arXiv: process and collaboration

Simeon Warner

simeon@cs.cornell.edu

XXVI Annual Charleston Conference:
Issues in Book and Serial Acquisition
Charleston, SC, USA, 8–11 November 2006

Process: Submission

- Users must create account with arXiv (verifies email contact)
- Must be **endorsed** for subject area
- Verify contact information
- Grant license
- Choose subject area

License click-through

A. Verify Your Contact Information

...explanation omitted...

First Name: Simeon
Last Name: Warner
Suffix: ('Jr.', 'II', etc; may be blank)
Affiliation: Cornell University
E-mail: simeon@cs.cornell.edu

☐ I certify that the above contact information is correct.

B. Legal Statement

- I grant arXiv.org a license to distribute this article.
- I certify that I have the right to grant this license.
- I understand that submissions cannot be completely removed once accepted.
- I understand that arXiv.org reserves the right to reclassify or reject any submission.

☐ I agree to the above terms.

- Self-certify author or proxy
- Enter separate metadata (Title, Authors, Comments, Journal-ref, etc.)
- Upload file(s)
- Automatic checks: metadata, TeX processing, format checks, size limits
- Notification of pending status and likely identifier
- Notification of any changes from **moderation**
- Submissions before 4pm published at 8pm (Cornell time)
- Submissions cannot be removed or changed when public; **'replacements' create a new version**

Process: Moderation

Q. Why moderation? Why isn't arXiv "open"?

A. arXiv would be less useful without moderation.

negative moderation — with no action arXiv runs entirely automatically, everything accepted.

Volunteer moderators used to implement policy:

1. **Not obviously wrong or inappropriate**

   "of refereeable quality"

2. **In correct subject area**

Most articles need no attention, a few cost a significant amount of time (1 admin, 200-300 submissions/day)

Process: Alerting and Reading

- Daily publication of new submissions
- Subject category based email alerts
  — 18k subscribers
- Worldwide availability, 17 mirror sites
- Web browse, web search engines, other portals
  — 20M downloads in 2005
- Long term support for access
  — have already done PS→PDF migration

Collaboration

Q. How can arXiv best serve its user community?

A. In part, by collaborating with others to make our data more useful.

Based on a vision of scholarly communication that relies upon a set of open access repositories providing an information substrate upon which other services are built.

Web search

Increasingly used by academics to find academic content — one stop shopping.

arXiv is part of the **deep web** $\longrightarrow$ need tailored crawl.

Google Google Scholar	Significant fraction of arXiv traffic. One crawler for both services.
Yahoo!	Second most popular web search.
MS Live Academic	New collaboration — work in progress.

Successful long running collaborations

Ahead of their time, necessarily custom.

SLAC SPIRES	daily 2-way reference exchange integrates valuable manual effort at SLAC
Front	metadata harvesting predating OAI-PMH popular specialized mirror/overlay for math

Heavyweight collaborations

Theme — **customized full data sharing**.

NASA ADS	almost complete astronomy/astrophysics coverage, share anonymized usage data
ISI	OAI-PMH metadata harvesting & PDF web citation indexing
IOP	full arXiv mirror → local access reference linking, institution identification

Citebase	take data from local mirror automated citation extracting and linking
CiteSeer	most popular CS portal (?)
OSTI	full-text harvesting for specialized search and alerting
Cornell CS	provide full-text search as part of search research

Lightweight collaboration

Theme — **open (OAI-PMH) metadata harvesting**, no customization.

Oaister	∼10 million records harvested via OAI-PMH
100+ services	harvesting metadata from arXiv each month
research support	interfaces to bulk-download data, arXiv is a popular test corpus

Plan to make full-text and plain-text more easily harvested.

Current work...

Service API — To allow services to be built on top of arXiv without the need for complete harvesting (c.f. NCBI Entrez)

Plain text export — To support remote full-text search and data-mining

Conference paper ingest — Collaboration with conference management systems for post-conference ingest and archiving of submissions

Dataset support — Support "small-data" science by allowing moderate amounts of data to be packed with submissions with submissions

That's all folks...