# Recap: SIGIR 2001 OAI Workshop

19 September 2001 -- OAI Provider Workshop,
University of Illinois at Urbana-Champaign

Thomas G. Habing *(thabing@uiuc.edu)*
University of Illinois at Urbana-Champaign

# Overview

- Eleven attendees (slightly over half of those originally scheduled)
- Broad interest in OAI from participants:
  - Only a few of the participants had actual experience implementing OAI, but most had potential OAI projects
  - Tech reports, NCSTRL, Physics E-Prints, West African Digital Library, National Gallery of Spoken Word, Bibliographies, Personal Archives (Kepler), Thesauri, etc.

- What exactly is an Open Archive?
  - How is it related to Digital Libraries? How is it related to traditional paper archives? Is its function preservation or access, both or neither?
  - Is it metadata only, or can it be full-text? How about non-textual data? Can a thumbnail of an image be considered metadata? Can OAI support non-textual data?
    - The OAI PMH seems to be fairly neutral on these issues. It can support any well structured data, including non-textual data if it is properly encoded and wrapped in XML.
  - The OAI definitions of terms such as Archive and Record may conflict with other usage. This needs to be made clear in the spec or FAQ.

- Dublin Core
  - Is it useful as a least-common-denominator?
  - Can service providers build useful, value-added systems with only DC metadata?
  - What about objects for which DC makes little sense, such as people?
    - The consensus seemed to be that DC should continue to be required even when the mappings were forced or contrived, as with people, but that some guidance or best practice for mapping these 'oddball' cases should be provided.

- Access and Authority Control
  - Which is the authoritative record, especially if brokering or mirror sites are developed?
  - How can you prove an item existed at a certain time in a certain repository?
  - Does the protocol need to support its own access controls, or will the HTTP(S) access and user authentication protocols suffice?

- Rights Management
  - We need machine readable policies for the <about> section of a record.
    - Enumerated list of values with pre-specified meanings
    - Hyperlinks to external rights management statements or systems

- Sets
  - Complex use of sets
    - Used for submitting general queries to a repository.
    - How should ListSets respond in these cases?
    - Could ListSets point to an external list such as PACS?
  - How to request the number of records per set?
  - How to lists the sets for a given record?
  - How to request records not belonging to any set?
  - The syntax for the setSpec should be expanded to support arbitrary Unicode (not just ASCII)
    - Not "([A-Za-z0-9])+(:[A-Za-z0-9]+)*"
    - But maybe "([^:])+(:[^:]+)*"
  - Issues with moving records between sets?

- XML Metadata
  - How to support multiple namespaces
  - How to handle schema versioning
  - What does "oai_dc" mean? Why not just "dc"?
  - How can some specific metadata fields, and not others, be requested?
    - Maybe by defining different metadata formats
  - How can single, invalid XML records be handled in the middle of a much larger response, without invalidating the entire response?
    - Possible treat the results as normal text (don't try to parse as XML) until the entire response is complete. Then try to validate records in a batch. The resumptionTokens could be extracted using common (non-XML) text parsing techniques.

- Datestamp
  - May want to add an optional time component to the datestamp to support finer granularity of harvesting, and more dynamic repositories.
    - Consensus is that currently a two-day harvest overlap is required to accommodate the timezone and datestamp granularity issues.
  - There may be collections for which an OAI Datestamp may not be readily available.
    - Could Datestamp be made optional?
    - Could another date be reasonably substituted, such as today's date or the creation or publishing date of the object itself, without harming the protocol?
    - Need to educate metadata creators and maintainers that a datestamp for the metadata itself is important.

- OAI Service Providers
  - Consensus is that harvesting records for local search is better/simpler than distributed search systems, such as Dienst.
  - Hybrid systems are possible, combining local search with distributed search.
  - Distributed architectures:  OAI metadata brokering or mirror sites may be used to improve the performance of the overall system
    - Deduping, handling duplicate records may become an issue
  - Workflow Systems
  - Citation Linking

- Communities
  - Community-based OAI registries for providers.
  - Need best practice guidelines for utilizing OAI for different communities, such as for traditional archives (EAD), museums, publishers, etc.
  - Need tools to make support of OAI easy.
  - Community-based OAI working groups, possibly affiliated with the DublinCore.org, might be useful.
  - Develop application profiles for different communities: best practice guidelines, custom XML metadata schemas, RDF schemas, standard thesauri, XSLT stylesheets for transforming metadata, etc. - all collected together in one place.

- Internationalization, multilingual metadata
- Best practices for handling deleted records
- Should protocol support an explicit expiration mechanism for resumptionTokens
- Errors
  - Does the protocol need standard or suggested 'reason' phrases to use for the different HTTP 400 errors? Possibly.
  - Should error handling be done in the XML body of the HTTP response, instead of via the HTTP status code? There was some sentiment that it should.
  - Should an effort be made to de-couple OAI from the HTTP protocol?
- Should SOAP be explored as a wrapper for OAI?