Revealing a New Dynamic: Interaction in an Open Access Archive

Steve Hitchcock

The Open Citation Project (OpCit), Southampton University

These slides prepared for the 1st Workshop of the Open Archives Forum, Pisa, 13-14th May 2002

OpCit is a joint JISC-NSF

International Digital Libraries Project 1999-2002

OAF: what we have in common

- An international group
- Want to promote and support better, more efficient access to scholarly resources via digital libraries
- Support for the Open Archives Initiative (OAI), and use of its protocol for metadata harvesting

OAF: what we *might* have in common

- OAI participants:
    - data providers (e.g. an institution)
    - service providers (e.g. Arc, Torii, OpCit)
- A wish for open access to complete resources, e.g. eprint archives, as promoted by the Budapest Open Access Initiative (BOAI)

We have no mandate to change the system of scholarly publication. We have to make the case and persuade authors and users of the advantages of Open Archives.

This presentation

Shows that open access works for authors and users. Reveals some new aspects of the *social life* of an eprint archive.

Illustrating software and services developed as part of the **Open Citation Project (OpCit)**, and using data from our associated studies of arXiv user behaviour, it will be shown that **a new 'dynamic', the speed of interaction between users, becomes evident when access to full resources is free, open and unrestricted**.

This is important for all those who are building open archives, and for those who are tentatively moving towards building open archives (e.g. the biomedical community).

Key characteristics of eprint archives

- *Very* low cost to maintain (est. > $5/paper, see Ginsparg*)

- Free to users

- Rapid dissemination of preprints *and* postprints

- Fully automated (light moderation, no peer review)

The best solution is *author self-archiving*. This was the original focus of OAI.

Not all disciplines will adopt this approach. In biomedicine, the Public Library of Science advocates *publisher archiving* within six months - two years *after* journal publication.

* Creating a global knowledge network. *Second ICSU-UNESCO International Conference on Electronic Publishing in Science*, Paris, February 2001
http://associnst.ox.ac.uk/~icsuinfo/ginspargfin.htm

Budapest Open Access Initiative supports self-archiving

- Launched February 2002
- Promoting free access to research literature through self-archiving *and* alternative publishing models
- Over 2000 individuals and 130 organizations have signed the initiative, including Library of Congress, the Association of Research Libraries, the Canadian Association of Research Libraries, the Australian Vice Chancellors Committee, and a growing number of individual universities
- Backed by the Soros Open Society Institute

Important requirements of open access archives

Access critical for users

Impact critical for authors

Quality important to research

"Articles freely available online are more highly cited" – Lawrence

Nature, May 2001

http://www.nature.com/nature/debates/e-access/Articles/lawrence.html

Characterising open access

All the Refereed Literature,

Freely Accessible Online,

for Anyone,

Anytime,

Anywhere

This creates equality of access between institutions, countries, developed vs developing

"In an open system we compete with our imagination, not with a lock and key" – Negroponte, *Being Digital* (1995)

Benefits of freeing the refereed literature

- Online Academic CVs linked to full-texts in institutional eprint archives
- Universal searching
- New impact indicators (search ranking)
- New digitometric analyses
- Continuous research assessment

OpCit: how it can help you

The Open Citation project is developing software and services to support OAI and BOAI through the promotion of eprint archives. OpCit can help OAI data providers and service providers:

• **EPrints.org software**: free software to build and manage OAI-compliant eprint archives

• **Citebase**: citation-ranked search

EPrints.org software

Generates eprints archives that are compliant with the Open Archives Protocol for Metadata Harvesting. EPrints is free (GPL) software. It is aimed at organisations and communities.

