Open Access: From Myth to Paradox

Paul Ginsparg

CIS and Physics, Cornell University

Abstract: True open access to scientific publications not only gives readers the possibility to read articles without paying subscription, but also makes the material available for automated ingestion and harvesting by 3rd parties. Once articles and associated data become universally treatable as computable objects, openly available to 3rd party aggregators and value-added services, what new services can we expect, and how will they change the way that researchers interact with their scholarly communications infrastructure? I will discuss straightforward applications of existing ideas and services, including citation analysis, collaborative filtering, external database linkages, interoperability, and other forms of automated markup, and speculate on the sociology of the next generation of users.

c.f. “Quarks: From Paradox to Myth”, K.G. Wilson (Cornell Univ), in Erice 1975, Proceedings, New Phenomena In Subnuclear Physics

Next-Generation Implications of Open Access

Scholarly Publication in the Sciences, what you've missed:

I. How did we get here?

II. Where are we?

III. Where are we going?

Next-Generation Implications of Open Access

Scholarly Publication in the Sciences, what you've missed:

I. How did we get here?

II. Where are we?

III. Where are we going?

NIH Public Access Policy Becomes Mandate in 2007

http://info-libraries.mit.edu/scholarly/open-access-initiatives/

On December 26, 2007, President Bush signed a spending bill that requires the US National Institutes of Health (NIH) to mandate open online access to all research it funds.

This is the first mandate for a major public funding agency in the US that requires research to be openly available; it changes the 2005 NIH Public Access Policy, which requested, but did not require, open access to NIH-funded research.

The new language stipulates that investigators funded by the NIH submit their peer-reviewed manuscripts to the National Library of Medicines open access repository PubMed Central when the manuscript is accepted for publication [additional > 70k/year]. The manuscript would then become openly available via PubMed Central within 12 months of publication in a journal. The policy will be implemented "in a manner consistent with copyright law."

Public Access Policy Made Permanent

Washington, D.C. March 12, 2009 — President Obama yesterday signed into law the 2009 Consolidated Appropriations Act, which includes a provision making the National Institutes' of Health (NIH) Public Access Policy permanent. The NIH Revised Policy on Enhancing Public Access requires eligible NIH-funded researchers to deposit electronic copies of their peer-reviewed manuscripts into the National Library of Medicines online archive, PubMed Central (PMC). Full texts of the articles are made publicly available and searchable online in PMC no later than 12 months after publication in a journal.

The NIH policy was previously implemented with a provision that was subject to annual renewal. Since the implementation of the revised policy the percentage of eligible manuscripts deposited into PMC has increased significantly, with over 3,000 new manuscripts being deposited each month. The PubMed Central database is a part of a valuable set of public database resources at the NIH, which are accessed by more than 2 million users each day.

Harvard To Collect, Disseminate Scholarly Articles For Faculty

http://www.fas.harvard.edu/home/news_and_events/releases/scholarly_02122008.html

Cambridge, Mass. - February 12, 2008 - In a move to disseminate faculty research and scholarship more broadly, the Harvard University Faculty of Arts and Sciences voted today to give the University a worldwide license to make each faculty member's scholarly articles available and to exercise the copyright in the articles, provided that the articles are not sold for a profit.

In proposing the legislation, Professor Stuart M. Shieber said, ". . . **scholarly journals have historically allowed scholars to distribute their research to audiences around the world. But, the scholarly publishing system has become far more restrictive than it need be. Many publishers will not even allow scholars to use and distribute their own work. And, the cost of journals has risen to such astronomical levels that many institutions and individuals have cancelled subscriptions, further reducing the circulation of scholars' works."**

MIT faculty open access to scholarly articles

http://web.mit.edu/newsoffice/2009/open-access-0320.html

Cambridge, Mass., March 20 2009 — In a move aimed at broadening access to MIT's research and scholarship, faculty at the Massachusetts Institute of Technology have voted to make their scholarly articles available to the public for free and open access on the Web.

Under the new policy, faculty authors give MIT nonexclusive permission to disseminate their journal articles for open access through DSpace, an open-source software platform developed by the MIT Libraries and Hewlett Packard and launched in 2002. . . . Authors may opt out on a paper-by-paper basis.

MIT's policy is the first faculty-driven, university-wide initiative of its kind in the United States. While Harvard and Stanford universities have implemented open access mandates at some of their schools, MIT is the first to fully implement the policy university-wide as a result of a faculty vote. MIT's resolution is built on similar language adopted by the Harvard Faculty of Arts & Sciences in 2008.

. . . potentially thousands of papers published by MIT faculty each year will be added to DSpace and made freely available on the web and accessible through search engines such as Google.

L'Europe veut ouvrir l'accès à ses articles scientifiques

[09/09/08]
http://www.lesechos.fr/info/innovation/4768682-l-europe-veut-ouvrir-l-acces-a-ses-articles-scientifiques.htm

La Commission envisage de mettre en ligne gratuitement les articles issus de certains projets européens.

Les scientifiques qui travailleront sur un projet financé dans le cadre du 7e PCRD (2007–2013) devront désormais lire soigneusement leur contrat. Une clause va en effet spécifier qu'un article scientifique écrit dans le cadre d'un projet européen devra être librement accessible sur Internet, aprés une période d'embargo de six à douze mois selon les secteurs. . . . "En outre, il s'agit pour le public d'un juste retour de la recherche financée par des fonds publics", insiste la Commission.

. . .

Archives ouvertes

. . .

Lui milite pour une ouverture plus rapide et radicale, sur le modèle de ce qui se pratique aux Etats-Unis avec la base ArXiv. Développée suivant le concept d'archives ouvertes (imaginé par le physicien américain Paul Ginsparg), elle met en ligne gratuitement tous les articles scientifiques dès leur publication.

. . .

Qui va payer ?

. . . d'ailleurs un appel d'offres pour créer un portail unique. . . . Un projet bien long aux yeux: "On va dépenser beaucoup d'argent pour arriver à des résultats qu'ArXiv et Hal font depuis longtemps."

Not always Binding

(proposed May 2006, get faculty buy-in? c.f. dspace)

WHEREAS

the Cornell Faculty Senate on 11 May 2005 passed a resolution on scholarly publishing, according to which "The Senate strongly urges all faculty to negotiate with the journals in which they publish either to retain copyright rights and transfer only the right of first print and electronic publication, or to retain at a minimum the right of postprint archiving"; and

. . .

THEREFORE BE IT RESOLVED THAT

The Senate urges faculty members to attach the SPARC Authors Addendum to publishing contracts that they sign unless they arrange to retain copyright itself and transfer only the right of first print and electronic publication.

SPARC Author's Addendum to Publication Agreement

http://www.arl.org/sparc/author/addendum.html

1. **Authors Retention of Rights.** In addition to any rights under copyright retained by Author in the Publication Agreement, Author retains:

(i) the rights to reproduce, distribute, publicly perform, and publicly display the Article in any medium for non-commercial purposes;

(ii) the right to prepare derivative works from the Article; and

(iii) the right to authorize others to make any non-commercial use of the Article so long as Author receives credit as author and the journal in which the Article has been published is cited as the source of first publication of the Article. For example, Author may make and distribute copies in the course of teaching and research and may post the Article on personal or institutional Websites and in other open-access digital repositories.

2. **Publishers Additional Commitments.** Publisher agrees to provide to Author within 14 days of first publication and at no charge an electronic copy of the published Article in Adobe Acrobat Portable Document Format (.pdf). The Security Settings for such copy shall be set to NoSecurity.

I. How did we get here?

Personal Prehistory

- **1969: 100 baud teletype, paper tape**
- **1971: keypunch cards**
- **1973: e-mail**
- **1981: thesis typed**
- **1982: more e-mail, still no spam, global village problem**
- **1984: TeXnical typesetting**
- **1985–1988: Harvard gets wired; email addresses common**
- **1989: critical mass**
- **1990: phototropism $\longrightarrow$ 25 MHz CPU,105 Mb hd, 16 Mb RAM**
- **1991: hep-th solves the “Harvard preprint problem”**

PUPT-1084
SLAC-PUB-4515
HUTP-87/A085

$\hat{c} = 1$ Superconformal Field Theory

L. Dixon[1], P. Ginsparg[2], and J. Harvey[3]

[1,3]Physics Dept.
Princeton University
Princeton, N.J. 08544

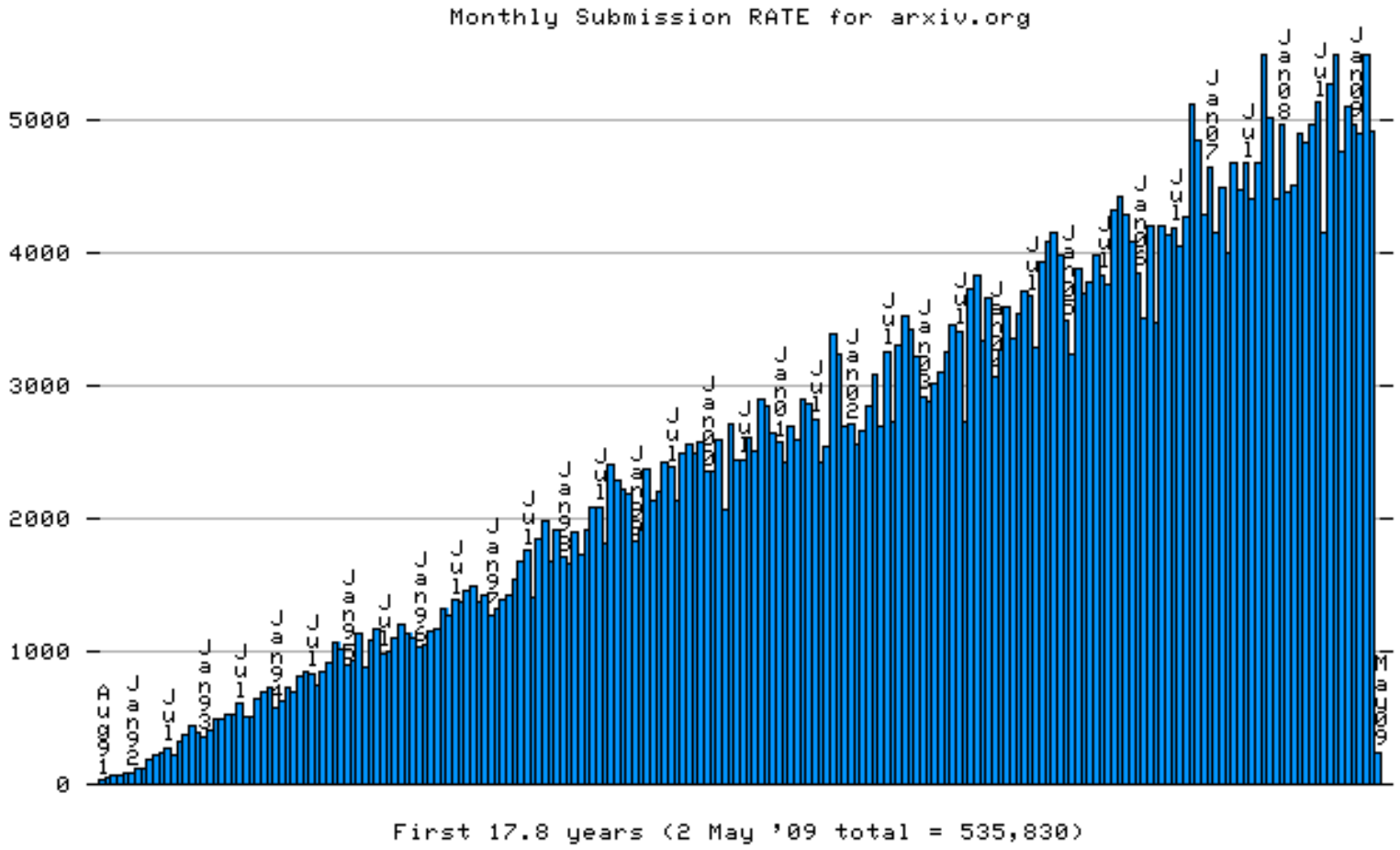
[2]Stanford Linear Accelerator Center
Stanford, CA 94305

We consider superconformal field theories with central charge $\hat{c} = \frac{2}{3}c = 1$. We find five continuous one-parameter families of theories all interconnected via a set of multicritical points that are reached by modding out theories with enlarged symmetries. We find as well 6 theories that have no integrable marginal operators and thus constitute isolated points of superconformal invariance in the $\hat{c} = 1$ moduli space. We briefly discuss $c = 3/2$ conformal theories that contain a twisted superconformal algebra, including 3 isolated theories with a twisted N=3 superconformal algebra, and theories constructed as the tensor product of the $c = 4/5$ and $c = 7/10$ minimal theories.

