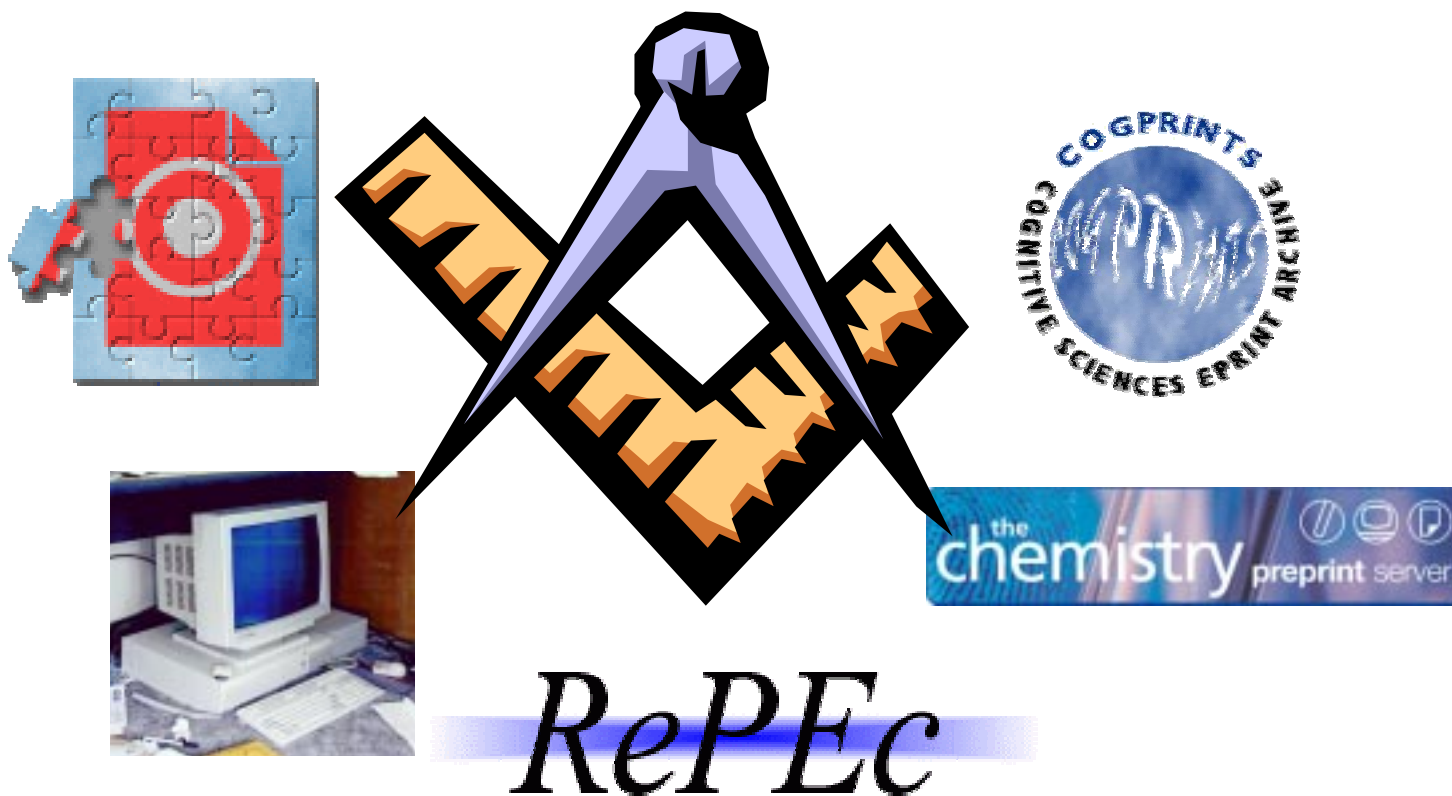# Distributed Eprints Archives and Scientometrics

Resolving an Anomaly
www.eprints.org

H. G. Wells, *World Brain: The Idea of a Permanent World Encyclopaedia*
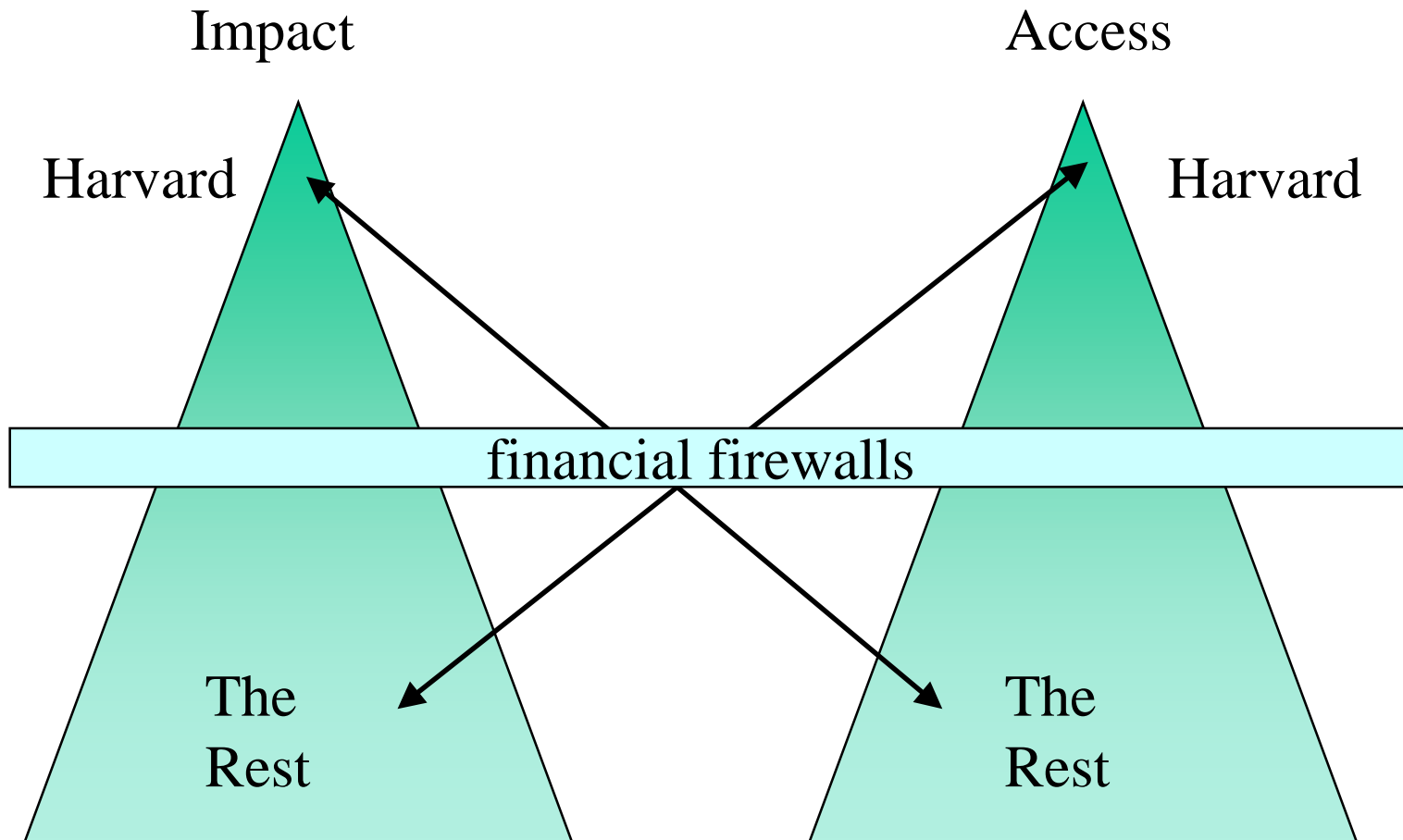## Encyclopédie Française, August, 1937