EPrints v. 2.0 released February 2002 (now on v 2.0.1, which fixes bugs and typos). Features:

• Internationalised metadata stored as Unicode

• Support for multiple archives on one server

• Improved user interface

Citebase search engine

"Google for the refereed literature"

Citebase is based on an open citation database

- Harvests metadata using OAI-PMH

- Extracts reference lists from arXiv papers

- Provides impact (and other)-ranked search based on reference data

- Re-exports metadata + references

Growth of arXiv

- 155,000+ papers submitted
- 30,000+ new submissions in 2000
- Nearly linear growth in submission rate
- Over 99% of submissions are entirely automated
- Serves 70,000+ users in over 100 countries
- 13 million papers downloaded in 2000
- 110,00 – 130,000 visits daily

– Luce, R. E., E-prints Intersect the Digital Library: Inside the Los Alamos arXiv. *Issues in Science and Technology Librarianship*, Winter 2001 http://www.library.ucsb.edu/istl/01-winter/article3.html

Revealing more about arXiv user behaviour

The following results are taken from

Mining the Social Life of an Eprint Archive
http://opcit.eprints.org/tdb198/opcit/

This Web site reports the raw data from the study. We have yet to publish these results formally, but plan to do do. The data are offered openly for analysis by others. We would be interested to hear from anyone who wishes to comment on these results.

arXiv site hits

(based on UK mirror for August 1999 to May 2000)

Site Hit Distribution

Browsing Help 1.16%
Multifind Form 1.97%
Browsing Year 0.61%
Multifind POST 3.67%
Browsing Archive 3.20%
Other 2.57%
Formater downloads 1.94%
Paper Downloads 27.94%
Find POST 9.52%
Find GET 7.56%
Picture Downloads 7.60%
Index Accesses 8.26%
Browsing Listings 2.16%
Browsing Recent 3.68%
Browsing New 6.88%
Abstract Downloads 11.27%
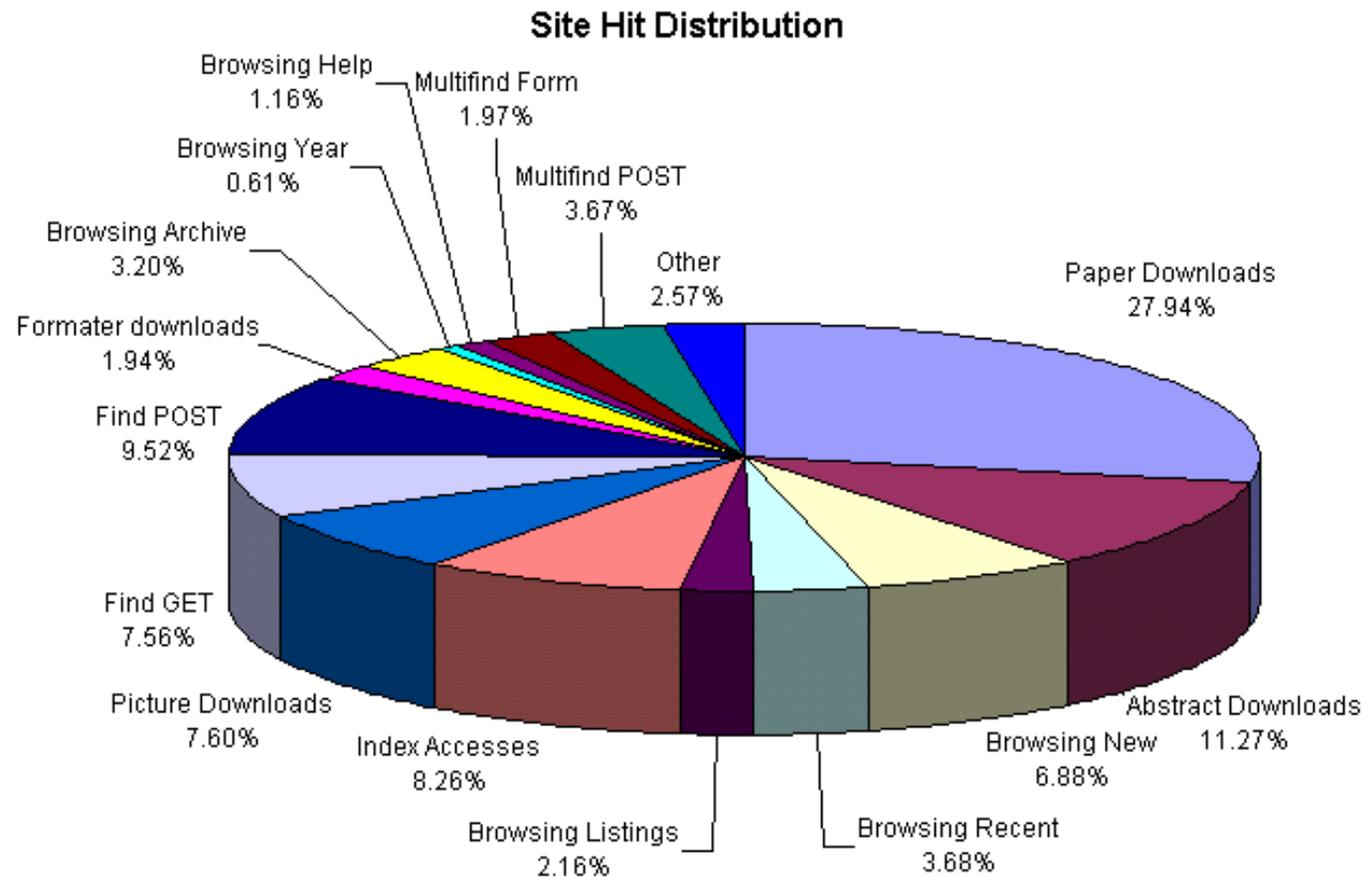
28% of downloads are papers, 11% are abstracts, the rest are browse and search

