LANL or through the more generalized and integrated interface of the Networked Computer Science Technical Reference Library (NCSTRL)[8].

The LANL Archive represents a substantial body of literature in physics, mathematics and computer science, but the full texts are archived in a variety of forms, from HTML to TeX to PDF to PS (Figure 1), and the first problem that needs to be solved is designing a way to integrate and navigate them seamlessly.

One especially important feature of full texts – their reference list – is arguably the most natural and powerful way of interconnecting and navigating this literature. The 'links' are already provided by the authors themselves, and users already have a long, skilled tradition of navigating with them 'offline' (looking up the references in paper).

The Open Journal Project[9,10] and CogPrints (http://cogprints.soton.ac.uk) successfully used citation linking to interconnect a small but *interdisciplinary 'seed' database of full texts in the cognitive sciences* with a much larger 10-year set of abstracts and their reference lists from a subset of the ISI (http://www.isinet.com/prodserv/citation/citsci.html) journal citation database in the cognitive sciences (psychology, neurobiology, computer science, linguistics, philosophy). This work went some way toward solving the problem of automatically recognizing and linking (within and between texts) the finite but noisy set of existing citation formats[11–17]. The reaction of users was exhilaration with citation-based navigation, but frustration at accessing only abstracts. The obvious conclusion to be drawn from this was that the real power of citation linking can only be realized with full-text linking. OpCit is now doing this with the LANL Archive.

3.0 Citation analysis

Citation-linking offers further benefits over and above natural navigability for users and a natural unifying con-

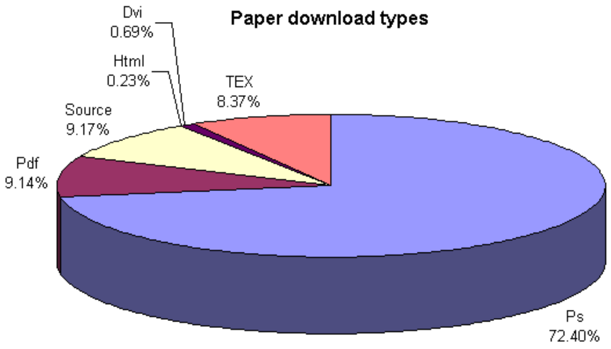

Figure 1. File formats in LANL Archive.

straint on interoperability and metadata processing in interconnecting and integrating the literature[18]. It also offers new ways of analysing and understanding how the literature is used: Author-end citation analysis[1] already reveals the offline lineages and dynamics in the growth of research knowledge[19]. Reader-end citation analysis can now be used to analyse online usage patterns in a way that raw 'hit' rate (download frequency) cannot do (Figure 2). (The fact that you got from article A to article B by the series of citation links k, l, m . . . is informative whether or not you actually stopped to read the full text of each link along the way: It reveals which paths are used in the citation forest, and how well-travelled they are.)

The LANL Archive has an additional interest from the standpoint of usage and citation analysis. It consists of both *unrefereed preprints* and *refereed reprints*. Within one year of being deposited, about 60% of the papers in LANL are updated to include the full reference to the journal in which they have been accepted for publication. It is not yet clear how many preprints are updated to append the full final text of the refereed reprint, but LANL papers are being updated as many as five times (Figure 3).

A new form of citation has also appeared (in both the paper and the online literature): citing the LANL preprint number[20]. This will no doubt become a standard practice and must hence be covered as a special case of citation linking, so that links can be dynamically updated and aliased to the common source as soon as the reprint takes the place of the preprint.

The emerging patterns of preprint vs. reprint citation and use are also a natural object of analysis in their own right; they represent a revealing microcosm of the overall transitional process that is taking place as this new medium evolves its own niche in scholarly and scientific research practice – the 'scholarly skywriting' continuum – to reveal previously hidden embryological stages of learned inquiry and interaction[21].

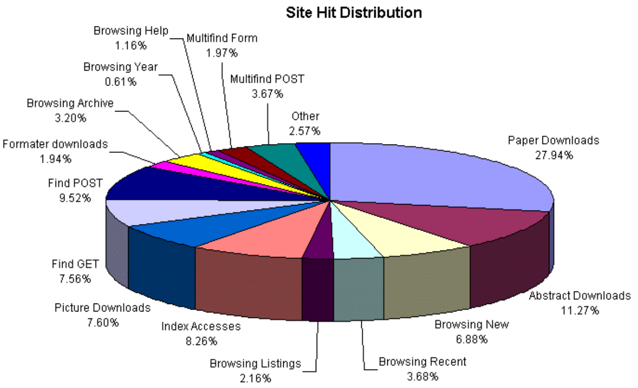

Figure 2. Distribution of 'hits' of different kinds.

4.0 Eight components of the OpCit linking project

The OpCit linking project has 8 components which are being pursued partly in parallel, because some aspects of them are independent of one another, and partly serially, because some aspects of later components depend on the outputs from earlier ones. In addition, there is an open-ended 9th category of further planned enhancements.

