


Introduction to the OAI Protocol for Metadata Harvesting **Version 2.0**



Hussein Suleman
Virginia Tech DLRL
25 March 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 ?

- ❖ Scheduled for release in mid-2002
- ❖ Detailed schedule
 - Alpha test - since 1 March 2002
 - Beta test - 1 April 2002
 - Expected release - 1 June 2002
- ❖ Anyone want to be an alpha or beta tester ?

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
- ✿ Stay tuned...

Some New Features

- ✿ Provenance containers
- ✿ Granularity of harvesting
- ✿ Orthogonality of setSpecs
- ✿ More flow control
- ✿ XML error codes
- ✿ More consistent schemas

Provenance Containers

- ✎ `<about>` is now a repeatable field
- ✎ Special provenance container for historical information - esp. w.r.t hierarchical harvesting

- `<provenance>`
 - `<originDescription>`
 - `<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
- ✿ Full range: YYYY, YYYY-MM, YYYY-MM-DD, YYYY-MM-DDThh:mmZ, YYYY-MM-DDThh:mm:ssZ, YYYY-MM-DDThh:mm:ss.sZ
- ✿ Must support coarser granularities
- ✿ Error on finer granularities

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>
```

# More Flow Control

- resumptionTokens now take additional optional parameters for better flow control
  - retryAfter ? resumeAfter ?
  - completeListSize
  - cursor
  - expirationDate

# XML Error Codes

- ❧ `<error code="badArgument"/>`
- ❧ Errors within the regular responses - no more HTTP 404s
- ❧ Illegal verb results in an Identify response
- ❧ Philosophy: separation of PMH and HTTP

# More Consistent Schemas

- ✿ Use standard schemas where available e.g., Dublin Core
- ✿ Use single namespace for all requests/responses
- ✿ No more arbitrary restrictions on format of metadataPrefix, setSpec, etc.

# Now What ?

- ✿ Anyone want to try it ?
- ✿ Some features are still hotly debated
- ✿ Any and all feedback will be appreciated
- ✿ Future Work
  - Experiments
  - Extensions
  - SOAPification
  - Other protocols