~1/88 *(submitted to Nucl. Phys. B)*

[1] (ljd@pupthy.princeton.edu or ljd@pucc.bitnet)
[2]Current address: Lyman Lab. of Physics, Harvard Univ., Cambridge, MA 02138
(ginsparg@huhepl.hepnet, ginsparg@huhepl.bitnet, or ginsparg@huhepl.harvard.edu)
[3] (jah@pupthy.princeton.edu or jah@pucc.bitnet)

Submissions per month, ’91 – ’08



Total $\gtrsim$ 535,000 (May 2009)

". . .We have mentioned that scientists themselves create elements to fulfill information needs that are not being satisfied by existing media. These newly created elements affect other elements in the system by changing the scientist's information-seeking and information-disseminating behavior. . . .

The chain of events, in a fast-moving research area, may begin with publication lag becoming so great that current information needs are unsatisfied. As a result, the exchange of preprints among scientists working in this area will increase. At some point the exchange of preprints becomes unmanageable on an individual basis and it becomes necessary to organize a more formal preprint-exchange mechanism. Often this new mechanism is a preprint-exchange group, organized by an elite few concerned with a single specialty, who invite other active researchers in the field to join the group. As this information medium grows it takes on more and more of the attributes of its formal counterpart — the scientific journal — and it begins in many ways to serve as a substitute for the journal. . . . some of the practices associated with the traditional formal media are adopted by the members of the group. For example, within the group strict enforcement of priority of information disseminated by way of preprint exchange may be established. This process of formalization may continue to evolve until someone realizes that an institution has emerged which has most of the characteristics of an archival journal: a large and increasing input of manuscripts, an existing gatekeeping group, an eager and expanding audience, and growing economic problems. And thus a new journal – and possibly a scientific society – is born."

More Prehistory (1967)

excerpted from section on "Social Dimensions", p. 1012, from:

"Scientific Communication as a Social System",
William D. Garvey[1] and Belver C. Griffith[2],
Science, New Series, Vol. 157, no. 3792, pp. 1011-1016 (1 Sep 1967)

(in turn adapted from an address delivered at "Communication in Science: Documentation and Automation" symposium, sponsored by the Ciba Foundation, London, 22 Nov 1966)

[1] Professor of psychology and director of the Center for Research in Scientific Communication, Johns Hopkins University

[2] Director of the Project on Scientific Information Exchange in Psychology of the American Psychological Association, and associate in communications, Annenberg School of Communications, University of Pennsylvania

arXiv.org e-Print archive

Automated e-print archives [physics] Search | Form Interface | Catchup | Help

11 Nov 2004: New CoRR interface introduced for our cs users.
29 Sep 2004: Search engine for user help pages installed.
For more info, see cumulative "What's New" pages.
Robots Beware: indiscriminate automated downloads from this site are *not* permitted.

Physics

- Astrophysics (**astro-ph** new, recent, abs, find)
- Condensed Matter (**cond-mat** new, recent, abs, find)
  includes: Disordered Systems and Neural Networks; Materials Science; Mesoscopic Systems and Quantum Hall Effect; Other; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, abs, find)
- High Energy Physics - Experiment (**hep-ex** new, recent, abs, find)
- High Energy Physics - Lattice (**hep-lat** new, recent, abs, find)
- High Energy Physics - Phenomenology (**hep-ph** new, recent, abs, find)
- High Energy Physics - Theory (**hep-th** new, recent, abs, find)
- Mathematical Physics (**math-ph** new, recent, abs, find)
- Nuclear Experiment (**nucl-ex** new, recent, abs, find)
- Nuclear Theory (**nucl-th** new, recent, abs, find)
- Physics (**physics** new, recent, abs, find)
  includes (see detailed description): Accelerator Physics; Atmospheric and Oceanic Physics; Atomic Physics; Atomic and Molecular Clusters; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Physics and Society; Plasma Physics; Popular Physics; Space Physics
- Quantum Physics (**quant-ph** new, recent, abs, find)

Mathematics

- Mathematics (**math** new, recent, abs, find)
  includes (see detailed description): Algebraic Geometry; Algebraic Topology; Analysis of PDEs; Category Theory; Classical Analysis and ODEs; Combinatorics; Commutative Algebra; Complex Variables; Differential Geometry; Dynamical Systems; Functional Analysis; General Mathematics; General Topology; Geometric Topology; Group Theory; History and Overview; K-Theory and Homology; Logic; Mathematical Physics; Metric Geometry; Number Theory; Numerical Analysis; Operator Algebras; Optimization and Control; Probability; Quantum Algebra; Representation Theory; Rings and Algebras; Spectral Theory; Statistics; Symplectic Geometry

Nonlinear Sciences

- Nonlinear Sciences (**nlin** new, recent, abs, find)
  includes (see detailed description): Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons

Computer Science

- Computing Research Repository (**CoRR** new, recent, abs, find)
  includes (see detailed description): Architecture; Artificial Intelligence; Computation and Language; Computational Complexity; Computational Engineering, Finance, and Science; Computational Geometry; Computer Science and Game Theory; Computer Vision and Pattern Recognition; Computers and Society; Cryptography and Security; Data Structures and Algorithms; Databases; Digital Libraries; Discrete Mathematics; Distributed, Parallel, and Cluster Computing; General Literature; Graphics; Human-Computer Interaction; Information Retrieval; Information Theory; Learning; Logic in Computer Science; Mathematical Software; Multiagent Systems; Multimedia; Networking and Internet Architecture; Neural and Evolutionary Computing; Numerical Analysis; Operating Systems; Other; Performance; Programming Languages; Robotics; Software Engineering; Sound; Symbolic Computation

Quantitative Biology

- Quantitative Biology (**q-bio** new, recent, abs, find)
  includes (see detailed description): Biomolecules; Cell Behavior; Genomics; Molecular Networks; Neurons and Cognition; Other; Populations and Evolution; Quantitative Methods; Subcellular Processes; Tissues and Organs

---

About arXiv

- some related and unrelated servers (including arXiv **mirror** sites)
- RSS feeds are now available for individual archives and categories.
- today's usage for arXiv.org (not including mirrors)
- some info on delivery type [src] and potential problems
- arXiv Advisory Board
- available macros and brief description
- available help on submitting and retrieving papers
- some background blurb, including invited talk at UNESCO HQ (Paris, 21 Feb '96), update Sep '96
- some info on hypertex

Cornell University Library

arXiv is an e-print service in the fields of physics, mathematics, non-linear science, computer science, and quantitative biology. The contents of arXiv conform to Cornell University academic standards. arXiv is owned, operated and funded by Cornell University, a private not-for-profit educational institution. arXiv is also partially funded by the National Science Foundation.

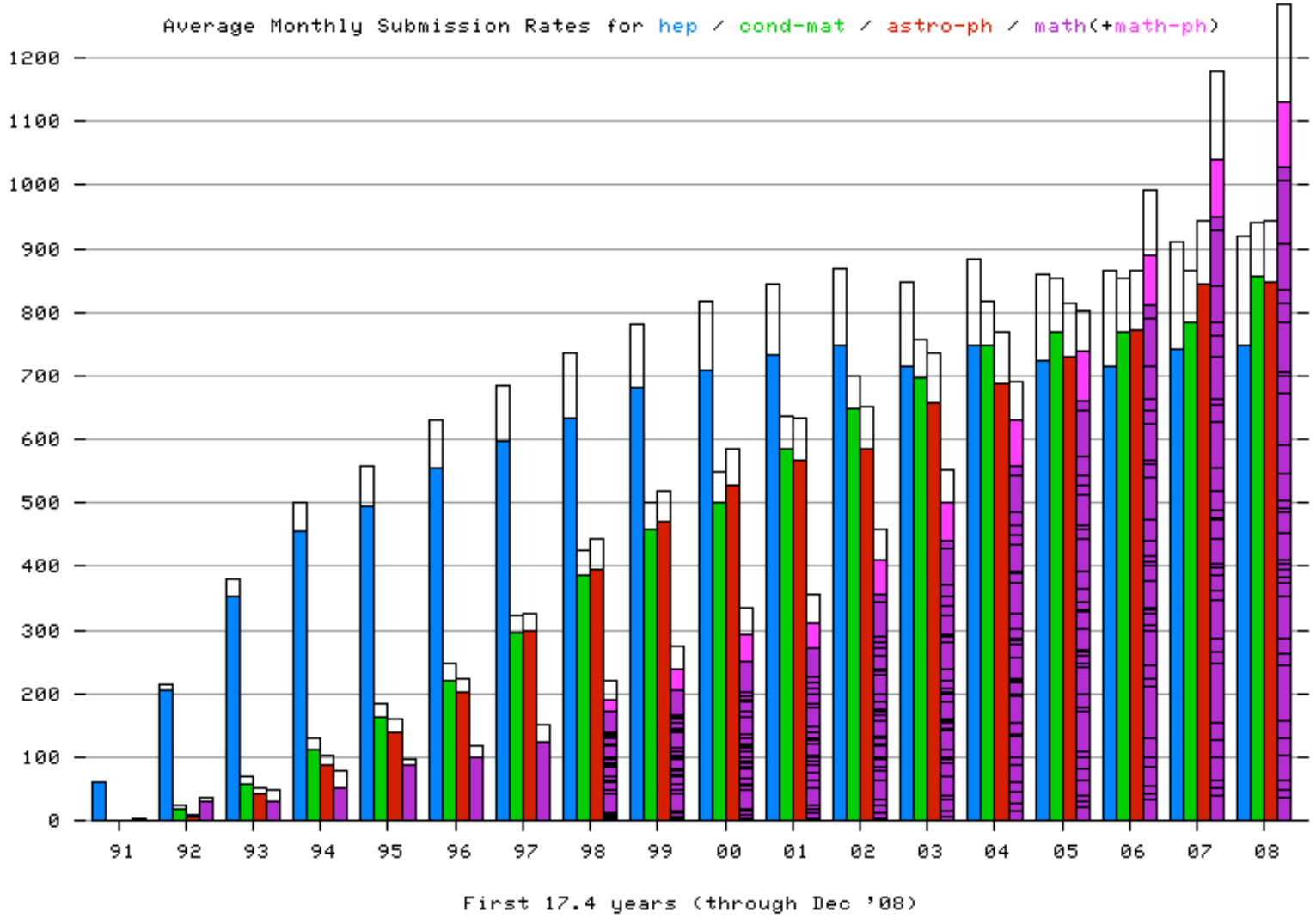
The Cornell University Library acknowledges the support of Sun Microsystems and U.S. Department of Energy's Office of Scientific and Technical Information (providers of the E-Print Alert Service, which automatically notifies users of the latest information posted on arXiv and other related databases).