- Encyclopaedias of the past sufficed for the needs of a cultivated minority
  - universal education was unthought of
  - gigantic increase in recorded knowledge
- Discontent with the role of universities and libraries in the intellectual life of mankind
- Universities multiply but do not enlarge their scope
  - thought & knowledge organization of the world
- **No obstacle** to the creation of an efficient index to *all* human knowledge, ideas and achievements

The Optimal and Inevitable for Researchers

All of this will come to pass. The only question is "How Soon?"

- The entire full-text refereed corpus online
- On every researcher's desktop, everywhere
- 24 hours a day
- All papers citation-interlinked
- Fully searchable, navigable, retrievable
- For free, for all, forever

Globalizing Research

Impact

Access

Harvard

Harvard

financial firewalls

The Rest

The Rest

Resolving an Anomaly
www.eprints.org

The Subversive Proposal:

Sufficient to free entire refereed corpus forever, immediately:

1. **Universities install off-the-shelf, OAI-compliant Eprint software**
2. **Authors self-archive (preprints & postprints)**
3. **Institutions subsidize first start-up wave of self-archiving**
4. **The Give-Away corpus is freed**

Hypothetical Sequel:

5. *Users prefer free version?*
6. *Publisher S/L/P revenues shrink, Library S/L/P savings grow?*
7. *Publishers downsize to QC/C service-providers + optional add-ons?*
8. *QC/C service costs funded by author-institution out of reader-institution S/L/P savings?*

Resolving an Anomaly
www.eprints.org

Five Essential PostGutenberg Distinctions:

(if you don't make them, none of this will make sense)

1. Distinguish the **non-give-away** *vs*. **give-away** literature
   Litmus test: "Does the author seek a royalty/fee?":
   books (<u>*yes*</u>*) vs. refereed journal papers (*<u>*no*</u>*)*

2. Distinguish **income** (from paper sale) *vs*. **impact** (from paper use)
   (and distinguish give-away-author *imprint-income* [0] vs. *impact-income* [??])

3. Distinguish give-away author copyright protection from:
   theft-of-authorship (wanted) *vs*. **theft-of-text** (unwanted)

4. Distinguish **self-publishing** (vanity press) *vs*. **self-archiving**
   (of published, refereed research)

5. Distinguish unrefereed **preprints** *vs*. refereed **postprints**
   "eprints" = preprints + postprints

Zeno's Prima-FaQs

"I worry about self-archiving because…":

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
23. (your prima-FaQ here…)

Answers available at < http://cogsci.soton.ac.uk/~harnad/Tp/resolution.htm >
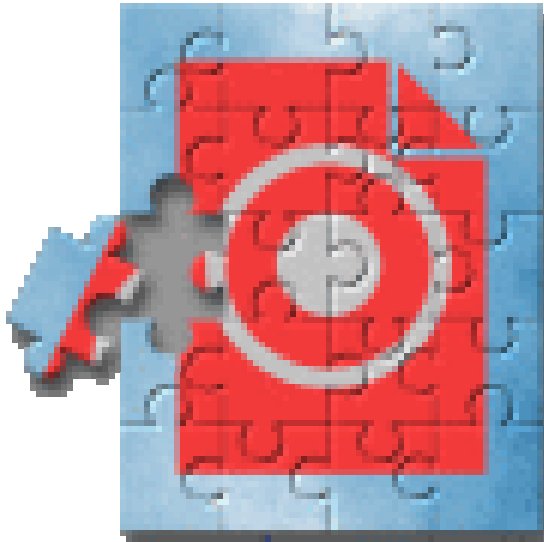
www.eprints.org

Eprints < **www.eprints.org** >   is dedicated to freeing the research literature, pre- and post-refereeing, through author/institution *self-archiving* in interoperable

Open Archives < **www.openarchives.org** > To help the self-archiving initiative quickly gain momentum, *archive-creating software,* compliant with the OAi protocol, hence fully interoperable with all other Open Archives, has been developed at the University of Southampton.

Eprints is designed to be as flexible and adaptable as possible, so that all universities world-wide can immediately adopt and configure it with minimal effort for all their disciplines' self-archiving needs.

The *Eprints* software, has been available (*for free, of course*) from **eprints.org** since December 2000.

From Linear Growth to Exponential

Monthly Submission RATE for arXiv.org

Last 111 months (25 Oct '00 total = 144,914)

