Georgia Roidouli

Agents & Multimedia Group (IAM) Electronics & Computer Science Department University of Southampton - UK

Georgia Roidouli, Dr. Leslie Carr, Prof. Wendy Hall

Intelligence, Agents & Multimedia Croup (IAM) Electronics & Computer Science Department University of Southampton - UK

How E-Prints can help the refereed research literature to be freed!

ABSTRACT

Refereed journals will be available online, very soon. This means that anyone could be able to access them from any connected pc to a network world-wide. The literature will be interconnected by citation, author, and keyword/subject links, allowing users to access and navigate online open archives. E-prints will help the scholarly and scientific literature, eventually, to be free from cost barriers and institutions will be able to create E-print archives in which their authors can self-archive all their refereed papers for free, for all and forever!

The paper will review libraries' long perspective of adopting e-print archives as well as outline the most important issues revolving around the e-prints evolution. In addition, the author would like to introduce you to an existing e-print archive known as arXiv and describe her research topic; how is related to e-prints and what her study could possibly examine.

Key - words: *e-prints, self-archive, scholarly communication, scientific literature, open archives, arXiv.*

Biography

Georgia Roidouli has graduated from Liverpool John Moores University by obtaining a bachelor BSc(Hons) in Information and Library Management and an MScin Information Systems. She worked for the academic library of T.E.I, of Larissa for a year as a web-based resources co-ordinator, including responsibility for the Grey Literature Project. She is currently a research student in IAM group at the University of Southampton and she works on her MPhil thesis. Her research interest revolves around 'knowledge mining'. This could be translated into how scholars use e-print archives, how they decide to download particular pages and why documents are popular. More specifically, these could attribute towards an understanding of user behaviour. This is the first time for Georgia Roidouli to participate in a Greek conference and she is delighted to be invited by the T.E.I. of Larissa.

E-prints

΄E-prints' are electronic copies of academic research papers. They can take the form of 'pre-prints' or 'post-prints'. They could be journal articles, conference papers, book chapters or any other form of research output. An 'e-print archive' is simply an online repository of these materials. Typically, an e-print archive is normally made freely available on the web with the aim of ensuring the widest possible dissemination of its contents.

Pre - print

The digital text of a paper that has not yet been peer-reviewed and accepted for publication by a journal.

Post - print

The digital text of an article that has been peer-reviewed and accepted for publication by a journal. This includes:

- the author's own final, revised, accepted digital draft
- the publisher's, edited, marked-up version, possibly in PDF
- any subsequent revised, corrected updates of the peer-reviewed final draft.

E-print Archive

An E-print Archive is a collection of digital documents. More specifically, it could be an online archive of pre-prints and post-prints. OAI-compliant E-print Archives means that they share the same metadata and make their contents interoperable with one another. Their metadata can be harvested and they can be navigable by any user.

OAI

The Open Archives Initiative (OAI) "develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content." The OAI Metadata Harvesting Protocol creates the potential for interoperability between e-print archives by enabling metadata from a number of archives to be harvested and collected together in a searchable database. The metadata harvested is in the form of Dublin Core and normally includes information such as author, title, subject, abstract, and date.

Self-archiving

To self-archive is to deposit a digital document in a publicly accessible website, preferably an OAI-compliant E-print Archive. Depositing involves a simple web interface where the depositor copy/pastes in the "metadata" (date, author-name, title, journal-name, etc.) and then attaches the full-text document. The purpose of self-archiving is to make the full text of the peer-reviewed research output of scholars/scientists and their institutions visible, accessible, harvestable, searchable and useable by any potential user with access to the Internet.

Benefits of self-archiving

Scholars, researchers should self-archive in order to maximize the visibility, accessibility, usage and impact of their work. The big picture of self archiving is the aim of freeing up research output and thereby improving research communication. With online archives, all papers can be located by anyone quickly and easily and at no cost. Authors can put draft copies and successive updates up for public view, until the final, peer-reviewed (published) version appears. The eprints.org software has a self - archiving facility which creates online archives.

