Open Archives Initiative protocol development and implementation at arXiv

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

OAI Open Day, Washington DC
23 January 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  (‘e-prints’ may be unpublished works, pre-prints or published works.)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.

Background

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.

arXiv **software**

The software running arXiv comprises of the order of 30,000 lines of Perl running under Linux with numerous other programs (TEX, ghostscript, tar, gnuzip,...).

- Software has evolved over past 9 years.
- Currently tidying and rewriting Perl code in modular fashion.
- Insufficient resources for complete off-line rewrite.
- Pressure to change underlying database and identifier structure.

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 15th 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 (see the `description` section of the Identify verb included in the protocol specification as an example).

OAI protocol v1.0 implementation

- arXiv is a *data provider*
- Concepts in protocol: identifiers, datestamps, sets, deleted records, and metadata formats.
- The verbs: Identify, ListSets, ListMetadataFormats, GetRecord, ListIdentifiers, ListRecords
- Flow control
- Cost

Identifiers

Internal arXiv identifiers have the form:

```
arch-ive/YYMMNNN
arch-ive.SC/YYMMNNN
```

OAI protocol restricts identifiers to follow the URI syntax. Appendix 2 describes the `oai-identifier` type which we use. In practice this means we prepend `oai:arXiv:` to our internal identifiers.

`hep-th/9901001`	`oai:arXiv:hep-th/9901001`
`quant-ph/9912010`	`oai:arXiv:quant-ph/9912010`
`math.SG/0001001`	`oai:arXiv:math.SG/0001001`
`cs.SE/0101002`	`oai:arXiv:cs.SE/0101002`

Datestamps

arXiv keeps logs of the date of:

- submission of different revisions, and
- dates of cross-listings.

However, no logs have been kept of by-hand administrator modifications or the addition of journal-references.

For OAI, we need a datestamp that reflects *any* change in the metadata.

⇒ Use the file modification date extracted from the file modification time.

⇒ Need to build indexes to support ListIdentifiers and ListRecords requests.

⇒ Indexes must be updated daily to avoid missing updates.

Sets

The OAI protocol characterizes sets as “an optional construct for grouping items in a repository for the purpose of selective harvesting of records”, and they may be arranged in *zero or more* hierarchies.

Note:

- OAI *data providers* don’t have to implement sets.
- Even when harvesting from *data providers* that implement sets, *service providers* don’t have to use sets (i.e. they never specify a `setSpec`).

Sets (cont'd.)

arXiv has a two- or three-level (depending on subject area) grouping hierarchy based on the subject of the e-print. The three levels are:

group There are four groups: physics, math, cs, nlin

archive The physics group has many archives (e.g. `hep-th`, `astro-ph`, `cond-mat` and even `physics`). Group and archive may be identified for the three other groups.

subject-class Some archives have subject-classes and individual e-prints may belong to one or more subject-class.

arXiv declares four sets which correspond to the four groups listed above.

Deleted records

arXiv does not allow e-prints to be removed once submitted. We do permit authors to submit a withdrawal notice as a new revision. Our OAI interface exports metadata for the withdrawal notice in the same was as for any other e-print.

There are also a very small number of e-prints that have been deleted.

- Currently 9 deleted e-prints in arXiv.
- Return the OAI `status="deleted"` attribute.
- Implemented using a lookup table.

Metadata formats

We disseminate metadata for e-prints 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.

Many mappings are obvious, e.g. Title → Dublin Core ‘title’, and Abstract → Dublin Core ‘description’.

Identify verb

e.g. `http://arXiv.org/oai1?verb=Identify`

Trivial to implement, simply writes a number of configuration variables out in an XML record. arXiv returns two `description` containers:

`oai-identifier` Declares our use of the OAI identifier scheme defined by `oai-identifier.xsd`, and gives a sample identifier `oai:arXiv:quant-ph/9901001`.

`eprints` As we are part of the e-prints community, we include a description container that follows the `eprints.xsd` schema.

ListSets verb

Again, very straightforward to implement. The code simply extracts the group names (long and short) from configuration variables and writes out as XML:

```
<set>
 <setSpec>nlin</setSpec>
 <setName>Nonlinear Sciences</setName>
</set>
<set>
 <setSpec>math</setSpec>
 <setName>Mathematics</setName>
</set>
...
<set>
 <setSpec>cs</setSpec>
 <setName>Computer Science</setName>
</set>
```

Small number of sets ⇒ no need to implement the partial response and acceptance of a `resumptionToken`.

GetRecord verb

The majority of the effort involved in implementing the GetRecord verb is in performing the metadata format conversion and mapping from our internal format to the format requested. Four cases to consider:

1. Item does not exist → no `<record>` container returned.
2. Item is ‘deleted’ → `<record status="deleted">` container with `<header>` block returned.
3. Item exists but can not be disseminated in the requested metadata format → `<record>` container with `<header>` but no `<metadata>` block returned.
4. Item exists and can be disseminated in the requested metadata format → `<record>` container with `<header>` and `<metadata>` blocks returned.

Chose not to implement `<about>`.

ListIdentifiers verb

This verb is essentially a search by datestamp, the optional `from` and `until` parameters specifying the datestamp range, and the optional `setSpec` parameter limiting the archives searched.

- Don't want to return all 150,000 identifiers at once.

⇒ Implement partial response, supply `resumptionToken` as necessary.

We *choose* to build `resumptionToken` from the `from`, `until` and `setSpec` parameters for new request.

ListRecords verb

Essentially a combination of the ListIdentifiers and GetRecord verbs:

- Implement search by datestamp and set using ListIdentifiers code.
- Implement metadata dissemination using GetRecord code.

Flow Control

The main arXiv site, `http://arXiv.org/`, is heavily used and has various automated scripts to prevent badly written and non-conforming (i.e. not obeying `/robots.txt`) robots from loading the server to the point where there is denial-of-service to other users. arXiv is particularly vulnerable 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). Flow control is thus essential to avoid legitimate OAI *service providers* getting blocked.

- Implemented with HTTP 503 and `resumptionToken`.
- Minimum delay between requests from any given site, else Retry-After.
- Successfully avoids compliant harvesters from getting blocked (e.g. arc).

Implementation cost

‘Santa Fe convention’, OAI Dienst Subset, December 1999 ≈2-3 weeks effort which included time to understand Dienst, writing utility routines such as an XML writing library (now available in standard libraries), coding a search by date, and coding conversion of our metadata to RFC1807.

OAI v0.2, October 2000 ≈2 days which included writing TEX→UTF-8 conversion code for the special characters which are currently mostly TEX encoded in our metadata. Also included rewriting parsing code for simplified syntax.

OAI v1.0, January 2001 Many small changes during protocol development, most of which involved simplifications in the code!

The majority of the effort necessary to create a new implementation is likely to be in routines to implement metadata format conversion; a search to find records by datestamp; and perhaps flow control through partial responses and Retry-After returns.

What happens now?

- Widespread adoption of OAI protocol by e-print and other communities.
- Development of enhanced metadata set specific to e-print community.

$\Longrightarrow$

- Generic OAI tools using Dublin Core metadata.
- E-print specific tools using community specific metadata set.

$\Longrightarrow$

- Better environment for research.
- Increased visibility for work submitted to arXiv.

That’s all folks...