
Proceedings of the IRCS Workshop on Open Language Archives



10-12 December 2002
University of Pennsylvania,
Philadelphia, USA



Edited by Steven Bird and Gary Simons
Funded by the National Science
Foundation

[Program](#) | [Participants](#)

Proposed Standards

- [OLAC Metadata](#) (Gary Simons and Steven Bird)
- [OLAC Process](#) (Gary Simons and Steven Bird)
- [OLAC Repositories](#) (Gary Simons and Steven Bird)

Recommendations

- [Recommended metadata extensions](#) (Gary Simons and Steven Bird)

- [OLAC Language Vocabulary](#) (Gary Simons and Anthony Aristar)
- [OLAC Access Vocabulary](#) (Heidi Johnson)
- [OLAC Linguistic Field Vocabulary](#) (Helen Aristar Dry and Michael Appleby)
- [OLAC Role Vocabulary](#) (Heidi Johnson)
- [OLAC Linguistic Type Vocabulary](#) (Heidi Johnson and Helen Aristar Dry)
- [OLAC Discourse Type Vocabulary](#) (Heidi Johnson and Helen Aristar Dry)

Notes

- [A Gentle Introduction to Metadata](#) (Jeff Good)
- [Third party extensions](#) (Gary Simons and Steven Bird)
- [A query facility for selective harvesting of OLAC metadata](#) (Gary Simons)
- [Viser: A virtual service provider for displaying selected OLAC metadata](#) (Gary Simons)
- [Specifications for an OLAC metadata display format and an OLAC-to-OAI_DC crosswalk](#) (Gary Simons)

Presentation Slides

- [The OLAC Process and the Protocol for Metadata Harvesting](#) (Gary Simons)
- [OLAC Metadata Format and Extension Mechanism](#) (Steven Bird)

- [OLAC-Language: Linguist Codes](#) (Anthony Aristar)
- [OLAC-Linguistic-Type, OLAC-Linguistic-Field, OLAC-Discourse-Type](#) (Helen Aristar Dry and Gayathri Sriram)
- [OLAC-Language-Technology-Fields](#) (Baden Hughes)
- [OLAC-Role](#) (Heidi Johnson)
- [OLAC-Rights](#) (Heidi Johnson)
- [Notes on Linguistic Type](#) (Helen Aristar Dry)
- [Revised OLAC Vocabulary for Language Technology](#) (Baden Hughes)
- [State Of The Archives Address](#) (Joan Spanne)
- [Outreach](#) [Jeff Good]
- [Integrating ELRA/LDC Metadata into an OLAC Repository](#) (Andrew Cole and Khalid Choukri)
- [IMDI & Endangered Languages Archives](#) (Heidi Johnston)
- [Language Technology WG Work In Progress Report](#) (Baden Hughes)
- [Notes on 2003 Agenda](#)

[Steven Bird](#), University of Melbourne and University of Pennsylvania

[Gary Simons](#), SIL International

sb@ldc.upenn.edu, Gary_Simons@sil.org

[Open Language Archives Community](#)

IRCS Workshop on Open Language Archives

10-12 December 2002

University of Pennsylvania,
Philadelphia, USA



Organized by Steven Bird and Gary
Simons

Funded by the National Science
Foundation

WORKSHOP PROGRAM

Tuesday Morning: Standards

8:15 *BREAKFAST*

8:55 Welcome and Local Arrangements [Steven Bird, Laurel Sweeney]

9:00 Introductions: 1-2 minute introductions to each archive and service [ALL]

9:30 The OLAC Process and the Protocol for Metadata Harvesting [Gary Simons] [[ppt](#) | [pdf](#)]
OLAC Metadata Format and Extension Mechanism [Steven Bird] [[ppt](#) | [pdf](#)]

10:30 *BREAK*

11:00 Group discussion of OLAC Standards
Plenary discussion and resolution

12:30 *LUNCH*

Tuesday Afternoon: Vocabularies

2:00 *Each presentation to give an overview of the vocabulary (explaining a document draft which has been circulated), and identify specific outstanding issues for discussion.*

1a. OLAC-Language: Ethnologue [Gary Simons]

1b. OLAC-Language: Linguist Codes [Anthony Aristar] [[ppt](#) | [pdf](#)]

2. OLAC-Linguistic-Type, OLAC-Linguistic-Field, OLAC-Discourse-Type [Helen Aristar Dry, Gayathri Sriram] [[ppt](#) | [pdf](#)]

3. OLAC-Language-Technology-Fields [Baden Hughes] [[ppt](#) | [pdf](#)]

4. OLAC-Role [Heidi Johnson] [[ppt](#) | [pdf](#)]

5. OLAC-Rights [Heidi Johnson] [[ppt](#) | [pdf](#)]

3:30 *BREAK*

4:00 Group discussions of outstanding vocabulary issues:

1. Rights/Role [Heidi Johnson, chair]
2. Linguistic Field/Type [Helen Aristar Dry, chair]
3. Language Technology [Baden Hughes, chair]

5:30 *CLOSE*

Wednesday Morning: Vocabularies

8:15 *BREAKFAST*

9:00 Working group chairs report back [Heidi Johnson, Helen Aristar Dry [[ppt](#)][pdf](#)] , Baden Hughes [[ppt](#)][pdf](#)]

9:15 Archivists/Librarians Panel: guidance on vocabulary discussions [Joan Spanne, Diane Hillmann, Elaine Westbrooks]

9:45 Further working group discussion

10:30 *BREAK*

11:00 Further working group discussion

12:30 *LUNCH*

Wednesday Afternoon: Archives and Services

2:00 State Of The Archives address [Joan Spanne] [[ppt](#)][pdf](#)
A report on our [archive survey](#)

3:00 Live review of OLAC website

3:30 *BREAK*

4:00 Open Forum for New Initiatives, Subcommunities, ...

1. Outreach [Jeff Good] [[ppt](#)][pdf](#)]
2. Net-DC [Andrew Cole, Khalid Choukri] [[ppt](#)][pdf](#)]
3. IMDI [Heidi Johnston] [[ppt](#)][pdf](#)]
4. Viser [Gary Simons] [[html](#)]

5:15 Presentation of Revised Standards [Gary Simons, Steven Bird]

5:30 *CLOSE*

Thursday Morning: Resolution

8:15 *BREAKFAST*

9:00 Working Group Reports:

1. Language Technology [Baden Hughes] [[ppt](#)][pdf](#)]
2. Rights/Role [Heidi Johnson]
3. Type/Field [Helen Aristar Dry]

10:30 *BREAK*

11:00 Plenary Discussion:

Vocabularies, Revised OLAC Standards, Website

12:30 *LUNCH*

Discussion of 2003 agenda
New Tasks [[ppt](#)][pdf](#)]

2:00

CLOSE

[Steven Bird](#) , [Gary Simons](#)

sb@ldc.upenn.edu, Gary_Simons@sil.org

[Open Language Archives Community](#)

IRCS Workshop on Open Language Archives

10-12 December 2002

University of Pennsylvania,
Philadelphia, USA



Organized by Steven Bird and Gary
Simons

Funded by the National Science
Foundation

WORKSHOP PARTICIPANTS

This list gives the following information:

participant name - archive/service/project (affiliation).

