arXiv, the OAI, and peer review

Simeon Warner
(arXiv, Los Alamos National Laboratory, USA)
(simeon@lanl.gov)

Workshop on OAI and peer review journals in Europe,
Geneva, 22–24 March 2001

What is arXiv?

- `http://arXiv.org/`
- aka ‘Los Alamos e-print archive’, formerly ‘xxx’.

  (‘e-prints’ may be unpublished works, pre-prints or published works.)
- Unrefereed author self-archiving.
- No-fee retrieval by users worldwide.

Evolution

Aug 1991	Physics e-print archive started: `hep-th` archive with email interface.
1992	ftp interface added. `hep-ph` and `hep-lat` added locally; `alg-geom`, `astro-ph` and `cond-mat` added remotely.
Dec 1993	Web interface added.
Nov 1994	Data at some remote archives (using the same software) moved to main site, the remote sites become mirrors.
Jun 1995	Automatic PostScript generation from TEX source.
Apr 1996	PDF generation added.
Jun 1996	Web upload facility added.
from 1996	Worldwide mirror network grows.
from 1999	arXiv involved in the OAI.

The present

- Covers physics, math, computer science, and non-linear systems.
- Serves over 70,000 users in over 100 countries.
- Estimated 13 million downloads in 2000.
- Over 30,000 new submissions in 2000, over 150,000 e-prints total (approximately linear growth in submission rate, $\approx$3500 extra each year).
- $>$99% of submissions entirely automated.
- Submission via web (68%), email (27%) and ftp (5%).
- Some journals now accept and arXiv identifiers instead of requiring direct submission (e.g. APS: Phys. Rev. D, Elsevier: Phys. Lett. B).
- Los Alamos site funded by DOE and NSF; mirror sites funded locally.

Involvement of arXiv in the OAI

The meeting held in Santa Fe in 1999, from which OAI has emerged, was organized by Paul Ginsparg (arXiv), Rick Luce and Herbert Van de Sompel. arXiv has continued to be actively involved in both management and technical development.

The subset Dienst protocol resulting from the Santa Fe meeting was implemented at arXiv by 15^{th} February 2000.

Initial focus of the OAI was **e-print archive interoperability**. While the scope of OAI has expanded considerably, the e-print community has led the protocol development.

The e-print community is, so far, the only community to have defined community specific formats for use in the OAI (e.g. `description` section of the Identify verb).

OAI protocol v1.0

- Protocol for **metadata** harvesting
- *data providers* e.g. arXiv
- *service providers* e.g. arc
- 5 verbs: Identify, ListSets, ListMetadataFormats, GetRecord, ListIdentifiers, ListRecords
- Concepts in protocol: identifiers, datestamps, sets, deleted records, metadata formats, and flow control.

Metadata formats

OAI supports **parallel metadata sets**; arXiv disseminates metadata in the following formats:

`oai_dc` Dublin Core encoded in XML.

`oai_rfc1807` RFC1807 encoded in XML.

`arXiv` Test-bed for new internal XML metadata format.

`arXivOld` XML encoded version of current internal metadata format.

`amf` Academic Metadata Format (draft by Krichel and Warner).

Flow control

- Avoid ‘accidental’ DoS attack
- arXiv particularly vulnerable (on-the-fly PS/PDF generation)

arXiv implementation:

- Implement partial response `resumptionToken`.
- Implement delay with HTTP 503 and `Retry-After`.
- Successfully avoids compliant harvesters from getting blocked (e.g. arc).