Moreover, the benefits could be described as the following:

- "Lowering 'impact barriers'. E-print archives make papers more visible. Papers are freely available for others to consult and cite.
- Ease of access. This is the other side of the coin of lowering impact barriers. It means that access to the literature should be freed up, in contrast to the current system where most of the research literature is not easily available to most researchers.
- Rapid dissemination. Depending on what document types are accepted in the archive (pre-prints or post-prints) online repositories can really speed up the process of dissemination of research findings. In certain fast-moving disciplines this can be an attractive prospect.
- Raising the profile of the institution. Ensuring that the research output of the institution is widely disseminated. This helps to enhance its reputation and thus its ability to attract high quality researchers and further research funds.
- Long term cost savings. These savings will result in reducing outlays for periodical subscriptions". [1]

Libraries' Perspective

In order libraries to facilitate self-archiving they need to:

- Administer e-print archives, mandate them and help in author start up
- Offer trained digital librarian help in showing faculty how to self-archive their papers in the university E-print Archive
- Offer trained digital librarian help in doing "proxy" self-archiving, on behalf of any authors who feel that they are personally unable (too busy or technically incapable) to self-archive for themselves. Authors need only supply their digital full-texts in word-processor form: the digital archiving assistants can do the rest.
- Digital librarians, collaborating with web system staff, should be involved in ensuring the proper maintenance, backup, mirroring, upgrading, and migration that ensure the perpetual preservation of the university E-print Archives. Mirroring and migration should be handled in collaboration with counterparts at all other institutions supporting OAI-compliant E-print Archives.

Addressing concerns about self-archiving

Stevan Hamad [2] pointed out 22 different kinds of concerns that might be interesting to people who worry about self-archiving. These are:

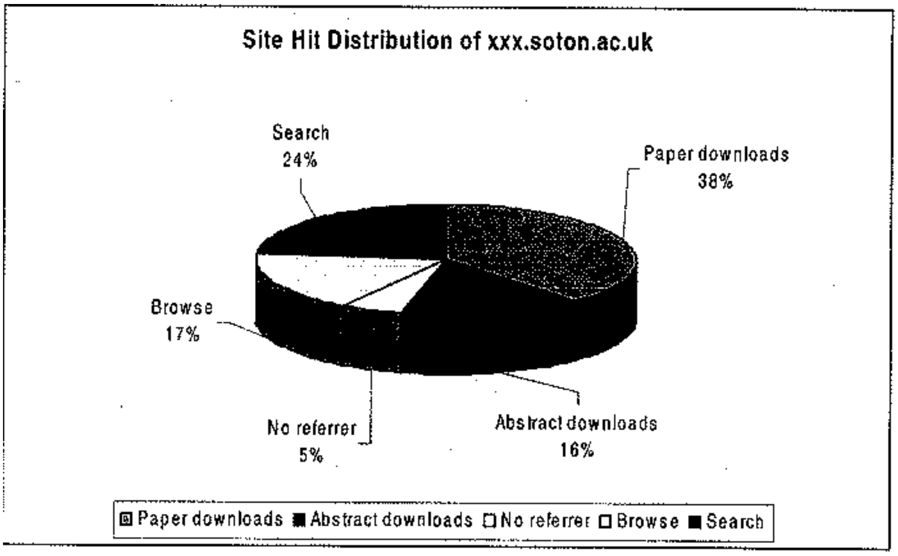
1. Preservation
2. Authentication
3. Corruption
4. Navigation (info-glut)
5. Certification
6. Evaluation
7. Peer review
8. Paying the piper
9. Downsizing
10. Copyright
11. Plagiarism
12. Priority
13. Censorship
14. Capitalism
15. Readability
16. Graphics
17. Publishers'future
18. Libraries' future
19. Learned Societies' future
20. University conspiracy
21. Serendipity
22. Tenure/Promotion

