Evidence of Hypertext in the Scholarly Archive

Tim Brody
IAM Research Group
University of Southampton
United Kingdom
+44 2380 594059
tb01r@ecs.soton.ac.uk

Leslie Carr
IAM Research Group
University of Southampton
United Kingdom
+44 2380 594479
lac@ecs.soton.ac.uk

Stevan Harnad
IAM Research Group
University of Southampton
United Kingdom
+44 2380 592582
harnad@ecs.soton.ac.uk

ABSTRACT
This paper attempts to substantiate recent observations about the development of hypertext rhetoric in scholarly archives by reporting the results of some simple quantitative studies of the use by researchers of a major scholarly archive.

Keywords
Textuality, web, hypertext rhetoric, scholarly and scientific communication, navigation.

1. INTRODUCTION
Dalgaard's recent article [3] argues that the part of the Web that constitutes the scientific literature is composed of increasingly linked archives. He describes the move in the online communications of the scientific community towards an expanding zone of second-order textuality, of an evolving network of texts commenting on, citing, classifying, abstracting, listing and revising other texts. In this respect, archives are becoming a network of texts rather than simply a classified collection of texts. He emphasizes the definition of hypertext as multi-linear text, in contrast to the simple definition of a hypertext as 'a document with links in'.

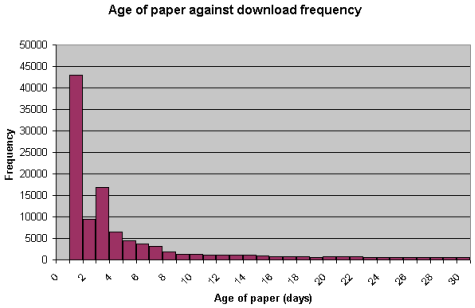
The HEP archive (`www.archiv.org`) is one of the pre-eminent examples of a scholarly archive, consisting of the pre- and reprint articles (submitted by the authors) in the area of High Energy Physics. Its user community is technically sophisticated, having been on the 'cutting edge' of web dissemination of research communications (in fact, it was for just this community that the Web was developed), and has long had a culture of sharing prepublication documentation, even before the widespread use of personal computers.

The OpCit project (`opcit.eprints.org`, funded under the NSF/DLI 2 programme) started as a project to interlink the texts stored in the HEP archive and is now involved in providing more general metatextual services for such archives. As part of its remit, it has been investigating the way in which researchers and users of the archive have been using it to deposit and read new research results, and to try to understand the influence of the archive on the way that scientific publication has been changing over the last decade.

2. Evidence of Textualities in `archiv.org`
This considered the articles that have been submitted to the archive from the point of view of the reading process (user as surfer) and the writing process (user as author and depositer) and then consider how the two processes combine to form the standard lifecycle of academic publishing (reading, writing, citing, reviewing, revising, publishing).

The archive is a publicly accessible store of published and unpublished papers submitted by scientists from around the world. First established in the early 1990s, it has grown to contain 130,000 papers and to receive over 30,000 "hits" per day (at the time when this study began in August 2000). To alleviate pressure on the main archive there are a number of mirror sites around the world, including UK one at the University of Southampton; it is the data held at this mirror that we have analysed [2].

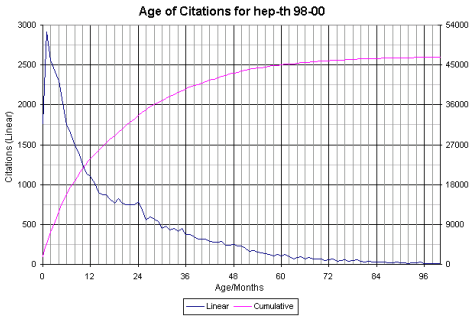
2.1 Most Requested Text Type
The most frequent kind of information downloaded from the archive is the full-text article (28%) rather than the paratext elements (title, abstract, keywords at 11%), archetexts (classification listings, 13%) or search requests (23%). The point of the paratext is to "mediate the book [or article in this case] to the reader", but it appears that many readers come to the text 'pre-mediated'—this may be due to the email alerts which are not represented in the usage logs.

2.2 Reading Habits


One of the obvious advantages of electronic- over print-media is speed of delivery and speed of production. This is matched by the speed of consumption from the archive: the most-read articles are those that have been most recently deposited. In fact, most of the downloads at any time are due to articles under a week old.