www-admin@arxiv.org

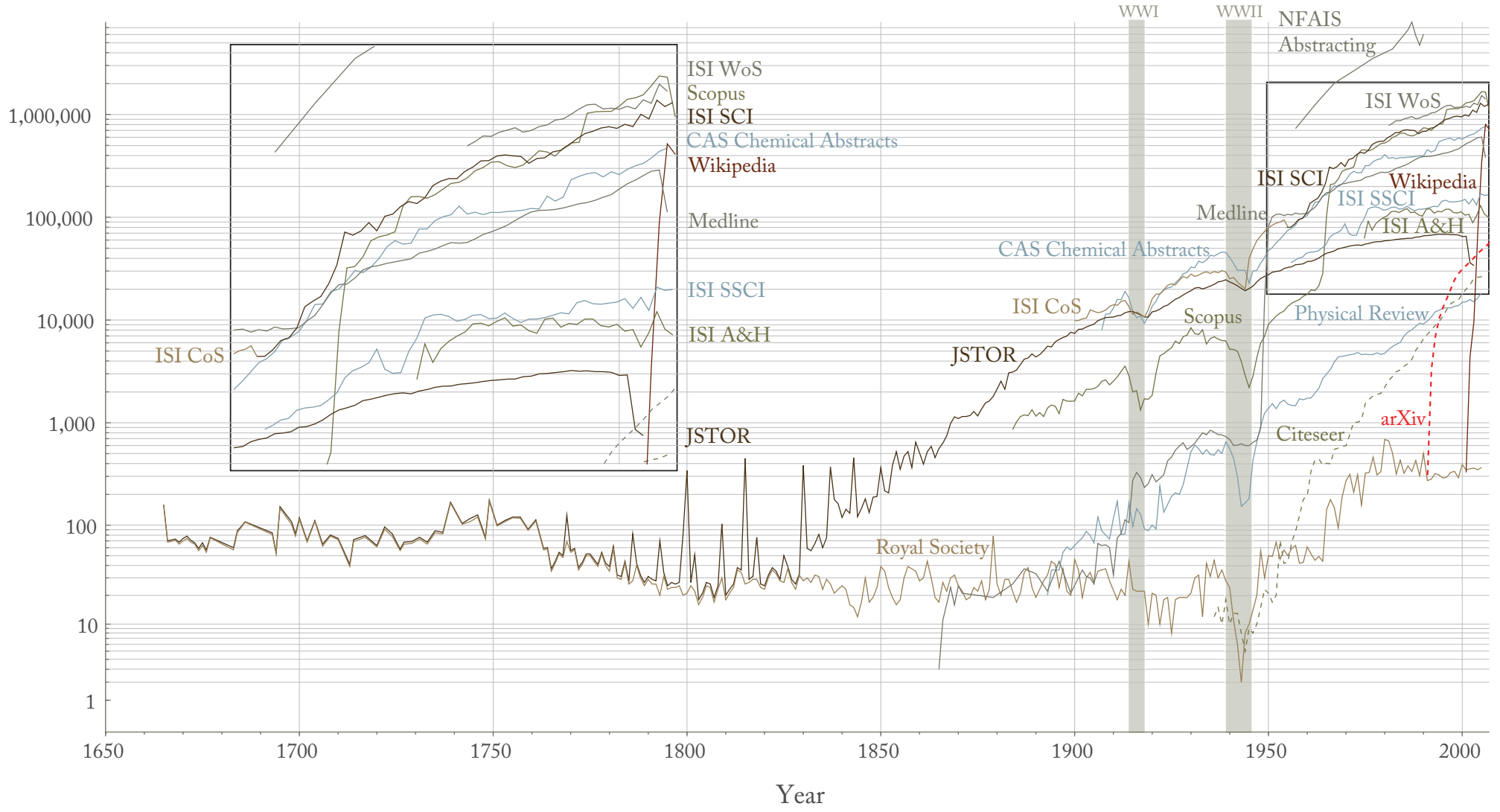
arXiv.org

- **e-mail interface started August 1991**
  - **download data available from start**
  - **WWW usage logs starting from 1993**
- **$\sim$535,000 full text documents (with full graphics), early May 2009**
  - **physics, mathematics, q-bio, non-linear, computer science**
  - **growing at 64,000 new submissions per year (est. 2009 $\Rightarrow$ $>$ 580,000 at end of year)**
  - **20 references per article (over 11 million total)**
- **over 55 million full text downloads during calendar year '08**
  - **over 600 downloads per article from '96-'07 ($>$250M total)**
- **overall: 15k ingested links (5.5k urls) to 10k articles (1.9% of 535k)**
  '08: 3.2k ingested links (1.6k urls) to 2.3k articles (4% of 59k)
  '09 (so far): 1.1k ingested links (.5k urls) to .9k articles (4.4% of 20.8k)
- **hydrophilia: Now managed by CU library (starting roughly 2001)**

Top four subject areas

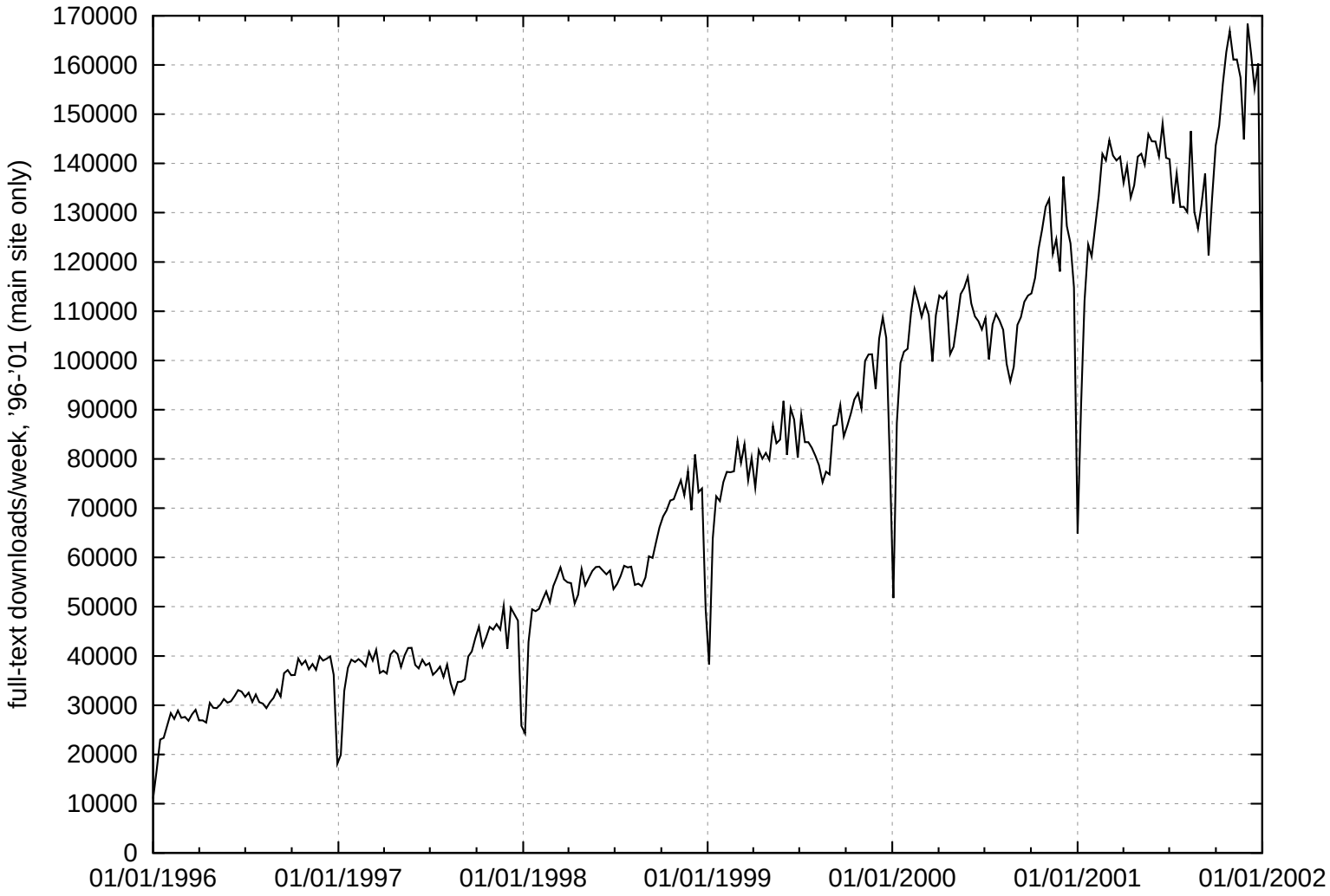


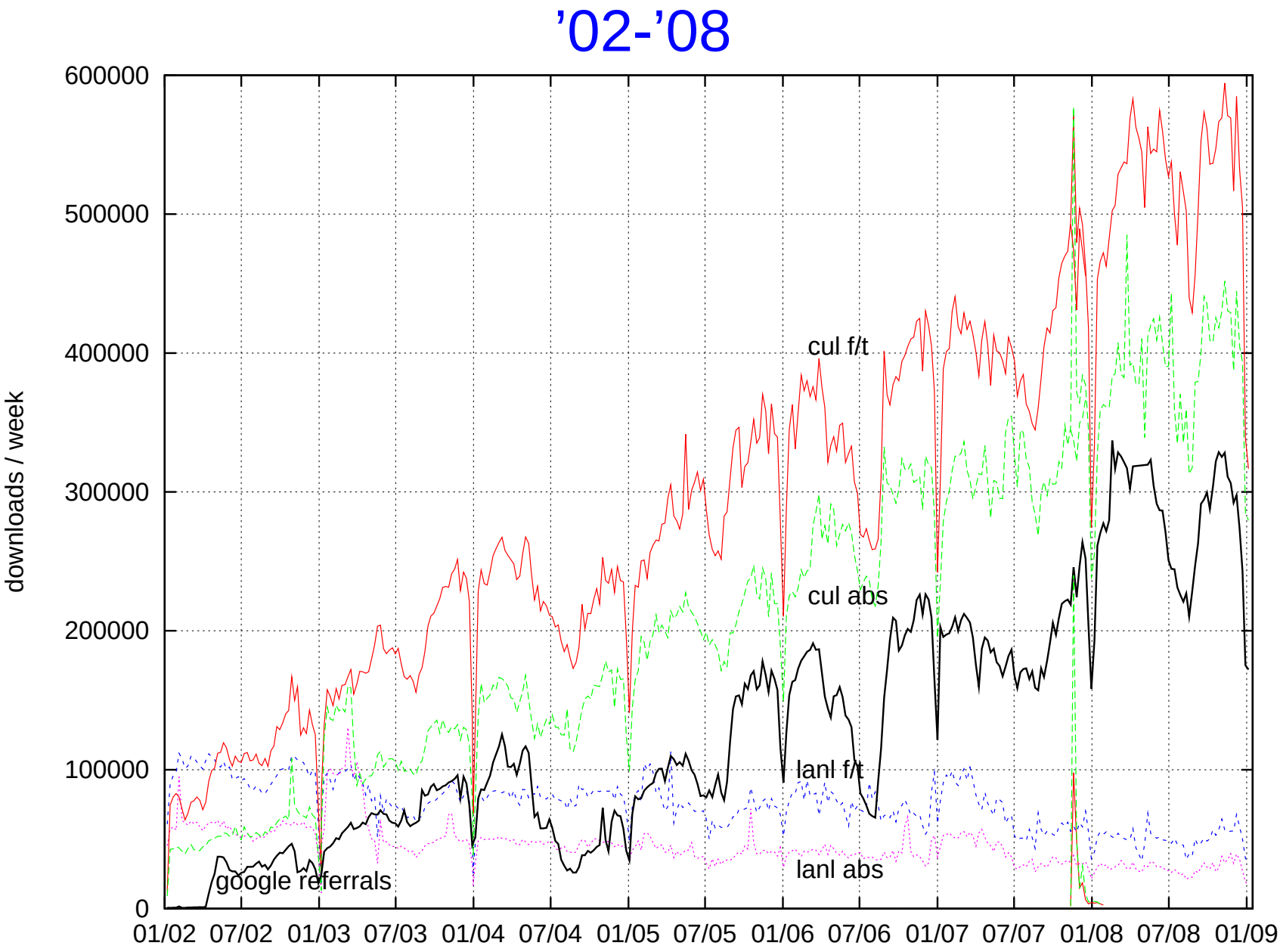
Papers & Wikipedia Entries



"Atlas of Science: Guiding the Navigation and Management of Scholarly Knowledge",
Part I: The Rise of Science and Technology. (2009)
Chart showing the number of papers/wikipedia entries for different databases and publication years.
Contact Katy Börner <katy@indiana.edu> or Elisha Hardy <efhardy@indiana.edu> for details.