OAI repositories as of 8 March 2001

“arXiv” (Simeon Warner)
- 155522 identifiers (duplicates, 1000 identifiers/block)

“OCLC Theses and Dissertations Repository” (Jeff Young)
- 102762 (100 identifiers/block)

“NACA” (Michael Nelson)
- 6352 identifiers (all in one reply)

“M.I.T. Theses”
- 5196 identifiers (all in one reply)

“The Oxford Text Archive”
- 1290 identifiers (50 identifiers/block)

OAI repositories as of 8 March 2001 (contd. 1)

“Perseus Digital Library” - 1030 identifiers (all in one reply)

“CogPrints” - 1028 identifiers (all in one reply, eprints.org s/w)

“NSDL at Cornell” - >870 identifiers (30 identifiers/blocks)

“PhysNet, Oldenburg, Germany” - 472 identifiers (200 identifiers/block)

“Humboldt University of Berlin” - 464 identifiers (200 identifiers/block)

“Resource Discovery Network” - 388 identifiers

“A Celebration of Women Writers” - 142 identifiers

“European Language Resources Association” - 183 identifiers

OAI repositories as of 8 March 2001 (contd. 2)

”Linguistic Data Consortium” - 216 identifiers

“University of Tennessee Libraries” - 201 identifiers (20 identifiers/block)

“The Natural Language Software Registry” - 78 identifiers

“CDLCIAS” - 15 identifiers (3 deleted, eprints.org s/w)

“CDLDERM” - 2 identifiers (eprints.org s/w)

“Tobacco Control Digital Repository” - 1 identifier (eprints.org s/w)

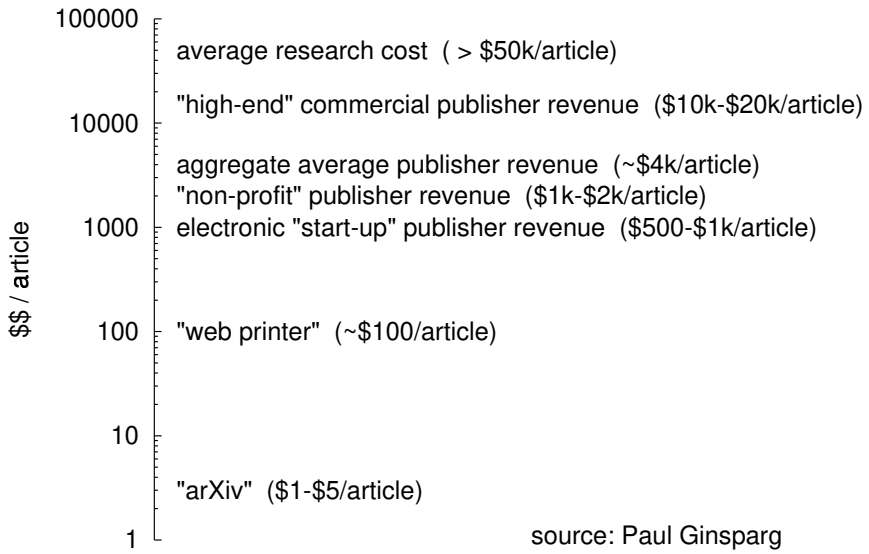
Who is using flow control?

- arXiv - partial response and retry-after
- OCLC TDR - partial response
- Oxford Text Archive - partial response and retry-after
- NSDL - partial response

Overlays to arXiv

- Advances in Theoretical and Mathematical Physics (ATMP)
  Subscriptions for paper copy, peer reviewed.
  http://pascal.intlpress.com/journals/ATMP/
- Geometry and Topology
  Subscriptions for paper copy, peer reviewed, also keeps local copies.
  http://www.maths.warwick.ac.uk/gt/gtmono.html
- JHEP - The Journal of High Energy Physics
  No formal duplication mechanism but almost all papers also on arXiv
  http://jhep.cern.ch/

Cost per article



$\Rightarrow$ arXiv is inexpensive

$\Rightarrow$ peer review adds $\approx$\$500 per article

What does arXiv provide?

- Minimal screening (we hope to improve moderator coverage)
- Low level of formatting control
- Size control (important for worldwide access)
- ‘free’ access
- ‘long term’ availability

arXiv and peer review

- Easy to think of arXiv as passively orthogonal to peer review (perhaps Elsevier does?).

- in some fields (notably `hep-th`), arXiv makes peer review obsolete for *scientific communication* because of the speed with which the field evolves.
- arXiv could support separation of 'publication' and peer review by storing certification information.

That’s all folks...