OLAC State of the Archives

A summary of implementation practices during the first year

Overview

- Review of archives descriptions
- Review of element usage
  - How it was used
  - Problem practices
  - Suggestions for improvement
  - Changes already anticipated
- Summary: recommendations for implementation aspects in need of guidance

Archives descriptions

- Some good, some really lacking (none in the middle for reviews submitted)
- Most often missing:
  - Curator
  - Contact information
  - Access terms and instructions
- More thorough completion needed as a requirement for registration?

The Elements

- 15 elements from DCMES
- 9 additional elements unique to OLAC

Contributor & Creator

- General meaning of elements clear
- Distinction between these two not consistent
- Problematic practices:
  - Multiple names in single element instance
  - Name entry form not ready for sort
  - Quotation marks enclosing corporate names "Institute for Slovene Language ""Fran Ramovs"", Slovene Academy for Sciences and Arts, Ljubljana, Slovenia"
  - Inconsistency in corporate name forms

Contributor & Creator (cont.)

- Is it a problem to have so much information loaded in one element content?
  "Alexandra Jarosov, Slovak Academy of Sciences, Bratislava (sasaj@juls.savba.sk) editorship, corrections Vladimir Benko; Comenius University, Bratislava (jazybenk@savba.savba.sk)."
- Suggestions
  - One name per element instance
  - Surname, firstname order; Main unit, subunit order
  - No quotes—if name is a translation from its usual form or not usually given in English, use the lang attribute
  - Means to identify the first author: can the order of instances of an element be significant?
- Creator and Contributor developments
  - OLAC Role as an extension applicable to Contributor element through a coded attribute

Coverage

- Used creatively by one archive for extent information
- Good potential for use of existing vocabularies as extensions to improve consistency

Date

- Lots of kinds of dates
- Problematic practices:
  - Refining terminology (“recorded on”, “donated on”) incorporated into the element text
  - Coded year value given then mm/dd/yy value given in element text
- Date developments
  - DCQ has 8 refining terms for Date: created, valid, available, issued, modified, dateAccepted, dateCopyrighted, dateSubmitted

Description

- Wide variety of use—a “catch all” concept:
  - Prose description of resource—an abstract
  - Lists of subject terms
  - Description of container/location
  - Extent
  - Condition
  - Access requirements and assistance

Case 1
<description>Telephone conversations Material type: 45 minute cassette Condition: good</description>

Description (cont.)

Case 2
<description>pronunciation</description>
<description>Hub5-LVCSR, EARS</description>
<description>1500</description>
<description>Number of CDs: 0</description>
<description>Recommended applications: speech recognition</description>
<description>Member license: [a URL]</description>
<description>Nonmember license: [a URL]</description>
<description>Online documentation: [a URL]</description>
<description>Readme file: [a URL]</description>

Description (cont.)

- Other perhaps more suitable elements:
  - Format (for extent)
  - Subject
- Description developments:
  - DCQ has 2 refining terms: tableOfContents, abstract

Format and its refinements

- Several different semi-controlled vocabularies in evidence
- Most often used for IMT (sometimes coded, but not clear if repository really meant a Type code, not an Internet Media Type code)
- Also for medium and extent information (however, no instance for either DC qualifier was actually specified)
- OLAC had 5 refinements (cpu, encoding, markup, os, sourcecode) but each of these was used very little, if at all
- Format developments:
  - OLAC extensions to Format: OS, CPU, Sourcecode,
  - Markup and character set encoding awaiting attention

Identifier

- Definition: An unambiguous reference to the resource within a given context
- Problem practices:
  - Many non-unique Identifier URLs:
    - In a few archives, multiple resources were 'identified' with the same URL, usually availability info or a further description, but not the resource itself
    - Often apparently mistaken as the only place one could put a URL associated with this resource
    - Sometimes incomplete (relative?) paths given
  - Some identifiers seemed useful only to the archives, but were not relevant for resource discovery or request for access

Identifier (cont.)

- Where does availability information belong?
  - It was placed here as well as in Description, Publisher, Relation, Source, and Rights elements by various archives
  - 'Available' as a refinement pertains to Date, not to other aspects of availability
- Identifier would probably benefit from a more thorough best practice document

Language and Subject.Language

(grouped here because of structural similarity)

- Two of the cleanest elements ☺
  - It contained language name or code
  - It was usually repeated for multiple languages
- Relatively low use of the attribute supplying OLAC language code 😐
- Clarification between these still needed for some archives

Publisher

- Usually a publisher or the archives itself, sometimes with URL, sometimes URL given in separate instance of element
- Problem practice:
  - One archive used it for host publication information, which should be in Relation

Relation

- IsPartOf and hasPart used most frequently, either through refinement code or noted in element content
- "Previously", "See" and "Recording on" most frequent of un-coded relationships ("Previously" could utilize "Replaces")
- DCQ offers many qualified terms for relation

Rights

- Not extensively used
- Not clearly understood

Definition: Information about rights held in and over the resource.
Comment: Typically, a Rights element will contain a *rights management statement* for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.

Rights (cont.)

- Problem practice:
  - Copyright statement should be in text of element, not in the 'code'
- Rights developments:
  - With the Access extension on Rights, OLAC is integrating access and permitted use
  - Leave the work to the content of the element or a referral to additional information
- Additional good practice guidance is needed regarding parameters of protection: duration of restriction, entity with authority to override, expectations placed on users

Source

- Should refer to another resource from which the described resource is derived
- Problem practices:
  - Identifier was used (repeated) for what clearly had to be a Source resource URL (based on contextual content of record)
  - Source was used to give information on the linguistic consultant, with lengthy description. The whole would have been more appropriate in Description element. (Contributor was used also)
  - Source was used to specify an entity responsible for development, creation, donation, etc. of the resource (in one a Ph.D. granting institution is named, another the gov. agency responsible, others, SIL is named)

Subject

- Problem practice:
  - Element should be repeated for multiple subject terms
- Subject developments
  - OLAC extensions for Language, Linguistic field, Discourse type
  - DCQ offers LCSH, MESH (vocabularies), DDC, LCC, UDC (classifications)

Type

- Exhibited perhaps the most different archive-specific interpretations of its use
- Evidence of different vocabularies for type used by numerous archives
- When coded, the codes were generally applied correctly
- Abused by some poor mappings
- Type developments
  - OLAC maintaining best practice application of DC Type vocabulary

Type.Linguistic

- Confusion in use evident
- Metadata better placed elsewhere
  - Description
    <type>'A Comparison of Poman and Yuman' (MA Thesis)</type>
  - Subject
    <type>Grammar, morphology, verbal suffixes</type>
- Type.Linguistic developments
  - OLAC Linguistic Type extension for Type significantly changed

Type.Functionality

- Not used
- Highly desired—metadata placed in:
  - Type
    <type>Speech analysis, Speech editing, Speech processing</type>
  - Description
    <description>Recommended applications: speech recognition, spoken dialogue systems</description>
- Functionality development: suggestion for a new element

MORE WORK

- More thorough best practice guidelines for:
  - Description
  - Identifier
  - Rights
  - Subject qualifiers and extensions (dealing with overlap, use of multiple schemes)
  - Type and its extensions
- The definitions and controlled vocabularies have to be in order FIRST