1. Helen Aristar Dry - LINGUIST, OLAC Advisory Board
(Eastern Michigan University, USA)
2. Anthony Aristar - LINGUIST (Wayne State University, USA)
3. Steven Bird - OLAC Coordinator (University of Melbourne,
Australia, and Linguistic Data Consortium, USA)
4. Laura Buszard-Welcher - SCOIL (University of California,
Berkeley, USA)
5. Ru-Yng Chang - AS (Academia Sinica, Taiwan)

6. Chin-chuan Cheng - OLAC Advisory Board (City University of Hong Kong)
7. Khalid Choukri - ELRA (European Language Resources Association, France)
8. Chris Cieri - LDC (Linguistic Data Consortium, USA)
9. Andy Cole - LDC (Linguistic Data Consortium, USA)
10. Alexis Dimitriadis - LTRC (Utrecht University, Netherlands)
11. Jeff Good - CBOLD (University of California, Berkeley, USA)
12. Erik Grostic - AILLA (University of Texas, Austin, USA)
13. Diane Hillman - DCMI (Cornell University, USA)
14. Gary Holton - ANLC (University of Alaska, Fairbanks, USA)
15. Chu-Ren Huang - AS, OLAC Advisory Board (Academia Sinica, Taiwan)
16. Baden Hughes - UQ Flint (University of Queensland, Australia)
17. Heidi Johnson - AILLA, IMDI (University of Texas, Austin, USA)
18. Steven Krauwer - ELSNET (Utrecht University, Netherlands)
19. Terry Langendoen - OLAC Advisory Board (University of Arizona, USA)
20. Haejoong Lee - LDC (Linguistic Data Consortium, USA)
21. Mark Liberman - LDC (Linguistic Data Consortium, USA)
22. Jim Mason - Rosetta (Long Now Foundation, USA)
23. Gary Simons - OLAC Coordinator (SIL International, USA)
24. Joan Spanne - SIL-LCA (SIL International, USA)
25. Gayathri Sriram - LINGUIST (Eastern Michigan University, USA)

26. Zina TucsnaK - ATILF (Analyse et Traitement Informatique de la Language Francaise, France)
 27. Elaine WestbrookS - Virtual Linguistics Lab (Cornell University, USA)
 28. Christopher York - Perseus (Tufts University, USA)
-

[Steven Bird](#), [Gary Simons](#)

sb@ldc.upenn.edu, Gary_Simons@sil.org

[Open Language Archives Community](#)

OLAC Process and OLAC Protocol: A Guided Tour

Gary F. Simons
SIL International

*OLAC Workshop
10 Dec 2002, Philadelphia*

Our original schedule

- Dec 2000: Founding of OLAC
- 2001: Develop standards with alpha test sites
- 2002: Launch to wider community while freezing proposed standards for one year
- 2003: Revise standards and adopt them as version 1.0

2

Give us your feedback

- The workshop notebooks are loose-leaf for a reason!
- If you see anything in a document
 - That is wrong
 - That doesn't make sense
 - That seems like a bad idea
 - Etc.
- Mark it on the page and give the page to one of the editors of the document.
- We'll circulate updates to some documents.

3

OLAC Process

- A "Candidate Standard"
- The document summarizes the governing ideas of OLAC (i.e. the purpose, vision, and core values) and then describes how OLAC is organized and how it operates.

4

Mission statement (p. 3)

- OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by:
 - developing consensus on best current practice for the digital archiving of language resources, and
 - developing a network of interoperating repositories and services for housing and accessing such resources.

5

Vision statement (p. 3)

- The "seven pillars" of language archiving.
- Any user on the Internet should be able go to a single GATEWAY to find all the language resources available at all participating institutions, whether the resources be DATA, TOOLS, or ADVICE. The community will ensure on-going interoperation and quality by following STANDARDS for the METADATA that describe resources and services and for processes that REVIEW them.

6

Core values (pp. 3-4)

- Openness
- Consensus
- Empowering the players
- Peer review

7

Organization (pp. 4-5)

- Coordinators
- Advisory board
- Participating archives and services
- Working groups
- Participating individuals

- See last three sheets in notebook for relevant web site pages

8

Types of documents (pp. 5-6)

- Standards — Obligatory compliance
- Recommendations — Optional compliance
 - About application of standards
 - About other digital archiving practices
- Notes
 - Experimental
 - Informational
 - Implementation

9

Status of documents (pp. 6-7)

- Draft
- Proposed
- Candidate
- Adopted
- Retired
- Withdrawn

10

The document process (pp. 7-8)

- Intellectual property rights: OLAC documents are published under OPL.
- Review: this process establishes if a document is ready to advance in status
- Voting: a means of measuring consensus in a distributed community
 - Release, Revise, Resubmit, Reject

11

Document life cycle (pp. 8-9)

<i>Phase</i>	<i>Status</i>	<i>Promoted by</i>
Development	Draft	Working group
Proposal	Proposed	Community
Testing	Candidate	Implementers
Adoption	Adopted	Community
Retirement	Retired	

12

Working group process (pp. 10-11)

- Anyone can set up a working group. It takes:
 - A purpose germane to mission of OLAC
 - One or more planned documents
 - At least three members representing at least three institutions
 - A designated chairperson
- OLAC provides a web page and an open mailing list

13

Proposed changes for 1.0 (p. 1)

- Section 2: OLAC also solicits anonymous peer review on conformance to standards/recs
- Section 3: "Prospective Participants" removed
- Section 4: Two types of recommendations
- Section 5: Only participating institutions vote on recommendations about application of standards, while participating individuals vote on other recommendations.

14

Open issues

- We probably need to add a section on the "Registration process"
- We have never performed the community voting process, but must in order to move our 1.0 proposals to Adopted status.
 - Is there existing "community ware"?
 - Can someone develop a web app for us?
- A bootstrapping problem with the Process document. Solution: Get consensus of this workshop and the Advisory Board.

15

OLAC Protocol for Metadata Harvesting

- It was a "Proposed Standard"; this new version is a "Draft Standard".
- This document defines the protocol OLAC service providers use to harvest metadata from OLAC data providers. It defines the responses that OLAC data providers must make to the requests of the protocol.
- Based on the Open Archives Initiative protocol for metadata harvesting

16

OLAC-specific requirements

- `Identify` response must have:
 - `<oai-identifier>` description (Section 2)
 - `<olac-archive>` description (Section 3)
- `ListMetadataFormats` must include `olac`
- `ListIdentifiers` with `olac` must return at least one record
- `GetRecord` and `ListRecords` with `olac` must return records that conform to OLAC schema

17

The main changes

- Based on OAI-PMH, version 2.0
 - Only 4 participants need to upgrade from 1.1
 - The rest get a free ride when Vida upgrades
- OLAC-PMH specifies an OAI minimal repository implementation minus one feature:
 - Supporting `oai_dc` is not required
 - OLACA gives a free ride for `oai_dc` support

18

Other changes for OLAC-PMH 1.0

- Add shortLocation to <olac-archive>.
- All repository identifiers must change to be based on a registered Internet domain name.
- Open issue (see To Do). How do we handle repository identifiers where there is no institutional domain name?
 - oai:ore.language-archives.org:repo:id ?
 - oai:repo.language-archives.org:id ?
 - Other?

19

Possible change to PMH document

- The original OAI model: a data provider must implement the protocol interface
 - = A dynamic repository
- A newly specified OAI model (inspired by OLAC's Vida): a data provider may generate an XML document of all items
 - = A static repository
- The PMH document could shift to a focus on repository implementation (incl. both)

20

OLAC Metadata

Steven Bird
University of Melbourne /
University of Pennsylvania

OLAC Workshop
10 December 2002

OLAC Metadata

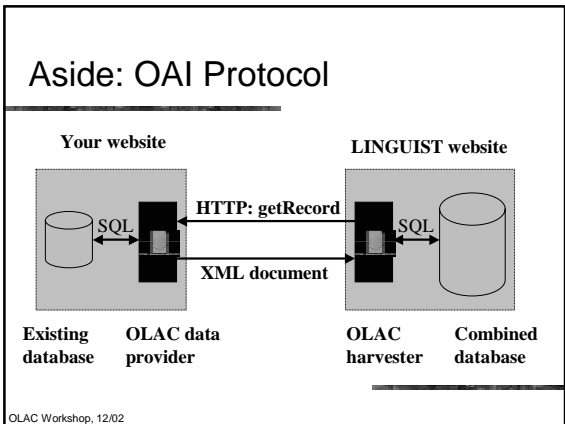
- OLAC Metadata - Simons & Bird
<http://www.language-archives.org/OLAC/metadata.html>
- Draft standard
- Purpose:
 - Define the metadata format
 - Define the extension mechanism

OLAC Metadata

1. Introduction
2. Metadata elements
3. Metadata format
4. OLAC extensions
5. Defining a third-party extension
6. Documenting an extension

1. Introduction

- XML
- OAI framework
- From data provider to service provider
 - How we ship the metadata around
 - Data is stored/presented in other ways



2. Metadata Elements

- 15 DC elements - dublincore.org
- Need to describe language resources with greater precision
- Follow DC recommendation for qualifying elements
 - *Dublin Core Qualifiers*
<http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/>
 - Refinements: meaning of element is narrower, more specific
 - Encoding schemes: controlled vocabularies and standardized formats

Community-specific qualifiers aka "OLAC Extensions"

- Access rights
dc:rights
 - Discourse type
dc:type
 - Language identification
dc:language
dc:subject
 - Linguistic field
dc:subject
 - Linguistic data type
dc:type
 - Participant role
dc:creator
dc:contributor
- Vocabularies to be discussed this afternoon...*

