Usage analysis reveals CoRR's role in publishing

A Usage Based Analysis of CoRR

Les Carr, Steve Hitchcock, Wendy Hall and Stevan Harnad
The Open Citation Project, IAM Group
Department of Electronics and Computer Science
University of Southampton
Highfield, Southampton SO17 1BJ, UK
lac@ecs.soton.ac.uk

Abstract

Based on an empirical analysis of author usage of CoRR, and of its predecessor in the Los Alamos eprint archives, it is shown that CoRR has not yet been able to match the early growth of the Los Alamos physics archives. Some of the reasons are implicit in Halpern's paper, and we explore them further here. In particular, we refer to the need to promote CoRR more effectively for its intended community—computer scientists in universities, industrial research labs and in government. We take up some points of detail on this new world of open archiving concerning central versus distributed self-archiving, publication, the restructuring of the journal publishers' niche, peer review and copyright.

H.4.3 *Communications Applications—*

Keywords*: Computing Research Repository, CoRR, eprint archives, open archives*

CoRR leads the way to open archives

The architects of the Computing Research Repository (CoRR) were remarkably prescient but have not yet been rewarded in terms of the growth of the archive and the number of archived papers.

CoRR's "hybrid" architecture described by Halpern, based on the repository software used by the Los Alamos National Laboratory (LANL) eprint archives but with access to distributed archives using the protocol and common interface developed for NCSTRL, has since been largely adopted by the Open Archives initiative (OAi). News of this must have reached Halpern as he completed his paper as it elicits a mention at the very end. Like CoRR, the initiative envisages a "global multi-disciplinary research collection," and the vital component in this vision is the open interface that allows third-party services to enhance access to the contents of these distributed archives, for example, the reference linking service for archives that we are developing in the Open Citation project (Harnad et al., 1999).

Submission rates to CoRR

Not surprisingly, this open architecture has not yet brought many new computer science authors to CoRR, judging by the rate of submission of papers. Nor, it appears, has it brought publishers other than the ACM, its main initial sponsor, to publicly support and work with the archive. This contrasts with PubMed Central (http://pubmedcentral.nih.gov/), the new eprint archive from the National Institutes of Health serving fields in biomedicine and life sciences, which is predicated on the support of publishers—yet to be realized on a large scale—as it only allows materials to be posted by cooperating publishers, not by authors.

Formally, CoRR was launched in September, 1998, but folded into the archive are papers posted to the LANL computation and language (cmp-lg) archive that began in 1994, and papers

ACM Journal of Computer Documentation, 2000,24:54-59
ISSN 1527-6805

from the electronic *Journal of AI Research* (JAIR) which are archived according to date of publication (not date of archiving). Hence CoRR appears to have a history of posting prior to its launch.

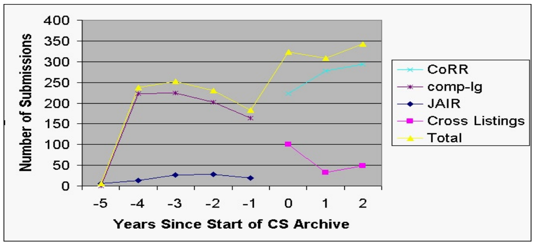

Figure 1. Annualized rate of submissions since 1993 to the LANL computer science archive, which became CoRR in 1998. Cross-listings are papers that have their main listing in other non-CoRR LANL archives.

Submission rates to CoRR from 1993, including the prior archives, can be seen in Figures 1 and 2. Figure 1 shows submissions on an annualized basis (with the data for this year loosely extrapolated from data for the first two calendar months only). Figure 2 shows the monthly data to date. The lines for CoRR post-September 1998 subsume the total postings to cmp-lg and JAIR.

