The role of metadata in the infrastructure for archival interoperation

Gary F. Simons
SIL International and
Graduate Institute of Applied Linguistics
Co-coordinator,
Open Language Archives Community

LSA Workshop on Sociolinguistic Archive Preparation
Portland, 4-5 Jan 2012
The problem: Sharing

► Sociolinguists are asking each other:
  ▪ How do we archive our corpora so that they can be shared?

► We need to be able to
  ▪ Compare current findings with previous findings to describe change over time
  ▪ Compare findings from multiple speech communities to describe synchronic differences
  ▪ Study someone’s data to confirm their findings
With sustainability

► And we want to keep doing these things far into the future.

► But given the relentless:
  ▪ Entropy that degrades digitally stored information
  ▪ Innovation that obsoletes hardware and software
  ▪ Discovery that provides new ways of doing things

► How do we keep our corpora from
  ▪ Falling into disuse, then
  ▪ Slipping into oblivion?
1. Foundational concepts:
   - Five *necessary conditions* for the sustainable sharing of sociolinguistic corpora
   - Four *key players* in the infrastructure of sustainable sharing
   - Three terms: *archive, metadata, interoperate*

2. Corpus-level metadata and OLAC as a global infrastructure for corpus sharing

3. Observation-level metadata as the basis for data interoperation between corpora
Necessary conditions

► In order for a corpus to be shared today, it must be:
  ▪ Discoverable
  ▪ Available
  ▪ Interpretable
  ▪ Portable

► And for this to continue far into the future, it must also be:
  ▪ Preserved
1. Discoverable

- A corpus cannot be used unless the prospective user is able to find it.
- The key is descriptive metadata:
  - The description of the corpus must be published in such a way that the user to whom it is relevant is able to discover its existence when searching.
  - The description of the corpus must be done in such a way that the user to whom it is relevant is able to judge it as being relevant without having to first obtain a copy.
2. Available

► A corpus cannot be used unless it is available to the prospective user.

► Availability has two major facets:
  ▪ User must have the right to access and use the corpus; the rights must be sorted out when the corpus is created and clarified when it is archived
  ▪ User must know the procedure for gaining access

► Open Access fosters the most widespread use

► Long term access requires persistent URIs
A corpus cannot be used if the user is not able to make sense of the content.

OAIS standard (ISO 14721) states that:

- Archives must ensure that resources are “independently understandable” by the designated user community (*i.e.*, no need to consult producer)

*E.g.*, Document the context of the study, the methodology, terminology, abbreviations, markup conventions, character encodings
4. Portable

- A corpus cannot be used if it does not interoperate in user’s working environment.

- A corpus must work with:
  - User’s hardware and operating system
  - Software tools available to the user
  - Best practices of the designated user community

- Maximizing portability means:
  - Formats that are open and transparent (not proprietary)
  - Following best practice markup and terminology
Use of a corpus cannot be sustained if a faithful copy of the original resource ceases to exist.

Archiving institution must follow procedures to:

- Ensure that resources are preserved against all reasonable contingencies (e.g., offsite backup)
- Ensure periodic migration to fresh and current media
- Ensure that all copies are authenticated as matching the original
- Keep preservation metadata (provenance, fixity)
It takes an infrastructure

- **Sociolinguists** can create corpora that are *portable* and *interpretable*.
- They cannot *preserve* them long term or provide the means of *access* to all users.
  - That’s what **Archives** do.
- They cannot make them *discoverable*.
  - That’s what **Aggregators** do (e.g., Google).
# The key players

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>A person who creates language resources</td>
</tr>
<tr>
<td>Archive</td>
<td>An institution that curates language resources for long-term preservation</td>
</tr>
<tr>
<td>Aggregator</td>
<td>An institution that makes resources from many archives interoperate</td>
</tr>
<tr>
<td>User</td>
<td>A person who wants to use language resources</td>
</tr>
</tbody>
</table>
The big picture

- Creator
- Resources
- Archive
- User
- Requests
- Aggregator
The term is polysemous in common usage.

- *E.g., Wikipedia:* An **archive** is a collection of historical records, or the physical place they are located.
- In “Workshop on sociolinguistic **archive** preparation”, the first sense is in focus; but the new emphasis on archiving in the linguistics community, puts the focus on the second.

