OLAC, EMELD, & “Us”

Helen Dry & Anthony Aristar
LINGUIST List: http://linguistlist.org
LSA Symposium:
The Open Language Archives Community
4 January 2002

Who is “Us”?

• The community of academic linguists
  • who produce data & documentation on languages
  • who use language data & documentation in their research
• Includes most subscribers to The LINGUIST List

The LINGUIST List

• 15,200 subscribers
• 105 different countries
• 4 European mirror sites:
  Tübingen | Stockholm
  Edinburgh | Moscow
• Current project: EMELD . . .

What is E-MELD?

• “Electronic Metastructure for Endangered Languages Data”
• 5 year collaborative project, begun Sept. 2001
• Participants:
  • The LINGUIST List (Eastern Michigan University, Wayne State University, University of Arizona)
  • The Linguistic Data Consortium (University of Pennsylvania)
  • The Endangered Languages Fund (Yale University, Haskins Laboratories)
• Funded by NSF

E-MELD Objectives:

To aid in ...
• the preservation of Endangered Languages (EL) data and documentation
• the development of infrastructure for linguistic archives

The Problem with ALL archives:

• Lack of interoperability < many different procedures and data formats
• Lack of permanence < use of proprietary tools & standards
• Inadequate input from linguists into the standards-setting enterprise
Result:

Endangered Languages  
*plus*

Endangered data

---

**EMELD Components**

- Catalog of language resources on the Internet
- Promotion of community consensus about best practice in:
  - Language identification
  - Resource description
  - Markup or annotation
  - "Showroom of Best Practice"

---

**Languages**

<table>
<thead>
<tr>
<th>Language</th>
<th>Speakers</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mocovi (Guaicuruan)</td>
<td>7000 speakers</td>
<td>[EMU]</td>
</tr>
<tr>
<td>Blao Min (Mienic)</td>
<td>21,000 speakers</td>
<td>[WSU]</td>
</tr>
<tr>
<td>Ega (Kwa)</td>
<td>300 speakers</td>
<td>[LDC]</td>
</tr>
<tr>
<td>Cambap (Mambiloid)</td>
<td>30 speakers</td>
<td>[LDC]</td>
</tr>
<tr>
<td>Lakota (Macro-Siouan)</td>
<td>30 speakers</td>
<td>[LDC]</td>
</tr>
<tr>
<td>Tofa (Turkic) [ELF]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Two from: Alamblak, Dadibi, Mapos Buang, Takaulu Kalagan, Tuwali Ifugao - [SIL]
- Two from Post-Docs as yet to be determined.

---

**OLAC & EMELD:**

- Common Goals
- Needed: Collaboration!

---

**OLAC-related Components**

1. Catalog of resources → *OLAC Service Provider*
2. Promotion of community consensus about best practice in:
   1. Resource description → *OLAC metadata*
   2. Markup → propose as *OLAC best practice*
   3. Language identification → Ethnologue /LINGUIST language codes proposed as *OLAC best practice*
OLAC Launch, LSA-02

LINGUIST = Gateway to information on best practice

LDC = Repository of Standards & Software
SIL = Vocabulary Server for Languages

OLAC Launch, LSA-02

What you need to know to … Understand Metadata

- Is it really as simple as it sounds? Yes
- Is it really important? Yes
- Why??
  a) Standardization is power
     (for Computers)
  b) Standardization is hard
     (for People)

OLAC Launch, LSA-02

Metadata

- Data about data, e.g., cataloguing information
- Facilitates resource description, including summarization
- Enables search and retrieval

OLAC Launch, LSA-02

How LINGUIST will use Metadata

- Harvest metadata from OLAC archives
- Collect metadata from individual linguists
- Provide a searchable database of information (metadata) on
  - Language data & documentation
  - Software & tools
  - Standards & formats

OLAC Launch, LSA-02

An Example

<olac xmlns="http://www.language-archives.org/OLAC0.3/"
<creation>Derbyshire, Desmond C.</creation>
<date code="1986"></date>
<title>Topic continuity and OVS order in Hikuryana</title>
<relation refine="isPartOf">In Joel Sherzer and Greg Urban (eds.), Native South American discourse, 237-306, Berlin: Mouton.</relation>
</olac>
OLAC Metadata . . . built on Dublin Core set of 15 elements:

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Title
- Type

Added for Language Resources:

- **Subject.language**
  - A language the resource is about
  - E.g. A Grammar of Russian written in English has Subject.language = Russian
- **Type.linguistic**
  - The nature of the content from a linguistic point of view
  - E.g. transcription, annotation, description, lexicon

Important for LL Searching

```
<olac xlink:href="http://www.language-archives.org/OLAC/0.3"/>
<creator>DeRose@Kim_Dimson @C</creator>
<title>Topic:personality and OVS order in Hkilyana</title>
<description>See also: Native South American discourse, 237–306, Berlin: Mouton</description>
<type linguistic="description/grammatical"/>
<subject language="x-sil-HIX"/>
</olac>
```

What’s been done so far:

- OLAC harvester on the LINGUIST site:
  - prototype: http://saussure.linguistlist.org/olac/
- Language identification:
  - Code list for ancient languages, constructed languages, and language families to complement the Ethnologue code list
  - Everything on LINGUIST site (not just harvested metadata) categorized according to these codes: see Directory of Linguists

What needs to be added? . . . to LINGUIST Gateway

- Advice about software, tools, formats
- User reviews of archives, software
- Look up for
  - Controlled vocabularies
  - OLAC best practice

What needs to be done? . . . on Language Codes

- Mechanism ensuring community input into system
- Establishment of working group using OLAC process
- Promotion of code use among OLAC data providers
What needs to be done? . . .on Markup

- Finish knowledge base for markup (U. of Arizona)
- Input needed from linguists
- Sample annotation schemas
- Feedback on proposed KB content
- Contact Terry Langendoen [terry@linguistlist.org]

Outcome?

Improved

- Data Access
- Data Permanence
- Accuracy of language representation