4.1 *Universalizing the author self-archiving software: Eprints*

The LANL Archive's present author self-archiving system is a robust and successful one; it is what has made LANL the essential resource it has become. Authors in physics are comfortable with it, and it is clear that it is extremely effective. But now that the success of LANL's physics sector has made it apparent that self-archiving should be extended to other disciplines[22], the more general needs and practices of those other disciplines need to be taken into account in designing new, generalized archive software suitable for all fields.

LANL itself is upgrading to keep pace with new technical developments and with evolving practices among physicists. But now the adaptation has to go beyond the physics community, which is already accustomed to the present LANL self-archiving interface and procedure, to other disciplines that are familiar neither with LANL nor with *eprint* (= electronic preprint and reprint) self-archiving.

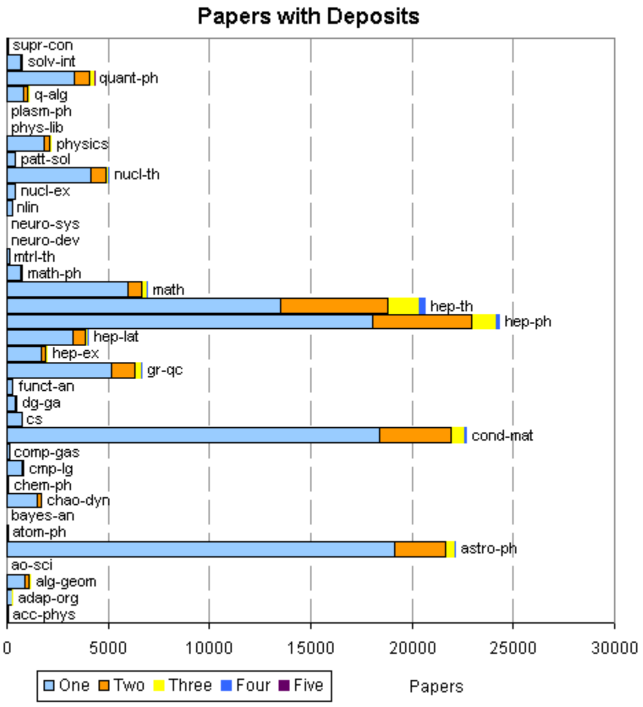

Figure 3. Multiple updates by LANL subfield.

The open archiving initiative (http://www.openarchives.org) has recently formulated the Santa Fe Convention[18] agreeing to adopt a subset of the Dienst Protocol[23] for tagging and sharing metadata to ensure that all compliant archives are interoperable. Eprints (http://eprints.org/), in collaboration with OpCit (http://opcit.eprints.org/), has accordingly designed and released (free) open-archiving software for creating, customizing and maintaining Santa Fe-compliant open archives. Eprints is designed for adoption by all universities and research institutions worldwide, so they can collectively become a distributed, interoperable, universal archive of the research literature.

The constraint of citation linking itself provides a shared skeletal structure to constrain the design and to unify the format of open archiving: The full range of variation in citation formats exists in all disciplines; hence this skeletal structure must be extractable from all texts, in a form that can be used for hypertext linking. The adaptations (in both the author interface and the infrastructure for depositing texts) dictated by the need to extract and link citations in the texts, their reference lists, and the texts they cite, for all formats, will then constrain the drafting of future texts, the formats in which authors are encouraged to submit them, and the way those formats are processed by the Eprints software. In other words, whatever it takes to make all deposits interoperable, specifically for citation extraction and linking will also help to make them interoperable in other respects, because citation linking is a representative microcosm of text interlinking in general.

To the extent that these citation-specific adaptations influence author practices, they should also help to speed the standardization of formats and procedures that will eventually converge on the optimal online universal resource for the learned research community[24].

4.2 *Optimizing the Eprints reader interface for open archive navigation*

The present LANL reader interface is designed for retrieving papers as one would from a bibliographic database: top-down, using titles, author names or keywords (as well as less general classifiers such as year, subject area, etc.). Once the paper is retrieved, the duties of the interface are done. Apart from any hyperlinks in the paper itself – if it happens to be available in a readily linkable format such as HTML (for the retrieved paper might instead be in a variety of other formats, including TeX or Postscript – see Figure 1) – the present LANL interface and infrastructure offer no further navigational possibilities; the only way to retrieve another paper is by going back up for another top-down search.

OpCit is making it possible to retrieve LANL papers in a citation-linked format (currently HTML or PDF), so that once a user has retrieved an entry-level paper, navigation of the entire archive can continue via citation-links, with no need to launch another top-down search (although the top-down capabilities – keyword, author, and eventually even full-text search of the archive – continue to be available at the paper level). Heuristics and algorithms for content classification have been developed to do the citation linking[12,16].

4.3 *Extracting citation data, generating hypertext links, and automatic addition of hypertext links in the archive*

With author and user interfaces adapted for open archiving and navigation by Eprints (section 4.1–4.2), author and reader practices will evolve, but until they do, all papers must now be converted into a citation-linkable form, and then linked[25]. Many details of document format conversion and bibliographic formatting need to be addressed in order to optimize this automatic reference detection and linking capability. Some approximate solutions have already come from the Open Journal Project and CogPrints[9,14], which worked on texts in PDF and HTML formats from various different journal publishers and from the Psycoloquy (http://www.princeton.edu/~harnad/psyc.html) BBS (http://cogsci.soton.ac.uk/bbs), CogPrints (http://cogprints.soton.ac.uk) Archives (plus a *Cognitive Science Subset* of the ISI abstracts/citations database http://journals.ecs.soton.ac.uk/TryOJ.htm).