OLAC Workshop, 12/02

Refinements vs encoding schemes

Refinement:

- Role vocabulary, e.g. annotator; translator
role of contributor is more specific

Encoding scheme:

- Linguistic data type, e.g. lexicon; dataset
free-text description is summarized with a restricted term, facilitating precision and recall

Both:

- Subject language, e.g. es; x-sil-BAN
subject is more specific (about language) restricted vocabulary

OLAC Workshop, 12/02

3. Metadata format

- Follows guidelines for DC/DCQ in XML
 1. *Guidelines for implementing DC in XML*
<http://dublincore.org/documents/2002/09/09/dc-xml-guidelines>
 2. *Recommendations for XML Schema for DCQ*
<http://www.ukoln.ac.uk/metadata/dcmi/xmlschema/20021007/>
- Application profile
 - Metadata schema
 - Combines elements from multiple sources
- OLAC = DC application profile for LRs
 1. DC: dc.xsd
 2. DCQ: dcterms.xsd
 3. OLAC extensions

OLAC Workshop, 12/02

Tour of an OLAC record

```
<olac:olac
  xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
  xmlns="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="
    http://www.language-archives.org/OLAC/1.0/
    http://www.language-archives.org/OLAC/1.0/olac.xsd">
  <creator>Bloomfield, Leonard</creator>
  <date>1933</date>
  <title>Language</title>
  <publisher>New York: Holt</publisher>
</olac:olac>
```

OLAC Workshop, 12/02

(1) Container and namespace

```
<olac:olac
  xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
  xmlns="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="
    http://www.language-archives.org/OLAC/1.0/
    http://www.language-archives.org/OLAC/1.0/olac.xsd">
  <creator>Bloomfield, Leonard</creator>
  <date>1933</date>
  <title>Language</title>
  <publisher>New York: Holt</publisher>
</olac:olac>
```

OLAC Workshop, 12/02

(2) XML Schema information

```
<olac:olac
  xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
  xmlns="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="
    http://www.language-archives.org/OLAC/1.0/
    http://www.language-archives.org/OLAC/1.0/olac.xsd">
  <creator>Bloomfield, Leonard</creator>
  <date>1933</date>
  <title>Language</title>
  <publisher>New York: Holt</publisher>
</olac:olac>
```

OLAC Workshop, 12/02

(3) DC namespace & content

```
<olac:olac
  xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
  xmlns="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="
    http://www.language-archives.org/OLAC/1.0/
    http://www.language-archives.org/OLAC/1.0/olac.xsd">
  <creator>Bloomfield, Leonard</creator>
  <date>1933</date>
  <title>Language</title>
  <publisher>New York: Holt</publisher>
</olac:olac>
```

OLAC Workshop, 12/02

Using DC Qualifiers

- Extra namespace declaration:
xmlns:dcterms="http://purl.org/dc/terms/"
- Qualified element:
<dcterms:created
 xsi:type="dcterms:W3C-DTF">
 2002-11-28
</dcterms:created>
- "created" is a refinement of date
 - refinement relationship is represented in the dcterms schema ("substitutionGroup")

OLAC Workshop, 12/02

xml:lang attribute

- the language of the *element content*
- expressed using RFC 1766

```
<title xml:lang="x-sil-LLU">
  Na tala 'uria na idulaa diana</title>

<dcterms:alternative xml:lang="en">
  The road to good reading</dcterms:alternative>
```

- no need to declare xml namespace

OLAC Workshop, 12/02

4. OLAC extensions

- xsi:type - a feature of XML Schema
- ... xsi:type="olac:language" ...
 - xsi = namespace for XML Schema Instance
 - value = complex type
 - overrides the type declared for the element
 - new type must be validly derived from the overridden type
- optional code attribute
- element content for comments

OLAC Workshop, 12/02

Example: Language

1. <subject>Dschang</subject>
2. Refinement only:
<subject xsi:type="olac:language">
 Dschang
</subject>
3. Refinement and encoding scheme:
<subject xsi:type="olac:language"
 code="x-sil-BAN"/>

OLAC Workshop, 12/02

Example: Language

```
<xs:complexType name="language">
  <xs:complexContent mixed="true">
    <xs:extension base="dc:SimpleLiteral">
      <xs:attribute name="code"
        type="olac-language" use="optional"/>
    </xs:extension>
  </xs:complexContent>
</xs:complexType>
```

OLAC Workshop, 12/02

Example: Language

```
<xs:simpleType name="olac-language">
  <xs:restriction base="xs:string">
    <xs:enumeration value="aa"/>
    <xs:enumeration value="ab"/>
    <xs:enumeration value="ae"/>
    <xs:enumeration value="af"/>
    <xs:enumeration value="am"/>
    <xs:enumeration value="ar"/>
    ...
  </xs:restriction>
</xs:simpleType>
```

OLAC Workshop, 12/02

Example: Language

```
<subject
  xsi:type="olac:language"
  code="x-sil-BAN"
/>
```

OLAC Workshop, 12/02

5. Defining a third-party extension

- OLAC records can use extensions from other namespaces
 - sub-communities develop/share extensions
 - use xsi:type to extend OLAC metadata
 - no need for them to modify OLAC schema

```
<contributor xsi:type="myolac:role" code="commentator">
  Sampson, Geoffrey
</contributor>
```

OLAC Workshop, 12/02

Schema for a 3rd-party extension

```
<xs:schema xmlns="http://www.example.org/myolac/"
  targetNamespace="http://www.example.org/myolac/">
  <xs:complexType name="role">
    <xs:complexContent mixed="true">
      <xs:extension base="dc:SimpleLiteral">
        <xs:attribute name="code" type="my-role" use="required"/>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
  <xs:simpleType name="my-role">
    <xs:restriction base="xs:string">
      <xs:enumeration value="calligrapher"/>
      <xs:enumeration value="censor"/>
      <xs:enumeration value="commentator"/>
      <xs:enumeration value="corrector"/>
    </xs:restriction>
  </xs:simpleType>
```

OLAC Workshop, 12/02

Augmenting OLAC extensions

- some third-party extensions:
 - add terms to an existing OLAC vocabulary
- two methods:
 1. 3rd-party extension includes OLAC vocabulary
 2. 3rd-party extension only has new terms
- recommend latter, for benefit of service providers & end-users

OLAC Workshop, 12/02

Harvesting third-party extensions

- OLAC service providers harvest:
 - tag name
 - element content
 - value of xsi:type
 - value of code attribute
- Third-party extensions may define other attributes
 - ignored by standard OLAC service providers
 - can be used by subcommunity service providers

OLAC Workshop, 12/02

6. Documenting an extension

- All extensions should be documented
 - in human-readable form
 - at a web-accessible location
- The XML schemas for extensions should also contain machine-readable documentation
 - name, version, description, DC element, documentation URL

OLAC Workshop, 12/02

olac-extension element

```
<olac-extension xmlns="http://www.language-archives.org/OLAC/1.0/olac-extension.xsd">
  <shortName>role</shortName>
  <longName>Code for My Specialized Roles</longName>
  <versionDate>2002-08-16</versionDate>
  <description>A hypothetical extension for an individual archive,
  defining specialized roles not available in the OLAC Role
  vocabulary.</description>
  <appliesTo>creator</appliesTo>
  <appliesTo>contributor</appliesTo>
  <extensionDoc>http://www.my.org/roles.html</extensionDoc>
</olac-extension>
```

OLAC Workshop, 12/02

Summary

- XML format follows DC recommendations
 - new DC qualifiers automatically adopted
 - other communities can use OLAC qualifiers
- Limited change from version 0.4:
 - subject.language becomes
subject xsi:type="olac:language"
- Flexible: optionality, free-text content
- Extensible: mix in third-party extensions

OLAC Workshop, 12/02

LINGUIST Codes for Ancient and Constructed Languages

Ethnologue Codes

- Consistently apply an operational definition of language so that all entities for which an identifier is assigned are of a comparable nature
- Encompass all of the languages of the world,
- Clearly document the speech variety that each identifier denotes
- Maintain and update the system on an on-going basis
- Make the system freely and readily accessible to the public over the Internet

For every language description:

- The countries the language is spoken in
- The alternate names that refer to the language
- The number of speakers of the language
- The classification of the language

Mutual Unintelligibility

- Varieties of language are only assigned a code if they are mutually unintelligible with varieties of any language to which a code has already been assigned.