’96-’01


full-text downloads/week, '96-'01 (main site only)
170000
160000
150000
140000
130000
120000
110000
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
01/01/1996
01/01/1997
01/01/1998
01/01/1999
01/01/2000
01/01/2001
01/01/2002


’02-’08
downloads / week
600000
500000
400000
300000
200000
100000
0
01/02
07/02
01/03
07/03
01/04
07/04
01/05
07/05
01/06
07/06
01/07
07/07
01/08
07/08
01/09
cul f/t
cul abs
lanl f/t
lanl abs
google referrals

Study of multi-muon events produced in p-pbar collisions at sqrt(s)=1.96 TeV

Authors: **CDF Collaboration**
Comments: 70 pages, 46 figures, 11 tables. Submitted to Phys. Rev. D
Report-no: FERMILAB-PUB-08-046-E

Subj-class: High Energy Physics - Experiment (hep-ex)

Blog/News Links for 0810.5357:

1. Not Even Wrong » Blog Archive » Discovery of a New Particle? [Not Even Wrong @ www.math.columbia.edu/~woit] [Thu Oct 30 22:23:00 2008]

2. CDF publishes multi-muons!!!! « A Quantum Diaries Survivor [A Quantum Diaries Survivor @ dorigo.wordpress.com/2008] [Fri Oct 31 22:30:00 2008]

3. Uncertain Principles: Fermilab Discovers... Something. Maybe. [scienceblogs.com/principles] [Fri Oct 31 08:52:00 2008]

4. RESONAANCES: On CDF Anomaly [resonaances.blogspot.com/2008] [Fri Oct 31 18:52:00 2008]

5. CDF Ghost Muons | Cosmic Variance [cosmicvariance.com/2008] [Sun Nov 2 19:50:01 2008]

6. News from the CDF and PAMELA experiments [Theorema Egregium @ egregium.wordpress.com/2008] [Fri Oct 31 19:41:00 2008]

7. Fermilab 'ghosts' hint at new particles - physicsworld.com [Physics World @ physicsworld.com/cws] [Mon Nov 3 20:52:09 2008]

8. At Tevatron, spectral particles spook Fermilab physicists [Physics Today News Picks @ blogs.physicstoday.org/newspic...] [Tue Nov 4 10:59:03 2008]

9. Interpretation of multi-muons! « A Quantum Diaries Survivor [dorigo.wordpress.com/2008] [Wed Nov 5 11:13:46 2008]

10. Nima Arkani-Hamed's letter on multi-muons - and my reply [A Quantum Diaries Survivor @ dorigo.wordpress.com/2008] [Wed Nov 5 17:11:31 2008]

11. CDF multi-muons, hmm « Charm &c. [superweak.wordpress.com/2008] [Wed Nov 5 11:14:47 2008]

12. Fermilab's CDF Result Sparks Rumors of New Physics [physorg.com @ www.physorg.com/news145029766....] [Wed Nov 5 11:16:07 2008]

13. Ghost in the Machine? Physicists May Have Detected a New Particle at Fermilab | Discover Magazine [Discover @ blogs.discovermagazine.com/80b...] [Wed Nov 5 21:56:04 2008]

II. Where are we?

Fantasy

Reflect for a moment

Current practice:

- **free access articles, background material from authors, slide presentations, video, related software, on-line animations, blog discussions, 3rd party notes, microblogged seminars, captured video feed, random factoids, collective wiki-exegesis**
- **course websites, e-mail, course blogs, wiki for notes**

New expectations (harvest all related, activity maps, concept browse).

Collapse internet resources to subset of unique ideas, authenticated.

Marketplace for preresearch barter of tools, resources, capabilities.

Authoring tools.

New generation of users.

So what's wrong with this picture?

NIH Public Access Policy Becomes Mandate in 2007

On December 26, 2007, President Bush signed a spending bill that requires the US National Institutes of Health (NIH) to mandate open online access to all research it funds. . . . The policy will be implemented "in a manner consistent with copyright law."

Harvard To Collect, Disseminate Scholarly Articles For Faculty

Cambridge, Mass. - February 12, 2008 - In a move to disseminate faculty research and scholarship more broadly, the Harvard University Faculty of Arts and Sciences voted today to give the University a worldwide license to make each faculty member's scholarly articles available and to exercise the copyright in the articles . . .

Web 2.0?

Congress Passes Law Requiring Users to Post to Youtube, Flickr, ...

Open Access (OA)

- **inevitable? possible? sensible? promising? threatening?**
- **OA "supports the principle that the published output of scientific research should be available, without charge, to everyone" (UK House of Commons Science and Technology Committee, 2004)**
- **self-evident from public policy standpoint? $\Rightarrow$ legislated?**
- **endorsed by Nobel laureates, library associations, and US Chamber of Commerce.**
- **published research: share knowledge + author recognition + some specious arguments**

OA $\neq$ “free access”

- **OA: authors can retain copyright and give license under to permit future uses (frequently prohibited when copyright transferred)**
- **OA: can be deposited in central server, available in searchable “information space” in perpetutity**
- **OA: Any third party can aggregate and datamine, articles treated as arbitrarily computable objects, linkable and interoperable with associated databases**

Tautology

- costs real money to do quality control via time-honored methodology
- how much varies from publisher to publisher
    - so does the profit
- to persist in that methodology, must keep the funds flowing:
    - if not to the usual suspects, different set of suspects
    - if not at all, then different methodology for quality control and authentication (scalable / sustainable)

Two ways to reduce the overall amount of money flowing into the journal publication system:

a) reduce the average profit margin

b) reduce the average cost of publication (as opposed to revenue)

Finances

- **globally $8B/year for 1.5–2M STM articles/year**

  $\Rightarrow\sim$$4500/article aggregate revenue (researchers unaware)

- **Large hierarchies in revenues ($1k – $15k / article)**

- **and large hierarchies in costs (Jul 04 data):**

  - **APS: editorial = $1000/ published article, + production = minumum $1800/article**
  - **science= $12000, nature = $18000, ACS = $2500**
  - **PNAS: 1/6 acceptance rate, $3600/article, $2800 w/o print**
  - **J Cell Biology = $8000/ published article, 15–20% acceptance rate (just editorial and production, not print)**
  - **selective journals cost more to produce?**
  - **Blume: more peremptory editorial rejection to reduce costs**

Will OA reduce costs? or just shift point at which funds enter system?

Are all disciplines created equal?

OA costs $<$ 1% of research budget?

NIH: $\sim$60,000 NIH funded articles, research budget $\sim$\$20B

$\Rightarrow$ public funding $>$ \$300k/article

Typical "well-funded" discipline:

Theoretical HEP: DOE + NSF funding $<$ \$40M/year,

$>$ few thousand articles / year (primary US authors)

$\Rightarrow$ public funding $<$ \$20k/article

And the rest . . . ?

(e.g. J. Ewing: $>$ 2/3 of mathematicians have no grant funding at all)

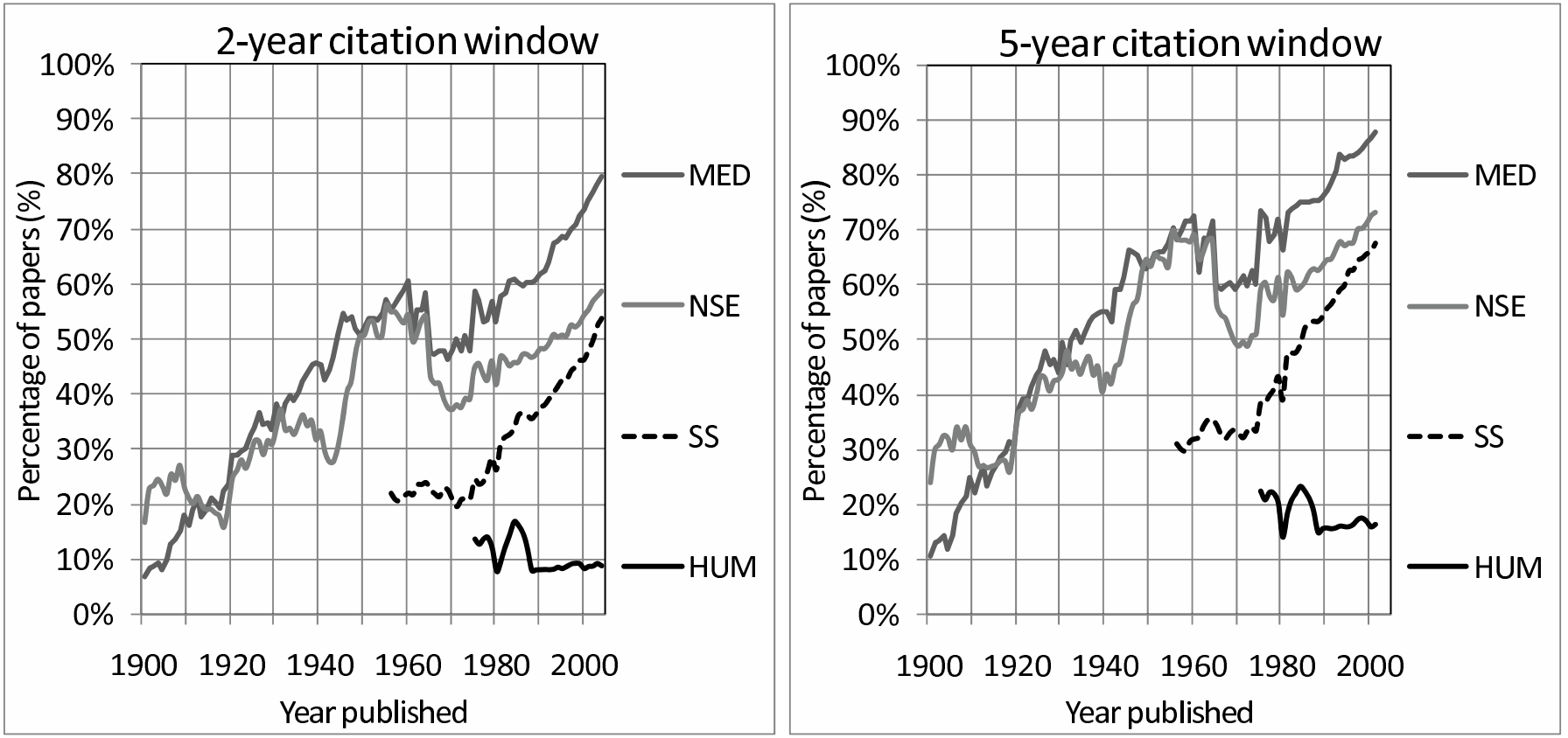


Figure 1. Percentage of papers that received at least one citation, two- and five-year citation windows, by field, 1900–2005 and 1900–2002

From "The decline in the concentration of citations, 1900-2007",

V. Lariviere, Y. Gingras, E. Archambault [arXiv:0809.5250]

Backdoor Route to Open Access

- **More than one-third of the high-impact journal articles in a sample of biological/medical journals published in 2003 were found at nonjournal Web sites (Wren, 2005).**
- **Unsystematic (Ginsparg, 2006, "As We May Read"): 75% of publications from 2000 or later posted at web site of incoming president of Society for Neuroscience available without subscription (preprints, open-access journal sites, copies at nonjournal web sites).**
  Perhaps already farther along than most realize?
- **Expectations of next generation independent of outcome of government mandate debate**

Past Confusion

Still no wysiwig?

Metastable co-existence state?

Efficacy of search engines?

Other fields? (not just information processing...)

Wikipedia?

Caution: new developments no longer academic-centric

Present Confusion

More than a new means of distribution?