One of the many advantages of extending this work to LANL is that partial results can be used to hasten progress toward fuller results. A subset of the LANL Archive can already be fully citation-linked immediately, using the current linking tools, namely, those papers that have correctly specified, well-formed bibliographic citations and have been typeset by software which maintains the textual contents of the page. That subset has now been fully linked to all papers it cites that are likewise in the Archive (including those that are not yet themselves part of the fully interlinkable subset because their own references cannot be further linked outwards: their titles, author-names, abstracts and keywords are nevertheless enough to find and link inward into them).

Users will accordingly have a chance to experience and compare functionality under two conditions: when they retrieve papers whose full texts are also linked (http://arabica.ecs.soton.ac.uk), and when they retrieve papers that are dead-ends, like the abstracts that frustrated the users of the ISI/Open Journal Cognitive Science database (http://journals.ecs.soton.ac.uk/TryOJ.htm).

[Heavily used open archives like LANL allow author and user culture to evolve and converge very rapidly under the pressure of collective feedback about impact barriers. An attempt to retrieve a dead-end paper can be made to trigger an automatic email to the author of that paper indicating that a user has tried to 'cite-visit' it, but that this was not possible because the author has not yet provided a version from which linkable citation data could be derived. This could be accompanied by clear instructions on how to self-archive such a version now; authors could indicate whether they wanted to see such access-failure reports for their deposited work instantly, weekly, monthly, semi-annually, or never; they could of course also request successful 'hit' statistics too, thereby self-monitoring their brainchildren from their earliest embryological stages onward; Figure 4.]

This double inducement – (i) from experience as a user, able to fully cite-navigate some papers but unable to do so with other papers because they had not yet been archived in a form that could be citation linked, and (ii) from experience as an author, learning of unsuccessful attempts to cite-navigate through one's work via citation links – should help to accelerate and focus changes in author practices that will in turn increase the ratio of useable documents even while OpCit is still working directly on extending the reference link extraction tools beyond the immediately linkable subset in the current corpus.

4.6 *Optimizing the deposit procedures and formats and upgrading the citation-navigating capabilities*

The developments in section 4.3 also feed back on 4.1, the author self-archiving procedure, which can be continuously upgraded and optimized in accordance with what proves to be best for the success of section 4.3. The high level of use of the LANL Archive will help to accelerate convergence on author practices and standards that are optimal for the user community. Similarly, developments in section 4.5 will lead to upgrades of the user interface and its capabilities (section 4.2).

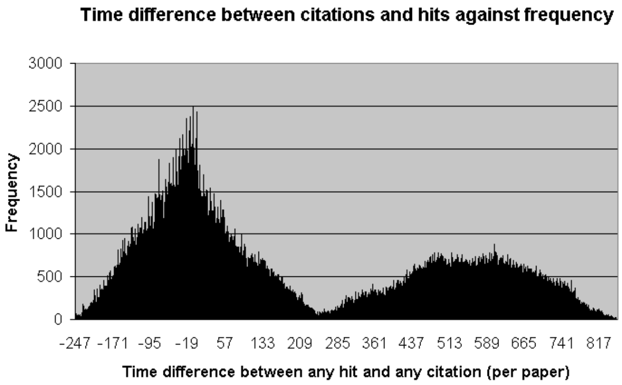

Figure 4. Time-course of hits and citations. The graph shows 2 peaks, one just before zero and one at about 550 days. The peak just before zero is from people reading a paper X before a new paper, Y, is deposited citing paper X. The reader of paper X could be either the author of paper Y or (more likely) another archive user. The peak at around 550 days will be due to people finding a paper, Y, that cites paper X, then, because of that citation, finding paper X in the archive.

4.5 *Informetric analysis of citation and usage: the digital embryology of knowledge*

A citation-linked online literature makes new forms of usage[26] 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 (Figure 4).

For example, author-end citation patterns are being analysed to determine the scope of the LANL Archive. What proportion of citations point to current papers that are in LANL? What proportion of citations point to current papers that are not in LANL? or to papers that predate LANL (Table 1)? to books? to papers in their unrefereed preprint form?, to papers in their final published form (Figure 5)? How do these patterns change as the archive's holdings grow, as its user-base grows, as its years of coverage grow?

Reader-end citation-based navigation patterns are being analysed to determine how open archives are used. This is entirely new informetric territory, because citation searching could only be done off-line until now, so there was no automatic way to analyse how readers actually go about it.