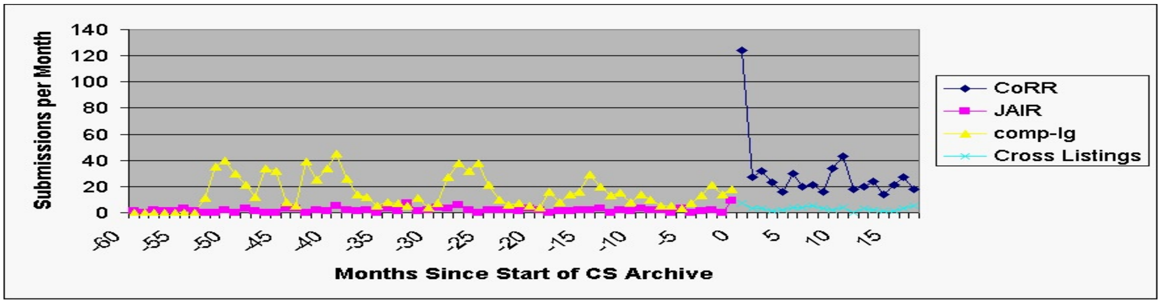

Figure 2. The same data as Figure 1 shown on a monthly basis to February 2000 (the total is omitted for clarity).

It should be noted that submission to the Los Alamos archives requires authors to specify a principal archive for the paper, but allows papers to be cross-listed between archives, again as the author directs. Our graphs show these different categories of papers, i.e. those native to CoRR, and those which presumably are primarily physics or math papers but which are also indexed within CoRR. While there were very few cross-listed papers pre-CoRR (less than ten), the wider scope of CoRR has enabled a more significant number of papers to be cross-listed since.

So what is the effect of CoRR? There was a noticeably high level of activity immediately before and after the launch of CoRR, a one-off blip shown in Figure 2. The total number of papers accessible on CoRR is just over 1600. These data, taken directly from the archive, do not quite match Halpern's quoted figures which indicated about 1800 papers in the archive at the time of his writing. Approximately 1000 papers in total were submitted between 1994 and the start of CoRR and, so far, about 600 (both native and cross-listed) since. Due to the way CoRR was filled out just prior to the official launch, we can't be precise about this breakdown.

We can also look at the data from the perspective of submitting authors. In this case we have derived our results from analysis of a copy of the Los Alamos archive that we hold locally for the purposes of the Open Citation project. These figures are only until October. Until that date we find, from a list of corresponding authors (not all named authors) based on e-mail addresses, that nearly 800 unique authors have submitted to the archive. One author has posted 103 papers, another has posted 55 papers. Of all authors, 220 have only posted to the archive since CoRR was launched. Given the shape of the curves it could be assumed that many of these "new" authors were attracted to post in the archive by the establishment of CoRR, so this is some measure of the impact of CoRR in its main constituency.

The CogPrints archive (http://cogprints.soton.ac.uk/) for cognitive sciences has similar numbers so far (900 submitted papers, with no previous LANL archive) but such figures are rather eclipsed by LANL's 128,000 papers. In economics, RePec (http://netec.mcc.ac.uk/RePEc/) is rather more advanced with 71,000 papers, but this is a distributed archive, with papers actively harvested rather than depending on central deposits (Van de Sompel et al., 2000). The Open Archives initiative may well be what puts all the self-archiving approaches, central and distributed, over the top by integrating them all through the glue of interoperability (Van de Sompel and Lagoze, 2000).

In computer science the totals for CoRR are small compared with over 27,000 papers quoted to be on the institutional servers that comprise NCSTRL. For another reference point, the Collection of Computer Science Bibliographies (http://liinwww.ira.uka.de/bibliography/) claims to contain "more than one million references (mostly to journal articles, conference papers and technical reports)," of which more than 90,000 link to a full-text version of the paper.

Of course, CoRR is far behind the number of papers now held across all the LANL archives, but a fairer comparison is with the early years of the LANL from 1991. At an equivalent stage, 18 months after launch, LANL was attracting about 400 papers/month and held 3700 papers in total. Nearly half of these papers were being submitted to the original, and largest, of the LANL archives hep-th (high-energy physics-theoretical). It was eight months or so before LANL began to develop other archives of significant size (Figure 3). The number of submissions per month in hep-th reached a plateau of 200-300 papers/month from about month 40 (Figure 4). Growth in the rate of submissions to the archives since then can be shown to be effectively linear, much of this growth coming from the addition of new archives. The implications of this linearity for the entire physics journal literature were commented on by Smith (1999a) of the American Physical Society:

> LANL has grown at a pretty much linear rate, handling probably 25,000 new submissions this year, 20,000 last year, 15,000 the year before, etc. Projecting this linear growth forward it will take another 10 or so years to capture all articles published in pure physics (currently something like 1/3 of papers we receive also appear on the archive), roughly 50 years to capture both pure and applied physics, and at least 200 years to capture most of scientific publishing.