Deposit
Rates

Disciplines

Time
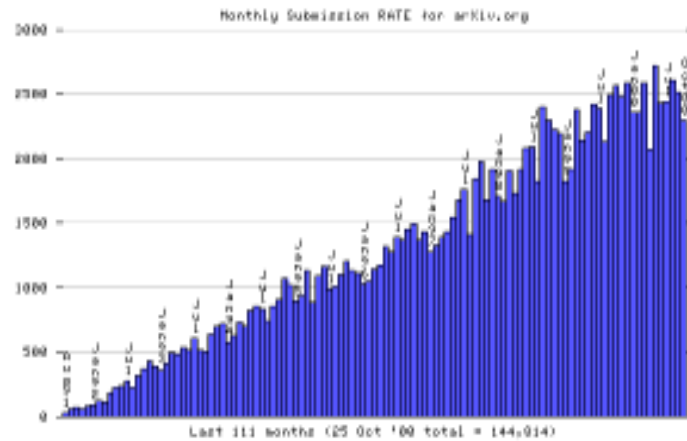
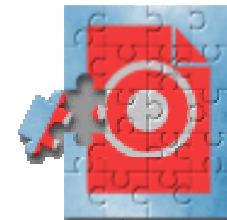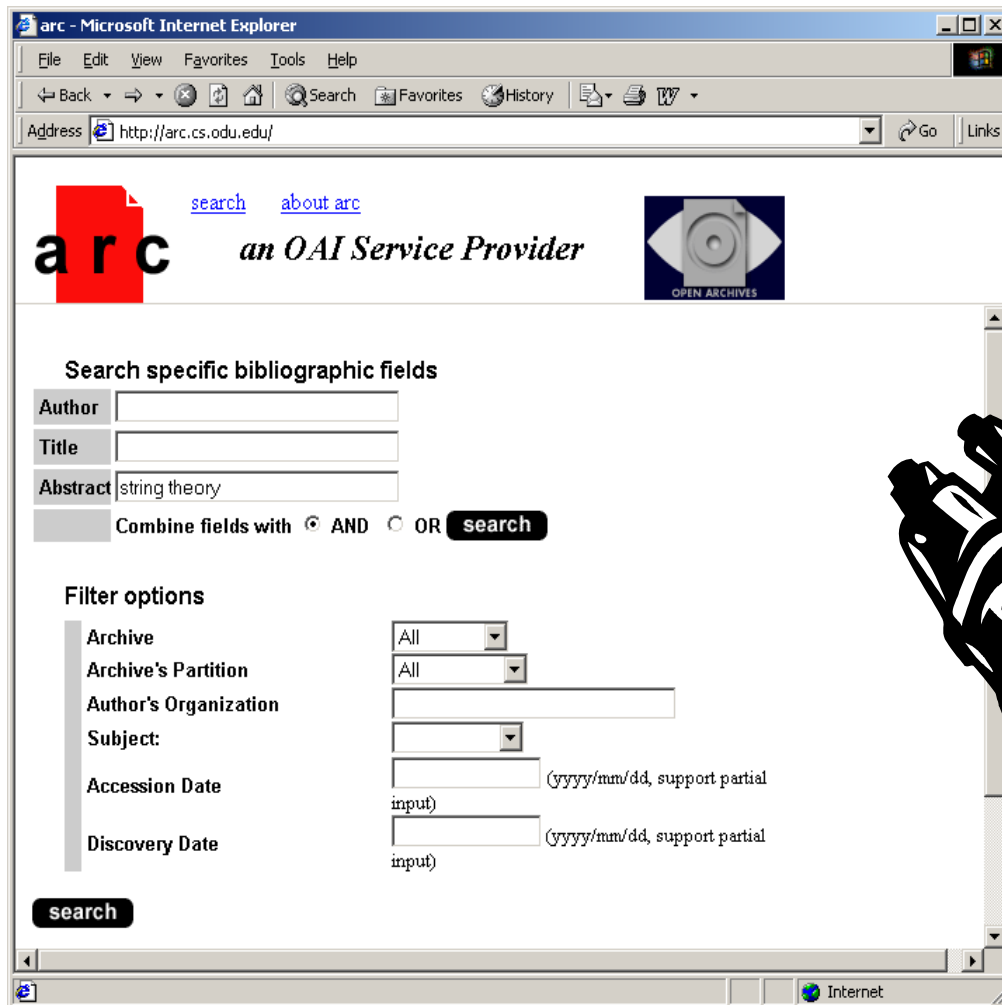arXiv submission rates
- *linear growth only*

30% of citations to
papers deposited in
arXiv
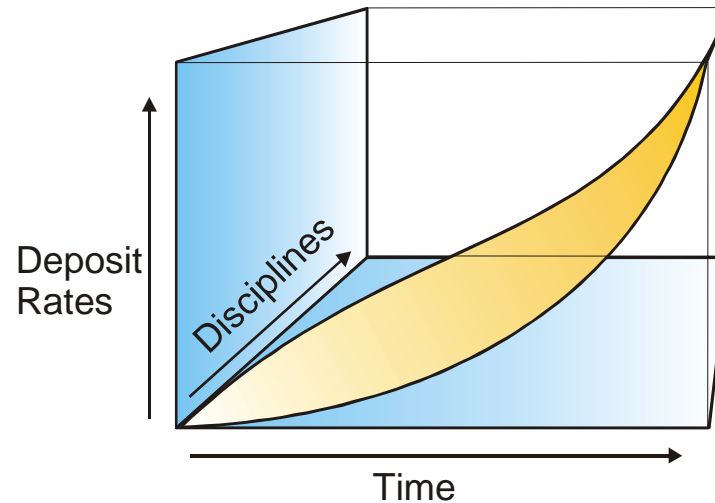
Exponential growth in
archiving to catch up with
paper-based research
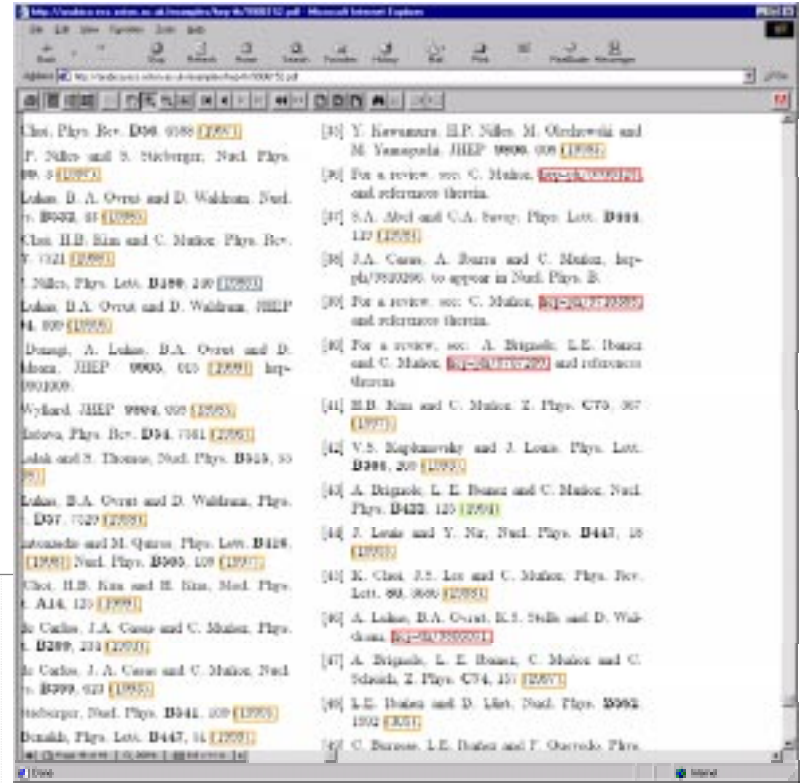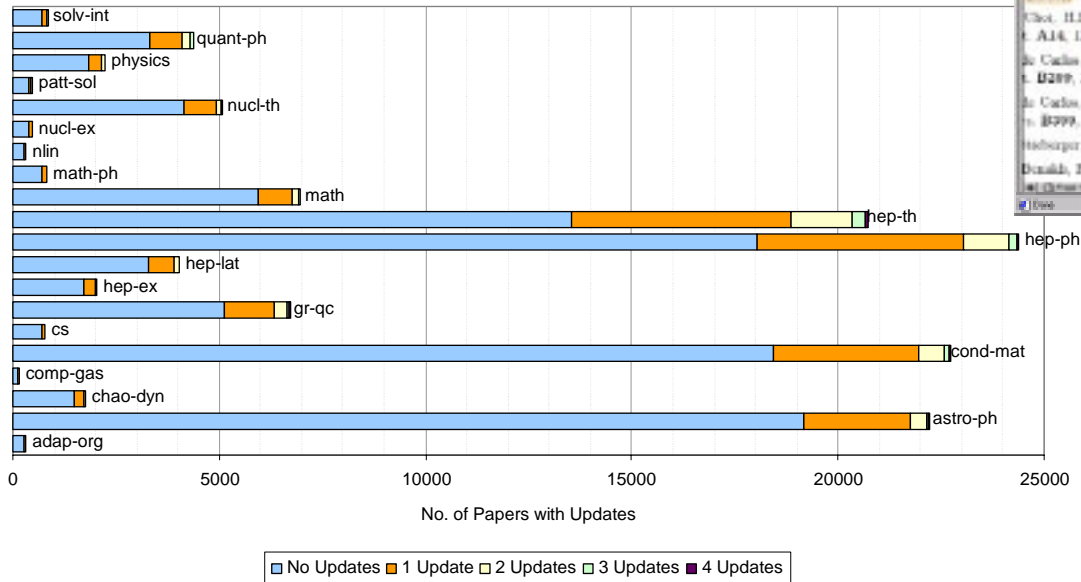
100% of papers archived, in all
disciplines