Crippled by document format? (TeX, Word $\longrightarrow$ PDF, 70's methodology)

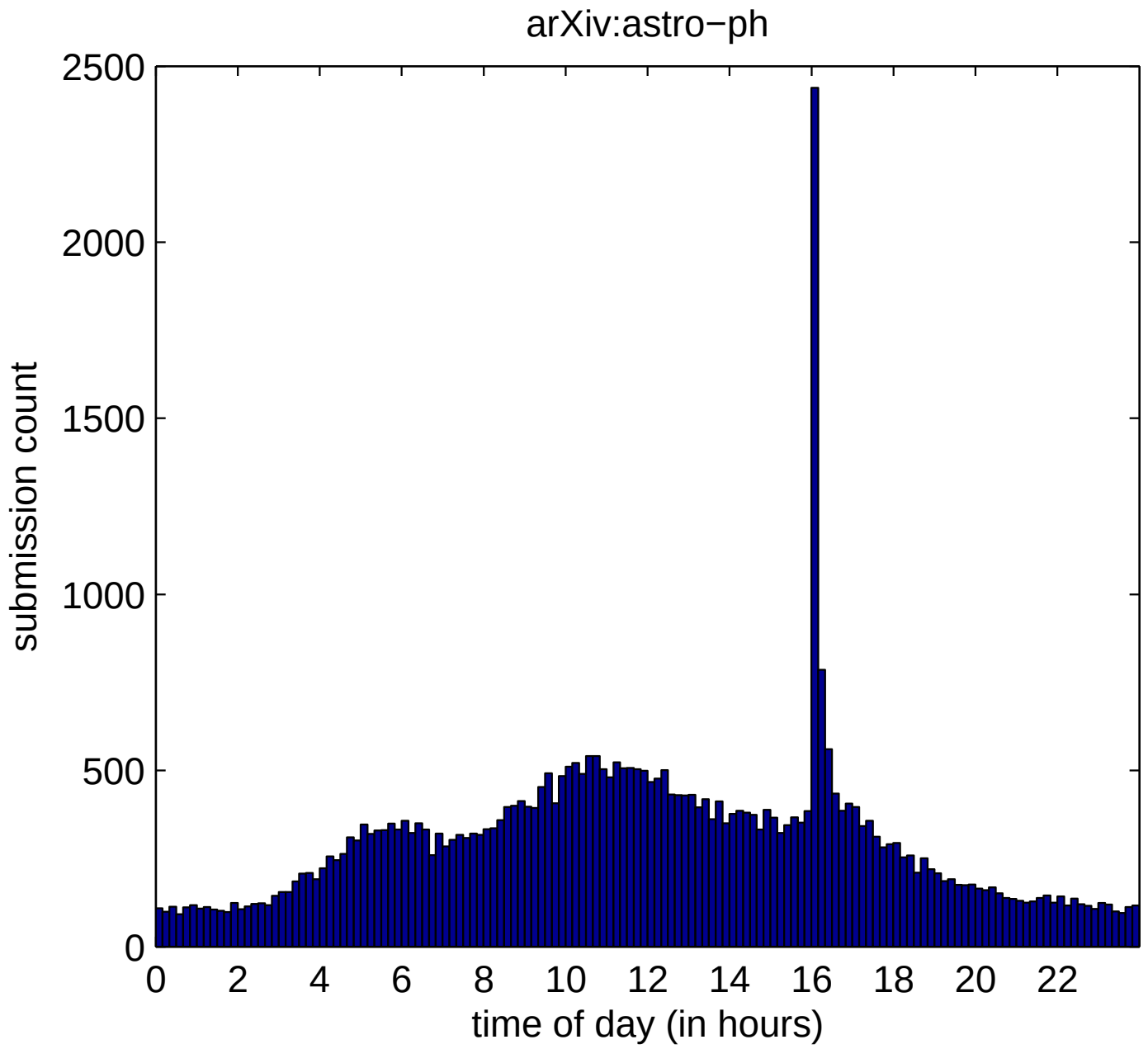
Implications of next generation open XML document format [.docx, . . .] not yet appreciated.

(Commercial tools for authoring in NLM/NCBI DTD?

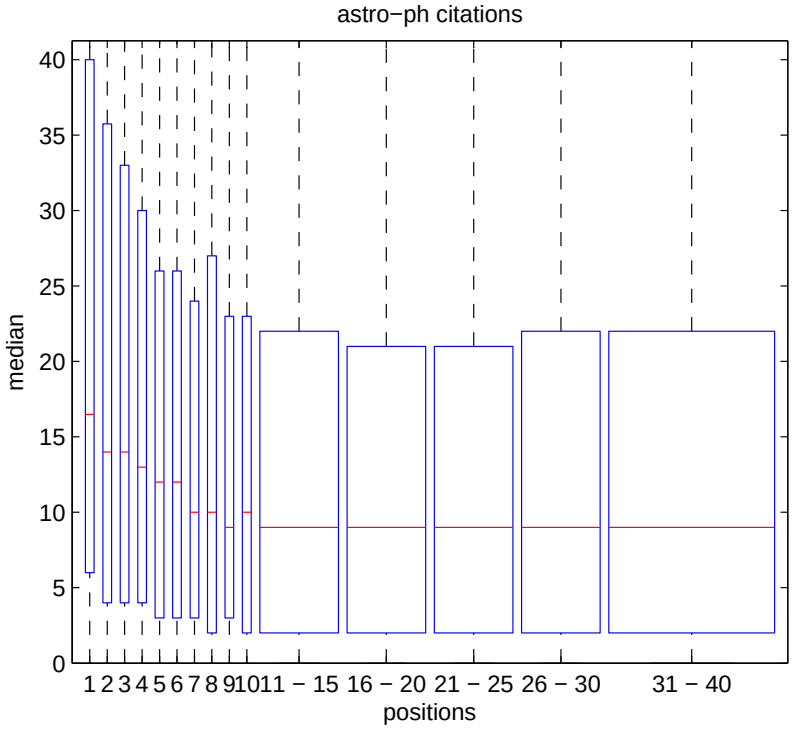
Article authoring add-in for MS Word 2007)

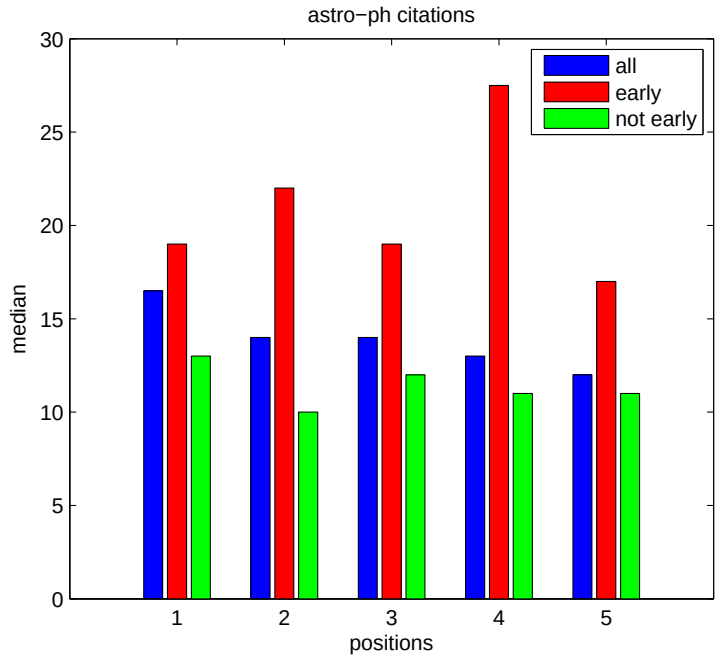
Paradox of physics: some well-established areas could fit into a semantic web context, amenable to a "commons" approach via open ontologies and sets of relationships

(more generally, tie semantic content in existing centralized literature databases to distributed network databases using relevant ontologies and machine-readable document standards)


arXiv:astro−ph
2500
2000
1500
1000
500
0
submission count
0
2
4
6
8
10
12
14
16
18
20
22
time of day (in hours)

astro-ph citations

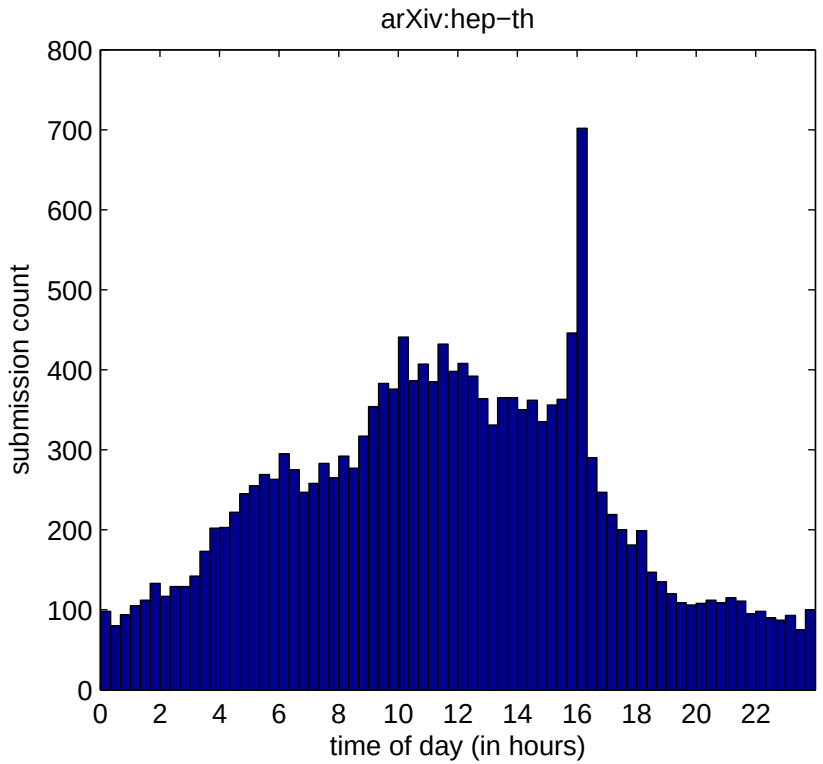


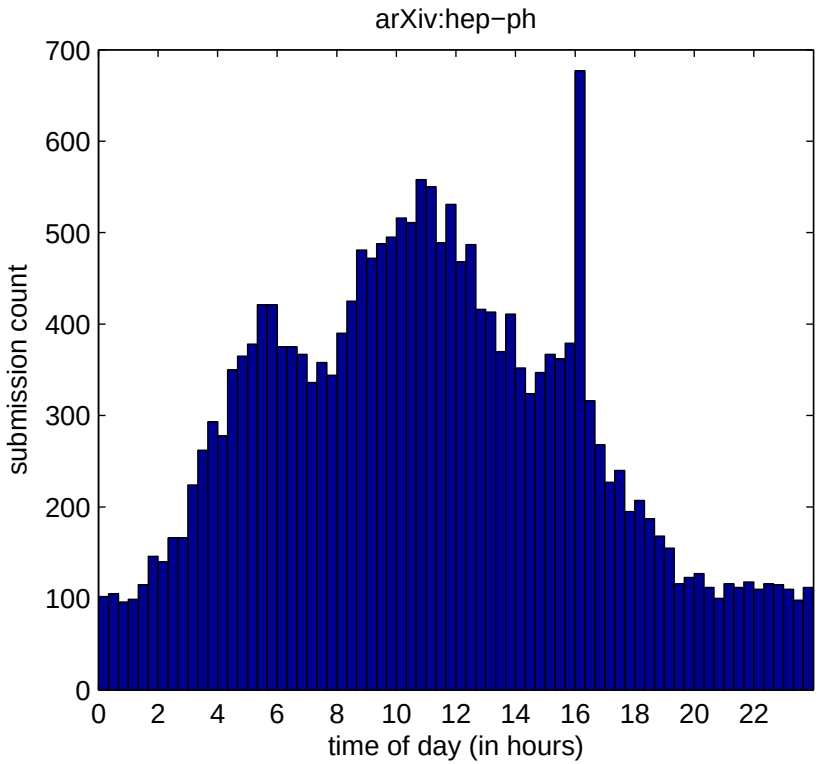


Dietrich (2008a,b), then Asif-ul Haque + PG,

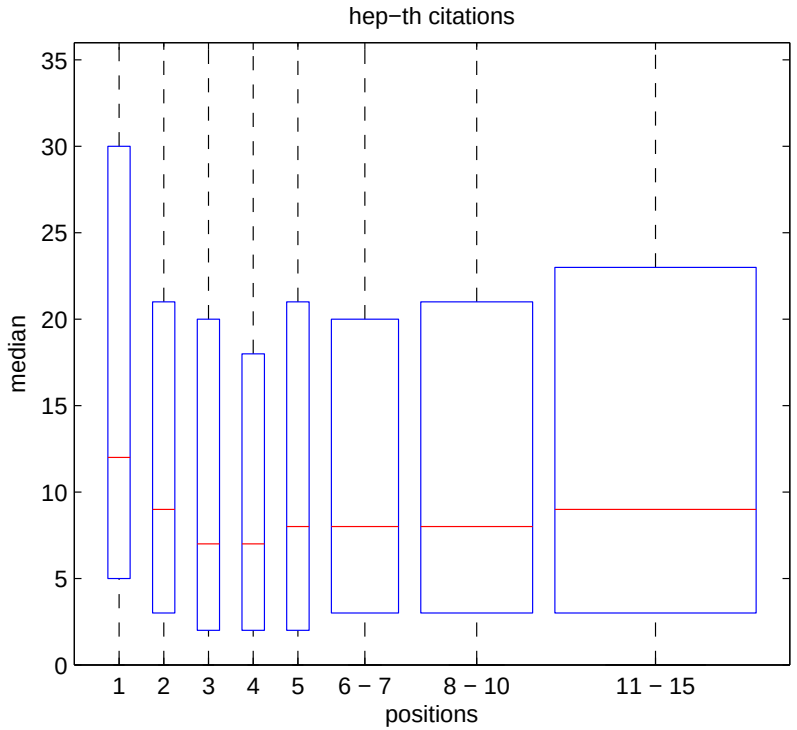
“Positional Effects on Citation and Readership in arXiv” (to appear)