The "new paper rush"

Age of paper against download frequency
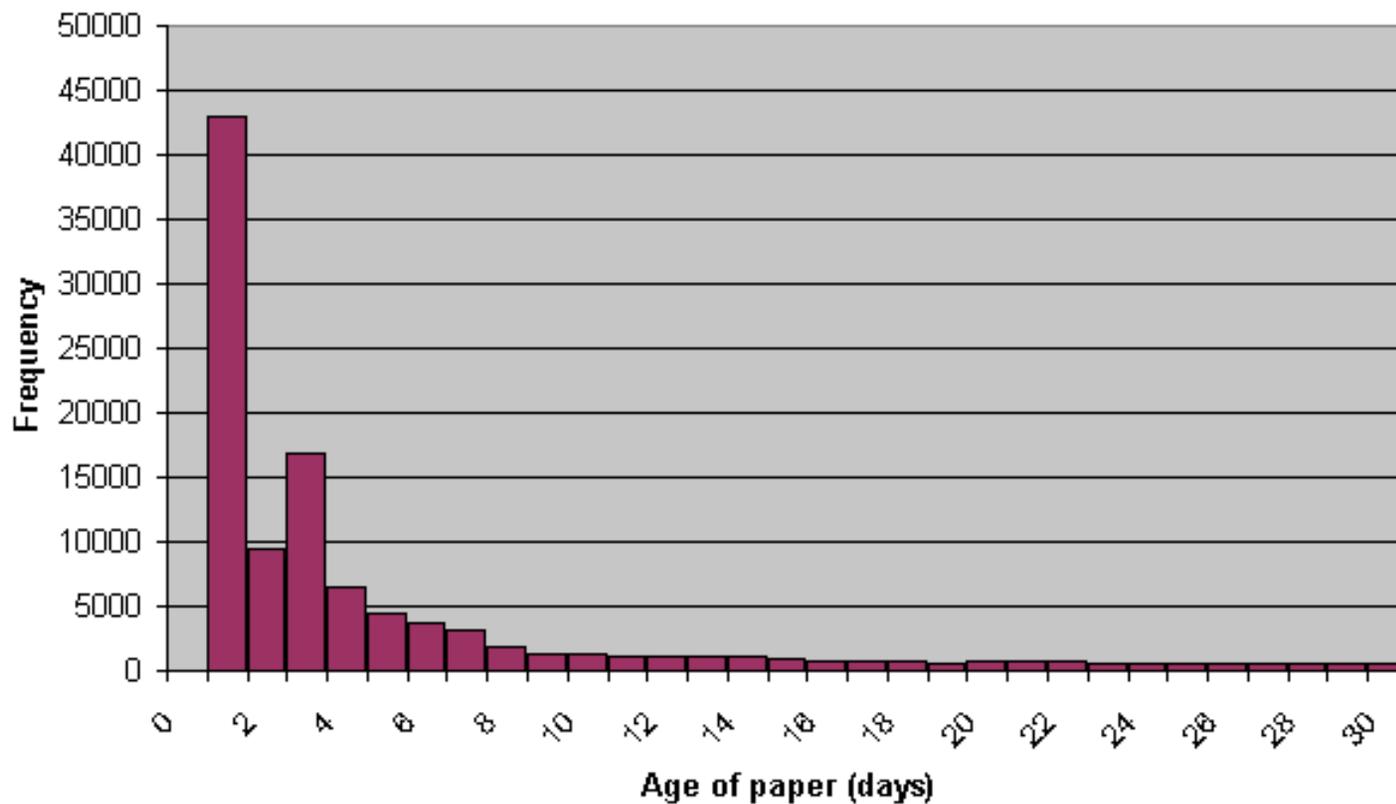
86.3% of papers in arXiv are hit during the first month in the archive

