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

Abstract

I sketch the development of the Los Alamos e-print archive (arXiv). I then highlight some concepts of the Open Archives Initiative (OAI) protocol and briefly comment on the relation of arXiv to peer review. [1]

1 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  ('e-prints' may be unpublished works, pre-prints or published works.)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.
- E-mail notification of new submissions in chosen subject areas.

1.1 Evolution

The Los Alamos e-print archive, now called arXiv [1] and formerly known as 'xxx', was started by Paul Ginsparg in August 1991. It allows unrefereed author self-archiving of research papers and no-fee retrieval by users worldwide. Initially arXiv was limited to high-energy theoretical physics, operated over email only, and served ≈200 users. The number of users grew to over 1000 in a few months and has grown to over 70,000 now. The scope of archive has expanded and new facilities have been added steadily since 1991 [2]. The following is a list of some of the more significant developments:

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.

[1] Slightly expanded version of talk presented at the "Workshop on OAI and peer review journals in Europe", Geneva, 22–24 March 2001.

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.

1.2 The present

arXiv is the dominant means of information dissemination within certain areas of physics, notably high-energy physics, and has growing significance in many others [3]. It has rendered the paper distribution of pre-prints obsolete. The following is a list of some key statistics:

- Covers physics, mathematics, 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).
- $>$98% 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[2] and NSF; mirror sites funded locally.

If each submission required just 15 minutes of arXiv staff effort then we would require 7 full-time staff just to deal with the current submission volume. It is thus very important that most submissions require no manual intervention.

1.3 Involvement of arXiv in the OAI

The Open Archives Initiative (OAI) [4] developed from a meeting held in Santa Fe in 1999 which was initiated by Paul Ginsparg (arXiv, Los Alamos National Lab.), Rick Luce (Los Alamos National Lab.) and Herbert Van de Sompel (University of Ghent, Los Alamos National Lab.). arXiv has continued to be actively involved in both management of the initiative and technical development of the protocol.

[2] arXiv has direct DOE grant funding and also support from the 'Library Without Walls' project (LWW/STB-RL) within Los Alamos National Lab.

The protocol that resulted from the Santa Fe meeting [5] was a subset of the Dienst protocol [6] developed at Cornell University. While the syntax has changed significantly, the philosophy remains similar and our implementation has developed from that written to comply with the Santa Fe Convention and announced on 15^{th} February 2000.

The initial focus of the OAI was author self-archived scholarly literature — e-prints, as they are often known. While the scope of the OAI has expanded considerably, the e-print community has led the protocol development. An example of this lead is the `eprints.xsd` schema (appendix 2 in the protocol specification [7]) which defines an e-print community specific `description` section for the response to the Identify verb.

2 OAI

- Protocol for **metadata** harvesting.
  This could be extended after the initial 12–18 month test period for v1.0.
- *data providers* e.g. arXiv.
- *service providers* e.g. arc.
- 6 verbs: Identify, ListSets, ListMetadataFormats, GetRecord, ListIdentifiers, ListRecords.
  The simplicity of the protocol is provides a low barrier for implementation.
- Concepts in protocol: identifiers, datestamps, sets, deleted records, metadata formats, and flow control.

Of the concepts listed above I will comment further on sets, metadata formats and flow control.

2.1 Sets

arXiv currently exposes 4 sets which correspond to the 'groups' we use to categorize submissions: `physics`, `cs`, `math` and `nlin`. It is not yet clear whether sets will be much used as implemented within OAI. One should note that sets are designed solely for *selective harvesting* and their meaning should not be overloaded with information that would be better placed in the metadata. To date we know of one example of the use of sets: Before the creation of OAI the computer science (`cs` set) papers in arXiv were harvested by the NCSTRL [8] system using the Dienst [6] protocol.

2.2 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.

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

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

`amf` Test-bed for Academic Metadata Format (draft by Krichel and Warner [9]).

Involvement in the OAI has highlighted the need for arXiv to collect better metadata.

2.3 Flow control

- Avoid 'accidental' denial-of-service attack.
- arXiv particularly vulnerable (on-the-fly PS/PDF generation) because of the fact that most papers are stored as TEX source and processed to produce PostScript or PDF on demand (with a large cache).

arXiv implementation:

- Implement partial response and `resumptionToken`.
- Implement delay with HTTP 503 and `Retry-After`.
- Successfully avoids compliant harvesters from getting blocked (e.g. arc [10]).