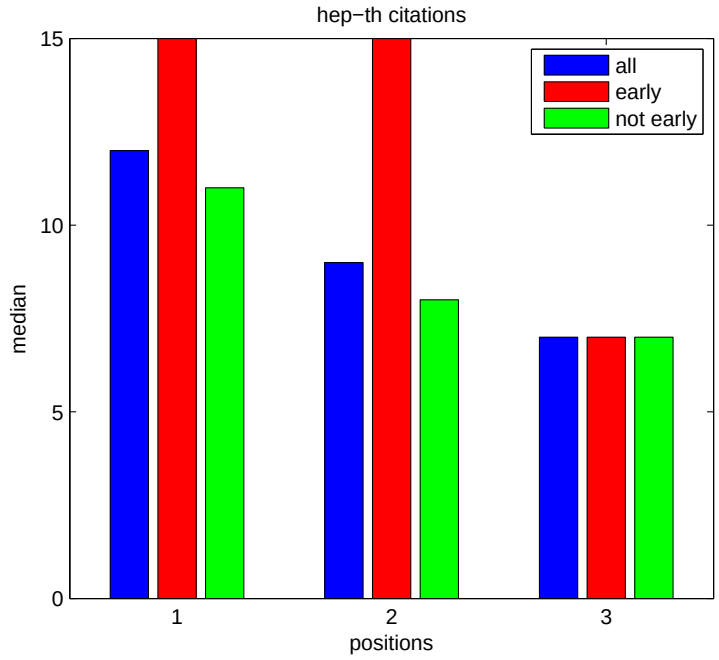
hep-th / hep-ph submissions



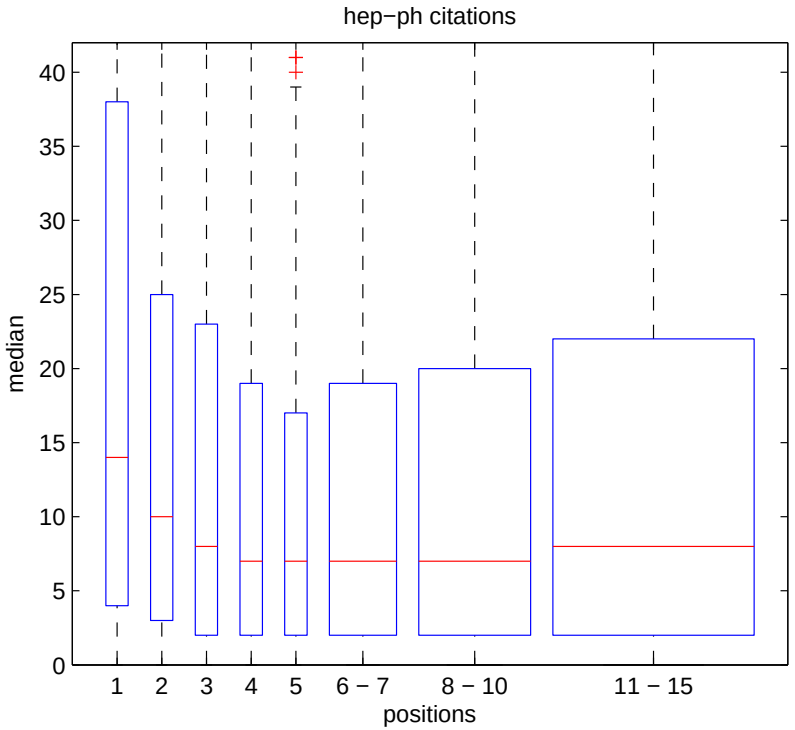


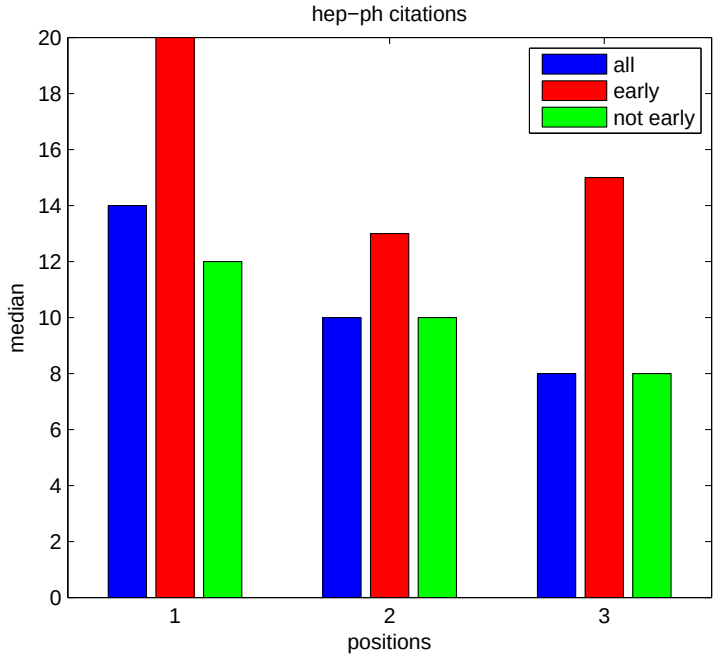
hep-th citations





hep-ph citations





Game Theory

A few percent submitted in first 60 seconds (increasing with time...).

Citation data for ’02–’04 submissions:

astro-ph median citations: pos 1= 16.5, pos 10–40= 9 (83%)

NE pos 1= 13 (44% visibility boost)

hep-th median citations: pos 1= 12, pos 11–15= 8 (50%)

NE pos 1= 11 (38% visibility boost)

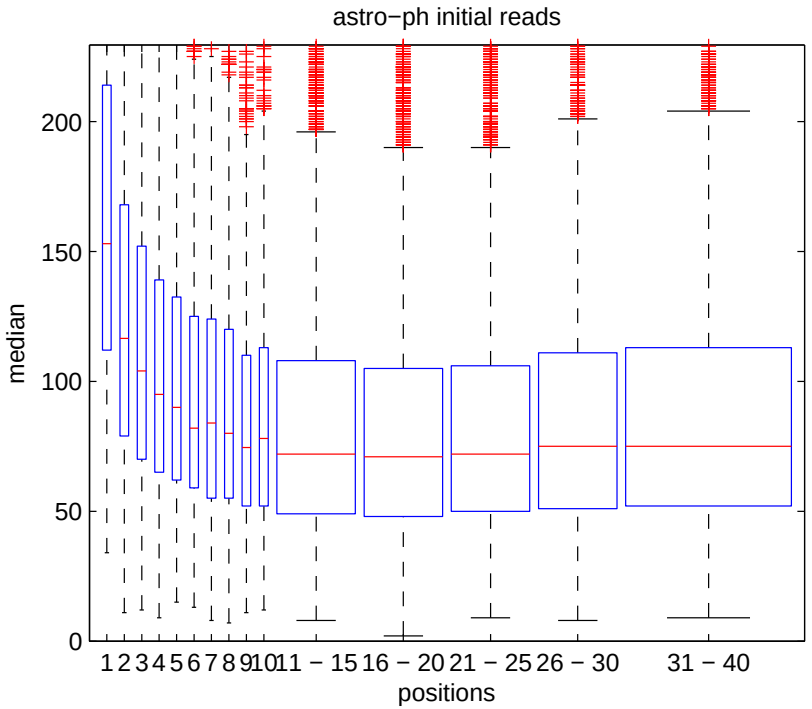
hep-ph median citations: pos 1= 14, pos 11–15= 7 (100%)

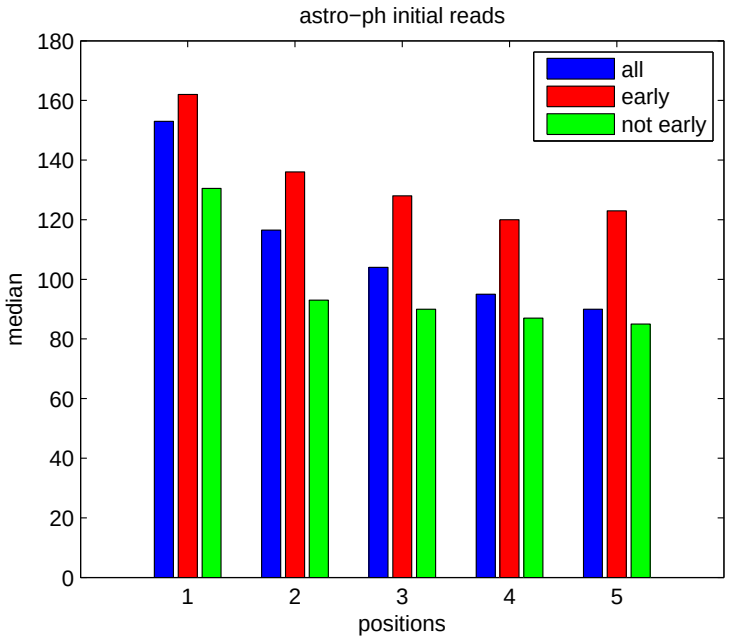
NE pos 1= 12 (71% visibility boost)

- **Visibility Bias**
- **Self-promotion Bias**

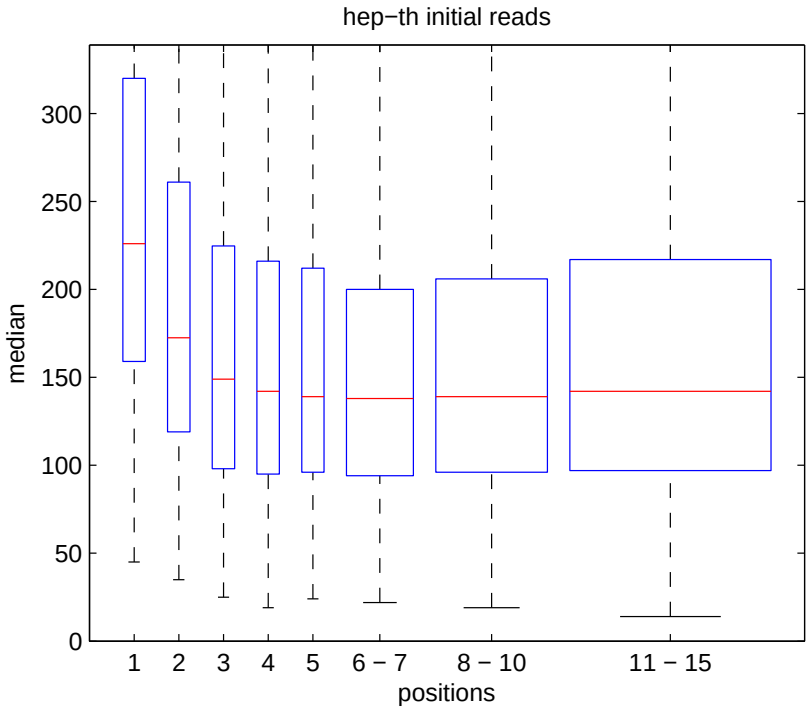
SP highest, but also difference definite VB effect (cited not necessarily due to inherent quality)