Resolving an Anomaly
www.eprints.org

Well's Global Research Database?

Resolving an Anomaly
www.eprints.org

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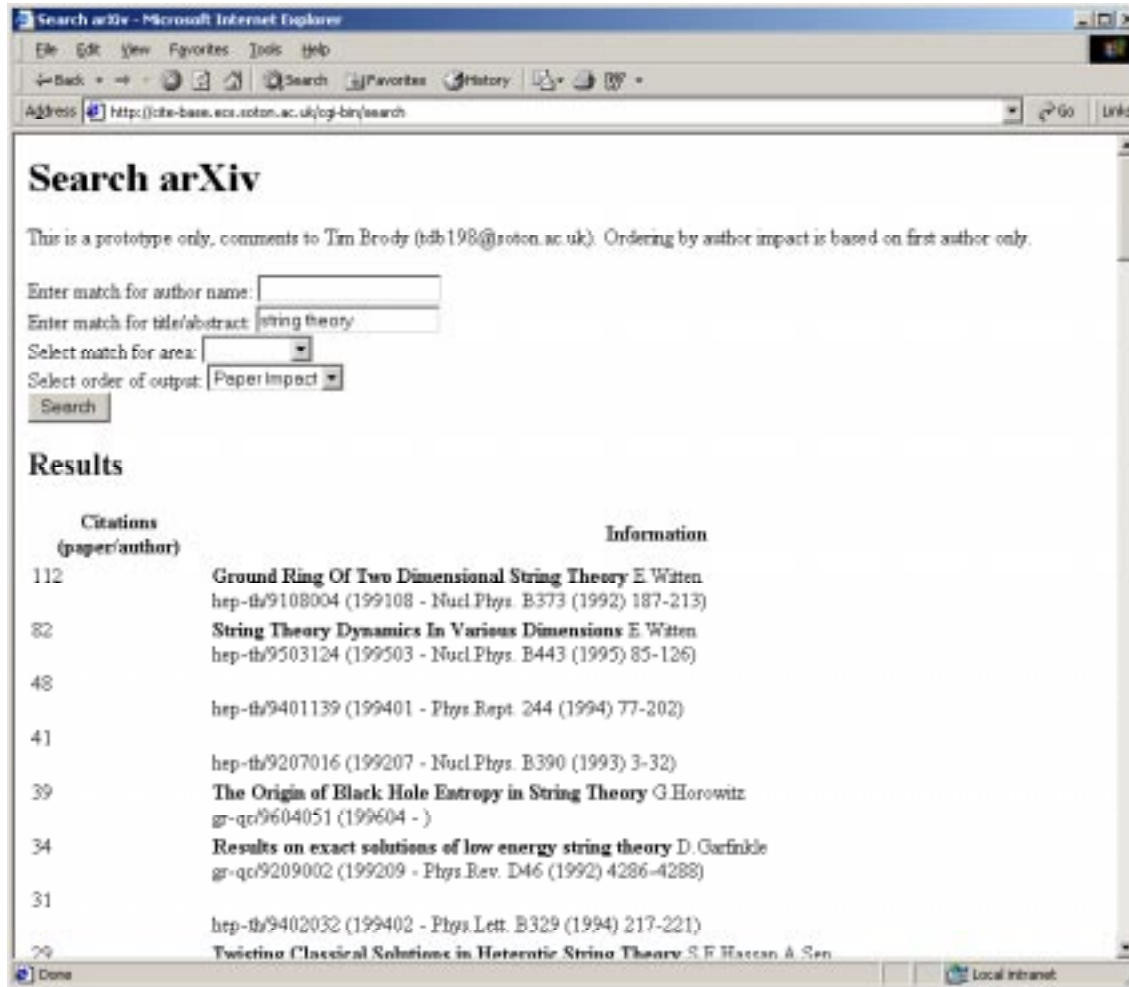
Multiple Updates by LANL Subfield
(based on LANL meta-data)

No. of Papers with Updates

Legend: No Updates | 1 Update | 2 Updates | 3 Updates | 4 Updates

Citation Linking & Scientometric Analysis

Resolving an Anomaly
www.eprints.org

Citation-Ranked Searches

Resolving an Anomaly
www.eprints.org

Citation-based Visualisation

ACM Hypertext Author Co-Citation Map (1989-1998)