Current Use

- The Ethnologue system is intended to encompass only those languages of the world in current use. Thus the Ge'ez (Ethnologue code GEE) and Sanskrit (Ethnologue code SKT) languages both appear in Ethnologue
- Most ancient languages are thus absent

Shortcomings in Ethnologue

- Every language in Ethnologue is documented to a greater or lesser degree. But we usually do not have a clear idea of the evidence upon which it was decided to assign the language a unique code. Nor does the system allow for conflicting language classifications. For example, there is disagreement amongst scholars as to the classification of Low German dialects. This is not indicated in Ethnologue.

Criteria for Ancient and Constructed languages

- Conform as closely as is reasonable to the standards set by Ethnologue

But...

- The criterion of mutual intelligibility has to be abandoned

e.g. Anglo-Norman, which was an aberrant dialect of Old French. However, it evolved independently, and has a literature distinct from that of Old French. This scholars treat separately. Thus it must be assigned a distinct code so that work on it can be discriminated from work on Old French.

Mutual intelligibility breaks down in another way

- Ancient languages often have a diachronic dimension that can usually be ignored with modern languages

e.g. Old Latin gave rise to Classical Latin, which in turn gave rise to Late Latin, which in turn gave rise to Vulgar Latin or Proto-Romance...

- It is likely that no two adjacent stages of this complex process would have been mutually incomprehensible, had there been any speakers who could speak the two versions. How many codes do we assign here on the basis of mutual intelligibility?

Undeciphered Scripts

- Ancient languages in scripts which have as yet not been deciphered, e.g. Minoan

Conclusion

- Codes should be assigned to ancient languages which are treated distinctly by the scholarly community.
- The standard of mutual intelligibility should apply as far as possible. All apparently mutually intelligible ancient languages spoken at approximately the same period should be assigned one code, unless this conflicts with scholarly usage.
- In cases where the level of mutual intelligibility cannot be clearly ascertained, separate codes should be assigned.
- Codes should be assigned to undeciphered scripts, and to uninterpretable ancient languages in known scripts.

Conclusion (cont)...

- The system should be as complete as possible. Ancient languages should not be excluded simply because they are obscure.
- All alternate names of ancient should be listed, even those which are deprecated by scholars
- To integrate with Ethnologue codes, the primary geographic categorization of ancient languages should be by the modern countries in which they once existed
- All codes should have provenance information
- Committees of specialists will provide the provenance information

Constructed Languages

- Constructed languages cannot be treated by the criterion of mutual intelligibility, since they are almost never actually spoken, and are as much cultural objects as linguistic. In some cases there exist variants of originally identical constructed languages which have begun evolving independently. Esperanto (Ethnologue code ESP) and Ido (LINGUIST code CIDO) are instances of this phenomenon. These should be assigned distinct codes.
- No attempt should be made to assign constructed languages to geographical regions, since they do not exist in the real world.

The Canary Agreement

- All languages which require codes and which became extinct before 1950 should become the responsibility of LINGUIST. All languages after 1950 will be in the purview of Ethnologue.
- The two code sets will be unified into one three-letter code-set.

Linguistic Data Types



& Discourse Types & Linguistic Fields

Helen Aristar-Dry & Gayathri Sriram
LINGUIST List / Eastern Michigan U.
OLAC Workshop, Dec 10-12, 2002

Outline



- Motivate the creation of 3 different vocabularies-
-review Metadata List discussion
- For each vocabulary (linguistic data type, discourse type, linguistic field):
 - Explain codes (vocabulary items)
 - Review results of “translation experiment” mapping the codes to existing resource descriptions
 - Suggest possible vocabulary revisions for discussion

OLAC Workshop, Dec 10-12, 2002

2

“Translation” experiment



- Mapped controlled vocabulary items (plus synonyms used in the document descriptions and examples) to the existing resource descriptions.
- Fields searched:
 - Type
 - Type.linguistic
 - Description(The only fields containing the search terms.)

OLAC Workshop, Dec 10-12, 2002

3

“Translation” experiment



- Intended to find out:
 - Are there other data types, discourse types, and linguistic fields that need to be included?
 - Do the terms used in the definitions and examples reflect common usage?
 - Ex: we use Corpus to exemplify Dataset. Is it being used by archives to describe datasets or single texts?
- Results:
<http://linguistlist.org/olac-translation.html>

OLAC Workshop, Dec 10-12, 2002

4

“Translation” experiment



Possible practical application:

We wanted to assess the degree of automation possible, based on string search for related terms:

- for service providers: to use the new codes for searching, and “translate” existing descriptions into new codes behind the scenes.
 - See: <http://linguistlist.org/olac/search-demo.html>
- for archives: to “translate” existing resource descriptions into new terminology.

OLAC Workshop, Dec 10-12, 2002

5

Linguistic Data Types



- Describe the resource as representing a recognized structural type of linguistic information
- Types:
 - Lexicon
 - Dataset
 - Primary text
 - Description

OLAC Workshop, Dec 10-12, 2002

6

Previous Draft

- 6 data types: transcription, annotation, lexicon, dataset, description, text
- 64 subtypes
- Problems:
 - transcription & annotation not “data types”
 - subtypes repeated linguistic fields
 - subtypes inconsistent in classifying principle: “apples & oranges”

OLAC Workshop, Dec 10-12, 2002 7

Repeat of Linguistic Field

dataset	description
dataset/phonetic	description/phonetic
dataset/phonological	description/phonological
dataset/prosodic	description/prosodic
dataset/orthographic	description/orthographic
dataset/gestural	description/gestural
dataset/kinesic	description/kinesic
dataset/morphological	description/morphological
dataset/part-of-speech	description/part-of-speech
dataset/syntactic	description/syntactic
dataset/semantic	description/semantic
dataset/discourse	description/discourse
dataset/musical	description/pedagogical
	description/comparative

OLAC Workshop, Dec 10-12, 2002 8

Inconsistent Classification

lexicon	text
lexicon/dictionary	text/narrative
lexicon/wordlist	text/oratory
lexicon/wordnet	text/dialogue
lexicon/thesaurus	text/singing
lexicon/terminology	text/drama
lexicon/proper-names	text/formulaic
lexicon/frequency	text/procedural
lexicon/bilingual	text/report
lexicon/etymological	text/ludic
lexicon/phonetic	text/unintelligible speech
lexicon/analytical	

OLAC Workshop, Dec 10-12, 2002 9

Current Revision:

3 Different Vocabularies

- Linguistic Data Types: dataset, lexicon, description, primary text
- Discourse Types: narrative, oratory, dialogue, report, procedural, etc.
- Linguistic Fields: phonetics, syntax, phonology, morphology, etc.

OLAC Workshop, Dec 10-12, 2002 10

Sample Descriptions

- A Kuna narrative text:
 - Linguistic Type: primary text
 - Discourse Type: narrative
 - Subject Language: Kuna
- A Quechua phoneme chart:
 - Linguistic Type: dataset
 - Linguistic Field: phonology
 - Subject Language: Quechua

OLAC Workshop, Dec 10-12, 2002 11

Sample Descriptions

- A videotape of an interview
 - Linguistic Type: primary text
 - Discourse Type: dialogue
 - Format: videotape
- A dictionary of French medical terms
 - Linguistic Type: lexicon
 - Subject: medical terminology
 - Subject Language: French

OLAC Workshop, Dec 10-12, 2002 12

“Translation” experiment

- Searched Type, Type.linguistic, and Description for linguistic data types + related terms taken from the document descriptions and examples
 - Primary text: text, translation, song, transcription, story, narrative
 - Lexicon: dictionary, vocabulary, terms, word list, word, lexicon, terminology
 - Dataset: graphs, set, data, chart, file card, slip, corpus
 - Description: grammar, note(s), paper, manuscript, thesis, chapter, description

OLAC Workshop, Dec 10-12, 2002 13

What they put in Type.Linguistic

1. index to tapes
2. catalog of JPH materials
3. Focal person ranking
4. roots/affixes, grammatical phenomena
5. -a-: plural theme
6. hache, ?freeze, frozen' etc.: notes, use, examples
7. plants with ethnomedicinal uses
8. two note cards, attached
9. Grammar: 2 ring binders (1-2 of 4) of notes on misc. topics for dissertation
10. Misc. notes
11. Notes on numerals?
12. A Chimariko song
13. texts; notebook 24
14. Dialogue, texts (transcribed from reel tape 9:2, part b)
15. rehearing of early Esselen and Rumsen vocabularies; ?Medicine practices of Mrs Ascencion Solorsano'
16. unknown