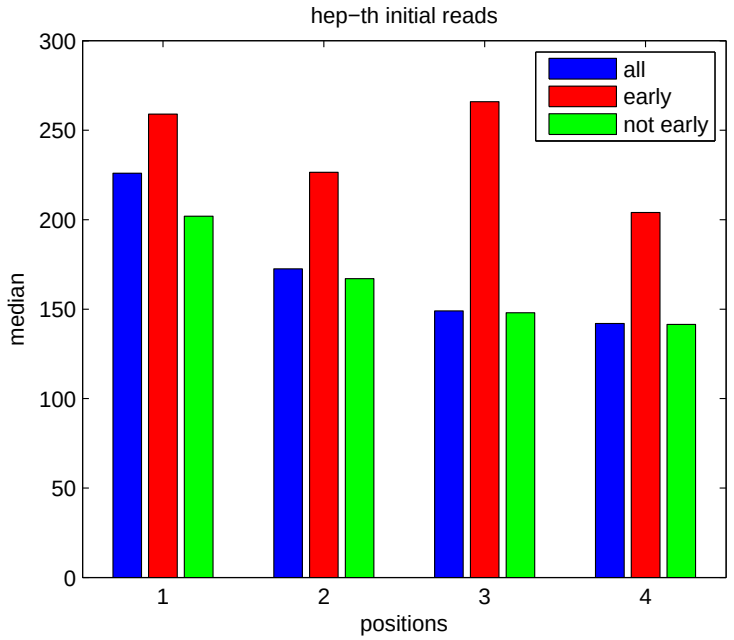
astro-ph active reads



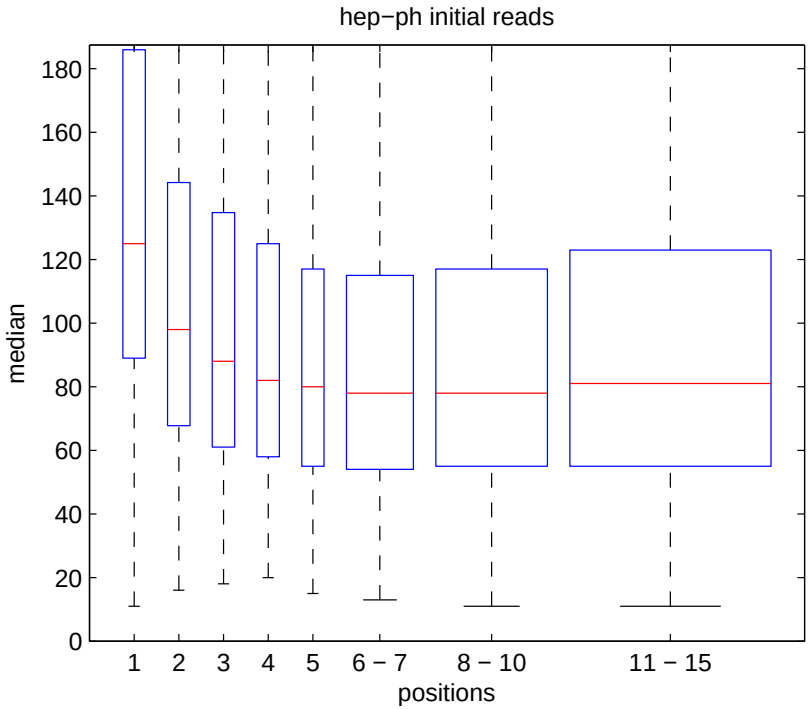


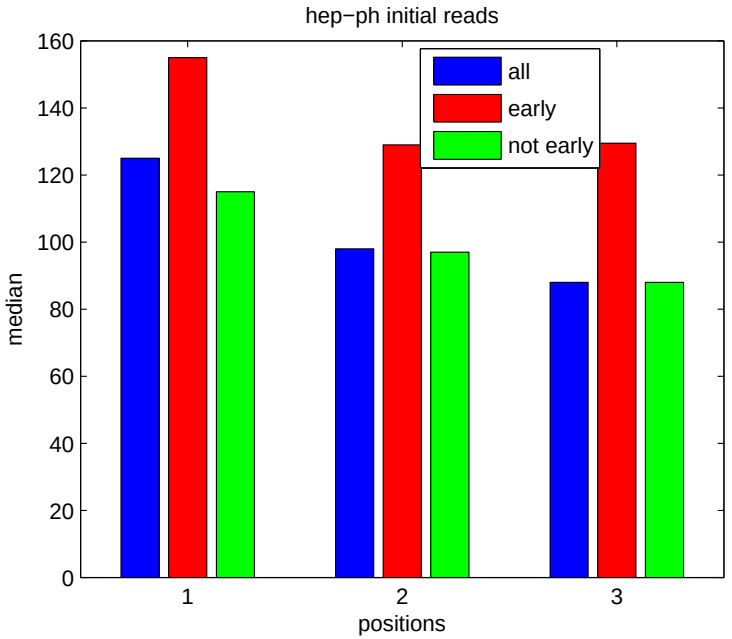
hep-th active reads





hep-ph active reads





Readership

Median number of full text downloads during initial active periods ('02–'07 submissions):

astro-ph pos 1 = 105, pos 5–15 =73 (82%)

NE pos 1=112 (53%)

hep-th pos 1 = 226, pos 5–15 =140 (61%)

NE pos 1=202 (44%)

hep-ph pos 1 = 125, pos 5–15 =79 (58%)

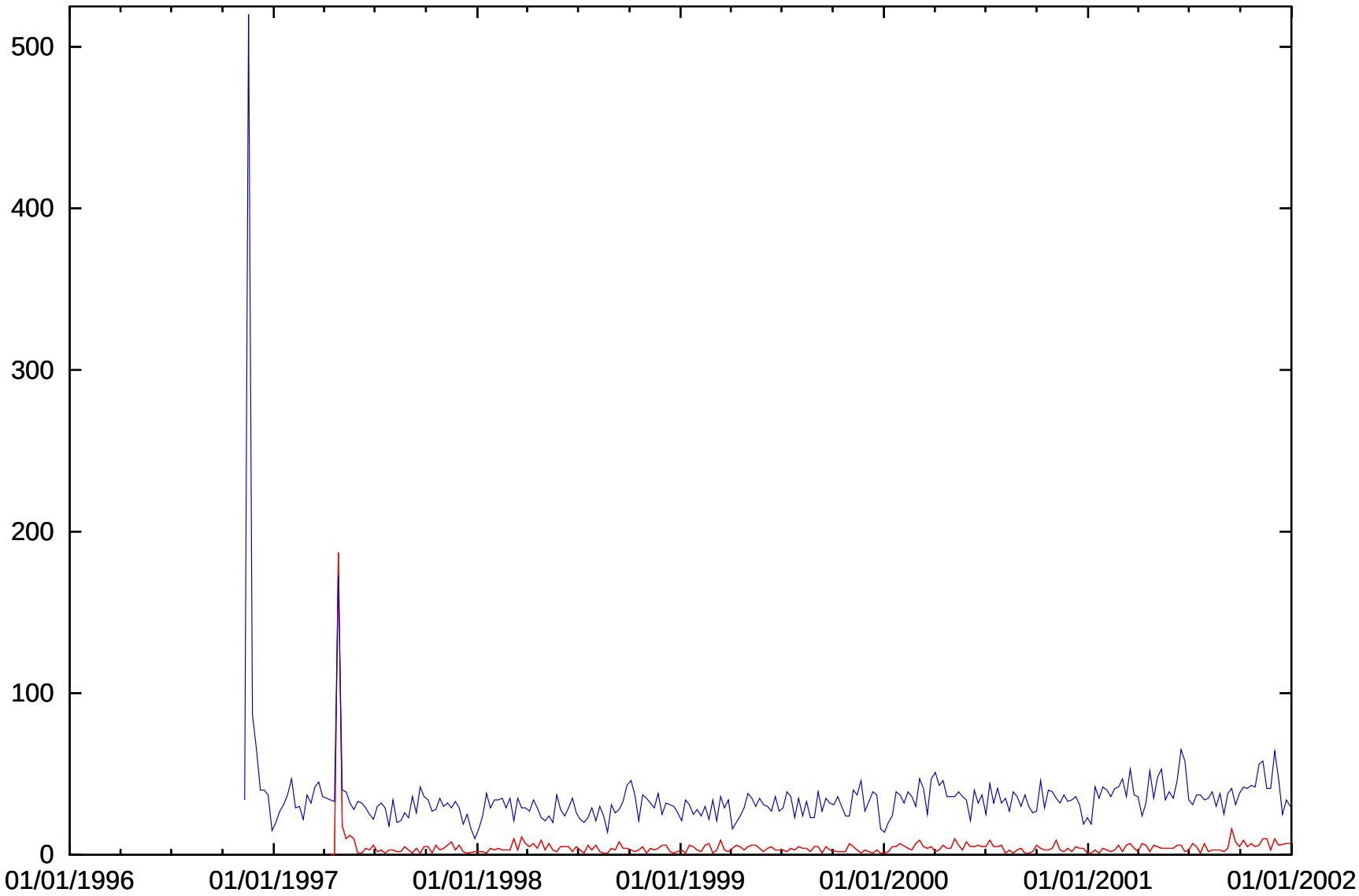
NE pos 1=115 (46%)

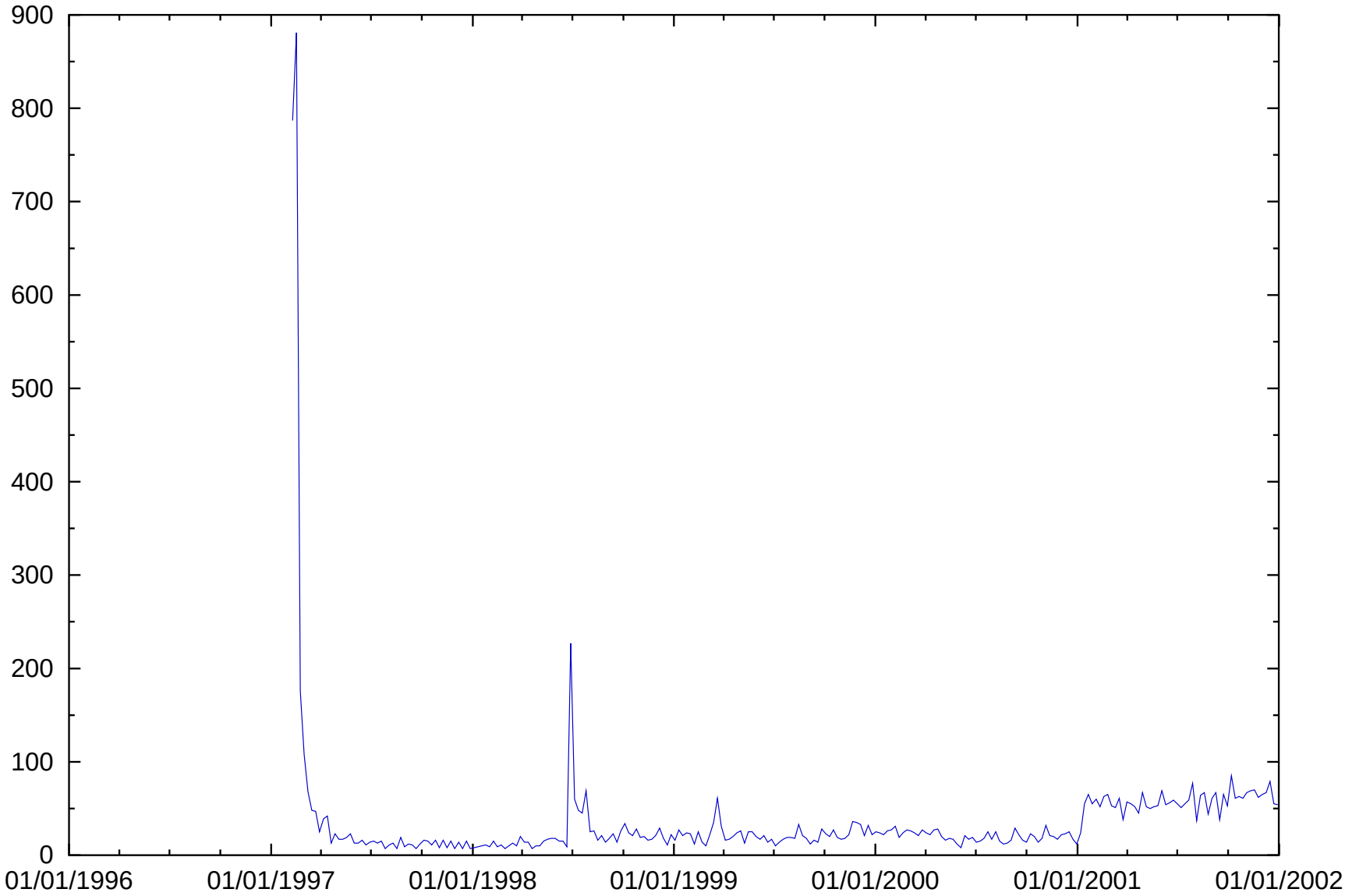
Never before a global system in resonance.