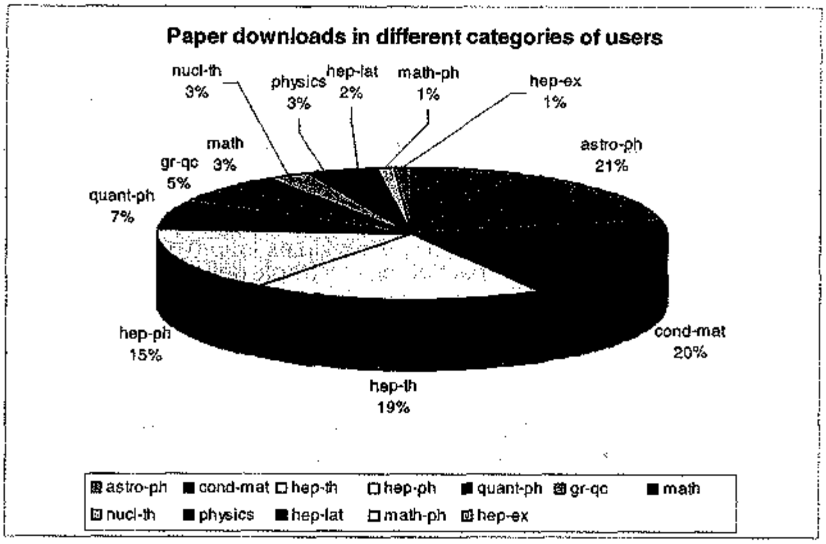
An e-print archive example: arXiv

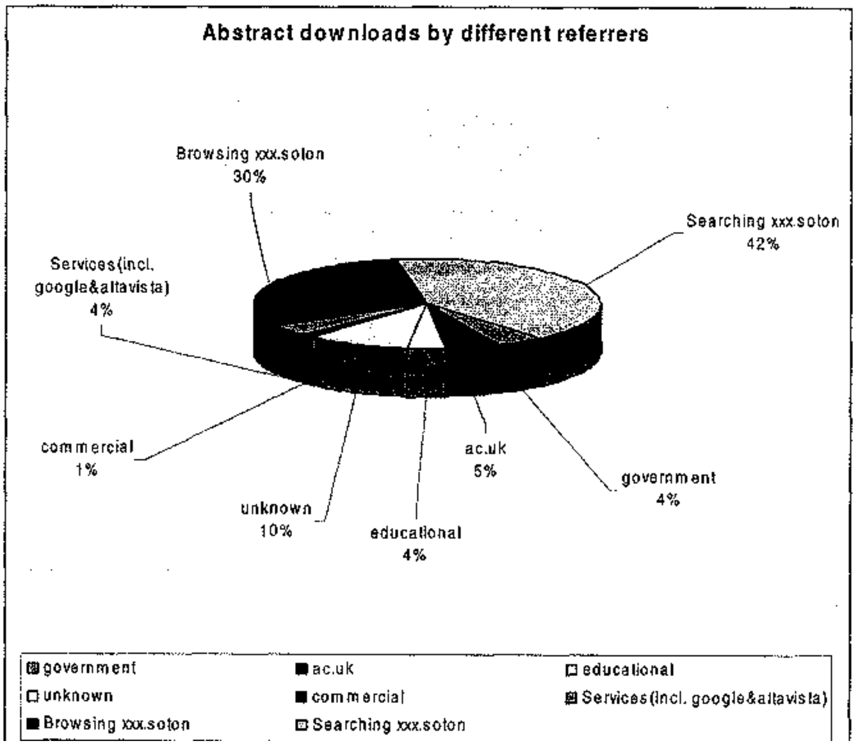
arXiv is the abbreviation of the Los Alamos Physics e-print archive. It has been developed 11 years ago by P. Ginsparg [3] to create a method for organising and distributing scientific information via a computerised network. It covers most of physics and has expanded to include repositories for non-linear sciences, mathematics, computation and language. It is mirrored in 16 countries including the UK, has reasonable search facilities, and offers services such as email notification of new submissions of interest.