OLAC Workshop, Dec 10-12, 2002 14

What they put in Type

1. Annotation Tools , Development Tools , Corpus Analysis , Lexicon Management , Part-of-Speech Tagging , Partial Parsing , Shallow Parsing , Terminology Extraction
2. Morphological Analysis , Part-of-Speech Tagging
3. Speech Synthesis , Spoken Dialog Systems , Spoken Language Generation , Text-to-Speech Synthesis
4. Electronic text
5. corpus [for an electronic text, Orosius]
6. TERMINOLOGY
7. lexicon
8. dataset
9. poetry
10. SPEECH: TELEPHONE
11. WRITTEN: MONOLEX
12. CHAT
13. recordings
14. two note cards, attached

OLAC Workshop, Dec 10-12, 2002 15

What they put in Description

- a. (found in survey office desk drawer, 2000)
- b. (relocated)
- c. 1 of 18 notebooks
- d. Also Miami
- e. condition: Fair. Written on yellow paper? Many smudges and smears. Edges are yellowing and becoming frayed. Dark pencil is still very legible, though
- f. incomplete
- g. labeled 'Reel 1'
- h. No spool; BAE 647
- i. original folder labeled 'N Afx'
- j. published?
- k. some material probably from much earlier
- l. spool missing

OLAC Workshop, Dec 10-12, 2002 16

Search of field: type

Records with values for type	2007
Classified as Primary Text	1340
Classified as Lexicon	162
Classified as Dataset	212
Classified as Description	12
Other	411

OLAC Workshop, Dec 10-12, 2002 17

Search of field: type.linguistic

Records with values for type.linguistic	8202
Classified as Primary Text	5811
Classified as Lexicon	1868
Classified as Dataset	80
Classified as Description	443
Other	299

OLAC Workshop, Dec 10-12, 2002 18

Search of field: Description

Classified as Primary Text	2179
Classified as Lexicon	2844
Classified as Dataset	3960
Classified as Description	1505
Other	18307

OLAC Workshop, Dec 10-12, 2002 19

Results: Linguistic Data Types

- <http://linguistlist.org/olac-translation.html>
- Found 2 linguistic data types unaccounted for:
 - Index (Dataset? Lexicon?)
 - Paradigm (Dataset)
- “Corpus” used for Primary Text, not Dataset
- Discovered problem with Tools
 - Not listed as “Software” in Type
 - So misclassified in our mapping

OLAC Workshop, Dec 10-12, 2002 20

Results: Linguistic Type

- Want to reserve “Description” for description of some aspect of a language. Do not want analytical papers & books classified as “Description.”
- Want to be able to identify “Tools” and “Advice” related to each of the data types, e.g., software for building a lexicon should be related to “Lexicon.”

OLAC Workshop, Dec 10-12, 2002 21

Tools & Advice

Solution 1:

- a. Call the extension “OLAC Types” rather than “Linguistic Data Types”
- b. Add “Analysis,” “Tools,” and “Advice”
- c. Objections:
 - a. “Apples and oranges”: datasets, lexicons, primary texts, description, tools, advice
 - b. Still doesn’t tell us that the software tool is a lexicon tool.

OLAC Workshop, Dec 10-12, 2002 22

Tools & Advice

Solution 2:

- a. Revise Linguistic Data Type definition to say “represents or is relevant to” a data type
- b. Classify “Tools” and “Advice” according to the type of data they relate to:

Ex: software for building lexicons would be classified as:

Linguistic Type: Lexicon
Type = Software
- c. Objection: Some tools aren’t software but services

OLAC Workshop, Dec 10-12, 2002 23

Discourse Type

- Describes the content of the resource as representing a particular kind of discourse
- Types:

Dialogue	Narrative
Drama	Procedural
Formulaic	Report
Ludic	Singing
Oratory	Unintelligible Speech

OLAC Workshop, Dec 10-12, 2002 24

Mapping: Discourse Types

- Searched Type, Type.linguistic, and Description for discourse type & related terms taken from the document descriptions and examples

Dialogue	Conversation, Interview, Correspondence, Consultation, Greeting, Leave-taking, Dialogue
Drama	Play, Skit, Scene, Drama
Formulaic	Prayer, Curse, Blessing, Charm, Curing ritual, Marriage vow, Oath
Ludic	Play language, Joke, Secret language, Humor, Speech disguise, Game
Oratory	Sermon, Lecture, Political speech, Invocation, Oratory, Oration

OLAC Workshop, Dec 10-12, 2002 25

Mapping: Discourse Types

Vocabulary items & synonyms:

Narrative	Narrative, Myth, Folktale, Fable, Story, Stories
Procedural	Recipe, Instruction, Plan, Procedure
Report	News report, Essay, Commentaries, Report
Singing	Chant, Song, Chorus, Singing
Unintelligible Speech	Sacred language, Speaking in tongues, Singing syllable, Unintelligible

OLAC Workshop, Dec 10-12, 2002 26

Search of field: type.linguistic

Records with values for type.linguistic	8202
Classified as Narrative	18
Classified as Dialogue	29
Classified as Procedural	6
Classified as Formulaic	2
Classified as Singing	7
Classified as Report	4
Classified as Oratory	3
Other	8199

OLAC Workshop, Dec 10-12, 2002 27

Search of field: Type

Records with values for Type	2008
Classified as Narrative, Dialogue, Ludic, Procedural, Report, Singing, etc.	0
Other	2008

OLAC Workshop, Dec 10-12, 2002 28

Search of field: Description

Classified as Narrative	134
Classified as Drama	371
Classified as Dialogue	627
Classified as Procedural	62
Classified as Ludic	23
Classified as Singing	19
Classified as Report	9
Classified as Oratory	3
Other	8585

OLAC Workshop, Dec 10-12, 2002 29

Results: Discourse Type

- Add "Poetry"
- Add "relevant to" discourse type (for resource about DT)
- "Dialogue" suggests 2 speakers.
 - Change to "Conversation"?
 - To "Interactive Discourse"?
- "Formulaic," "Ludic," "Procedural" = adjs.
 - Change to "Formula," "Language Play," "Procedural Discourse"?

OLAC Workshop, Dec 10-12, 2002 30

Linguistic Field

- Describes the resource as relevant to a particular subfield of linguistic science
- Fields:
 - anthropological linguistics
 - applied linguistics
 - cognitive science
 - computational linguistics
 - discourse analysis
 - general linguistics
 - historical linguistics
 - history of linguistics

OLAC Workshop, Dec 10-12, 2002 31

Linguistic Field

- Fields (cont):
 - Language Description
 - Lexicography
 - Linguistics and literature
 - Linguistic theories
 - Morphology
 - Neurolinguistics
 - Philosophy of science
 - Phonetics
 - Phonology
 - Pragmatics

OLAC Workshop, Dec 10-12, 2002 32

Linguistic Field

- Fields (cont):
 - Psycholinguistics
 - Semantics
 - Sociolinguistics
 - Syntax
 - Text and corpus linguistics
 - Translation
 - Typology
 - Writing systems

OLAC Workshop, Dec 10-12, 2002 33

Results: The the The if the Linguistic Field

- Add “Language Acquisition”?
 - Definition: The study of the process of acquiring human language.
 - Comment: Language Acquisition may be used to describe materials relating to either adult or child language acquisition, and to either first or later language acquisition. However, if the materials deal specifically with language teaching, or with the process of language learning from a pedagogical point of view, they may be best classified as Applied Linguistics.
 - Examples: Studies of first language acquisition, audio or video tapes of language acquisition experiments, and guides to experimental techniques in eliciting acquisition data.

OLAC Workshop, Dec 10-12, 2002 34

Problems w/ Linguistic Field

- Add “Forensic Linguistics”?
 - Definition: Applications of linguistic science to the domain of law
 - Comment: Forensic linguistics refers to the use of linguistic methodology to make legal determinations. Analyses of courtroom language are best classified as Discourse Analysis.
 - Examples: Papers on issues in dispute in court cases, e.g., authorship identification, assessment of ambiguity in texts, voice attribution.

OLAC Workshop, Dec 10-12, 2002 35

Search for Linguistic Fields

Demo page:
<http://linguistlist.org/olac/search-demo.html>

OLAC Workshop, Dec 10-12, 2002 36

OLAC Vocabularies and Schemas for Language Technology Fields

Baden Hughes
baden@compulinq.net
OLAC '02 Philadelphia

Language Technology (LT) Fields Needs Analysis

- Needs analysis based on ordinary end user interaction requirements
- *Possibility*: Can I use this software ?
- *Probability*: How much effort will it take for me to be able to use this software ?
- *Functionality*: Does this software do what I want ?