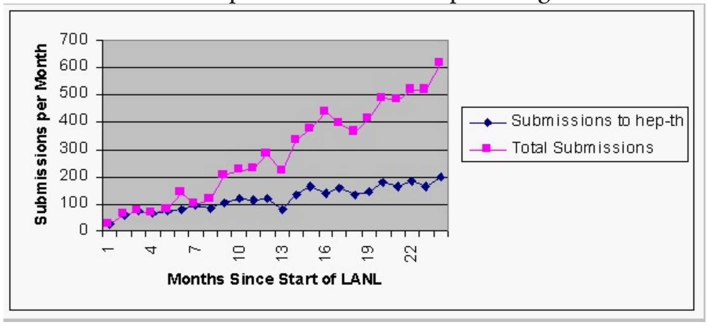

Figure 3. Submissions/month to LANL archives, also showing submission rate to the archive for high-energy physics-theoretical (hep-th).

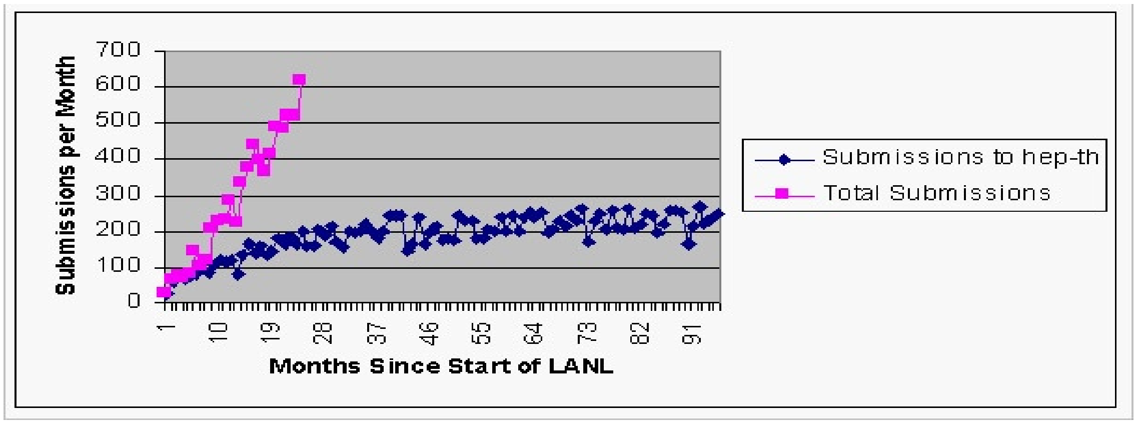

Figure 4. Submissions/month to LANL archives across an extended timescale compared with Figure 3, showing the plateau in rate of submissions to the hep-th archive.

Care has to be exercised in comparing LANL and CoRR directly. Internet services were different then, and certain sectors in physics always had a culture that is more amenable to circulating preprints (Odlyzko, 1997). The equivalent prepublication form in computer science is conferences and technical reports.

> there is an underlying lack of confidence that publishers ...will in general respond positively to the prospect of freely accessible archives...

Building a community of authors

Why does CoRR not yet appear to have caught the imagination of its key constituency? Halpern suggests that source formats, fears of plagiarism, and LANL's unfriendly user interface for submissions are among the possible restraining factors. For users relief is on the way. The CogPrints archiving software (modeled on LANL) is being redesigned to make it compliant with the OAi's Santa Fe convention (Van de Sompel and Lagoze, 2000) as well as generic, so that it can be used by all universities (free) to set up theirs own open archives in all of their own disciplines—and every effort has been made to make it as simple and useful as possible, for both author and reader (Harnad, 1999).