Scholars use the arXiv for all of their research communications. They check it every day for new information. They post all their papers there, cite references by archive number, use the search engine to find other papers, and need little or no other publication services. It costs the users nothing and it is self organising. Physicists all over the world can post their research results without being hassled by grumpy editors and referees. Also, they do not need to be part of some inner circle of accepted colleagues to be on the preprint mailing lists, they can find out what's new on the archive just as soon as everyone else.

Simply, in order to emphasise the impact of the electronic delivery services provided by arXiv it is worth illustrating the following diagrams. These diagrams give the percentages of site hit distribution, paper downloads in different categories of users and abstract downloads by various referrers. All analyses are based on a data set of incremental changes that has been gathered over a period of two and a half years approximately. To determine users' behaviour we analysed web-logs of the UK mirror from August 1999 to May 2002.




Paper downloads in different categories of users
nucl-th 3%
physics 3%
hep-lat 2%
math-ph 1%
hep-ex 1%
math 3%
gr-qc 5%
quant-ph 7%
astro-ph 21%
hep-ph 15%
cond-mat 20%
hep-th 19%
▨ astro-ph
■ cond-mat
□ hep-th
□ hep-ph
■ quant-ph
▨ gr-qc
■ math
▨ nucl-th
■ physics
■ hep-lat
□ math-ph
▨ hep-ex


Abstract downloads by different referrers
Browsing xxx.soton 30%
Searching xxx.soton 42%
Services(incl. google&altavista) 4%
commercial 1%
unknown 10%
educational 4%
ac.uk 5%
government 4%
▨ government
■ ac.uk
□ educational
□ unknown
■ commercial
▨ Services(incl. google&altavista)
■ Browsing xxx.soton
▨ Searching xxx.soton

Research Topic

In an ideal world of information provision we have various ways of accessing scientific information. We have the traditional way of visiting libraries; we have publishers, web servers and also e-print online archives. Since the Los Alamos E-print archive (arXiv) became a major source of information delivery we need to know more about how people seek and retrieve information in this particular digital environment. Easy access from one's desktop is leading to usage of serious scholarly material by a much wider audience, both of the scholars and the general population. Also easy electronic access to scientific information is changing the patterns of use. This study aims to investigate the nature, manifestations and behaviour of successive searching by users in digital online archives. More specifically, to derive conclusions about information retrieval patterns supporting successive searching behaviour in the Los Alamos E-print Physics Archive.

My research topic started by examining two questions: how do users obtain access to articles (abstracts and full papers) in the arXiv and why do they choose these access methods? The primary objectives for testing the above criteria were to measure the impact of electronic delivery services provided by arXiv and to check the differences in user behaviour over the years.

Steps and actions of the study

To start with we needed to focus on specific steps that will help us manipulate the data better. Those are:

- Identify all the users (look at session length, session size, No. of papers)
- Choose one particular user and look at the same parameters
- Identify all the possible browsing patterns
- Design diagrams that illustrate users' interactions during numerous sessions

More specifically, during this phase of the study we selected a particular user and we analysed all the interactions the user conducted in one session. Every session takes the form of a diagram that assists on identifying the sequence of the user followed in order to download a particular document. By applying this approach we can investigate the most typical and interesting navigations.

The diagrams below illustrate the interactions of a particular user in the arXiv that follows during different sessions (e.g. user A selected from ac.uk range), in order to download a particular page in the archive. The outcome of this approach certifies that the same user during numerous sessions behaves differently. Basically, from the diagrams we can see how each session (e.g.) starts off and then calls up each abstract, choosing only one or two to get the full text from. Each distinct document gets a new node so searches appear as if they are the same. Occasionally, we can also tell what has really been searched for because it appears as part of the URL.