Language Technology Vocabulary / Schema Implications

- LT archives are often very active software resource sites (esp. open source)
- Classification and description of software has practical implications for the end user
- LT has particular technical requirements for classification and description of software resources
- LT classification and descriptions can draw on wider IT vocabularies

Draft OLAC Vocabularies and Schemas ...

- OLAC-Functionality
- OLAC-OS
- OLAC-CPU
- OLAC-Sourcecode

OLAC-Functionality ...

- status: unreviewed draft
- "Controlled Vocabulary for Functional Classification"
- currently lists 17 core categories and 98 extended functional categories for LT
- based on HLT survey version 2 (from LT-World / DFKI)

OLAC-Functionality ... cont ...

- **Functionality Divisions:**
 - Information Extraction
 - Information Retrieval
 - Authoring Tools
 - Language Analysis
 - Language Understanding
 - Knowledge Representation and Discovery
 - Spoken Language Input
 - Written Language Input
 - Natural Language Generation
 - Spoken Output
 - Multilinguality
 - Multimodality
 - Coding and Compression
 - Mathematical Methods
 - Discourse and Dialogue
 - Language Resources
 - Evaluation

OLAC-OS ...

- Status: unreviewed draft
- "Controlled Vocabulary for Operating Systems"
- currently lists 41 operating systems
- based on industry standard IT classifications
- example

OLAC-CPU ...

- status: unreviewed draft
- "Controlled Vocabulary for CPU"
- currently lists 37 CPU types
- based on industry standard IT classifications
- example

OLAC-Sourcecode ...

- status: unreviewed draft
- "Controlled Vocabulary for Programming Languages"
- currently lists 286 programming languages
- based on industry standard IT classifications
- example

Issues ...

- Community review of drafts ?
- WG for Language Technology Fields ?
- Are OLAC-Functionality descriptions applicable to more resources than just language technology ?
- Should type be revised in OLAC Metadata document ?
- Proposal for OLAC-Sourcestatus ?

Issues ... cont

- interaction of these metadata elements with other related fields eg type ?
- service provider implementations for language technology resources ?

OLAC Role Vocabulary

Heidi Johnson / AILLA

IRCS Workshop on Open Language
Archives

1

Overview

- **Role is an attribute of both the Creator and Contributor elements**
- **Functional roles of people who contribute to the creation of an archive resource**
- **Name of role-bearer should appear in the element content**

IRCS Workshop on Open Language
Archives

2

Examples:

- `<creator xsi:type="olac:role" code="speaker">Olawituppini </creator>`
- `<creator xsi:type="olac:role" code="researcher">Joel Sherzer </creator>`
- `<contributor xsi:type="olac:role" code="sponsor">National Science Foundation</contributor>`

IRCS Workshop on Open Language
Archives

3

Vocabulary (sorted thematically)

- **Producers of written resources:**
- **Author**
- **Annotator**
- **Transcriber**
- **Translator**
- **Illustrator**

IRCS Workshop on Open Language
Archives

4

Vocabulary, Cont.

- **Producers of spoken resources (recorded, filmed):**
- **Speaker/signer**
- **Performer**
- **Interviewer**
- **"Participant": audience, interlocutor, observer, bystander...**
- **Recorder**
- **Photographer (filmer)**
- **(Transcriber)**

IRCS Workshop on Open Language
Archives

5

Vocabulary, Cont.

- **Producers of visual resources:**
- **Artist**
- **Photographer**

IRCS Workshop on Open Language
Archives

6

Vocabulary, Cont.

- **Producers of research & tools:**
- **Researcher**
- **Developer**
- **Respondent**

Vocabulary, Cont.

- **"Super" producers:**
- **Sponsor**
- **Supervisor (new)**
- **Compiler**
- **Editor**
- **Depositor**

Problematical terms: "Participant"

- **Conflicts with another document**
- **Intended for a passive function - "uh-huh" sayers, audience members**
- **Solutions:**
- **1. Leave it out - either they're Speakers or not**
- **2. Alternate term: observer, audience, bystander**

Problematical terms: "Compiler"

- **Redundant with Editor?**
- **Should it mean someone who produces a corpus, and/or someone who produces a single multi-part work?**
- **Examples of compiled resources:**
- **Book of articles or stories;**
- **CD w/many songs;**
- **Suite of tools**

New term: Supervisor

- **Definition: The participant supervised the creation of the resource.**
- **Examples: A thesis advisor; a project director; a program coordinator at a school.**

Questions

- **Do we have all the terms we need?**
- **What are the roles in other subfields of linguistics, esp. psycholinguistics and language acquisition?**
- **What supporting documents should we provide?**

More questions:

- **Do we need to define any other vocabularies for the Creator/Contributor elements?**
- **How can we advise people about using these elements in metadata definitions?**

Related Documents

- **Linguistic data type: what it is helps determine who to acknowledge**
- **Linguistic field: subfields have different conventional terms for roles**

OLAC Access Vocabulary

Heidi Johnson / AILLA

IRCS Workshop on Open Language
Archives

1

Overview

- **Attribute of the Rights element.**
- **Broadly classifies the way a resource may be used.**
- **Details, or reference to a document defining the details, should be given in the element content.**

IRCS Workshop on Open Language
Archives

2

Examples:

- `<rights xsi:type="olac:access" code="standard">OLAC Standard Use Guidelines</rights>`
- `<rights xsi:type="olac:access" code="restricted">Permission from the depositor is required. Log in to http://www.ailla.org and follow instructions for this resource.</rights>`

IRCS Workshop on Open Language
Archives

3

Vocabulary

- **4 levels of distinction:**
 1. Restricted
 2. Standard
 3. Non-profit
 4. Commercial

IRCS Workshop on Open Language
Archives

4

Vocabulary: Restricted

- **Access to the resource is restricted.**
- **This includes any kind of restriction beyond the standard "fair use" restrictions that apply to published materials in general.**
- **The vocabulary does NOT define the specific nature of the restriction.**

IRCS Workshop on Open Language
Archives

5

Vocabulary: Restricted, Cont.

- **Examples of restrictions:**
- **Permission required from someone just to access the resource.**
- **Time limits: the resource will be publically available on date X.**
- **Special conditions must be agreed to, e.g. keep speakers names anonymous.**

IRCS Workshop on Open Language
Archives

6

Vocabulary: Standard

- **Access to the resource is standard; that is, it can be used like any published work.**
- **Note: we need a document defining what this means for our users.**

Vocabulary: Standard, Cont.

- **Generally, standard use allows:**
 - Quotation of small portions;
 - Summaries, critiques, analyses
- **Standard use prohibits:**
 - Copying & redistribution;
 - Wholesale incorporation of the work;
 - Use without proper citation.

Vocabulary: Non-profit

- **The resource can be used for any non-profit purpose.**
- **Includes permission to**
 - Copy & distribute (free of charge);
 - Incorporate wholly into academic materials;
 - Create derivative works, such as translations.

Vocabulary: Commercial

- **The resource can be used for commercial purposes.**
- **Includes**
 - Copying & distributing for profit;
 - Creating and selling derivative works;
 - Incorporation into commercial products.

First question:

- **Is this enough? Are there other attributes we should develop for the Rights element?**

The short answer

- **Intellectual property rights are complex & solutions are highly varied & situation-dependent, so it is probably not possible for us to define controlled vocabularies that cover the subject.**

Remaining tasks

- **Make a list of the documents that we should provide to OLAC users concerning Rights & Access, and get people to volunteer to write them.**
- **First pass:**
 - **OLAC Guide to Standard Use**
 - **OLAC Citation Guidelines**
 - **Standard resource license agreement?**

Diane's Principles

- Differentiate “is-ness” and “about-ness”
 - Aboutness = subject property
- Don't duplicate terms already in the DC pantheon
- Elements can be omitted. Not all resources have to be described by a set of codes.
- Don't do any more work than you have to—e.g., why invent the wheel when you can steal one?