Labels: Carr, Pearl, Whitehead, *Open Hypermedia*, Schnase, Egan, *Classics*, Lucarella, Wright, Trigg, Moulthrop, Joyce, Streitz, Halasz, Schwabe, *Hypertext Writing*, Card, Engelbart, Nelson, Nielsen, *Visualisation*, *Design Models*, *User Interface*, Salton, Crouch, *Structural Analysis*, *Information Retrieval*, Chen, Rao, van Rijsbergen
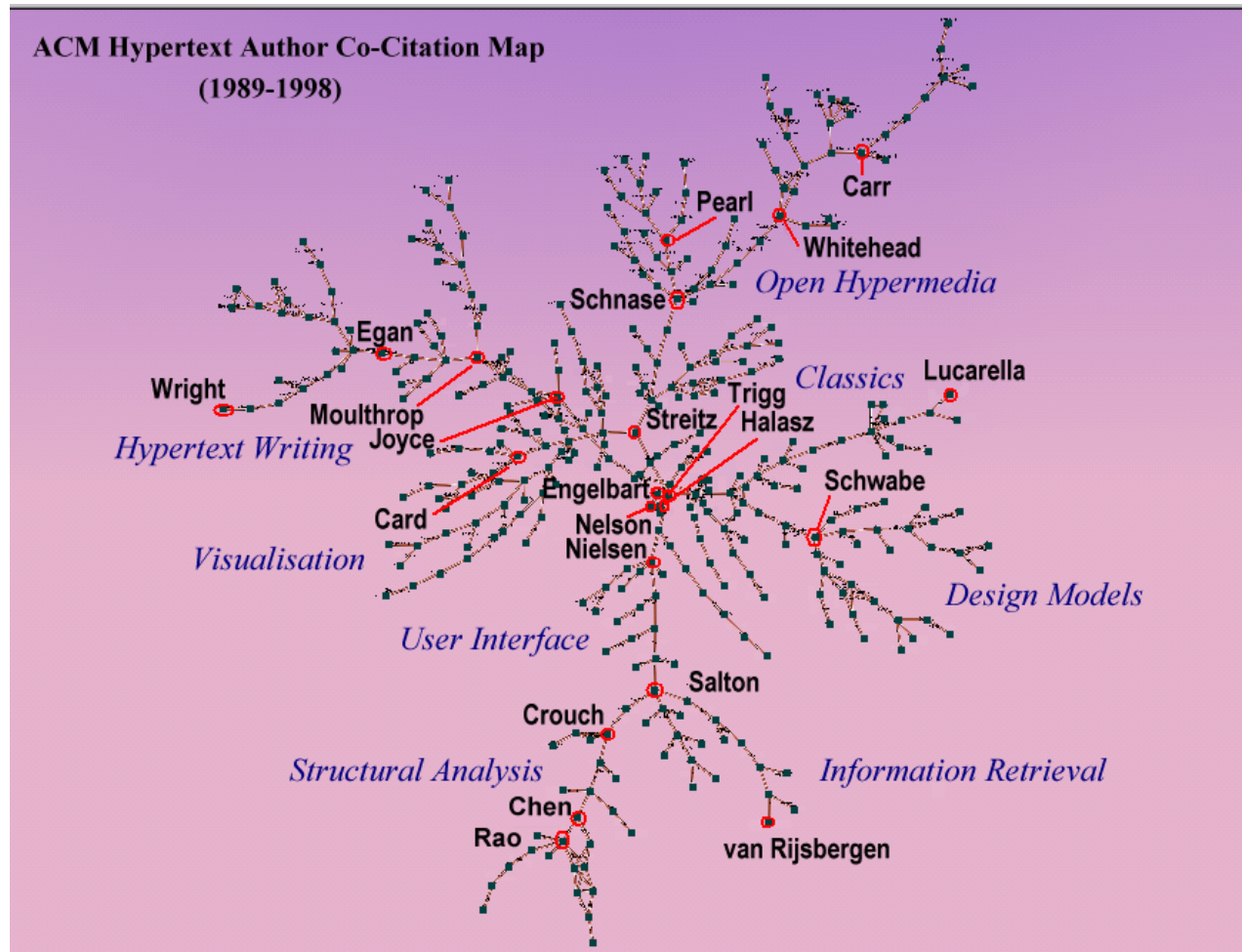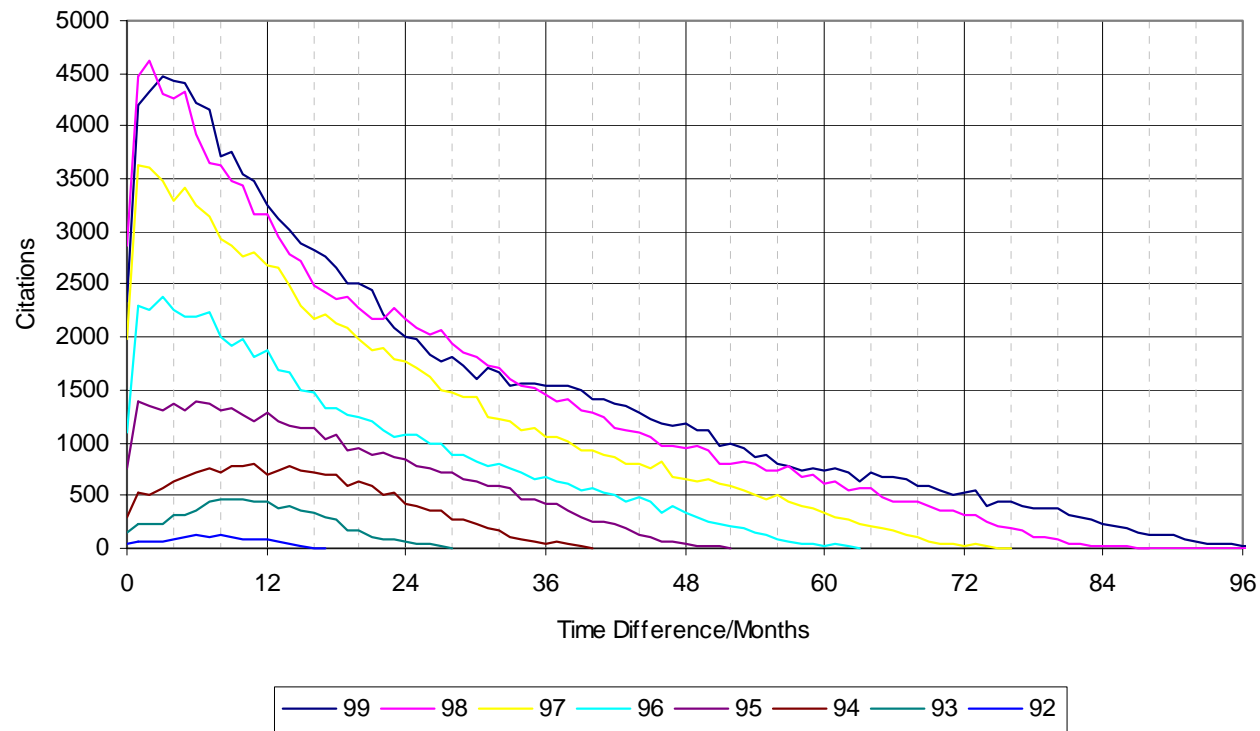
Resolving an Anomaly
www.eprints.org
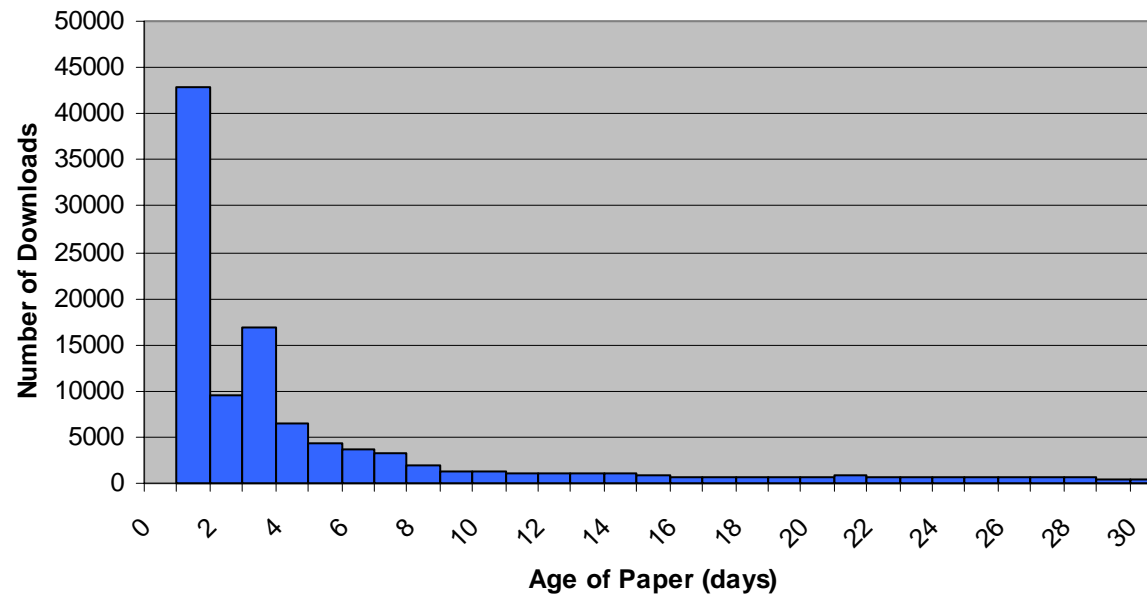
Decreasing Citation Latencies

