National Science Digital Library (NSDL)

Core Infrastructure

Metadata Repository ("union catalog")

Naomi Dushay
Cornell University

Aggregator Issues: Deleted Records

- indicated but transient
  - reharvested soon enough – no problem, mark **our** copy "deleted"
  - reharvested as "disappeared"
- not indicated
  - reharvested as "disappeared"

Solution?

- "Full reharvest"
  - Mark all the site's records in **our** repository "deleted"
  - Do a full harvest
  - Ingest each newly retrieved record into **our** repository, "un-deleting" if we over-write an old record

Aggregator Issues:
Poor Quality Harvested Metadata

What is poor quality?

- OAI protocol problems
- XML problems
- metadata "content" problems
- … it's a *knowledge gap*

Solutions?

- Clearer documentation
  - "OAI for Dummies" - details coming up
  - "XML for OAI Dummies" - details coming up
  - "Metadata for dummies" – details coming up
- More, better self-test tools for sites …
  - error messages for "dummies"
  - stricter, more thorough OAI validation checking
  - more XML schema validation of metadata
    - user friendly, extremely low entry
- OAI static repository
- Normalize metadata locally

“OAI for Dummies”

- identifiers (OAI vs. DC; the need for persistence)
- datestamps (<responseDate> vs. header <datestamp> vs. dc:date; format confusion)
- resumptionTokens (exclusive argument, stateless vs. stateful)
  - chunk size recommendation or rule of thumb
  - “stateless resumption token” general scheme for User Guidelines doc? (To be indicated via Identify response description?)
- about containers and their use (additional examples)
  - distinction between “about the metadata” and “about the resource” concepts (dc:rights vs. rights described in about)
- sets
- **multiple metadata formats are allowed** (many sites believe OAI means simple DC only)
  - **MUST** have valid XML schema
- Web service vs. flat file
  - HTTP vs. HTML

We offer:

- Donna Bergmark’s OAI validation tool (email me to get more info)

“XML for OAI Dummies”

- encoding
  - XML encoding
  - character encoding (UTF-8, UTF-16, etc.)
  - URL encoding
  - XML vs. URL vs. character
- Namespaces
  - what are they for? how are they used?
  - full syntax explanation
    - declaration, prefix, URI, scope, default, missing …
- XML schemas
  - what are they for? how are they used?
  - xsi:schemaLocation
  - validation – what it will and won’t find
  - validators – what’s there, what’s best for “my” site?

“Metadata for Dummies”

- simple DC vs. qualified DC
- What refers to metadata, what refers to resource?
  - Think identifiers
  - Think rights
- other …

We offer:

- Metadata Primer (currently being revised)
  - email me to get URL

Normalize Metadata Locally

- Aim to improve services (e.g. search results)
- Improve quality when possible
  - Supply missing information, if known
    - site is about Math; add “Mathematics” <dc:subject>
  - Correct wrong information, when possible
    - “text/pdf” → “application/pdf” in <dc:format>

- for further details, read our paper *Analyzing Metadata for Effective Use and Re-use,* submitted to DC 2003
    - email me to get URL for draft