## 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>
<relation refine=“isPartOf” >In Joel Sherzer and Greg Urban
(eds.), Native South American discourse , 237-306. Berlin:
Mouton.</relation>
<type code="Text" />
<type.data code="description/grammatical" />
<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

| Element | *Refine* attribute | *Code* attribute |
|---|---|---|
| Date | DC-Qualifiers | |
| Relation | DC-Qualifiers | |
| Type | | DC-Type |

## 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

| Element | *Refine* attribute | *Code* attribute |
|---|---|---|
| Contributor, Creator | OLAC-Role | |
| Format | | OLAC-Format |
| Format.cpu | | OLAC-CPU |
| Format.encoding | | OLAC-Encoding |
| Format.os | | OLAC-OS |
| Format.sourcecode | | OLAC-Sourcecode |
| Rights | | OLAC-Rights |
| Type.data | | OLAC-Data |
| Type.functionality | | OLAC-Functionality |