Table 1. LANL citations to pre-LANL ('antique') papers

Year	Papers deposited	Citations	Citations/ paper	Antique	Antique/ paper
91	305	19	0.0623	4296	14.085245900
92	2,891	1,291	0.447	26558	9.186440678
93	6,127	7,576	1.24	55509	9.059735597
94	8,901	19,171	2.15	76847	8.633524323
95	11,034	39,240	3.56	90397	8.192586551
96	13,709	61,019	4.45	101573	7.409220220
97	17,310	100,714	5.82	124926	7.216984402
98	21,040	132,096	6.28	142177	6.757461977
99	24,163	142,888	5.91	135084	5.590530977
00	10,460	93,674	8.96	79578	7.607839388

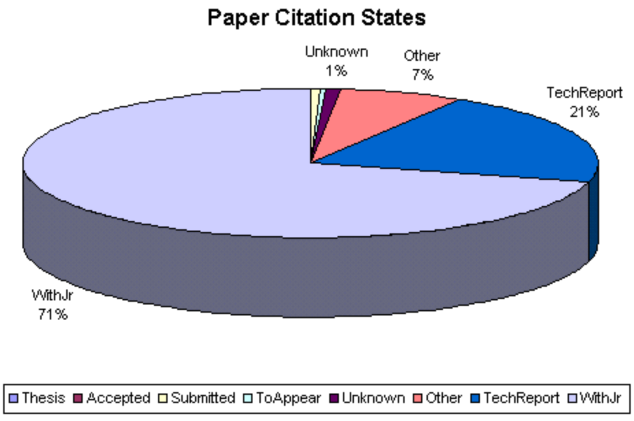

Figure 5. Citations to papers with and without journal reference (Hep-th 1998–2000).

Such data will be used to provide feedback for optimizing the features of the Eprints interface (4.2), to monitor and document open archiving and citation navigating practices, and to chart the course of both knowledge creation and use along the entire scholarly skywriting continuum (Figure 6).

5.0 Some metadata considerations

Although LANL is an Archive of unprecedented scope, covering a substantial specific technical literature, OpCit's primary objective is not to create an ultimate hypertext software resource, but rather to develop a family of generic tools based on current proposals in the metadata area[18,27] which can be applied to LANL as well as to other open archives – research, academic or commercial – that can benefit from this immediate functionality. The effect should be (i) to enhance substantially the power, scope and utility of the LANL Archive itself, (ii) to help solve online open archiving's prima facie problems of scale, compatibility and universality, (iii) to demonstrate, shape and focus user practices, and (iv) to hasten the growth and development of this unique and powerful new way of accessing, using and analysing the use of the scholarly/scientific literature.

The general applicability of these techniques to interoperable digital library architectures[28,29] is also being investigated. The Santa Fe Convention is establishing a set of standards for low-level interoperability, i.e. a means of communicating meta-data and meta-information not only between the existing mirror servers within the current network of online archives, but also between that network and other resources.

In particular, the problem of citations that are associated once and for all with destination URLs must be addressed. For practical flexibility, the recognition and analysis of citation information must be separate from document format convention or locus (e.g. centralized discipline-specific open archives like LANL or CogPrints, distributed university-based open archives like Eprints, and primary/secondary publisher archives or aggregation agents to which the citations may eventually be linked, perhaps with the help of the emerging scholarly publishing standard, SLinkS[30]). It is also currently impossible to convert mathematical markup to HTML; MathML (a realization of XML) is on the horizon and promises to improve this situation.

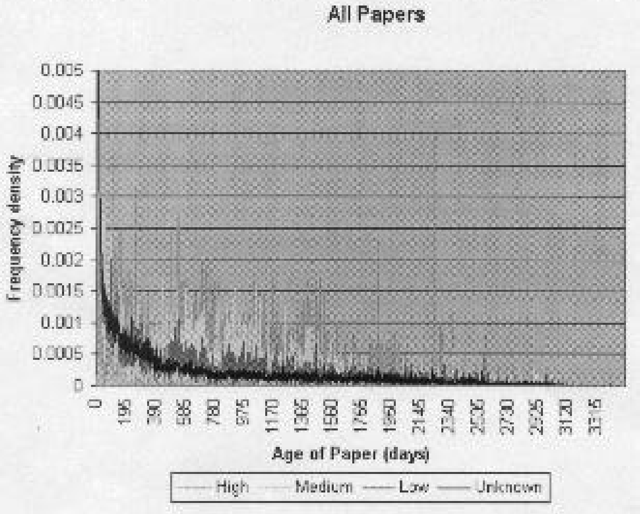

Figure 6. Download frequency vs citation frequency across time: After the initial download peak, the (eventually) more highly cited papers show higher and more sustained continuing download frequencies.

6.0 Citation tools developed by others

CiteSeer[16] is a prototype of an 'automatic citation indexing' system which is used to build a database of citations. Similar in function to the automatic citation linking of Open Journals (OJs) and OpCit, it has concentrated on linking in an unconstrained WWW environment, using a WWW crawler to gather PostScript (and latterly PDF) files from the public Web pages of research institutions. In contrast, the OJs' focus was on specific digital library resources, publishers' archives and existing citation databases which provide quasi-total coverage for chosen subject areas. OpCit's focus is on centralized discipline-based open archives like LANL and CogPrints (see also Van de Sompel and Hochstenbach[31] and http://ups.cs.odu.edu/) as well as distributed pandisciplinary institution-based Eprints open archives.