Frequency of Citation Latencies: 1992-1999

- The raw data show that the latency of the citation peak has been reducing over the period of the archive

Resolving an Anomaly
www.eprints.org
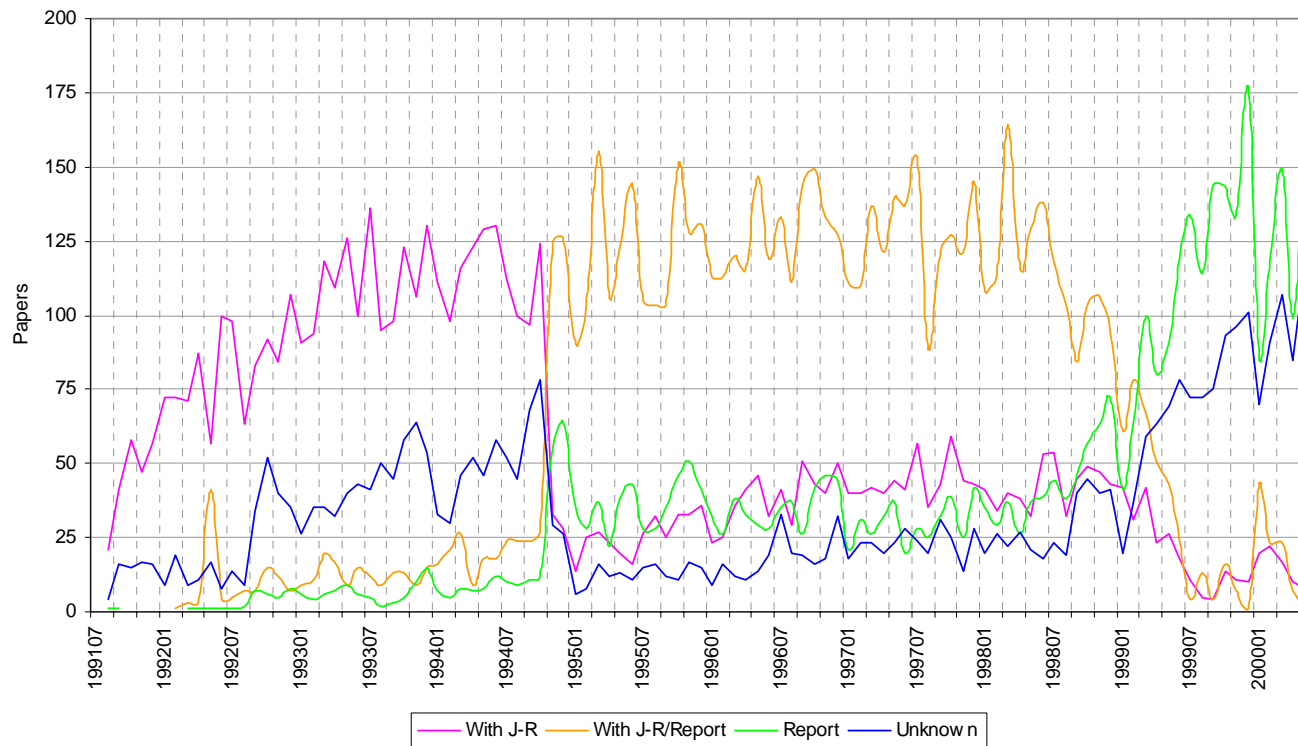
The "New Paper Rush"

Age of paper against number of downloads

- Users subscribe to an email alerting service that informs them of new papers.
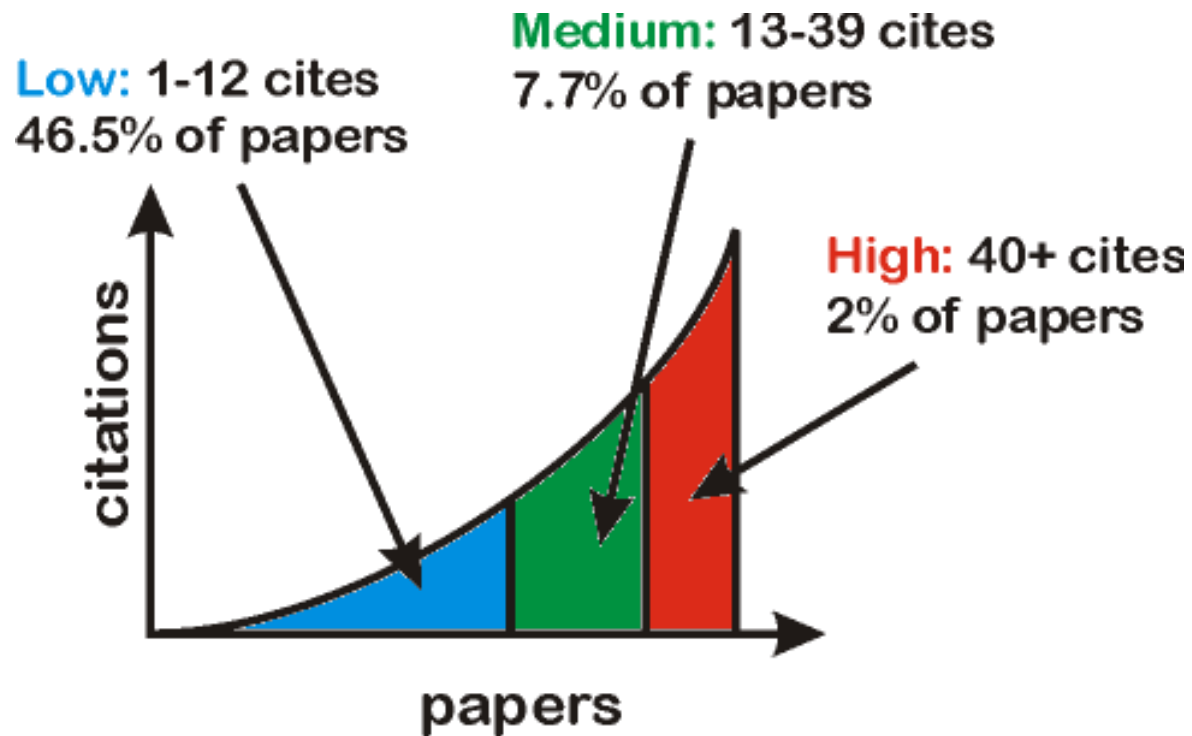
Article Embryology

hep-th

- Papers with a journal reference [J-R] cross papers without a J-R at an age of 13 months, suggesting a time difference of 13 months between pre-print and post-print