OLAC Workshop, Dec 10-12, 2002 1

Application

- Differentiate “is-ness” and “about-ness”
 - Reserve Linguistic Types for “is-ness”
 - Tools and Advice = “about-ness”
 - A lexicon tool should be classified as
 - Type: software
 - Linguistic Field (a subject extension): Lexicography

OLAC Workshop, Dec 10-12, 2002 2

Application

- Don't duplicate terms already in the DC pantheon
 - Don't call an attribute the same thing as an existing element.
 - Can't use Description as a code in Linguistic Type (it's already a DC element)
 - Changed to Language Description
 - Don't name a new attribute the same thing as an existing one.
 - Can't use Dataset as a code in Linguistic Type (it's already in DC Type)
 - Interface can still show Dataset, just have it write to Type behind the scenes

OLAC Workshop, Dec 10-12, 2002 3

Application

- Elements can be omitted. Not all resources have to be described by a given set of codes.
 - In Linguistic Types, don't need both Description and Analysis.
 - Can put in Language Description, define it clearly, and say that analytical papers and books aren't described by this element

OLAC Workshop, Dec 10-12, 2002 4

Application

- Don't do any more work than you have to
 - See if we can use Library of Congress terms for Linguistic Fields
 - Joan will bring them in

OLAC Workshop, Dec 10-12, 2002 5

Summary: Linguistic Types

- Lexicon
- Primary Text
- Language Description

OLAC Workshop, Dec 10-12, 2002 6

Summary: Linguistic Fields

- Added
 - Language Acquisition
 - Mathematical Linguistics
- Redefined
 - Cognitive Linguistics
- Checking examples to insure they reflect “about-ness”

OLAC Workshop, Dec 10-12, 2002

7

Revised OLAC Vocabulary for Language Technology

Basic Functional Classification

- Functionality Family
 - Data Collection
 - Data Management
 - Data Manipulation
 - Data Output

Data Collection

- Definition: Language technology resources which enable language resource creation.
- Based on Linguistic-Data-Type, DC-Type and OLAC-Format

Data Management

- Definition: Language technology resources which enable the management of language resources.
- Based on new Controlled Vocabulary

Data Manipulation

- Definition: Language technology resources which enables the manipulation of language resources.
- Based on new Controlled Vocabularies – Core and Extended

Data Output

- Definition: Language resources which result from manipulation of language resources.
- Based on Linguistic-Data-Type, DC-Type and OLAC-Format


Related Vocabularies

- *Data Management CV
- *Data Manipulation CV – Core
- *Data Manipulation CV - Extended
- OLAC-OS CV
- OLAC-CPU CV
- OLAC-Source-Code CV

* = new


Issues and Work Items

- resolved to form a working group
- interaction with Linguistic-Data-Type
- interaction with DC-Type
- interaction with orphan OLAC-Format
- unresolved OLAC-Source-Status




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 (jazbybenk@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

Outreach

Jeff Good
UC Berkeley

OLAC's Needs

- Maximal involvement from the whole community
 - The more data providers involved the more useful the services become
 - More data providers mean more input on how to improve standards and services

The average user's needs

- Documents making minimal use of technical vocabulary
- Services making OLAC participation straightforward
- A central location to access all OLAC services

Impediments from OLAC's perspective

- Old habits
- Lack of awareness of the basic issues
- Idea that it is "someone else's problem"

Impediments from user's perspective

- Technical issues not directly related to linguistics (or other fields of study)
- Time and work involved

The current state of outreach

OLAC's Needs

- Maximal involvement from the whole community
- Archives have largely been self-selecting, either by direct involvement with OLAC or by already making some sort of metadata available

The average user's needs

- Documents making minimal use of technical vocabulary
- "A gentle introduction to metadata"

Aside: A gentle introduction to XML (from the TEI)

XML is an extensible markup language used for the description of marked-up electronic text. More exactly, XML is a *metalanguage*, that is, a means of formally describing a language, in this case, a *markup* language. Historically, the word *markup* has been used to describe annotation or other marks within a text intended to instruct a compositor or typist how a particular passage should be printed or laid out. Examples include wavy underlining to indicate boldface, special symbols for passages to be omitted or printed in a particular font and so forth. As the formatting and printing of texts was automated, the term was extended to cover all sorts of special codes inserted into electronic texts to govern formatting, printing, or other processing.

The average user's needs

- Services making OLAC participation straightforward
- Vida, ORE, Viser
- EMELD

The average user's needs

- A central location to access all OLAC services
- OLAC? Linguist List?

Assessment of what is needed

Assessment of what is needed

- Contacting archives
 - Individually by members of OLAC
 - More broadly via public forums (LSA, Linguist List, . . .)

Assessment of what is needed

- Documents
 - Making the production of a non-technical documents part of the OLAC process
 - Published documents in journals giving an overview of OLAC for different communities (Language, IJAL...)
 - A FAQ on data archiving, annotation, and access in the digital age (with lots of examples)

Assessment of what is needed

- Services
 - Input from as wide a range of potential users as possible
 - Services not made public until "useful"
 - A group of official consultants from important linguistics subcommunities, preferably people not already closely involved with OLAC

Assessment of what is needed

- Central location
 - OLAC site for technical reference?
 - Linguist List for general community reference?

Overcoming the impediments

Impediments: OLAC's

- Old habits
- ??
- Lack of awareness of the basic issues
- Documents (online and in print)
- Idea that it is "someone else's problem"
- Endorsement (LSA and other high-profile organizations) (cf. 1992 *Language* article on endangered languages)

Impediments: User's

- Technical issues not directly related to linguistics (or other fields of study)
- Clear separation of general audience documents from technical documents
- Time and work involved
- Better tools, services
- Change of culture

Proposal: Outreach Working Group

Outreach working group

- Contact archives
- Oversee creation of non-technical documents
- Work towards gaining endorsements

Integrating ELRA/LDC Metadata into OLAC Repository

Andrew W. Cole **Khalid Choukri**

andrew.cole@ldc.upenn.edu choukri@elda.fr

Linguistic Data Consortium
University of Pennsylvania
www.ldc.upenn.edu

ELRA/ELDA
55 Rue Brillat-Savarin
F-75013 Paris, France
http://www.elda.fr

■ OLAC 2002 IRCS UPenn 1


Net-DC

- Net-DC : Networking Data Centers, an initiative funded by NSF and EC to coordinate activities of data providers – specifically LDC and ELRA but in ways that should encourage other centers to join
- Included a task for joint LDC/ELRA dissemination of information on resources being distributed
- LDC/ELRA concluded having NetDC fund the integration of their catalog into OLAC was the best solution.
- In the division of tasking, LDC agreed to write the converters for both LDC and ELRA/ELDA catalogs.

■ OLAC 2002 IRCS UPenn 2

ELRA Architecture

- System Overview



```

graph LR
    A[ELRA Catalog MS/Access Table] --> B[Visual Basic Conversion Module in MS/Access]
    B --> C[ELRA Catalog XML file]
  
```

- Access Table Match

<u>OLAC</u>	<u>OLAC/ELDA</u>
Subject.language	Subject.language
Type	Type
Type.linguistic	Not in ELDA Catalog
Coverage	Not in ELDA Catalog
Date	Not Applicable

■ OLAC 2002 IRCS UPenn 3

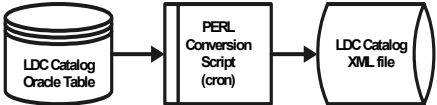
LDC/ELRA Code

- Coding/Knowledge Problems
 - Error in ELDA OLAC program, link to online description is between an <identifier> tag, <description> would be better.
- Nonetheless Access System is Simple and Robust
 - MS/Access Visual Basic Module of 280 lines.
 - Single MS/Access Table Converted to Single XML file.

■ OLAC 2002 IRCS UPenn 4

LDC Architecture

- System Overview



```

graph LR
    A[LDC Catalog Oracle Table] --> B[PERL Conversion Script (cron)]
    B --> C[LDC Catalog XML file]
  
```

- Oracle Table Problems

ldc_catalog_id:	LDC94817
name:	OGI Multilanguage Corpus
language:	English, Farsi, French, German, Hindi, Japanese, Korean, Chinese, Spanish, Tamil, Vietnamese,

■ OLAC 2002 IRCS UPenn 5

LDC PERL Coding

- Coding/Knowledge Problems

```

<record spec="lexicon">
<header>
  <recordId>olac:ldc:LDC94L2</recordId><datestamp>2002-10-16</datestamp>
</header>
<metadata>
<olac>
<identifier>LDC94L2</identifier>
<title>*COMLEX English Syntax Lexicon</title>
<type>lexicon</type>
...
  
```

- Nonetheless System is Simple and Robust
 - PERL Script of 150 lines (lots of comments).
 - Single Oracle Table converted to Single XML file.
 - Repairs taking less than a day, done by non-experts.