Some commercial publishers have now started to provide citation links between articles in their proprietary journal databases and online bibliographic services. Partners in the OJ project, BioMedNet[14] and ISI[15], were among the first to implement such links, along with The Institute of Physics, which developed its HyperCite service (IOP Publishing 1996). Centralized discipline-based open archives like LANL and CogPrints, however, do not have any of the financial firewall and access-barrier problems that arise between proprietary databases; and in physics LANL has much more comprehensive, self-contained coverage of the current literature. This will also be true of the distributed open archives created through the institutional author self-archiving initiative using Eprints[24].

7.0 Input to OpCit project from the OJ project

The aim of the OJ project was to interconnect and integrate a body of literature by automatically adding hypertext link overlays to a collection of existing documents served on the WWW[32,33]. This would allow navigating from paper to paper via citation links within or between archives, between documents with different text and reference styles and formats. The OJ project demonstrated this capability for archives consisting of both PDF and

HTML documents, representing display-based and document-based file formats[34].

The Distributed Link Service (DLS), which applied these links, was a WWW implementation of the hypertext techniques that had previously been demonstrated in Southampton University's microcosm research environment[32,35,36]. It made use of a WWW proxy environment to add links to HTML or PDF documents while they were delivered from a digital library (in plain, unlinked form) to a user's browser (with links integrated into them). The DLS software used various modules, called 'agents' (because they have an 'expertise' at automatically recognizing particular kinds of information in the document)[37].

7.1 *Keyword agent*

The keyword agent is very simple and uses various databases of stand-alone links (which can be keywords or other text strings), attaching them to the papers whenever those strings appear.

7.2 *Name agent*

The name agent looks for different appearances of a name [e.g. 'Eugene Garfield', 'Garfield, E.' or 'Garfield *et al.*'], possibly in a specified context.

7.3 *Citation agent*

The citation-agent recognizes occurrences of citations in academic papers in a large (and extendable) variety of formats and analyses their contents to determine author, year, publisher, page range and the like. It uses this information from each citation to perform a lookup in a bibliographic database and to add a link to either the online full text of the cited article (if the database shows that it is available) or to the bibliographic record consisting of abstract and citations (from the ISI database).

8.0 Application and further development of OJ software

The OJ project yielded a set of generalizable tools that were immediately applied to the LANL Preprint Archive. The following tools work on PDF or HTML documents:

Link harvester: A stand-alone program that extracts pre-existing links from a document and adds them into a database. A new link-free version of the document is also generated.

Link interpolater: A stand-alone program that inserts links from a database into a document. If different sets of databases are selected, the same document can be linked into different navigation strategies (e.g. citation, keyword, overlaid subject index).

Citation harvester: A stand-alone program that extracts citations from a paper's reference lists for storage in a database.

Citation interpolater: A stand-alone program that inserts links into a document based on the contents of a citation database. Links can be added to other documents in the same archive, to documents in other archives, or to generic bibliographic citation databases (such as ISI's Web of Science).

To view a fragment of the first page of an ACM DL library article with both keyword and person links added wherever interesting people and systems are mentioned, see http://www.staff.ecs.soton.ac.uk/~lac/somewords/image4.gif.

To view another fragment from an ACM DL article that has been automatically populated with links to the ACM library from any citation of CACM or an ACM Hypertext Conference, see: http://www.staff.ecs.soton.ac.uk/~lac/somewords/image5.gif

9.0 Further optimizing the optimal

9.1 *Other kinds of links*

Papers can be automatically provided with other kinds of links using distributed link overlays as demonstrated in the microcosm and DLS technologies[38,39]. http://www.mmrg.ecs.soton.ac.uk/publications.html. These overlays can include links based on keywords, author names (pointing to papers other than the explicitly cited ones), glossaries/indices, and even an inverted index for the corpus as a whole. Such services can be applied to the open archives data-bases (http://www.openarchives.org/sfc/data_provider_template.htm) by open archive service-providers (http://www.openarchives.org/sfc/service_provider_template.htm).

9.2 *Revision/update linking*

There is no reason a research report should remain in a 'frozen' state after it is published. The published version, suitably tagged, is a permanent formal milestone, especially for citation purposes, but an interlinked Archive also allows authors to deposit updated and revised drafts. The automatic linking system can be adapted to accommodate this, providing automatic forward and backward linking between versions.

9.3 *Commentary links*

For the same reason that links from unpublished preprints to refereed reprints to revised drafts of papers are of value, so are links to comments on papers, and authors' replies to comments, on the model of Behavioral and Brain Sciences (BBS) (http://www.princeton.edu/~harnad/bbs.html) and Psycoloquy (http://www.princeton.edu/~harnad/psyc.html)[40–42].

9.4 *Journal links*

There are several ways in which citation-linked open archives like LANL, CogPrints, and Eprints can be useful to the journals in which its papers are published. They can provide links to the version of a paper in the *journal's own official online archive*. They can also provide links to cited papers in the journal's online archive that do not appear in open archives. Authors might wish to have arrangements for official links with the published version in order to provide an authenticated draft, or one in which the paper page images can be viewed or cited by page and line.