2.4 OAI repositories as of 8 March 2001

The following is a list of results from ListIdentifiers requests made to all *registered* OAI repositories as of 8 March 2001. I note the total number of identifiers, and the number of identifiers returned in each block for repositories implementing partial responses. The list is currently dominated by archives of academic papers. Clearly some archives are in a test phase — the "Tobacco Control Digital Repository" returns only 1 identifier!

"arXiv" (Simeon Warner)
- `http://arXiv.org/oai1`
- 155522 identifiers (duplicates, 1000 identifiers/block)

"OCLC Theses and Dissertations Repository" (Jeff Young)
- `http://alcme.oclc.org:4342/etdcat/servlet/OAIHandler`
- 102762 (100 identifiers/block)

"NACA" (Michael Nelson)
- `http://naca.larc.nasa.gov/oai/`
- 6352 identifiers (all in one reply)

"M.I.T. Theses"
- `http://theses.mit.edu:80/Dienst/Index/2.0/OAI-1.0`
- 5196 identifiers (all in one reply)

"The Oxford Text Archive"
- `http://ota.ahds.ac.uk/cgi-bin/ota/oai.cgi`
- 1290 identifiers (50 identifiers/block)

"Perseus Digital Library"
- `http://www.perseus.tufts.edu/cgi-bin/pdataprov`
- 1030 identifiers (all in one reply)

"CogPrints" (Robert Tansley/Tim Brody/Stevan Harnad)
- `http://cogprints.soton.ac.uk/perl/oai`
- 1028 identifiers (all in one reply, eprints.org s/w)

"NSDL at Cornell"
- `http://nsdlib.nsdl.cornell.edu/nsdl/portal/oai`
- $>$870 identifiers (30 identifiers/blocks)

“PhysNet, Oldenburg, Germany”
- `http://physnet.uni-oldenburg.de/oai/oai.php`
- 472 identifiers (200 identifiers/block)

“Humboldt University of Berlin”
- `http://dochost.rz.hu-berlin.de/OAI-script`
- 464 identifiers (200 identifiers/block)

“Resource Discovery Network”
- `http://www.rdn.ac.uk/oai/nph-oai.cgi`
- 388 identifiers

“A Celebration of Women Writers”
- `http://digital.library.upenn.edu/webbin/OAI-celebration`
- 142 identifiers

“European Language Resources Association”
- `http://www.ldc.upenn.edu:85/OLAC/dp/elra.php3`
- 183 identifiers

”Linguistic Data Consortium”
- `http://www.ldc.upenn.edu:85/OLAC/dp/ldc.php3`
- 216 identifiers

“University of Tennessee Libraries”
- `http://helios.dii.utk.edu/cgi-bin/oai.cgi`
- 201 identifiers (20 identifiers/block)

“The Natural Language Software Registry”
- `http://www.ldc.upenn.edu:85/OLAC/dp/dfki.php3`
- 78 identifiers

“CDLCIAS”
- `http://eprints.cdlib.org/cias-perl/oai`
- 15 identifiers (3 deleted, eprints.org s/w)

“CDLDERM”
- `http://eprints.cdlib.org/derm-perl/oai`
- 2 identifiers (eprints.org s/w)

“Tobacco Control Digital Repository”
- `http://eprints.cdlib.org/tc-perl/oai`
- 1 identifier (eprints.org s/w)

2.5 Who is using flow control?

The OAI protocol flow control mechanism has two parts. First, responses may be split into a set of partial responses. Second, delays between requests may be implemented with the 503 ‘retry-after’ response. A few of the currently registered repositories use these mechanisms:

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

- NSDL - partial response

3 arXiv and peer review

3.1 Overlays to arXiv

- Advances in Theoretical and Mathematical Physics (ATMP)
  Subscriptions for paper copy, peer reviewed. This is very close to being a pure overlay.
  `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`

While not actually an overlay I'll also mention JHEP because of duplicated content:

- JHEP - The Journal of High Energy Physics
  No formal duplication mechanism but almost all papers also on arXiv
  `http://jhep.cern.ch/`