Diagram 1: *Session of User A at 25/08/1999 15:21*

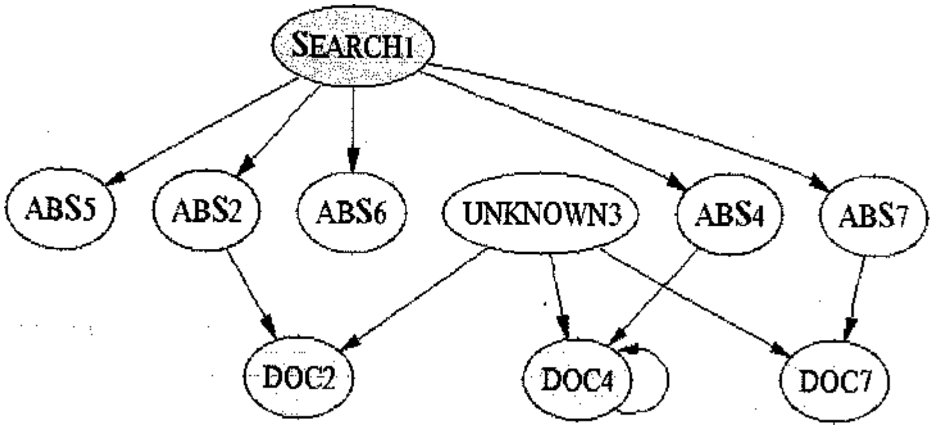


Diagram 2: *Session of User A at 30/08/1999 17:14*

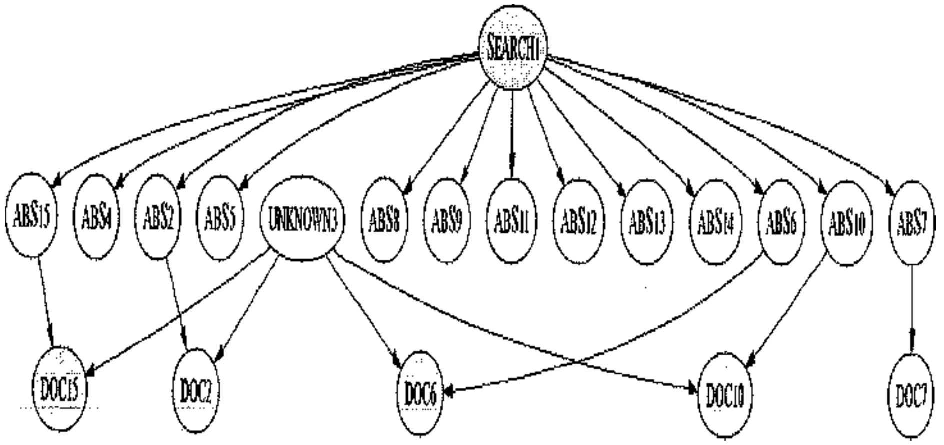


Diagram 3: *Session of User A at 14/10/1999 18:30:05*

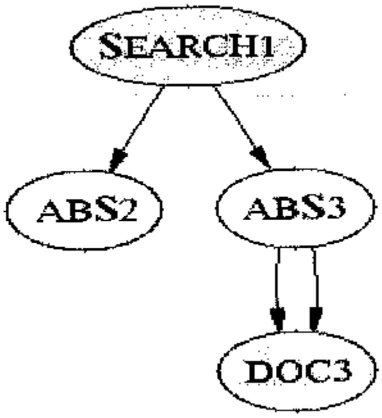


Diagram 4: *Session of User A at 22/09/2000 09:44:11*

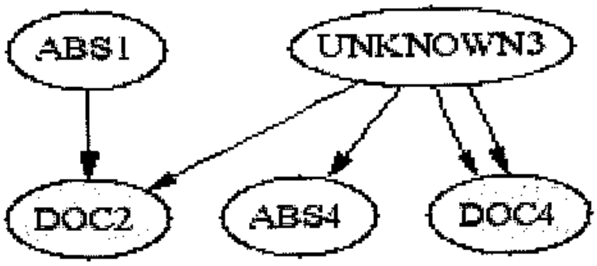


Diagram 5: *Session of User A at 13/10/2000 09:37:45*

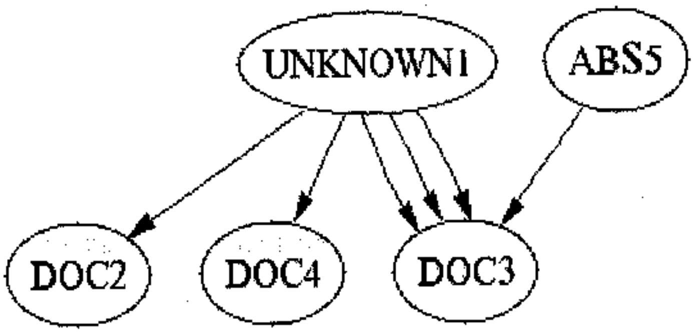
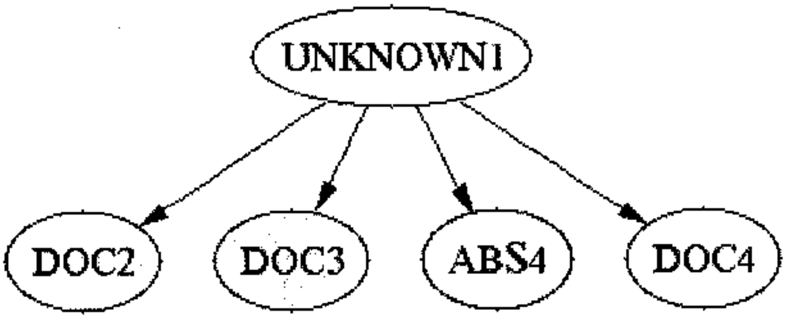


Diagram 6: *Session of User A at 27/02/2001 09:31:14*



Furthermore, in orderto derive conclusions about information retrieval patterns supporting successive searching behaviour in the Los Alamos E-print Physics Archive, we need to examine three dimensions for analysing the diagrams. These are:

- Where do users start from? (e.g. browse, search, come from unknown source)
- Do they read abstracts or full articles?
- Do they follow things up? (i.e. looking up citations, i.e. do they have deep branches on the diagrams?

Useful Internet Sources and Reading Material

[1] Stephen Pinfield, Mike Gardner and John MacColl. (2002). "Setting up an institutional e-print archive" http://www.ariadne.ac.uk/issue31/eprint-archives/