9.5 *Peer review*

Another useful service that open archives can provide to the journals in which their papers are published LANL is already beginning to provide the following: Authors submitting papers to the *American Physical Society Journals* (*APS*) can already do so by simply specifying the LANL version as their official submission. Referees can then be directed to that citation-enhanced draft in reviewing it. A password-controlled, non-public sector could also be created in LANL that would allow referee reports to be linked just as commentaries are in section 9.4 above, but under the control of each journal. This would effectively be the implementation of online peer review[42–44] for journals, and might be a model for the future relationship between refereed journals and open archives. Journals could also upload their final drafts to an open archive for distribution in their own formats with journal-specific identifying graphics, etc. The official journal version would then be part of the paper's overall revision 'history', which could continue with comments, responses and updates[21,22].

(Nor is there anything to prevent journals from using interoperable open-archiving software such as Eprints as 'closed' open archives, in that their metadata are open but their full texts are behind a financial firewall.)

9.6 *Links to proprietary databases*

Citation links leading out of open archives could also go to proprietary databases that charge for their services. These could include journal home archives, archives of scanned contents of back issues of journals, electronic books, and secondary publisher databases, such as INSPEC, MEDLINE or ISI. There are, however, strategic questions about whether OpCit should implement links that entail charges to the user[45–49].

9.7 *Links to other public archives*

Provisions could be made for citation links to papers in public archives other than open archives, but it may be more useful to make other public archives Santa Fe-compliant (as they are not competing in any sense, and only stand to benefit from interoperability, economies of scale, shared resources and development), perhaps through interfaces such as NCSTRL, into one seamless interconnected archive; this too would provide constraints to help guide convergence into a unified, distributed, global archive[13].

9.8 *Links to authors' home server archives*

Apart from mirroring, one useful form of redundancy that discipline-based open archives might encourage is that all their authors should also archive their papers on their home institutional servers, to which the central archives would also be linked. This is why the Eprints software has been developed. Links to the author's email address and URL are also standard components of the central version[44].

10.0 Overview

LANL is a public repository for a substantial and growing proportion of the current research literature in physics. It is becoming the primary way that the world physics community accesses its literature in a growing number of subfields. Citation-linking makes this resource still more powerful and useful for its current physicist users; this can be extended to all the rest of the disciplines in other online archives designed to be interoperable by complying with the Santa Fe Convention of the open archive Initiative.

The WWW is predicated on hypertext connections between documents, but for the scientific/scholarly world the scholarly link par excellence is formal citation of one paper by another. This is the way researchers have naturally been interconnecting their writings all along, but until now it has only been possible to follow those connections off-line, piece-wise, mediated by a great deal of real footwork in between. Now the entire corpus can be navigated via citations on-line. Commercial journal publishers, along with secondary indexing/abstracting services, are exploring ways of interconnecting the on-line journal literature, but those initiatives are intrinsically and severely limited by financial firewalls[45–51] that prevent free navigation across full texts and their citations until and unless the access fees for each 'hit' are first paid through subscription, site-license or pay-per-view (S/L/P). (To allow the full texts to be browsed for free would be equivalent to giving away the literature for free in the online medium.)

Open archives do not have this constraint; citation linking within the physics archive, some of whose subfields are virtually complete, yields seamless public access worldwide to the entire corpus. OpCit has the citation linking tools and has applied them to completely intralink the physics archive. A citation-linked online digital corpus also makes possible powerful new forms of online informetric analysis that go far beyond static citation analysis, measuring researchers' usage of all phases of the literature, from pre-refereeing preprint to post-refereeing postprint, from download to citation, yielding an embryology of learned inquiry.

The huge, international usership of LANL, extended still further by Santa Fe compliance, guarantees that the proposed enhancements will not only be widely tested, but that, if successful, they will strongly influence the evolution of open archiving of the rest of the refereed literature in all disciplines. There is no question that radical changes in scholarly/scientific publishing and communication practices are poised to take place (with teaching and learning practices ready to follow suit[52]). Citation linking will help to guide and hasten them in the right direction.

