 Metadata

Data about data, e.g., cataloguing info
Facilitates
- resource description, including summarization
- search and retrieval (by you and others)
- re-use and re-purposing of your data
Metadata is not the data itself. You can publish the metadata and still control access to your recordings and transcriptions

Examples of Metadata

- Bibliographical Entry:

- Same entry in OLAC format (XML)
  <olac xmlns="http://www.language-archives.org/OLAC/0.3/">
    <title>A New Approach to Primates</title>
    <creator>Monkeycrush, John</creator>
    <date code="2008"/>
    <type code="Text"/>
  </olac>

Examples of Metadata

<table>
<thead>
<tr>
<th>Access Database Record</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Book ID</strong></td>
</tr>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Topic ID</strong></td>
</tr>
<tr>
<td><strong>Copyright Year</strong></td>
</tr>
<tr>
<td><strong>ISBN Number</strong></td>
</tr>
<tr>
<td><strong>Publisher Name</strong></td>
</tr>
<tr>
<td><strong>Place of Publication</strong></td>
</tr>
<tr>
<td><strong>Translator</strong></td>
</tr>
<tr>
<td><strong>Purchase Price</strong></td>
</tr>
<tr>
<td><strong>Edition Number</strong></td>
</tr>
<tr>
<td><strong>Pages</strong></td>
</tr>
</tbody>
</table>
So what’s so hard about metadata?

Well…deciding what information to keep.

What metadata will you need
- For your own research purposes
- For an archive
- For a wider community
  - E.g., scholars or speakers
  - With search and retrieval based on a metadata standard

Assessing Uses
- What do you want to keep metadata on?
  - Documentation (recordings, field notes, transcripts, etc.)
  - Consultants (age, education, 1st & 2nd language, etc.)
  - Settings (place, date and time of recording, etc.)
- And what about:
  - Equipment (type of microphone, recorder, etc.)
  - Sources consulted (dictionaries of the language, encyclopedias of flora and fauna, etc.)
  - Transcribers and data entry persons (if not you)
  - Date, time, and place of transcription and data entry
  - Persons you’ve shown your work to, and their comments

Working with an archive
- When you deposit your work with an archive, the archive will maintain at least 3 types of metadata:
  - Descriptive metadata (information on title, topic, discourse type, language, consultant, setting, etc.)
  - Administrative metadata (accession information, physical location, cataloging info, access rights, etc.)
  - Technical metadata (information on original recording equipment, file size and format, data migration or transfer (date, equipment, audio engineer, etc.)
- Ask the archive for their guidelines
Download AILLA metadata forms & editors

- Depositor Packet
- OVLA Metadata Help
- AILLA Depositor Packet
- AILLA metadata templates
- AILLA Depositor Packet
- AILLA metadata templates

### Depositor Packet

- Provides instructions, licenses, agreements, and explanations of OVLA Terms and Conditions and access levels. You can access the digital version or send paper copies via the package with your materials.

- Types of data access a depositor can choose
- Explains the archive’s terms and conditions
- Explains what information to keep about contributors, data, equipment, etc., and why

### AILLA metadata templates

- In Excel
- Also available in Toolbox format

---

**Metadata**

You can keep all this information in a database on your computer while you’re working.

But you should plan to share the information:

- Publishing your metadata
  - Helps speaker communities and language teachers find information and recordings
  - Helps scholars find data and prior analyses
  - Helps funding agencies learn
    - what languages have been documented and to what extent
    - what languages are of scientific or community interest
  - Helps everyone make better decisions when allotting scarce resources for language documentation
Publishing your metadata requires some understanding of metadata standards. There are two metadata standards specific to digital language resources:

- OLAC (Open Language Archives Community)
- IMDI (ISLE Metadata Initiative)

**IMDI metadata**
- Developed by committees of linguists and language engineers involved in the ISLE (International Standards for Language Engineering) initiative
- Based on the session model; does a good job of bundling resources together (like METS)
- Session metadata defined by a 60-page document http://www.mpi.nl/IMDI/
- Cataloging standard defined by another 50-page document
- The IMDI metadata standard is extensive!

**IMDI metadata**
- Has its own free tools for metadata entry and retrieval:
  - Metadata editor
  - Controlled vocabulary editor
  - Metadata browser
  - http://www.mpi.nl/IMDI/tools/
- Sample IMDI metadata editor screen
Open Language Archives Community (OLAC): http://language-archives.org

- International partnership of institutions and individuals creating a virtual library of language resources
- Developing consensus on metadata standards for search and retrieval of language resources
- Developing a network of interoperating repositories and services for housing and accessing such resources

Participating Archives

- Aboriginal Studies Electronic Data Archive (ASEDA)
- Academia Sinica
- Alaska Native Language Center
- Archive of Indigenous Languages of Latin America (AILLA)
- ATILF Resources
- CHILDES Data Repository
- Cornell Language Acquisition Laboratory (CLAL)
- Dictionnaire Universel Boiste 1812
- Digital Archive of Research Papers in Computational Linguistics
- Ethnologue: Languages of the World
- European Language Resources Association (ELRA)
- LACITO Archive
- LDC Corpus Catalog
- LINGUIST List Language Resources
- Natural Language Software Registry
- Oxford Text Archive
- PARADISEC
- Perseus Digital Library
- Rosetta Project 1000 Languages
- SIL Language & Culture Archives
- Surrey Morphology Group Databases
- Survey for California and Other Indian Languages
- TalkBank
- Tibetan and Himalayan Digital Library
- TRACTOR
- Typological Database Project
- Univ. of Bielefeld Language Archive
- Univ. of Queensland Flint Archive