[2] Harnad, S. (2001) " For Whom the Gate Tolls? How and Why to Free the Refereed Research Literature Online Through Author/ Institution Self-Archiving, Now". http://www.ecs.soton.ac.uk/~harnad/Tp/resolution.htm

[3] Ginsparg, P. (2000) *"Creating a global knowledge network".* Electronic Publishing in Science, at UNESCO HQ, Paris, http://arxiv.org/blurb/pg01unesco.html

Harnad, S., Carr. L., jiao, Z., and Brody, T. *"Mining the Social Life of an E-Print Archive"* http://opcit.eprints.org/ijh198/

Harnad, S., Carr, L., (10 September 2000), *"Integrating, Navigating and Analyzing Open Archives Through Open Citation Linking (The OpCit Project)",* Current Science, Vol. 79, No.5, http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnadOO.citation.htm

Hitchcock, S. Carr, L, jiao, Z., Bergmark, D., Hail, W., Lagoze, C. & Harnad, S. (2000) *"Developing services for open eprint archives: globalisation, integration and the impact of links".* Proceedings of the 5th ACM Conference on Digital Libraries. San Antonio Texas June 2000.

http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnadOO.acm.htm

Pinfield, Stephen. (2001). *"How do physists use an E-print Archive?"* D-Lib IVlagazine 7. no. 12

E-Prints. http://www.eprints.org/

The Los Alamos E-Prints Archive, http://xxx.soton.ac.uk

The Open Citation Project, http://opcit.eprints.org/

OAI: http://www.openarchives.org/