As Halpern recognizes, the reluctance of authors to commit to CoRR probably has more to do with the nature of the formal publishing process. CoRR does not require any rights in a submitted work but recognizes that publishers may subsequently acquire all rights in refereed versions and can ask for the preprint to be withdrawn from freely accessible archives. This is a substantive issue. There has been no copyright challenge to LANL in its decade of existence, and a major publisher in physics, the American Physical Society (APS), officially allows self-archiving of both unrefereed preprints and refereed reprints. Hopefully all other journal publishers will soon follow suit.

In this respect, CoRR does not have a sufficiently robust stance in its relationship with journals or journal publishers. As Halpern notes, many of the major publishers in computer science allow preprints to be retained in eprint archives, and some learned society publishers among them allow the final refereed version to be posted too. Archives could indicate for the benefit of authors which publishers have so far agreed to work with them and under what conditions. Smaller electronic journals with advanced policies in particular might benefit from the recognition that this would bring. At this stage there would be significant, mutually realizable cross-promotional benefits between CoRR and the journals. CoRR does not need to favor any particular journals, simply list and link those that conform to its policies. This is one way of recognizing CoRR's contributory, not subsidiary role, in disseminating the refereed journal literature.

The launch of PubMed Central has been the catalyst for two new services providing free access to preprint and refereed papers by two well known publishers in the biomedical field:

- NetPrints in clinical medicine and health research (http://clinmed.netprints.org/) from the BMJ and HighWire Press.
- BioMed Central (http://www.biomedcentral.com/start.asp) from Current Science.

This may make a vital connection for authors, but there are other signs in Halpern's paper that CoRR may not yet have fully understood its authors. Kling and McKim (2000) argue that fields will continue to exhibit different communications practices during the shift to electronic communications and that we are unlikely to converge on a stable set of electronic forums. In which case, it cannot be assumed that the success of the Los Alamos physics archives will translate to computer science unless the approach is modified in some way.

The journal publishing culture within a discipline offers some pointers on how its electronic communications practices might develop. Despite the assertions of publisher support referred to above, there is an underlying lack of confidence that publishers of computer science journals will in general respond positively to the prospect of freely accessible archives, especially if these are to contain refereed versions of papers. This is evident from Halpern's inference that where there are many journals in a field, as in computer science, those journals can exert less influence over archives than in fields with fewer dominant journals, such as biomedicine. In other words, Halpern is hoping that CoRR will not depend on the support of journal publishers and editors, but fails to take account of the likely preferences of authors implied by their continued use and support for the journals.

A field with many journals is fragmented. In the case of computer science, continuing fragmentation of the field in terms of its journals suggests, until now at least, that its authors did not expect, or did not desire, wide recognition for the public manifestation of their work. As in other scholarly fields there is a demonstrated preference for publication—for scholarly works this is legitimized by inclusion in a collected work moderated by one's immediate peers—but in computer science the lesson of fragmentation is that this has been at the expense of communication among the widest possible audience. While the journals' control of these limited channels may be tenuous, bypassing the journals may not deliver the authors to CoRR, reinforcing the need for the archive to indicate its links with journals and publishers

An important potential misunderstanding has to be cleared up here: LANL is not a "fast publication"; it is not a publication at all, in academic terms, any more than a paper preprint or an oral conference report are publications. LANL contains both unrefereed preprints and refereed reprints. The latter are indeed publications, but only because they have been refereed and accepted for publication by a peer-reviewed journal.

Computer science is characterized as a dynamically changing field and it is to this that the continuing fragmentation of the field is frequently attributed. CoRR's interpretation of this fragmentation is shown in its use of the ACM classification system. Halpern acknowledges that this "does not map well to current major areas of academic computer science."

Nor does it help that the ACM is half-hearted in its endorsement of CoRR: a "two year experiment" (due to end this year) during which the ACM "is examining the impact that this will have on journal sales." It is reasonable to review the progress of a large initiative such as CoRR, but scholarly archiving projects require a long-term commitment without which there is a fundamental, possibly fatal, weakness. For the ACM's publishing activities, as for CoRR, the outcome of the review could be critical. Markets move very quickly on the Web and it would be risky for the ACM to revert to exclusive priority for its journals. Instead it should seek to continue asserting an influence on CoRR. Like Halpern, we hope for a positive decision from the ACM and a stronger statement of support for CoRR later this year.