3.2 Cost per article

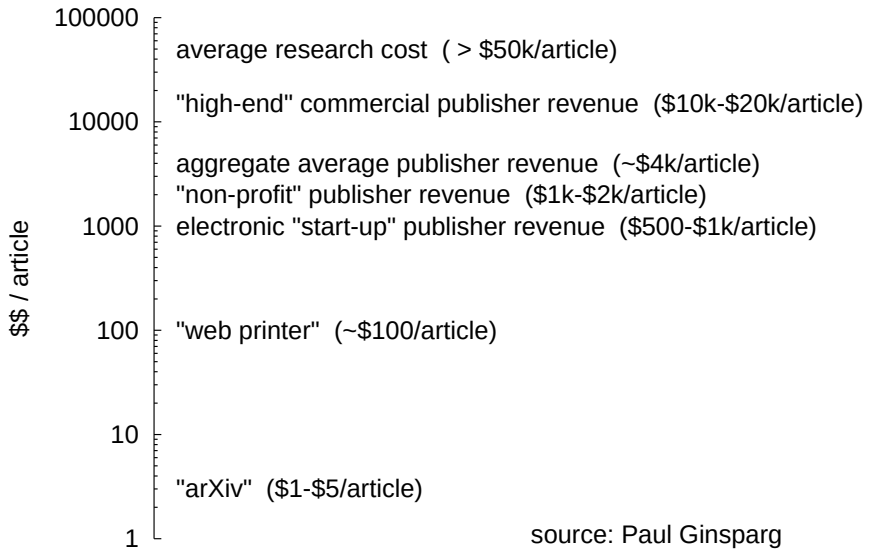

Figure 1: Cost per article for various publication types.

⇒ arXiv is inexpensive

⇒ peer review adds ≈$500 per article

If we increase the arXiv cost to ≈$10 per article then we include all R&D activity (including my attendance at this workshop!).

HighWire is the current "web printer" example. Their costs are slightly inflated because of the need to deal with legacy formats from publishers.

3.3 What does arXiv provide?

- Minimal screening (we hope to improve moderator coverage).
  The high profile of arXiv makes this an increasing problem but we feel that some screening is very important to maintain the usefulness of arXiv.
- Low level of formatting control.
  For example, we do not accept low-quality bitmap PDF.
- Size control (important for worldwide access).
- 'free' access.
  Which actually implies restriction of mass automated downloads.
- 'long term' availability.

3.4 arXiv and peer review

- It is 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. However, peer review is still required for tenure so most publications in `hep-th` do eventually appear in conventional journals.
- arXiv could support separation of 'publication' and peer review by storing certification information. How this should be achieved is an open issue and an important subject for this workshop to address.

References

[1] arXiv, the Los Alamos e-print archive. Web interface: `http://arXiv.org/` Help system, web: `http://arXiv.org/help` and via email: `help@arXiv.org`

[2] Messages announcing new features at arXiv are collected at `http://arXiv.org/new/`

[3] Further information relating to arXiv and electronic publishing can be found at `http://arXiv.org/blurb/`

[4] Open Archives Initiative (OAI) `http://www.openarchives.org/`

[5] Documents resulting from the Santa Fe meeting are available from the OAI website at `http://www.openarchives.org/sfc/sfc_entry.htm`
See also: Herbert Van de Sompel, Carl Lagoze *The Santa Fe Convention of the Open Archives Initiative* D-Lib Magazine 6 no 2
`http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html`

[6] The protocol resulting from the Santa Fe meeting was a subset of the Dienst 5 protocol described at
http://www.cs.cornell.edu/cdlrg/dienst/protocols/DienstProtocol.htm
The Open Archives subset is described at
http://www.cs.cornell.edu/cdlrg/dienst/protocols/OpenArchivesDienst.htm

[7] OAI protocol v1.0, released 18^{th} January 2001
http://www.openarchives.org/OAI/1.0/openarchivesprotocol.htm

[8] Networked Computer Science Technical Reference Library (NCSTRL)
http://cs-tr.cs.cornell.edu/

[9] Academic Metadata Format Thomas Krichel and Simeon Warner
http://http://openlib.org/home/krichel/ebisu.html

[10] arc - Cross Archive Searching Service, an OAI *service provider* developed at Old Dominion University, http://arc.cs.odu.edu/help/archives.htm