1. Garfield, E., *Science*, 1955, **122**, 108–111. http://www.garfield.library.upenn.edu/papers/science_v122(3159)p108y1955.html
2. Harnad, S., *Public-Access Comput. Systems Rev.*, 1991, **2**, 39–53. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad91.postgutenberg.html
3. Harnad, S., *Nature* [online] (c. 5 Nov 1998) 1998c. http://helix. nature.com/webmatters/invisible/invisible.html
4. Harnad, S., in *Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing* (eds Ann Okerson and James O'Donnell), Washington, DC, Association of Research Libraries, June, 1995. http://www.arl.org/scomm/subversive/toc.html
5. Campbell, R. D., A Universal Citation Database as a Catalyst for Reform in Scholarly Communication. Firstmonday, 27 April, 1997, 2. http://firstmonday.org/issues/issue2_4/cameron/index. html
6. Ginsparg, P., *Comput. Phys.*, 1994, **8**, 390–396. http://xxx.lanl.gov/blurb/
7. Ginsparg, P., Invited contribution, UNESCO Conference HQ, Paris, 19–23 February 1996. http://xxx.lanl.gov/blurb/pg96unesco.html
8. Davis, J. R. and Lagoze, C., *JASIS* (to appear). http://www2.cs.cornell.edu/lagoze/papers/NCSTRL–IEEE3.doc
9. Carr, L. and Hitchcock, S., The Open Journal Project, 1995. http://journals.ecs.soton.ac.uk
10. Evans *et al.*, 1998.
11. Hitchcock, S., Carr, L. and Hall, W., *Web Journals Publishing: A UK Perspective*, Serials, ISBN 0953-0460, 1997a, vol. 10, no. 3, pp. 285–299. http://www.mmrg.ecs.soton.ac.uk/publications/archive/hitchcock1997/
12. Hitchcock, S., Carr, L., Harris, S., Hey, J. and Hall, W., Proceedings of Second ACM Conference on Digital Libraries, Philadelphia, 1997b, pp. 115–122. http://www.mmrg.ecs.soton.ac.uk/publications/archive/hitchcock1997b/
13. Hitchcock, S., Quek, F., Carr, L., Hall, W., Witbrock, A. and Tarr, I., ICCC/IFIP Conference on Electronic Publishing 97: New Models and Opportunities, Canterbury, UK, April, 1997c. http://journals.ecs.soton.ac.uk/IFIP-ICCC97.html
14. Hitchcock, S., Carr, L., Harris, S., Probets, S., Evans, D., Hall, W. and Brailsford, D., *D-Lib Mag.*, December, 1998a.
15. Hitchcock, S., Quek, F., Carr, L., Hall, W., Witbrock, A. and Tarr, I., *Ser. Rev.*, 1998b, **24**, 21–33.
16. Giles, C. L., Bollacker, K. and Lawrence, S., Paper in the Third ACM Conference on Digital Libraries, ACM Press, 1998, pp. 89–98. http://www.neci.nj.nec.com/homepages/lawrence/citeseer.html
17. Bollacker, K. D., Lawrence, S. and Giles, C. L., CiteSeer: An Autonous Web Agent for Automatic Retrieval and Identification of Interesting Publications. Agents, 1998, 116–123. http://www.neci.nj.nec.com/homepages/lawrence/citeseer.html
18. Van de Sompel, H. and Lagoze, C., *D-Lib Mag*., 2000, **6**. http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html
19. Chen, C. and Carr, L., in Proceedings of the Tenth ACM Conference on Hypertext, Darmstadt, February, 1999.
20. Youngen, G., UIUC Physics and Astronomy library [online] (c. 5 November 1998). http://www.physics.uiuc.edu/Physics/library/preprint.html
21. Harnad, S., *Psychol. Sci.*, 1990, **1**, 342–343 (reprinted in *Curr. Contents*, 1991, **45**, 9–13). http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.skywriting.html
22. Harnad, S., Interactive Publication: Extending the American Physical Society's Discipline-Specific Model for Electronic Publishing, Serials Review, Special Issue on Economics Models for Electronic Publishing, 1992, pp. 58–61. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad92.interactivpub.html
23. Halpern, J. Y. and Lagoze, C., Paper submitted to Digital Libraries, the Fourth ACM Conference on Digital Libraries, Berkeley, CA, 1999. http://www2.cs.cornell.edu/lagoze/papers/dl99.pdf
24. Harnad, S., *D-Lib Magazine*, December, 1999, **5**. http://www.dlib.org/dlib/december99/12harnad.html
25. Hitchcock, S. Carr, L., Jiao, Z., Bergmark, D., Hall, W., Lagoze, C. and Harnad, S., Proceedings of the 5th ACM Conference on Digital Libraries, San Antonio, Texas, June 2000. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.acm.htm
26. Carr, L., Hitchcock, S., Hall, W. and Harnad, S., *ACM SIGDOC J. Comput. Doc.*, May, 2000. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.halpern.htm
27. Lassila, O. and Swick, R. (eds), Resource Description Framework (RDF) Model and Syntax Specification, W3C Proposed Recommendation, January, 1999. http://www.w3.org/TR/PR-rdf-syntax/
28. Lagoze, C. and Payette, S., Cornell University Computer Science, Technical Report TR98-1690, June, 1998. http://ncstrl.cs.cornell.edu/Dienst/UI/1.0/Display/ncstrl.cornell/TR98-1690
29. Leiner, B. M., *D-Lib Mag.*, December, 1998.
30. Hellman, E., Scholarly Link Specification Framework, 1998. http://www.openly.com/SLinkS/
31. Van de Sompel, H. and Hochstenbach, P., *D-Lib Mag.*, 1999, **5**. http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html
32. Carr, L., Davis, H., Hall, W. and Hey, J., in Proceedings of ELVIRA: The UK Digital Libraries Conference, De Montford University, UK, 1996a. http://www.mmrg.ecs.soton.ac.uk/publications/archive/carr1996c/
33. Carr, L. and Hall, W., Presentation at the First International Workshop on the Use of the WWW for the Public Understanding of Science, CERN November, 1998.
34. Probets, S., Brailsford, D. F., Carr, L. and Hall, W., in Proceedings of Seventh International Conference on Electronic Publishing, Document Manipulation and Typography, Springer-Verlag (Lecture Notes in Computer Science Series), April, 1998. http://www.mmrg.ecs.soton.ac.uk/publications/archive/probets1998/
35. Carr, L., De Roure, D., Hall, W. and Hill, G., *WWW J.*, 1995, **1**, 647–656. http://www.mmrg.ecs.soton.ac.uk/publications/archive/carr1995/