2.3 Writing Habits

The graph below shows the age of citations extracted from the articles deposited in a particularly active sub-area of the archive (hep-th, High-Energy Physics Theory) in 1998. It shows that citations within this area are very young (with a peak number of

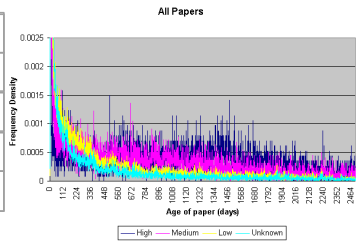


citations being just one month after the deposit of the cited paper); this speed of reference may not be obvious in the paper journal world, where it may take a paper up to two years to appear in print. (The time measurement has a granularity of months because of the way that deposits are arranged in monthly batches.)

2.4 Reading/Writing Cycle

It is possible to see a correlation between papers that are highly cited (and hence supposed highly influential) and those which are frequently downloaded. It is not known whether frequently downloaded papers lead to more citations, or whether the citations (and consequent linking by OpCit software) leads to more downloads. It is demonstrable that highly cited papers have a higher download longevity—they are downloaded more for longer.

Download type	r	n
AllPapers	0.11155	63671
High Impact Papers	0.27293	1981
MediumImpactPapers	0.01288	5937
Low Impact Papers	-0.01412	30163

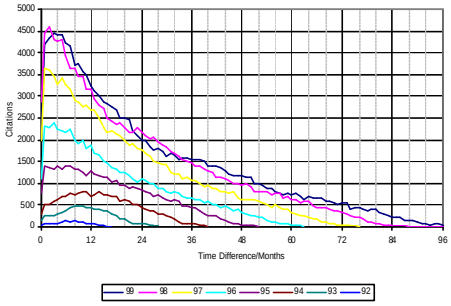


2.5 Intertextual Latency

The long-term effect of the archive seems to be to decrease intertextual latency: the period between an article appearing in the archive and it receiving its first citation has decreased over the period of the archive from about a year to about a month (see graph below). This implies that the speed of scientific communication – the rate of ideas affecting other researchers ideas – is increasing.

2.6 Evidence of Intertextuality

The accepted scholarly publishing model of course hinges on a work being published. Without publication there can be no dissemination; the work is effectively invisible. The archive exists by virtue of the fact that articles are deposited before they are published – before editors and referees have even decided whether they will be published. Although the majority (up to 80%) of deposited articles are eventually recorded as being published in journals (and subsequently read, cited, *etc*.), what happens to the content in the



remaining unpublished works? A study of 100 articles deposited in December 1998 showed that of the 19 that did not appear to have been published, 11 were cited by other (published) articles and 3 (conference papers) were subsequently rewritten to form new articles [3]. In other words, although the texts themselves did not feature in the printed record, they still retained an effective presence in the literature, either by direct inclusion of all or part of the text, or by intertextual reference.

3. Open Archive Initiative

The Open Archive Initiative (OAI [4]) is a recent interoperability initiative springing from the archive and which allows archives to share 'metadata' about their texts. This enables new classes of cross-archive information sharing services, and raises the interesting phenomenon of autonomous paratexts, which are traded, copied and processed independently of the original text which they advertise.

4. Concluding Remarks

The HEP archive is not only a repository of the scholarly literature, it is an embodiment of it and a focus of the process of scholarly communication between researchers. The role of the computer as evidenced in these results is to increase the interactivity of the scholarly process, diminishing the latencies and barriers between reading and writing and enhancing the scholarly community's ability to create a complex, multiply-branched (hyper) text.

5. ACKNOWLEDGMENTS

This work has been funded by JISC's OpCit project in the UK.

6. REFERENCES

[1] Brody, T and Hickman I. (2000) Mining the Social Life of an Archive. OpCit Internal Technical Report. `http://opcit.eprints.org/tdb198/opcit/`

[2] Carr, L. (2001) The Use of Open Archives: Who, How Often and Why. Presentation at Open Archives Workshop, European Conference on Digital Libraries 2001. `http://www.ecs.soton.ac.uk/~lac/opcit.whow`

[3] Dalgaard R. (2001) Hypertext and the Scholarly Archive: Intertexts, Paratexts and Metatexts at Work. In *Proceedings of ACM Hypertext 2001*, 175-184.

[4] Lagoze, C. and Van de Sompel, H. (2001) The open archives initiative: building a low-barrier interoperability framework. In *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*, 54 – 62. ISBN:1-58113-345-6