How OLAC works

- Based on a Digital Library Federation standard
  - Open Archives Initiative Protocol for Metadata Harvesting
  - Service providers use the protocol to harvest metadata from data providers
  - Data providers include archives and individuals.
  - To serve metadata, archives can:
    - Implement a dynamic interface to existing database
    - Map database to a static XML document
  - Individuals can use web form of OLAC Repository Editor: http://linguistlist.org/olac/ore/
Simple, by comparison to IMDI
Based on Dublin Core, a library metadata standard (inaugurated in Dublin, Ohio)
Contains 15 elements, with some extensions intended to customize the elements for description of language resources
Optimized for resource search and retrieval
Written in XML (eXtensible Markup Language)
OLAC Metadata standard

- OLAC adds extensions (with controlled vocabularies) to include information specific to linguistic resources:
  - **Subject** extensions specify
    - The language the resource is about
    - The linguistic subfield the resource treats
  - **Type** extensions specify
    - The linguistic type (lexicon, dataset, etc)
    - The discourse type (narrative, conversation, procedural discourse, etc.)

OLAC extensions:

- **Contributor** extensions specify the roles that a contributor may take, e.g., annotator, author, compiler, consultant, data_inputter, depositor, developer, editor, illustrator, interpreter, interviewer, participant, performer, photographer, recorder, researcher, research_participant, responder, signer, speaker, sponsor, transcription, translator

- **Language** extensions specify that languages are to be identified using the ISO 639-3 language codes, which are an international standard.

An OLAC Metadata Record (simplified)

```xml
<olac>
  <dc:creator>Bloomfield, L.</dc:creator>
  <dc:date>1933</dc:date>
  <dc:title>Language</dc:title>
  <dc:publisher>New York: Holt</dc:publisher>
</olac>
```

Note: dc: stands for Dublin Core. It tells us that the definition of the element, e.g., title, is part of the Dublin Core standard.

The record is in XML, which is:

- a markup language (like HTML)
- Text-based (an XML file is plain text)
- an important standard for data interchange

Unlike HTML, XML specifies content, not formatting

- XML has tags like `<publisher>` and `<creator>`
- HTML has tags like `<b>` "bold" and `<p>` "paragraph"

More on XML

- Later in the workshop
- XML 1.0 Specification: http://www.w3.org/TR/xml/
OLAC metadata in XML

<olac>
  <dc:creator>Derbyshire, Desmond C.</dc:creator>
  <dc:creator code="1986"></dc:creator>
  <dc:title>Topic continuity and OVS order in Hixkaryana</dc:title>
  <dc:date code="1986"></dc:date>
  <dc:type code="Text"/>
  <type xsi:type="olac:linguistic-type" olac:code="language_description"/>
  <dc:subject>Word order</dc:subject>
  <dcsubject xsi:type="olac:language" olac:code="hix"/>
</olac>

OLAC extensions important for searching

<olac>
  <dc:creator>Derbyshire, Desmond C.</dc:creator>
  <dc:creator code="1986"></dc:creator>
  <dc:title>Topic continuity and OVS order in Hixkaryana</dc:title>
  <dc:date code="1986"></dc:date>
  <dc:type code="Text"/>
  <type xsi:type="olac:linguistic-type" olac:code="language_description"/>
  <dc:subject>Word order</dc:subject>
  <dcsubject xsi:type="olac:language" olac:code="hix"/>
</olac>

Now the good news...

- You don’t have to write OLAC metadata in XML ‘by hand’
- You can use the OLAC repository editor: http://linguistlist.org/olac/ore/
- Or you can deposit your work with an archive that is a member of OLAC.
  - The archive will create metadata for its holdings in OLAC format and make this metadata available to the OLAC harvester
  - Metadata for your work will thus be made available to any OLAC search engine.
Plan to archive your work!

- An archive is a preservation repository, not just a website or a community center
- An archive will
  - Preserve your documentation, e.g., by providing Curation
  - Controlled storage conditions, as needed
  - A data migration plan that will ensure that your work does not become inaccessible because of technology changes
- Provide access to your work according to the conditions you have specified

Identifying and working with an archive:
- Archives that accept endangered language documentation:
  - AIATSIS, Australian Institute of Aboriginal and Torres Strait Islander Studies.
  - AILCA, Archive of the Indigenous Languages of Latin America, at the University of Texas at Austin.
  - ALNC, Alaska Native Language Center, at the University of Alaska Fairbanks.
  - AILPA, at the Max Planck Institute for Psycholinguistics, Nijmegen.
  - HRELP, Hans Rausing Endangered Languages Project, at the School of Oriental and African Studies, University of London.
  - LDC, Linguistic Data Consortium, hosted at University of Pennsylvania.
  - NWIC, Virtual Library, Northwest Indian College Oksale Program at the Graduate School of Library and Information Science, University of Texas at Austin.
  - PAKRARCH, Pacific and Regional Archive for Digital Sources in Endangered Cultures, University of Melbourne.
  - AROCA, Survey of California and Other Indian Languages, at University of California, Berkeley.
  - THDL, The Tibetan and Himalayan Digital Library, at the University of Virginia.

Classroom

- E-MELD School of Best Practices in Digital Language Documentation

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Language</th>
<th>Type</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-MELD</td>
<td>School</td>
<td>Best Practices</td>
<td>Digital Language Documentation</td>
<td>Introduction</td>
<td>Details</td>
</tr>
</tbody>
</table>