Problem and terminological solution

- If we call a collection of information an archive, linguists will think they’ve “archived” when they’ve created an “archive”.
- Rather we want them to create an **archivable corpus** and they’ve archived when they’ve placed that in an **archive**.
Terminology: metadata

► Literally, “data about data”

► This, too, has multiple meanings. Just as we have data at many levels, so also with metadata:
  - When librarians and archivists talk about metadata, they mean data about the items they are curating
  - When sociolinguists use the term, they often mean data about the individual observations they are taking

► To avoid confusion, I will speak of:
  - Corpus-level metadata vs. Observation-level metadata
Two or more systems *interoperate* when they can exchange information or services and then make satisfactory use of what is exchanged.

Two levels of interoperation (corresponding to corpus-level and observation-level) are distinguished:

- **macrointeroperation** — interoperation between archives to discover relevant corpora
- **microinteroperation** — interoperation between relevant corpora to compare their contents
1. Foundational concepts:
   - Five *necessary conditions* for the sustainable sharing of sociolinguistic corpora
   - Four *key players* in the infrastructure of sustainable sharing
   - Three terms: archive, metadata, interoperate

2. Corpus-level metadata and OLAC as a global infrastructure for corpus sharing

3. Observation-level metadata as the basis for data interoperation between corpora
OLAC is an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources by:

- Developing consensus on best current practice for the digital archiving of language resources
- Developing a network of interoperating repositories & services for housing and accessing such resources

Founded in 2000

- Now has a library of >100,000 items from 40 archives
Who’s involved?

- Aboriginal Studies Electronic Data Archive, Australia
- Academia Sinica, Taiwan
- African Language Materials Archive
- Alaska Native Language Center
- C'ek'aedi Hwnax Ahtna Regional Archive, Alaska
- Californiia Language Archive
- Central Institute of Indian Publications, India
- Centre de Ressources pour la Description de l'Oral
- CHILDES Data Repository
- Comparative Corpus of Spoken Portuguese, Brazil
- Cornell Language Acquisition Laboratory
- Ethnologue: Languages of the World
- European Language Resources Assoc., France
- Graduate Institute of Applied Linguistics
- Kaipuleohone, Univ. of Hawaii
- The Language Archive’s IMDI Portal, Netherlands
- Language Commons Language Corpora
- Linguistic Data Consortium Corpus Catalog
- LINGUIST List Language Resources
- Multi-Modal Media File Server, Switzerland
- Multimodal Teaching and Learning Corpora, France
- Natural Language Software Registry, Germany
- Online Database of Interlinear Text (ODIN)
- Oxford Text Archive, England
- PARADISEC, Australia
- Perseus Digital Library
- POLLEX Online, New Zealand
- Research Papers in Computational Linguistics
- Rosetta Project Library of Human Language
- SIL Language and Culture Archives
- Speech and Language Data Repository, France
- Surrey Morphology Group Databases, England
- TalkBank
- The Text Laboratory, Univ. of Oslo
- Tibetan and Himalayan Digital Library
- TST Centrale, Netherlands
- Typological Database Project, Netherlands
- University of Bielefeld Language Archive, Germany
- WALS Online, Germany
The community has defined standards for the encoding and exchange of corpus-level metadata to permit discovery and sharing:

- **OLAC Metadata** — XML format of metadata records
- **OLAC Repositories** — Protocol for metadata harvesting and requirements on compatible repositories
- **OLAC Metadata Usage Guidelines** — Explains the available metadata elements and how to use them
The 40 archives publish catalogs in a standard XML form … to be harvested by the OLAC aggregator … which supplies information to search services.

search.language-archives.org

Linguist List
The Sociolinguistics of Language

Language and society: The topic of sociolinguistics

Memories

Missionaries and Language

Second language proficiency report

Sociolinguistics 14:1 : Volume XIV number 1 spring/summer 1983
Taipale, Michele M.; Burnich, Joann. 1983. Chico, Cal : Research Committee on Sociolinguistics.

Sociolinguistics, current trends and prospects
Search for language resources

Results: 6

Showing hits 1 - 6 out of 6

BilingBank German-English Eppler Corpus

BilingBank Chinese-Hungarian Langman Corpus

SLX Corpus of Classic Sociolinguistic Interviews

Nationwide Speech Project
Cynthia G. Clopper and David B. Pisoni. 2007. The LDC Corpus Catalog.