Resolving an Anomaly
www.eprints.org

Effect of Paper Impact

Low: 1-12 cites
46.5% of papers

Medium: 13-39 cites
7.7% of papers

High: 40+ cites
2% of papers

citations

papers

- The papers were split into three sets based on the number of citations to them.
- There are an equal number of citations to the papers in the low, medium and high sets.

Resolving an Anomaly
www.eprints.org

Author Impact Quartiles

Quartile	Total	% Total	Citations	Papers	Citations/Author/Paper	Deposits	Mean Updates/Author
High 25%	798	2.09%	240,092	2,732	0.11	6,720	0.48
Med 50%	9,262	24.20%	733,272	37,318	0.00212	93,671	0.37
Low 25%	28,211	73.71%	251,925	67,951	0.000131	165,971	0.27

- High impact authors update more than medium or low
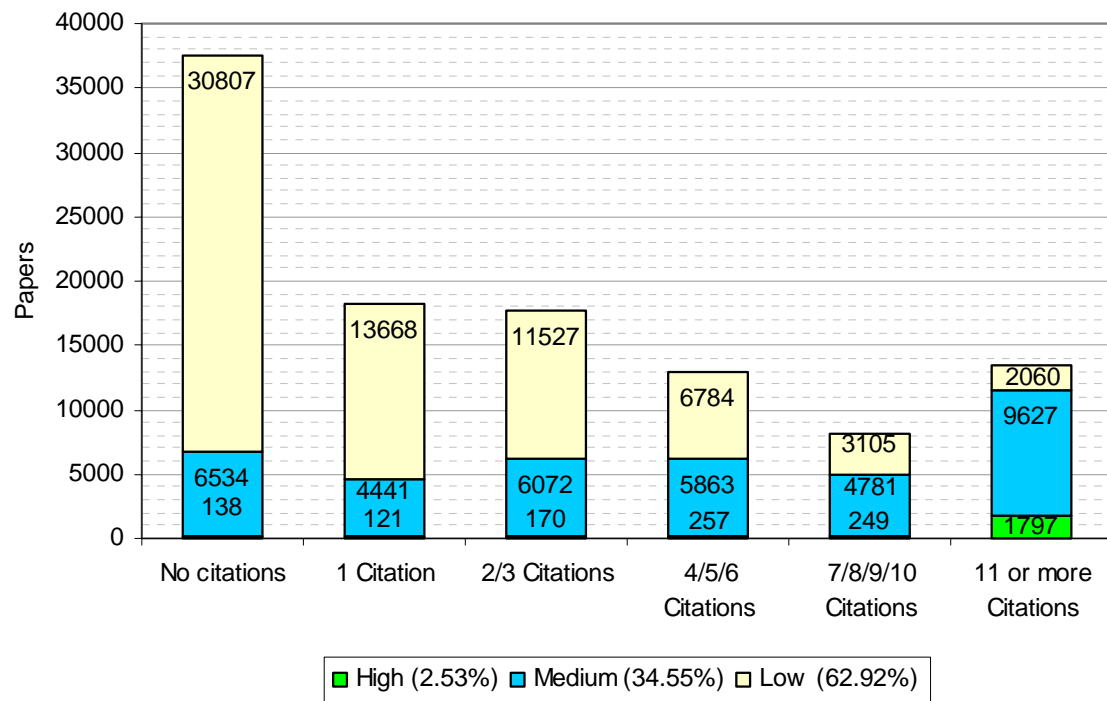- High and medium impact authors deposit more papers than low