36. Carr, L., Hall, W. and Hitchcock, S., in Proceedings of the Ninth ACM Conference on Hypertext, Pittsburgh, June, 1998b. http://www.mmrg.ecs.soton.ac.uk/publications/archive/carr1998a/
37. Carr, L., De Roure, D., Hall, W. and Hill. G., *WWW J.*, 1998a, **1**, http://www.mmrg.ecs.soton.ac.uk/publications/archive/carr1998b/
38. Carr, L., Davis, H., De Roure, D., Hall, W. and Hill, G., *Open Information Services, Computer Networks and ISDN Systems*, Elsevier, 1996b, vol. 28, pp. 1027–1036. http://www.mmrg.ecs.soton.ac.uk/publications/archive/carr1996b/
39. De Roure, D., Carr, L., Hall, W. and Hill, G. J., in Proceedings of the Third International Workshop on Services in Distributed and Networked Environments (SDNE96), Macau, 3–4 June 1996, IEEE Computer Society Press, 1996, pp. 156–161. http://www.mmrg.ecs.soton.ac.uk/publications/archive/deroure1996a/
40. Harnad, S., *Science*, 1979, **19**, 18–20.
41. Harnad, S., *Am. Psychol*., 1984, **39**, 1497–1498.
42. Harnad, S., Learned Inquiry and the Net: The Role of Peer Review, Peer Commentary and Copyright, Learned Publishing, 1998a, **4**, 283–292. http://citd.scar.utoronto.ca/EPub/talks/Harnad_Snider.html
43. Harnad, S. (ed.), *Peer Commentary on Peer Review: A Case Study in Scientific Quality Control*, Cambridge University Press, NY, 1982.
44. Harnad, S., in *Scholarly Publishing: The Electronic Frontier* (eds Peek, R. and Newby, G.), MIT Press, Cambridge MA, 1995, pp. 103–118. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad96.peer.review.html
45. Bachrach, S., Berry, S. R., Blume, M., von Foerster, T., Fowler, A., Ginsparg, P., Heller, S., Kestner, N., Odlyzko, A., Okerson, A., Wigington, R. and Moffat, A., *Science*, 1998, **281**, 1459–1460. http://www.sciencemag.org/cgi/content/full/281/5382/1459
46. Harnad, S., *Nature*, 1998b, **395**, 127–128. http://www.cogsci.soton.ac.uk/~harnad/nature.html
47. Harnad, S., For Whom the Gate Tolls? Free the Online-Only Refereed Literature, American Scientist Forum, September, 1998d. http://www.cogsci.soton.ac.uk/~harnad/amlet.html
48. Harnad, S., E-Knowledge: Freeing the Refereed Journal Corpus Online. Computer Law & Security Report, 2000, **16**, 78–87. [Rebuttal to Bloom Editorial in *Science* and Relman Editorial in *New England Journal of Medicine.*] http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.scinejm.htm
49. Harnad, S., Varian, H. and Parks, R., *Culture Mach. 2 (Online Journal),* 2000. http://www.cogsci.soton.ac.uk/~harnad/Temp/Varian/new1.htm
50. Odlyzko, A. M., *Int. J. Human-Comput. Stud*., 1995, **42**, 71–122. http://www.research.att.com/~amo/doc/tragic.loss.txt
51. Odlyzko, A. M., in *Technology and Scholarly Communication* (eds Ekman, R. and Quandt, R.), Univ. Calif. Press, 1998. http://www.research.att.com/~amo/doc/economics.journals.txt
52. Light, P., Light, V., Nesbitt, E. and Harnad, S., in *Rethinking Collaborative Learning* (ed. Joiner, R.), Routledge, London, 2000 (in press). http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.skyteaching.html

Further reading

1. Davis, H., Hall, W., Heath, I., Hill, G. and Wilkins, R., in Proceedings of the ACM Conference on Hypertext (ECHT'92), Milan, November, ACM Press, 1992, pp. 181–190. http://www.mmrg.ecs.soton.ac.uk/publications/archive/davis1992/
2. Hall, W., Davis, H. C. and Hutchings, G. A., *Rethinking Hypermedia: the Microcosm Approach*, Kluwer Academic Press, Boston USA, 1996, p. 195.
3. Okerson A. and O'Donnell, J. (eds), *Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing*, Washington, DC, Association of Research Libraries, June, 1995. http://www.arl.org/scomm/subversive/toc.html
4. Pope, S. and Miller, L., *Conserv. Ecol.* [online], (c. 5 November 1998). http://www.consecol.org/Journal/consortium.html

ACKNOWLEDGEMENTS. Many thanks to Ian Hickman http://fluffy.ecs.soton.ac.uk/ and Tim Brody for the ongoing citation analyses of which the figures in this paper are a sample.