The OLAC Metadata Set

Gary Simons

Workshop on The Digitization of Language Data: The Need for Standards
21-24 June 2001

What is metadata?

- “Structured data about data”
- Descriptive information about a resource whether it be physical or electronic
- Content designed for resource discovery
- Format designed for automated searching

An OLAC metadata description

<olac><date code="1986" />
<creator>Derbyshire, Desmond C.</creator>
<title>Topic continuity and OVS order in Hixkaryana</title>
<body>... in Joel Sherzer and Greg Urban (eds.), Native South American discourse , 237-306. Berlin: Mouton.</body><relation refine="isPartOf">
<title>Die Bremer Stadtmusikanten</title>
<identifier>http://www.ethnologue.com/show_work.asp?id=22059</identifier>
</relation><type code="Text" />
<subject>Word order</subject>
<subject>Topic</subject>
<subject>Typology</subject>
<subject.language code="x-sil-HIX" />
<identifier>http://www.ethnologue.com/show_work.asp?id=22059</identifier>
</olac>

Foundational design decisions

- We need a low overhead metadata set.
  - N.B. The Open Archives Initiative support for multiple metadata formats allows subcommunities to develop richer metadata sets.
- We should build on the Dublin Core metadata set.
- We should extend DC by using the qualification mechanisms recognized by DC.

The XML implementation

- All elements are optional and repeatable
- Use attributes for DC qualifications
  - Refinements: <relation refine="isPartOf">...
  - Encoded values: <date code="2001-06-22"/>
  - Language of element content: <title lang="de">Die Bremer Stadtmusikanten</title>
- Refinements with encoding schemes go in element name: <type.data code="lexicon/bilingual"/>

The fifteen Dublin Core elements

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Title
- Type
**Additional elements for DATA**

- Subject.language
  - A language the resource is about
  - Use `<Language>` for a language the resource is in
- Type.data
  - The nature of the content from a linguistic point of view
  - E.g. transcription, annotation, description, lexicon

**Additional elements for TOOLS**

- For matching DATA with TOOLS
  - Format.encoding
  - Format.markup
- For describing TOOLS
  - Format.cpu
  - Format.os
  - Format.sourcecode
  - Type.functionality

**Controlled vocabularies**

- Closed enumerations of allowed values for `refine`, `code`, and `lang` attributes
- To improve success of resource discovery
  - Recall – % of relevant resources that are found
  - Precision – % of found resources that are relevant
- Use element content as an escape hatch
  - When the right term is not in controlled vocabulary
  - When the term needs refinement or explanation

**Elements with DC vocabularies**

<table>
<thead>
<tr>
<th>Element</th>
<th>Refine attribute</th>
<th>Code attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>DC-Qualifiers</td>
<td></td>
</tr>
<tr>
<td>Relation</td>
<td>DC-Qualifiers</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td>DC-Type</td>
</tr>
</tbody>
</table>

**OLAC-Language**

- Used for
  - `Lang` attribute on all elements
  - `Code` attribute on `<Language>`
  - `Code` attribute on `<Subject.language>`
- Terms in the vocabulary follow RFC 3066
  - Unambiguous codes from ISO 639: en, fr, eng
  - All codes from Ethnologue: x-sil-HIX
  - Ancient languages at LINGUIST: x-LL-???

**Other OLAC vocabularies**

<table>
<thead>
<tr>
<th>Element</th>
<th>Refine attribute</th>
<th>Code attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributor, Creator</td>
<td>OLAC-Role</td>
<td>OLAC-Functionality</td>
</tr>
<tr>
<td>Format</td>
<td>OLAC-Format</td>
<td></td>
</tr>
<tr>
<td>Format.cpu</td>
<td>OLAC-CPU</td>
<td></td>
</tr>
<tr>
<td>Format.encoding</td>
<td>OLAC-Encoding</td>
<td></td>
</tr>
<tr>
<td>Format.os</td>
<td>OLAC-OS</td>
<td></td>
</tr>
<tr>
<td>Format.sourcecode</td>
<td>OLAC-Sourcecode</td>
<td></td>
</tr>
<tr>
<td>Rights</td>
<td>OLAC-Rights</td>
<td></td>
</tr>
<tr>
<td>Type.data</td>
<td>OLAC-Data</td>
<td></td>
</tr>
<tr>
<td>Type.functionality</td>
<td>OLAC-Functionality</td>
<td></td>
</tr>
</tbody>
</table>