Citation Quality

Do Papers Cite Papers of Like Impact
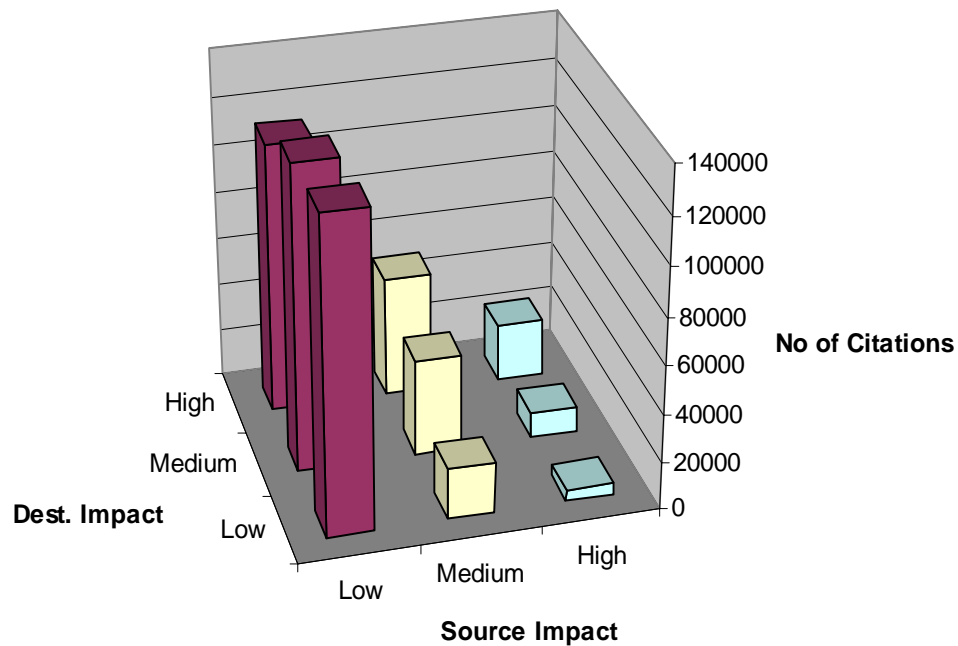
- Papers generally cite papers of like impact (χ^2 underway).
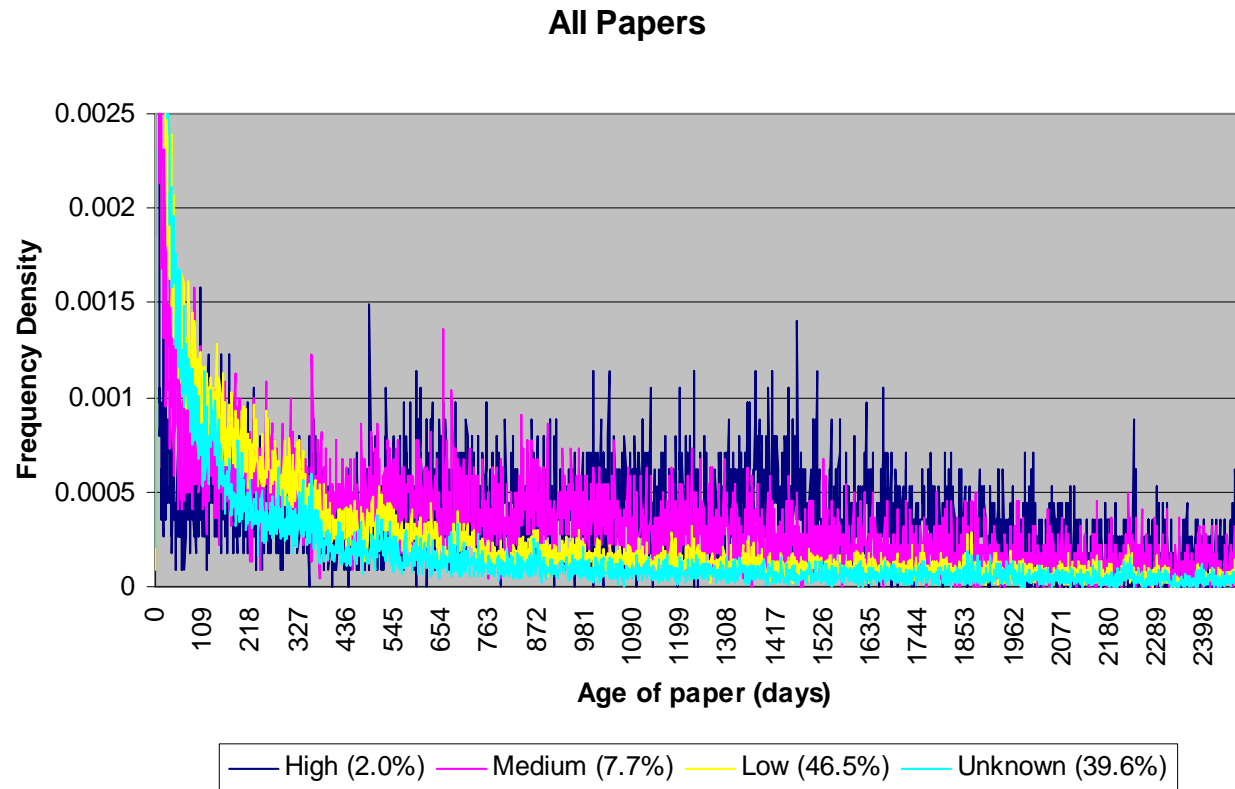
Resolving an Anomaly
www.eprints.org

Citation Spread

Histogram of Citations per Paper
(author impact) 30,000 papers were by authors with no citation

- A small number of papers receive a very large number of citations

Resolving an Anomaly
www.eprints.org

Effect of Paper Impact on Usage

All Papers

- Higher impact papers have a longer download life expectancy.

Resolving an Anomaly
www.eprints.org

Correlating citations and downloads

Download type	r	n
All Papers	0.11155	63671
High Impact Papers (2.0%)	0.27293	1981
Medium Impact Papers (7.7%)	0.01288	5937
Low Impact Papers (46.5%)	-0.01412	30163

- There is a significant positive correlation between citations and downloads for high impact papers.

Implementation Issues
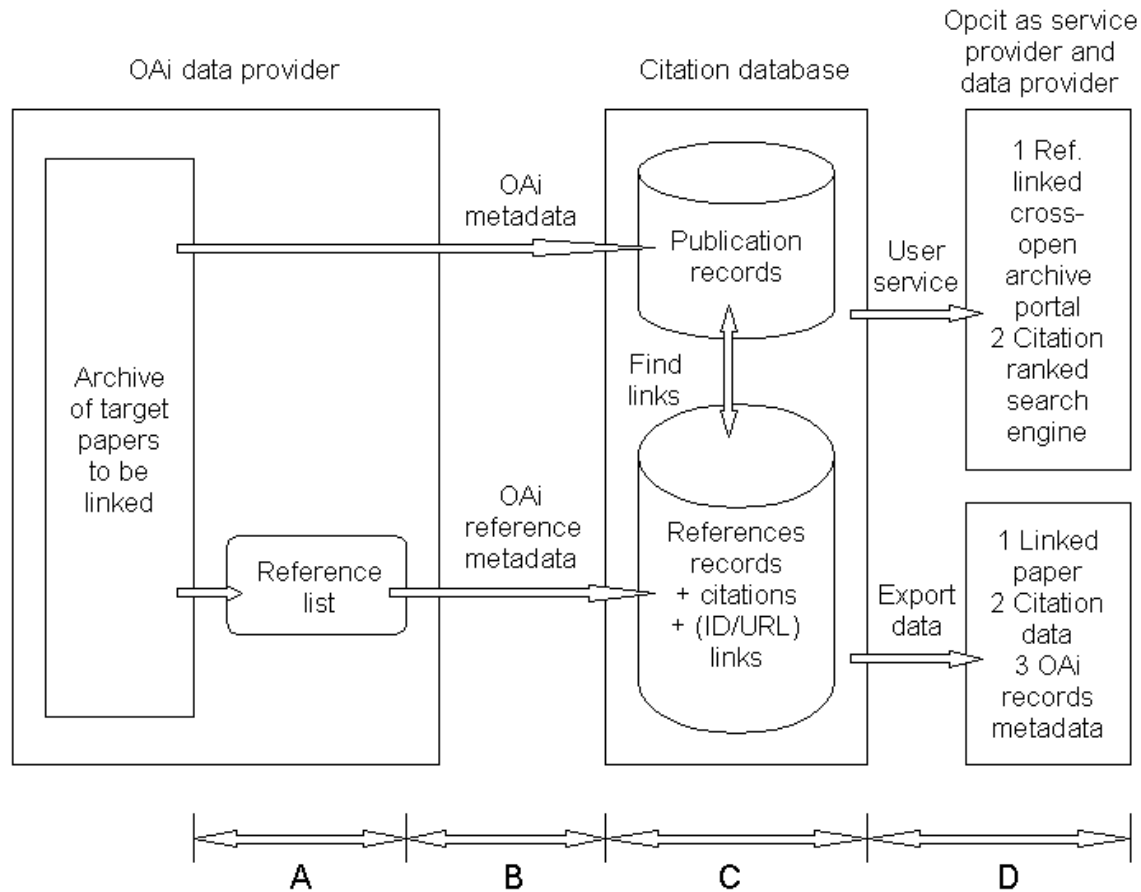
- Creating new metadata *vs*

- Creating new services

Resolving an Anomaly
www.eprints.org