■ OLAC 2002 IRCS UPenn 6

• **OLAC Issues.**

- Existing OLAC vocabulary assumes that linguistic data is for traditional linguistic research (ie. linguistic field) and that language technology developers are only interested in software not data.
- Difficult to determine/find the correct or applicable type and vocabulary from OLAC web site with unknowledgeable staff (eg., me, Andy).
- Need OLAC vocabularies encode information about pricing.
- Providers ramp-up to full meta-data compliance.

• **Web Links**

- **ELDA ECI** <http://www.elda.fr/cata/text/W0004.html>
- **LDC ECI** <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94T5>
- **OLAC LDC/ECI** <http://saussure.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search3.cfm?id=112715>
» (Bulgarian)
- **OLAC ELDA/ECI** <http://saussure.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search3.cfm?id=58000>
» (Turkish)

IMDI & Endangered Languages Archives

Heidi Johnson / AILLA

IRCS Workshop on Open
Language Archives

Acronyms & URLs

- IMDI = International Standards for Language Engineering MetaData Initiative: <http://www.mpi.nl/ISLE>
- MPI = Max-Planck Institute for Psycholinguistics: <http://www.mpi.nl/>
- DOBES = Documentation of Endangered Languages: <http://www.mpi.nl/dobes/>

IRCS Workshop on Open
Language Archives

Overview

- Goal: bottom-up design of a metadata schema for resources archived for DOBES.
- Considerations:
 - DC elements too shallow & fragmented.
 - Want to be able to "bundle" resources together.
 - Want to include all the information concerning a resource in its metadata schema.

IRCS Workshop on Open
Language Archives

Bundles of materials

- Multi-part resources:
 - Audio/video recording of a speech event; e.g. narration of a traditional myth;
 - Transcriptions, translations, & annotations;
 - Photographs, additional tracks, etc.;
 - Multiple formats are archived: .wav & .mp3; pdf & txt...

IRCS Workshop on Open
Language Archives

A problem for the DC/OLAC model:

How can we keep related resources together & make sure users get all the parts they need?

IRCS Workshop on Open
Language Archives

The IMDI Session Schema

- Describe a single time-bounded recording, plus derivatives (e.g. transcriptions).
- The schema is large & highly structured.
- Sub-schemas are "shareable" with other schemas, like the Written Resources Schema.
- Every sub-schema has a Description field
- Every sub-schema has customizable Key/Value pairs.

IRCS Workshop on Open
Language Archives

Session Schema: the big pieces

- Session info: Title, Abbr title, Date & Place.
- Project info: Title, Contact info.
- Depositor (Collector): Name & Contact info.
- Participants sub-schema
- Content sub-schema
- Resources sub-schema
- References

IRCS Workshop on Open
Language Archives

Participants sub-schema

- Name, nickname
- Role: = OLAC Role attribute
- Social/family role: parent, shaman...
- Age, sex, ethnicity, education level
- Place of origin
- Language(s): first given is native language.

IRCS Workshop on Open
Language Archives

Content sub-schema

- Modality: speech, writing, gesture
- Language(s) = Subject.language
- **Genre:** conversation, verbal contest, interview, meeting/gathering, riddling, consultation, greeting/leave-taking, humor, insult/praise, letter; procedure, recipe, description, instruction, commentary, essay, report/news; narrative, oratory, ceremony, poetry, song, drama, prayer, lament, joke; textbook, primer, workbook, reader, exam, guide, problem set; dictionary, word-list, grammar, sketch, field notes
- **Communication context:** elicited/non, planned/unplanned, etc.

IRCS Workshop on Open
Language Archives

Resources sub-schema

- Separate sub-schemas for different media. (AILLA conflates these.)
- All files:URL, size in bytes, format, access rules.
- Audio/video: quality, recording condition
- Text:
 - Character encoding, content encoding
 - Transcription & translation information
 - Language = DC Language.
 - Anonymous (use nicknames only)

IRCS Workshop on Open
Language Archives

MPI Implementation

- Hierarchical file system, XML files.
- Corpus Browser & Metadata Editor (PC)
- Elan: time-aligning annotation tool.
- Allows the researcher to create & manage a corpus in the field, & come home with ready-to-archive data.

IRCS Workshop on Open
Language Archives

AILLA Implementation

- Relational database.
- PHP Internet interface: metadata editor, search, display/download resources.
- Graded access system & user registration to protect resources.

IRCS Workshop on Open
Language Archives

IMDI - OLAC mapping

- OLAC terms are a subset: not everything has to be mapped
- Tricky part will be Genre: IMDI Genre conflates OLAC Linguistic data type & Linguistic discourse type
- Missing from IMDI: dataset, Linguistic field
- Missing from OLAC: teaching materials, literature (not strictly linguistic Types)

IRCS Workshop on Open Language Archives

Summary

- IMDI schema includes all the info that documentary linguists want.
- It doesn't need to cover other subfields, e.g. speech recognition.
- IMDI protocols support bundling, a key consideration for AILLA.

IRCS Workshop on Open Language Archives

Levels of description 1

OLAC			
Endangered language archives	Speech recognition data	Theor. papers	Language acquisition
AILLA		ROA	
DOBES		RRG	
Rausing?			

IRCS Workshop on Open Language Archives

Levels of description II

- Interoperability between AILLA ~ DOBES is desirable:
 - Common datatypes, resources
 - Overlapping pool of researchers (depositors)
- Interoperability between AILLA & every other linguistic archive on earth is unnecessary!

IRCS Workshop on Open Language Archives

The moral of the story

- Subfields can & should define metadata schemas that cover their subjects the way they want.
- Search engines should operate at different levels of compatibility:
 - coarse search across different subfields (OLAC)
 - fine search across similar archives (AILLA, DOBES)

IRCS Workshop on Open Language Archives

Language Technology WG Work In Progress Report

Agenda

- Remit
- Per CV and Schema Review
- Current CV and Schema Status
- WG Formation
- Timetable
- Issues

Remit: Vocabularies and Schemas

- Subject.Functionality
- Format
- Format.OS
- Format.CPU
- Format.Source-Code

- All are resource type independent

Subject.Functionality

- "the function of a resource" (doing-ness)
- LT domain focus
- Relocated from Type.Functionality
- CV and Schema
- 3 level CV hierarchy

Subject.Functionality Structure



Format

- "the physical or digital manifestation of a resource"
- eg. text/xml
- Domains beyond LT
- CV and Schema
- 2 level "Content Type" + "Subtype"
- Based on IETF RFC 1521/2077

Format.OS

- "the OS required to use a resource"
- eg. `unix/linux`
- LT domain focus
- CV and Schema
- re-usable by higher level ontologies

Format.CPU

- "the CPU required to utilise a resource"
- eg. `cpu/i386`
- CV and Schema
- reusable by higher level ontologies

Format.Source-Code

- "the programming language of a resource"
- eg. `sourcecode/perl`
- CV and Schema
- reusable by higher level ontologies

CV and Schema Status

CV/Schema	Status
Subject.Functionality	Editorial Process
Format	Under Development
Format.OS	Draft
Format.CPU	Draft
Format.Source-Code	Draft

WG Formation

- >3 Core Members
- OLAC WG web page
- WG document workspace
- LANGTECH Mailing List

Timetable ?

- Dec 2002: develop draft CV/Schema
- Jan 2003: review draft CV/Schema
- Feb 2003: proposals ?
- April 2003: recommendations ?

Issues

- Cross CV compatibility
 - Linguistic-Data-Type
 - Extensions for HLT Survey
- Orphan CV and Schema
 - Format.Encoding
 - Format.Markup
 - Format.Source-Status

2003 Activities

- Granularity recommendation (bundling) – Heidi, Alexis, Baden, Joan
- Outreach/User support? – Jeff, Laura, Heidi, Gary H, Joan
- DC usage recommendations – Elaine, Joan, Baden
- Digitization – Anthony, Jim, Christopher
- Text annotation – Mike, Terry, Steven, Gary S, Christopher, Laura, Alexis
- Pedagogy – Heidi, Laura, Gary H, Joan
- Language families / areal classification – Anthony, Jim, Chu-Ren (or colleague at AS)
- Best practice for character encoding – Peter Constable, Chu-Ren
- Format.markup – Baden,
- Personal repositories (identifiers, tools) – Baden
- Working group on IMDI-OLAC mapping? – Zina, Heidi, Daan?
- Registering OLAC application profile with DCMI – Elaine, Steven