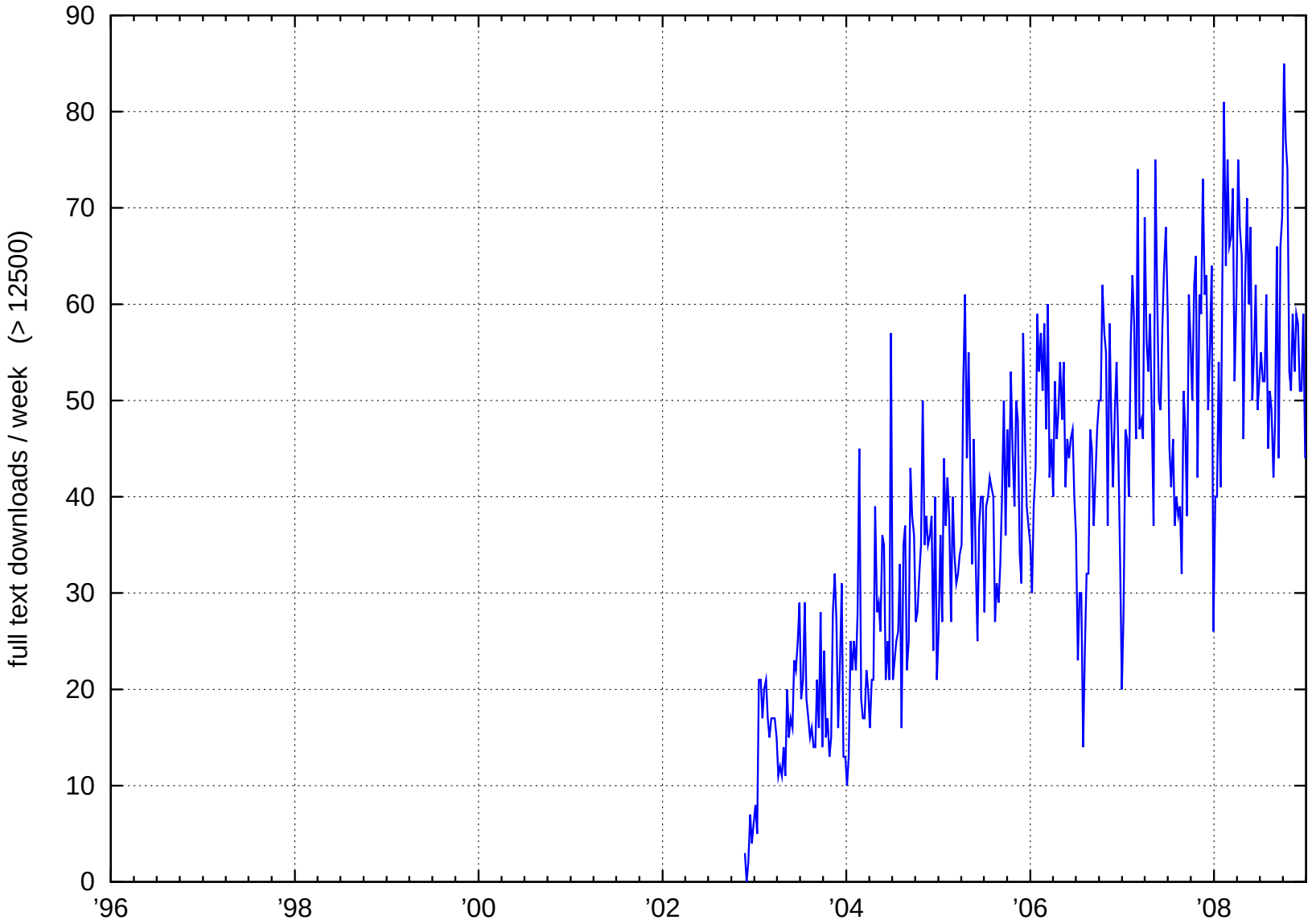
Information overload so some overlooked?

Need more tools to sort by relevance.

danger of recommender systems


500
400
300
200
100
0
01/01/1996
01/01/1997
01/01/1998
01/01/1999
01/01/2000
01/01/2001
01/01/2002


900
800
700
600
500
400
300
200
100
0
01/01/1996
01/01/1997
01/01/1998
01/01/1999
01/01/2000
01/01/2001
01/01/2002


full text downloads / week (> 12500)
0
10
20
30
40
50
60
70
80
90
’96
’98
’00
’02
’04
’06
’08

III. Where are we going?

What will Open Access Mean?

First survey current generic functionality:

- **PMC, ADS, Citeseer, PLoS, scholar.google, SLAC-Spires, . . .**
- **APS, ISI, Highwire, ScienceDirect, IoP, . . .**
- **nytimes, youtube, video.google, amazon, . . .**

Scholarship $\Longleftrightarrow$ Shopping $\Longleftrightarrow$ Entertainment

(sniff: we're no longer the bleeding edge)

Note importance of community building / social networking

Avoid emulating abacus

Watch out for interactions with blogspace/media

clearly no citation advantage ...

The mystery of Wikipedia

Same Old Questions

a) financial model for quality control

b) how pieces will merge into an interoperable whole

c) article of the future?

plus the role of text in a data-centric world

Web 3.0

Semantic enhancement: enhanced meaning, facilitates auomated discovery, enables linking to related data/ideas, access to actionable data within article, integration of data between articles and resources

Search Results: Display

- Summary
- Brief
- XML
- Taxonomy Tree
- Cited in Books
- CancerChrom Links
- Conserved Domain Links
- 3D Domain Links
- GEO DataSet Links
- Gene Links
- Genome Links
- Genome Project Links
- GENSAT Links
- GEO Profile Links
- HomoloGene Links
- Nucleotide Links
- OMIM Links
- Compound Links
- Substance Links
- PopSet Links
- Protein Links
- PubMed Links
- Cited Articles
- SNP Links
- Structure Links
- Taxonomy Links
- UniSTS Links

Article: Related Material

- PubMed record
- PubMed related arts
- PubMed LinkOut
- Gene
- HomoloGene
- Nucleotide
- Omim
- GEO Profiles
- Protein
- PubChem Compound
- PubChem Substance
- Taxonomy
- Taxonomy tree

Item specific

- standard metadata (title, author(s), submitter)
- browse related items, related keywords
  - ▷ local, in 3rd party (e.g., pubmed, ISI, scholar.google . . .)
- add tags, labels (**“flowering of the commons”**)
- more from this user
- rate this item
- save to favorites
- add to groups
- share, e-mail to friend
- blog this item
- post to 3rd party site (e.g., myspace)
- flag as inappropriate
- comments, responses, eletters (read, add)
- full text
- supplemental data
- show references, citations
- addenda, corrigenda
- related web pages
- export citation; cite or link using DOI
- alert when cited
- same object in 3rd party (e.g., pubmed citation)
- search 3rd party database (e.g., by same authors in scholar.google, h-index)
- flavors of relatedness by text, co-citation, co-reference, co-usage (also read)

Site specific

- subscribe
- alert to new issues
- upload
- personalization **(enhancement for other users, but privacy issues?)**
  - ▷ my articles (view collection)
  - ▷ add/subtract from private library

Browse

- groups, categories, subject area
- most recent
- recently featured
- most viewed
- top rated
- most discussed
- top favorites
- most linked
- most honored
- most shared
- most blogged
- most searched

Enhancement

Publisher:

e.g. Molecular BioSystems (RSC journal): enhanced html with terms highlighted, linked to chemical terminology databases (gene, sequence, and cell type ontologies) via combination of automated text mining and domain expertise of specialist editors

External:

e.g. Reflect (http://reflect.ws/): external service or browser plug-in tags gene, protein, small molecule names; linked to sequence, structure, interaction databases.

(Elsevier Grand Challenge 1st place)

Article of Future?

Shotton et al. (2009), Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Comput Biol 5(4): e1000361.

Off-the-shelf technologies:

semantic markup of textual terms; live linkages (DOIs, ...); re-orderable reference list; document summary (a study summary, document statistics, tagcloud and grouped tag trees of the marked-up named entities); citation analysis (within article); "Citations in Context" tooltip, with typing; downloadable spreadsheets for tables and figures, interactive figures; data fusion with results from other research articles and Google Maps. (+ fit screen, graphical abstract)

Structured Digital Abstract: machine-readable, summarizes key data and conclusions; all named entities with precise database identifiers; main results using controlled vocabularies; standard evidence codes for methodology.

Realization

- **author tools**
- **editor tools (value-added financial model)**
- **post-pub automated tools**

Very visible, encourage parity in the open sector (lag-time?)

Who hosts mark-up and data?

publishers for their articles, or independent SourgeForge-like data repositories?

How stored?

in document, outside, in triple-stores? author initiates, editor validates, later time-dependence peeled back? data changes?

Generic Database Interoperability

Example: Google Moupse (sic)

additional structure facilites natural language queries of databases.

catalyze locally written semantic markup interface locally (neo-cogito):

youtube-like benefits to going OA

Berners-Lee: linked data (well-defined tasks easier than complex A.I. to parse human ideas)

Wolfram|Alpha: curation, algorithms, interpreter, visualization

sets new bar, also manifests benefit of common semantic structure

The paradox of physics

Potential analogs abound: astronomical objects and experiments; mathematical terms and theorems; physical objects, terminology, and experiments in physics; chemical structures and experiments

Currently no coordinated effort to develop semantic structures for most areas of physics.

But are all fields equally amenable to ontological approach?

Field differences

Wally Gilbert: “no fundamental organizing principles in biology”?

Differences in the role played by data in different areas of science (e.g., Genbank + related, all federally federated)

Text decreasing in value compared to semantic services over next decade?

Article is more than just an impartial database entry: an exercise in rhetoric

John Wilbanks: find problem/question that would take weeks to solve/answer, multiple browser tabs, complicated graph traversal of database queries, other tools

$\longrightarrow$ new possibilities for community-driven scientic knowledge curation and creation

Future

Challenge from Word developers to Scientists:

Suggest 20 functions to provide optimal environment for scientific authorship (handshakes to networked databases, etc.)

Active + Passive user participation in bottom-up approach to QC

- **actively add tags, links; contribute to ontologies, correct wiki entries**
- **passively ingest readership, bookmarking, annotation behavior**

Incentive Question: expertise-intensive efforts beyond conventional journal publication (annotation, linkage, . . .) = scholarly achievement?

articles + blog commentary $\longrightarrow$ more modular objects

glue databases together into knowledge structure

Goal: semi-supervised, self-incentivized, self-maintaining knowledge structure, navigated via synthesized concepts, w/o redundancy/ambiguity, sourced, authenticated, highlighted for novelty

Network benefits to readers and authors

algorithms with access to personal and collective user behaviors

⇒ more comprehensive browsing

explanatory/complementary resources linked to words/eqns/figs/data

⇒ more incisive reading

Network-aware authoring tools analyze draft document content in progress, suggest links to external text and data resources (including semantic linkages)

Takes advantage of continued growth in distributed network databases, new interoperability protocols, machine-readable document standards, and relevant ontologies.

Neo-Minsky: “Can you imagine they used to have an internet in which authors, databases, articles, and readers didn’t talk to each other?”

Essential questions

How will the analog of NCBI/PubMedCentral be provided for other communities? (Who? With whose money?)

Common web service protocols, common languages (e.g., for manipulating, visualizing data), data interchange standards

Distributed version for other fields

networked resources $\Rightarrow$ new nonlinear reading strategies

ubiquitous mobile devices $\Rightarrow$ new usage of short-, long-term memory

Qualitatively new research and cognitive methodologies, transformation in the way we process scientific information, with academic community as role model for the creation and dissemination of knowledge to the public