Universal access for all and enhanced services

The most powerful argument for CoRR, for its authors and users, and ultimately for publishers, especially those that recognize this early, is the ability of free-to-post, free-to-view archives to transform access to the scholarly journal literature. Universal access will be the platform for enhanced services, many of which could be commercial services, briefly described by Halpern: "Publishers could provide value added to authors and readers such as summarization services, advanced searching tools, awareness, and filtering services that build on the content in CoRR-like repositories." It would also be possible to "experiment with other ways of filtering papers besides traditional peer review," one approach being the addition of comment facilities, Halpern argues. Currently, the primary example of an archive-based service is the "overlay" journal, such as implemented by the APS (Smith, 1999b).

For some the essential service is still quality-control/certification (and that is indeed a rather dynamic form of filtering) implemented by peer review. New add-on services will have to prove themselves in their own right. Will there be any demand for proprietary classifiers once the full, tagged texts are all archived, interoperable, and free? Will there be any demand for secondary publishers' services and products (abstracts, indices) or will autonomous citation indexing services such as NECI's ResearchIndex (http://www.researchindex.com/) or generic search engines do? Citation-linking is already on the way for free (http://opcit.eprints.org/). Should publishers be redirecting their efforts, or should they consolidate their traditional niche (quality-control/certification)? Just as self-archiving is a supplement rather than a substitute for publication, so peer commentary is a supplement rather than a substitute for peer review (Harnad, 1998). And before risking the quality of the research literature, hypothetical alternatives had best be rigorously tested.

Conclusion

The scope for CoRR is enormous, both within the NCSTRL network and within the Open Archives framework. The policy and design decisions on which CoRR is based are correct, and it is well positioned to serve the growing demand for free, universal access to the scholarly journal literature. To achieve its aims it needs more effective promotion, stronger support from its principal sponsors and a clearer relationship with refereed journals and publishers.

Acknowledgments

The Open Citation Project is funded by the Joint NSF-JISC International Digital Libraries Research Program.

References

Harnad, S. (1999). Free at Last: The Future of Peer-Reviewed Journals. *D-Lib Magazine*, 5(12), December 1999. http://www.dlib.org/dlib/december99/12harnad.html.

Harnad, S. (1998). Learned Inquiry and the Net: The Role of Peer Review, Peer Commentary and Copyright. *Learned Publishing*, 4(11), 283-292. http://citd.scar.utoronto.ca/EPub/talks/Harnad_Snider.html.

Harnad, S. et al. (1999). Integrating and Navigating Eprint Archives Through Citation-Linking. Proposal to NSF-JISC International Digital Libraries Research Programme. http://www.cogsci.soton.ac.uk/~harnad/citation.html.

Kling, R. and McKim, G. (2000). Not Just a Matter of Time: Field Differences in the Shaping of Electronic Media in Supporting Scientific Communication. *Journal of the American Society for Information Science*, to be published. http://xxx.lanl.gov/ftp/cs/papers/9909/9909008.pdf.

Odlyzko, A. M. (1997). The Slow Evolution of Electronic Publishing. In A. J. Meadows and F. Rowland (Eds), *Electronic Publishing '97: New Models and Opportunities* (pp. 4-18). Washington, D.C.: ICCC Press. http://www.research.att.com/~amo/doc/slow.evolution.txt.

Smith, A. (1999a). Re: The Forgotten Importance of Editors. *September98-Forum archives*, 6 August 1999. http://listserver.sigmaxi.org/scripts/wa.exe?A2=ind99&L= september98-forum&O=A&F=l&P=22568.

Smith, A. (1999b) The Journal as an Overlay on Preprint Databases. *Learned Publishing*, 13(1), 43-48. http://ridge.aps.org/APSMITH/ALPSP/talk1.html.

Van de Sompel, H., Krichel, T., Nelson, M.L., Hochstenbach, P., Lyapunov, V.M., Maly, K., Zubair, M., Kholief, M., Liu, X., O'Connell, H. (2000). The UPS Prototype: An Experimental End-User Service across E-Print Archives. *D-Lib Magazine*, 6(2), February 2000. http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html.

Van de Sompel, H. and Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine,* 6(2), February 2000. http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html.