Are preprints updated?

Multiple Updates by LANL Subfield
(based on LANL meta-data)

No. of Papers with Updates

Legend: □ No Updates ■ 1 Update □ 2 Updates □ 3 Updates ■ 4 Updates
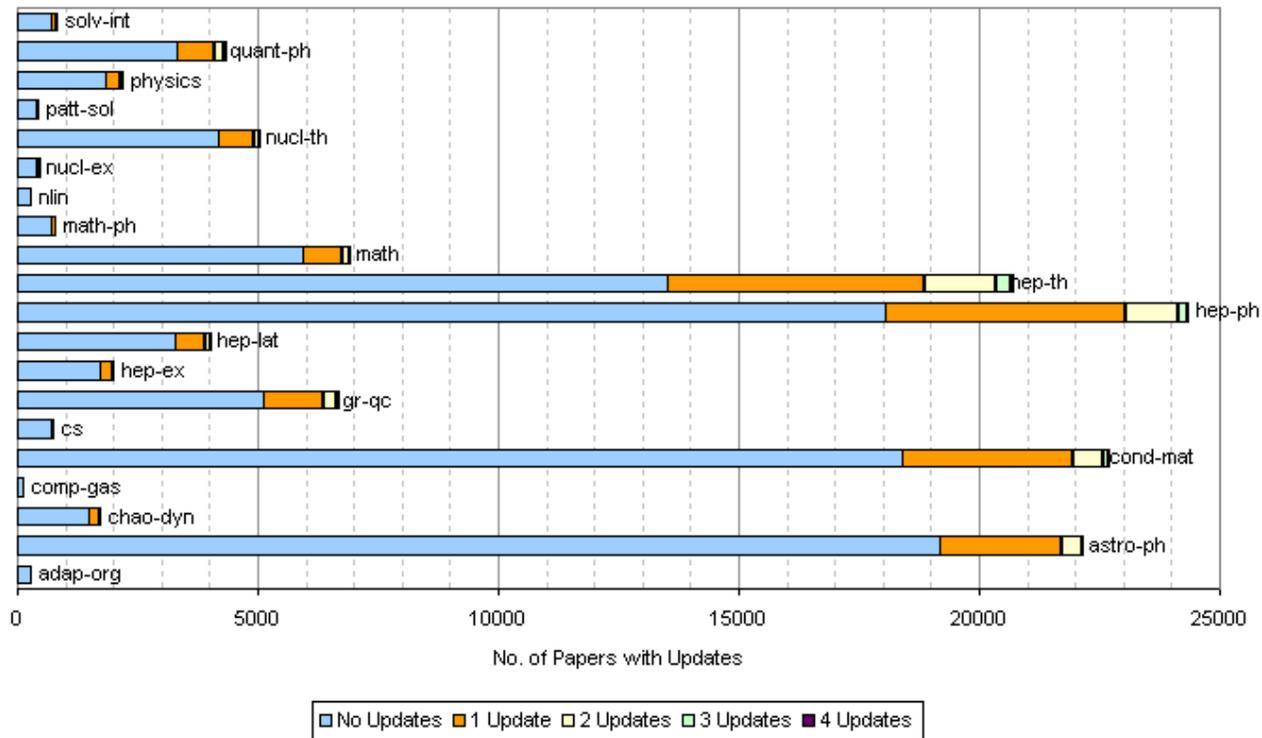
- 43% of arXiv papers are updated to include a Journal-Ref
- arXiv papers are updated as many as five times

