Introduction to the OAI Protocol for Metadata Harvesting

Version 2.0

Hussein Suleman
Virginia Tech DLRL
17 June 2002
Version 2.0

» Already ?
» Why ? What are you guys thinking ?
» But we didn’t implemented version 1.1 yet!
» OK, so what is it anyway ?
Already?

- Released 14 June 2002
- Detailed schedule
  - Alpha test - since 1 March 2002
  - Beta test - 1 April 2002
  - Expected release - 1 June 2002
- Many alpha and beta testers
Why?

- v1.0/v1.1 was experimental
  - v2.0 should be stable
- Some minor shortcomings in semantics of v1.1
- Encoding issues that could be fixed
But we didn’t implement v1.1 yet

That’s OK!

v2.0 is not very different and in fact simpler for some things

Existing archives need not upgrade as long as harvesters understand multiple versions of the protocol!
Some New “Features”

- Single schema
- ListIdentifiers changes
- Provenance containers
- Granularity of harvesting
- Orthogonality of setSpecs
- More flow control
- XML error codes
- Timezones
- resumptionToken idempotency
XML Error Codes

• <error code="badArgument">… </error>
  • badArgument, badResumptionToken, badVerb, cannotDisseminateFormat, idDoesNotExist, noRecordsMatch, noMetadataFormats, noSetHierarchy

 Errors within the regular responses - no more HTTP 404s

Philosophy: separation of PMH and HTTP
Single Schema

minated schemas where available e.g., Dublin Core from DCMI and MARC from LoC
Use single namespace and schema for all requests/responses
No more arbitrary restrictions on format of metadataPrefix, setSpec, etc.
New Response Format

- <OAI-PMH>
  <responseDate>...
  <request parm1="..." parm2="...">...
  <GetRecord>
    ...
  </GetRecord>
</OAI-PMH>

- <OAI-PMH>
  <responseDate>...
  <request parm1="..." parm2="...">...
  <error code=""/>
</OAI-PMH>
Orthogonality of setSpecs

Previously you could not find the sets that an item belonged to

Now, every record header must include setSpecs

- `<record>
  <header>
    <identifier>...</identifier>
    <datestamp>...</datestamp>
    <setSpec>cs</setSpec>
    <setSpec>math</setSpec>
  </header>```
ListIdentifiers Changes

- metadataPrefix is a required parameter
  - the records in different formats may be updated independently

- More information returned - complete headers instead of identifiers
  - datestamp, identifier, status, setSpec
Provenance Containers

- `<about>` is now a repeatable field
- Special provenance container for historical information - esp. w.r.t hierarchical harvesting
  - `<provenance>`
    - `<originDescription harvestDate="..." altered="true">`
      - `<baseURL>...</baseURL>`
      - `<identifier>...</identifier>`
      - `<metadataPrefix>...</metadataPrefix>`
      - `<datestamp>...</datestamp>`
      - `<harvestDate>...</harvestDate>`
    - `</originDescription>`
  - `</provenance>`
Granularity of Harvesting

- Now supports times in addition to dates!
- Finest granularity in Identify response
- Formats:
  - YYYY-MM-DD
  - YYYY-MM-DDThh:mm:ssZ
- Must support coarser granularities
- Error on finer granularities
More Flow Control

resumptionTokens now take additional optional parameters for better flow control

- completeListSize
- cursor
- expirationDate
Timezones

- Everything is now explicitly GMT
- If you use local dates, you may need to delay new entries to ensure stable harvesting
- If all sites use GMT, there will be no need to overlap dates by 1 day to cater for timezone differences
resumptionToken idempotency

- resumptionTokens can be sent repeatedly to a server and the response should be identical (or supersets)
- This helps with mid-stream errors in harvesting from large archives
Now What?

- Updated toolkits are gradually being announced on OAI-Implementers list
- Repository Explorer/ARC/OAIA were all compliant before release!

Future Work

- Experiments and Extensions
- SOAPification
- Other protocols