

# **EMELD Working Group on Resource Archiving**

Language Digitization Project,  
Conference 2003: Digitizing and  
Annotating Texts and Field  
Recordings



# Preamble

Sparkling prose that briefly explains why linguists should archive language documentation materials.

# Definitions

- **Archive:** a trusted repository created and maintained by an institution with a demonstrated commitment to permanence and the long-term preservation of archived resources.
- **Collection:** the body of documentary materials created by linguists and native speakers, that will be deposited in an archive.

# Best Practice Recommendations

- For archives: this [checklist] serves as a guide for the creation and maintenance of digital archives for language documentation resources.
- For individual researchers and projects: these [guidelines] will help you develop archive-ready collections.

# Checklist 1: Archive Management I

- Mission statement: Define clearly
  - ❖ the scope and scale of the collection (e.g. Algonquian languages);
  - ❖ where the resources will come from (e.g. legacy materials from researchers);
  - ❖ who will be the archive's primary users.

# Checklist 1: Archive Management II

Define procedures and policies for:

- acquisition of materials, including a triage strategy for prioritizing the digitization schedule;
- dissemination of materials, including access restrictions, interface languages, etc.;
- quality assurance;
- tracking digitization standards and forward migration to new digital formats;
- disaster recovery - backups, mirror sites, etc.

# Checklist 1: Archive Management - Readings

- OAIS Reference Model for Digital Libraries [website]
- EU-US Working Group on Spoken-Word Audio Collections [website]
- OLAC documents [http://www.language\_archive.org]

## Checklist 1: Archive Management - Tools

- MPI Corpus Browser [<http://www.mpi.nl/IMDI>]
- Greenstone Digital Library System [<http://www.gsdl.org>]
- DSpace [<http://www.dspace.org>]
- Dlese [<http://www.dlese.org>]



## Checklist 2: Intellectual Property

Develop policies that address:

- liability issues for the host institution;
- the copyrights of resource producers, both native speakers and researchers;
- access and use requirements for users.

Also provide guidelines for resource producers for eliciting consent to archive & publish.

## Checklist 2: Intellectual Property - Readings

- Lieberman article  
[<http://www.ldc.upenn.edu/exploration/expl2000/papers/liberman/liberman.html>]
- Copyright info from UT lawyer  
[<http://www.utsystem.edu/OGC/intellectualproperty/index.htm>]
- World Intellectual Property Organization  
[<http://www.wipo.int>]

## Checklist 2: Intellectual Property - Examples

- AIATSIS:  
<http://coombs.anu.edu.au/SpecialProj/ASEDA/ASEDA.html>
- AILLA:  
[http://www.ailla.utexas.org/site/use\\_conditions.html](http://www.ailla.utexas.org/site/use_conditions.html)
- OLAC: <http://www.language-archives.org/docs/license.html>

## Checklist 3: Metadata

- Metadata schema must be OLAC-compliant.
- Best practice is to adopt and customize an existing schema (OLAC, IMDI) to maximize inter-operability.
- Be an active participant in the international language archive community.
- Develop metadata for administration, content description, resource description, and IPR.

## Checklist 3: Metadata - Links

- OLAC [<http://www.languagearchives.org>]
- IMDI [<http://www.mpi.nl/IMDI>]
- Dublin Core [<http://dublincore.org/>]
- METS [<http://www.loc.gov/standards/mets/>]

## Checklist 4: Archival object definitions

- Definition of what constitutes an archival object must be clear and consistent:
  - ❖ digital objects correspond to original media;
  - ❖ digital objects correspond to documentary events (e.g. a recording session).
- Persistent identifiers should support
  - ❖ retrieval of the original (analog) medium;
  - ❖ matching related objects that reference original media (e.g. texts that refer to specific tapes);
  - ❖ correct citation of archived resources.

## Checklist 5: Formats I

- The archive must clearly distinguish master copy formats from presentation formats so that users understand that digital materials in presentation formats are not acceptable, archive-quality, materials.
- Archives should publish their digitization standards as guidelines for producers who wish to deposit digital materials.

## Checklist 4: Formats II

General requirements for archive-quality (master copy) formats:

- non-proprietary; that is, the encoding is in the public domain;
- portable, re-useable;
- best possible reproduction of the original.



## Checklist 4: Formats - Readings

- Links to the EMELD BP for Resource Creation, Transcription & Annotation, & Resource Conversion
- Links to sites with info about digitization standards
- Links to sites with info about digitization methods

## Checklist 4: Formats - Tools

- Links to sites/EMELD pages with recommendations for digitization equipment
- Links to software, e.g. Praat, or link to the Resource Creation Tools page

# Guidelines for Collections

- Intended for individuals and projects;
- Promote the production of archive-quality materials, and the preparation of existing corpora for archiving.

# Guidelines 1: Getting started

Search for an archive that covers your linguistic or geographical area:

- OLAC member archives [[link](#)];
- Relevant publications (e.g. SSILA newsletter);
- Other researchers in the same area;
- Funding agencies (e.g., Rausing Foundation).

## Guidelines 2: If there is a suitable archive

Go to their website and/or write to their contact person, and follow their guidelines for:

- metadata
- intellectual property:
  - ❖ consent
  - ❖ defining access restrictions
- formats & sorting materials into archive objects (e.g. session bundles)

## Guidelines 2: If there is no suitable archive I

- Choose a metadata schema [link] and create metadata for each item concerning:
  - ❖ IPR and access restrictions
  - ❖ content
  - ❖ creation of the original resource
- Choose a digital library system [link] that
  - ❖ runs on your platform and
  - ❖ supports your metadata.

## **Guidelines 2: If there is no suitable archive II**

- Define your policy concerning IPR and develop a consistent practice for obtaining consent (forms, recorded statements).
- Follow Best Practice Recommendations for creation and conversion of resources [links].
- Badger your research community into establishing a proper archive.

## Guidelines 2

# Resources for Language Documentation

- Links to readings & tools?
- Field reports from documentation projects?
- Anything else?