Metadata Standards & Applications

7. Approaches to Models of Metadata Creation, Storage, and Retrieval

Goals for Session

- Understand the differences between traditional vs. digital library …
  - Metadata creation
  - Storage options for metadata and content
  - Retrieval and discovery issues

Creating Metadata Records

- The "Library Model"
  - Trained catalogers, one-at-a-time metadata records
- The "Submission Model"
  - Authors create metadata when submitting resources
- The "Automated Model"
  - Automated tools create metadata for resources
- Combination approaches

The Library Model

- Records created "by hand," one at a time
- Shared documentation and content standards (AACR2, etc.)
- Efficiencies achieved by sharing information on commonly held resources
- Not easily extended past the "Granularity Assumptions" in current practice

The Submission Model

- Based on author or user generated metadata
- Can be wildly inconsistent
  - Submitters generally untrained
  - May be expert in one area, clueless in others
- Often requires editing support for usability
- Inexpensive, but not satisfactory as an only option

The Automated Model

- Based largely on text analysis; doesn't usually extend well to non-text or low-text
- Requires development of appropriate evaluation and editing processes to support even minimal quality standards
- Still largely research; few large, successful production examples ... Yet
- One simple automated tool to try: http://www.ukoln.ac.uk/metadata/dcdot/

"Like any other data management processes (such as data normalization or the control of information quality), the creation of metadata requires an investment of resources. However, the relationship between investment in metadata creation and the resulting level of resource discoverability is not linear. The more elements from a metadata set that are implemented, the greater the investment of resources that is required. In addition, the more data elements used, the greater the chances for error and divergence among record creators and implementations."

-- Norm Friesen, CanCore Guidelines: Introduction.

Combination Approaches

- Combination Machine and Human created Metadata
  - Ex.: INFOMINE (http://infomine.ucr.edu/)
  - Check out their tool: (http://assigner.ucr.edu/)
- Combination metadata and content indexing
  - Ex.: NSDL (http://nsdl.org)

Content “Storage” and Retrieval Models

- ‘Storage models’ in this context relate to the relationship between the metadata and content (not the systems that purport to ‘store’ content for various uses)
- This relationship affects how access to the information is accomplished, and how the metadata either helps or hinders the process (or is irrelevant to it)

Common ‘Storage Models’

- Content with metadata
- Metadata only
- Service only

Content with Metadata

- Examples:
  - HTML pages with embedded 'meta' tags
  - Most content management systems (though they may store only technical or structural metadata)
  - Text Encoding Initiative (TEI), a full-text markup language (as an example of an application, see the Comic Book Markup Language at http://www.cbml.org/)
- Often proves difficult to scale
- Not optimized to manage metadata well over time

Metadata only

- Library catalogs
  - Web-based catalogs often provide some services for digital content
- Electronic Resource Management (ERM) Systems
  - Provide metadata records for title level only
  - Usually intended to manage licensing and access to article level information
- Metadata aggregations (a.k.a. 'Digital Libraries' or 'Portals' linking to other people's content)
  - Using APIs or OAI-PMH for harvest and re-distribution

Service only

- Often supported partially or fully by metadata
- Google, Yahoo (and others)
  - Sometimes provide both search services and distributed search software
  - Using metadata from libraries as part of their large-scale digitization projects
- Electronic journals (article level)
  - Linked using 'link resolvers' or available independently from websites
  - Have metadata behind their services but don't generally distribute it separately

Common Retrieval Models

- Library catalogs
- Web-based ("Amazoogle")
- Portals and federations

“Old” Library Catalogs

- Based on a ‘Granularity Consensus’ increasingly mysterious to users
- Include expectations of uniformity of information content and presentation
- Designed to optimize recall and precision
- Addition of relevance ranking and keyword searching by vendor systems of limited value (the only ‘text’ used is the metadata itself, not the content)
- Retrieval options limited by LMS vendor ignorance of library data

“New” Library Catalogs

- ENDECA
  - North Carolina State University Libraries in 2006, was one of the first to experiment with new catalog technologies using legacy metadata
- eXtensible Catalog Project
  - University of Rochester attempting to provide a FRBR-ized catalog and integrated access to previously “silo-ed” data managed by libraries.

Web-based

- The "Amazoogle" model:
  - Lorcan Dempsey: "Amazon, Google, eBay: massive computational and data platforms which exercise strong gravitational web attraction."
  - Based primarily on full-text searching and link- or usage-based relevance ranking (lots of recall, little precision)
  - Some efforts to combine catalog and Amazoogle searches (ex.: collaborations with WorldCat)

Portals and Federations

- **Portals**: defined content boundaries
  - Some content also available elsewhere
  - ex.: Specific library portals, subject portals like Portals to the World (ex. http://www.loc.gov/rr/international/portals.html)
- **Federations**: protected content and services
  - Often specialized services based on specifically purposed metadata (ex.: BEN-http://www.biosciednet.org/portal/)

Information Discovery & Retrieval

- Z39.50
  - Basis for most federated search applications in current library software
- SRU (Search and Retrieval Via URL)
  - Seen as a replacement for Z39.50
  - To learn more about it see: http://www.loc.gov/standards/sru/index.html
- Federated search (Metasearch)
  - Simultaneous search multiple data sources
  - Not much uptake, seen as only as robust as its weakest link

Newer Possibilities

- RDF data is increasingly using options like the Simple Protocol And RDF Query Language (SPARQL)
  - Currently a W3C Recommendation
- Approaches using graphs, ontologies, topic maps, etc. seen as more attractive as Semantic Web technologies become more robust
  - These based more on "statements" than "records" …

Data Management Challenges for Libraries

- Moving from text to URIs for controlled values
  - Including personal and organization names as well as controlled concepts and topics
- Developing useful and efficient normalization and "smartening up" processes
- Ensuring that their changes are visible to downstream services

Can You Tell?

- Can you tell what's going on behind these sites?
- How are they organized?
- What creation and storage models are used?
  - *Plant and Insect Parasitic Nematodes:* *http://nematode.unl.edu/*
  - *Internet Movie Database:* http://www.imdb.com/