Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data Providers of Cultural Heritage Materials

Arwen Hutt, University of Tennessee
Jenn Riley, Indiana University

OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting
- Originally developed for sharing metadata about e-prints
- Two players
  - Data providers
  - Service providers
- Requires unqualified Dublin Core be exposed for all resources, but supplemental metadata formats are allowed

Dublin Core [Unqualified]

- Simple, flexible metadata format
- 15 elements
  - All repeatable
  - None required
- “Core” across all knowledge domains

“Cultural heritage” defined

- The intellectual creative and material output of society
- Libraries, museums and archives generally considered cultural heritage institutions
- Often primary source materials
- Tend to be older analog digitized for network access

Significant variability in OAI metadata

- Ward: found that only a small number of DC elements were used in the majority of OAI records
- Liu: Arc service provider studied controlled vocabulary usage in DC subject, type, format, language, and date fields
- NSDL: found errors missing data, incorrect data, confusing data, insufficient data
- UIUC: date, coverage, format, and type vocabulary varies significantly

Goals of the study

- Focus on cultural heritage community
- Examined 3 DC fields: date, creator, contributor
    - Semantic content
    - Syntactic form
- Results could inform community best practices
- One step towards improving the overall quality of OAI metadata

Harvesting statistics

- Successfully harvested metadata from 35 data providers
- 750,945 total records harvested
- 5% sample* from each data provider taken for analysis (37,564 records)

* Minimum of 1 record per provider, values rounded up to the nearest whole number

Processing steps

- Date, creator, contributor elements extracted into “silos”
- Repeated values grouped, keeping connections between elements and the records in which they appeared
- Certain characteristics tracked about each element
- Example

Characteristics recorded for all elements

- The presence of multiple discrete values in a single element
  <creator>Hutt, Arwen; Riley, Jenn</creator>
- The presence of pseudo-qualifiers within the value that refined the meaning of the element
  <creator>Berlin, Irving [composer]</creator>
- Whether the value was appropriate within the specified element based on DC rules and usage guidelines
  <date>Las Vegas, Nevada</date>

Additional characteristics of <date>

- The semantic type of the value (creation, copyright or digitization)

  <date>2000</date>

- The general specificity of the date (single date, range or period)

  <date>19th Century</date>

- Indication that a date is not definitive (that it is estimated or approximate)

  <date>ca. 1930</date>

- Whether the value is purely numeric or contains non-numeric text

  <date>March 18, 1902</date>

Additional characteristics of <creator> and <contributor>

- The semantic type of the value (personal name, corporate name or other)

  <creator>Newton, Isaac</creator>

- Whether the entity is known, unknown or ambiguous

  <creator>Vermeer, Johannes, 1632-1675 ?</creator>

- Whether the value is inverted or in direct order

  <creator>Charles Schultz</creator>

Strategies for categorization

- Automatic
  - Iteratively developed
  - Pattern matching
  - Identification of commonly occurring values
- Manual
  - Where feasible
- Not perfect!

Findings for <date>

- Values largely appropriate for element
- Few “pseudo-qualifiers”
- Different events represented
- Values mostly numeric
- Many dates not expressible in W3CDTF

Findings for <creator>

- Values largely appropriate for element
- Most were personal names
- Many “pseudo-qualifiers,” in comparison to other elements
- Often included information intended to disambiguate a name
- Some indication of the use of controlled vocabularies, but many different name forms present

Findings for <contributor>

- Used infrequently
- Many values inappropriate for element
- Majority personal names, but higher proportion of corporate names than occurred in <creator>
- Few “pseudo-qualifiers”

OAI DC record & intellectual object

- 1:1 principle – each DC record describes only one version of a resource

BUT

- Cultural heritage materials often digitized from analog originals, resulting in multiple versions of each intellectual object

OAI DC record & intellectual object

- Two choices for data providers
  - Adhere to 1:1 rule but omit pertinent information
  - Violate the 1:1 rule but create more complete records
- Many data providers in practice violate the 1:1 rule

OAI DC record & aggregated search environment

- Extraction of records from original collection context
- Aggregation with records from other collections

Moving towards better metadata – some possibilities

- Remove the OAI requirement for simple Dublin Core (or “the Nuclear Option”)
- Develop best practice documentation for cultural heritage materials that deviate from current DC best practice
- Combination of data provider education and service provider normalization
- Improved communication between data and service providers
- Encourage use of other metadata formats supplementing simple DC

Some other relevant initiatives

- Digital Library Federation and NSDL OAI and Shareable Metadata Best Practices Working Group
  - Development of general OAI best practices
  - Development of strategies for communication with vendors
- DLF Aquifer Metadata Working Group
  - Development of profile for DLF institutions (strong focus on cultural heritage)
  - Recommendations for specific metadata elements

Plans for extension of this research

- Primary analysis of the subject, coverage and publisher elements
- Analyze temporal information across date, subject and coverage elements
- Analyze geographic information across subject and coverage elements
- Analyze name information across creator, contributor and publisher elements

These presentation slides:

http://www.dlib.indiana.edu/~jenlrile/presentations/jcdl2005/jcdl2005.ppt

Arwen Hutt
Metadata Librarian
University of Tennessee Digital Library Center

ahutt@utk.edu

Jenn Riley
Metadata Librarian
Indiana University Digital Library Program

jenlrile@indiana.edu