Maximising impact: arXiv example

All Papers
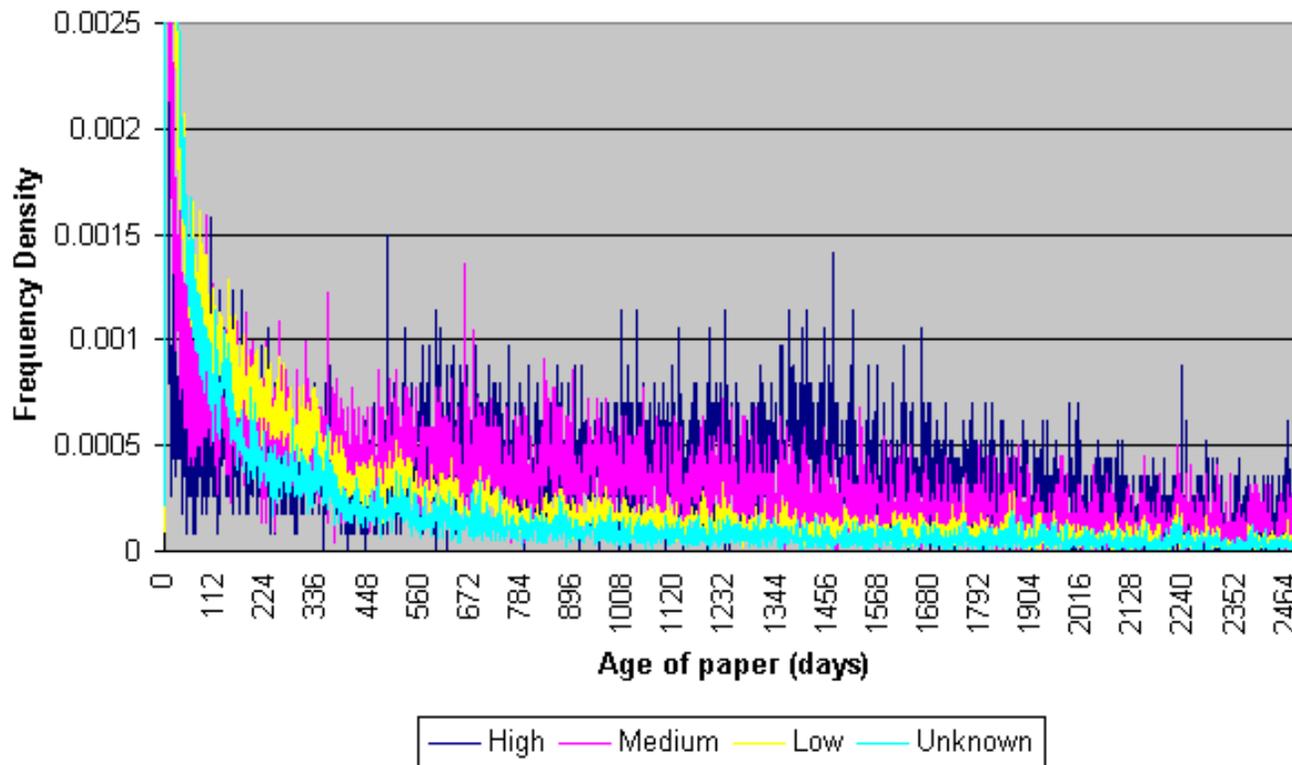
More highly cited papers show higher and more sustained download frequencies