CSLU: Kids' Speech Version 1.1

N4 NATO Native and Non-Native Speech
SLX Corpus of Classic Sociolinguistic Interviews

Title: SLX Corpus of Classic Sociolinguistic Interviews
ID: LDC2003T15
Online: No
Archive: The LDC Corpus Catalog (see archive description)
Contributor: Stephanie Strasssel, Jeffrey Conn, Suzanne Evans Wagner, Christopher Cieri, William Labov, and Kazuaki Maeda (author)
Date: 2003-11-25
Description: Release type: General
Non-member fee: 100.00 USD
Reduced-license fee: 100.00 USD
Extra-copy fee: 100.00 USD
Online documentation: http://www.ldc.upenn.edu/Catalog/docs/LDC2003T15
Application: sociolinguistics
Application: discourse analysis
Related research project: DASL
Related research project: Talkbank
Membership year: 2003
Data source: field recordings
Content language: English
<olac:olac>
  <dc:title>SLX Corpus of Classic Sociolinguistic Interviews</dc:title>
  <dc:creator xsi:type="olac:role" olac:code="author">Stephanie Strassel, Jeffrey Conn, Suzanne Evans Wagner, Christopher Cieri, William Labov, Kazuaki Maeda</dc:creator>
  <dc:date xsi:type="dcterms:W3CDTF">2003-11-25</dc:date>
  <dc:description>Application: sociolinguistics</dc:description>
  <dc:description>Data source: field recordings</dc:description>
  <dc:format>Sample rate: 22050Hz; Sample type: pcm</dc:format>
  <dcterms:extent>Corpus size: 1572864.000 KB</dcterms:extent>
  <dcterms:medium>Distribution: 1 DVD</dcterms:medium>
  <dc:identifier>LDC2003T15</dc:identifier>
  <dc:rights>Non-member license: http://www.ldc.upenn.edu/Catalog/nonmem_agree/generic.license.html</dc:rights>
  <dc:language xsi:type="olac:language" olac:code="eng"/>
  <dc:subject xsi:type="olac:language" olac:code="eng"/>
  <dc:type xsi:type="olac:linguistic-type" olac:code="primary_text"/>
  <dc:type xsi:type="dcterms:DCMIType">Sound</dc:type>
</olac:olac>
OLAC uses Dublin Core standard which has:

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

And adds extensions (with controlled vocabularies) specific to our community:

- Language Identification (ISO 639-3), Linguistic Data Type, Linguistic Field, Participant Role, Discourse Type
Corpus-level metadata for sociolinguistics

The OLAC standard provides a good starting point with an implemented infrastructure for discovery.

The sociolinguistics community could define further specialization for discovery across the community:

- Agree on a standard type label
  - E.g., `<dc:type>Sociolinguistic corpus</dc:type>`
- Use the OLAC extension mechanism to define a controlled vocabulary for relevant resource types
- Define standardized labels for standard formats and use them in `<dc:format>` elements
1. Foundational concepts:
   - Five *necessary conditions* for the sustainable sharing of sociolinguistic corpora
   - Four *key players* in the infrastructure of sustainable sharing
   - Three terms: *archive, metadata, interoperate*

2. Corpus-level metadata and OLAC as a global infrastructure for corpus sharing

3. Observation-level metadata as the basis for data interoperation between corpora
The data about the individual observations within a corpus is another kind of metadata, *e.g.*,:

- Coding of demographic characteristics
- Coding of social attitudes
- Coding of social situations

Interoperation over these requires definition of:

- Formats for marking up the structure of primary data and associated metadata (*e.g.* an XML schema)
- Controlled vocabularies for values of metadata elements
Automating microinteroperation

► When multiple corpora use the same markup format and controlled vocabularies
  ▪ Parsers can load them into a common database
  ▪ Search and aggregation of statistics across those corpora is then possible within that database

► Doing this on a large scale requires discovering all corpora that follow the supported standards
  ▪ Therefore, exploit macrointeroperation infrastructure
  ▪ Define standard labels for supported formats and vocabularies and use them in corpus-level metadata
Sociolinguists can share their corpora long into the future if they:

- Deposit them in archives that will preserve them, make them accessible to potential users, and make them globally discoverable through an aggregation infrastructure like OLAC

- Use community-wide standards of format for markup and controlled vocabularies for analysis to make them portable and interpretable, not only for stand-alone use but also for automated interoperation across multiple corpora