Maximising access: arXiv example

Rise and Fall of Citation Latencies: 1992-1999
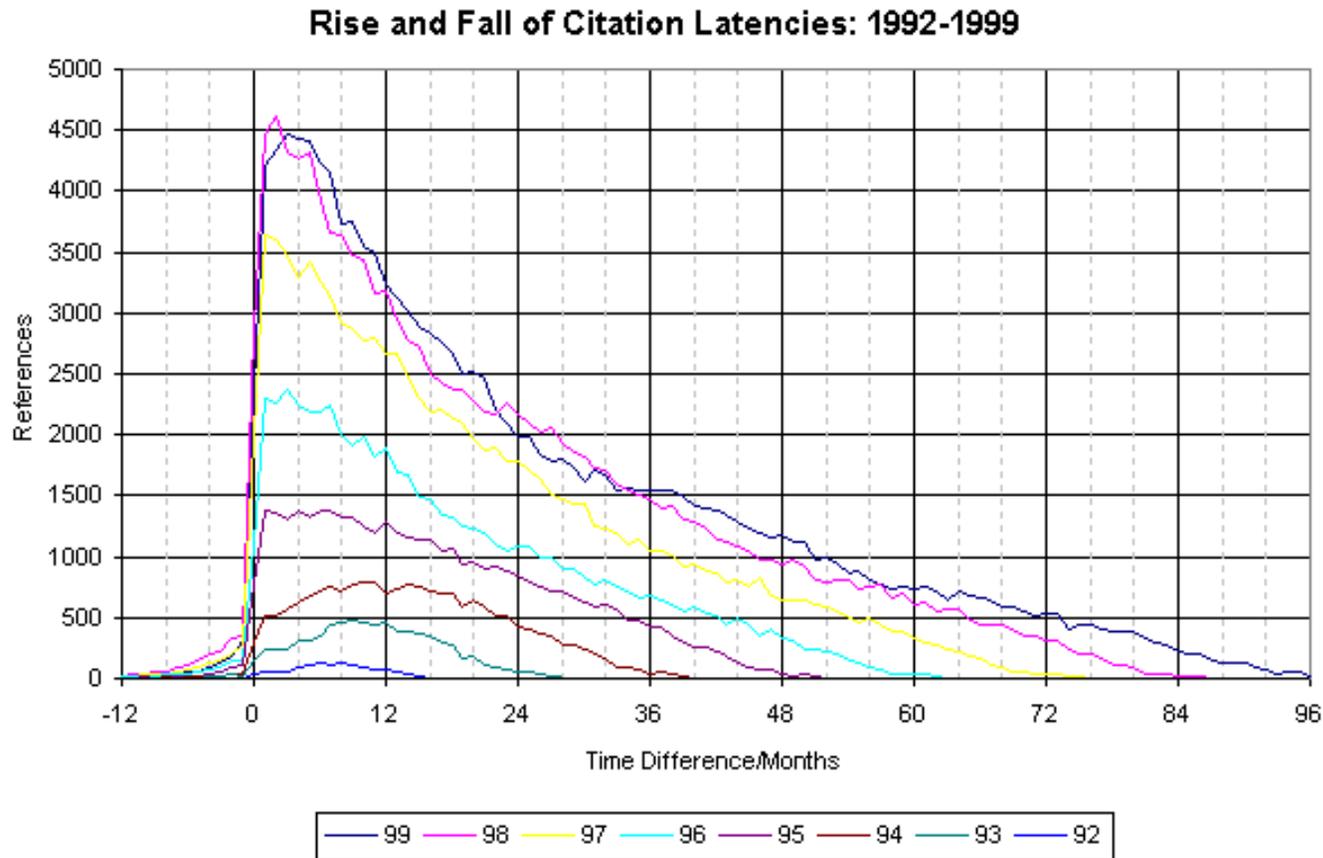
Decreasing citation latencies: The latency of the citation peak has been reducing over the period of the archive, i.e. each year papers are cited sooner and more often

Maximising interfaces

Citebase, a new interface to the scholarly literature

A maximising strategy

Results from the Open Citation Project show that authors who self-archive their papers in OAI-compliant institutional or discipline-based eprint archives will:

- **Maximise interfaces to their work**
- **Maximise access to their work**
- **Maximise impact of their work**

Credits

The Open Citation project is a collaboration between Southampton University, Cornell University and arXiv

- The project leaders are **Stevan Harnad** and **Carl Lagoze**
- Technical development at Southampton is directed by **Les Carr**
- EPrints.org software is being developed by **Chris Gutteridge**
- CiteBase is produced and managed by **Tim Brody**

A copy of these slides can be found on the OpCit Web site http://opcit.eprints.org/. Look for *Papers and Presentations*

Contact **Steve Hitchcock**: sh94r@ecs.soton.ac.uk