Some challenges ahead for the Open Language Archives Community

Gary F. Simons
*SIL International*

*Co-coordinator with Steven Bird, Open Language Archives Community*

Workshop on Data Archives and Languages of the Americas, LDC, University of Pennsylvania, 9 February 2018
Roadmap

1. What we are

2. How we obtain data and how users access it

3. The current challenges we face
   - Increasing coverage, relevance, sustainability

4. The envisioned way forward
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 and services for housing and accessing such resources

Founded in 2000

- Now has a catalog of \(~335,000\) items from 60 archives
Partial list of participants
(> 500 items; see complete list)

- Aboriginal Studies Electronic Data Archive
- Alaska Native Language Archive
- C'ek'aedi Hwnax Ahtna Regional Archive
- Califronia Language Archive
- COollections de COrpus Oraux Numeriques
- Crúbadán Projec
- Ethnologue: Languages of the World
- European Language Resources Association
- Glottolog 2.7
- Graduate Institute of Applied Linguistics
- Kaipuleohone, Univ. of Hawaii
- The Language Archive’s IMDI Protal
- Language Documentation and Conservation
- Linguistic Data Consortium Corpus Catalog
- LINDAT/CLARIN Digital Library, Prague
- LINGUIST List Language Resources
- Living Archive of Aboriginal Languages,
- Online Database of Interlinear Text (ODIN)
- Oxford Text Archive
- PARADISEC
- Pacific Collection, U of Hawai'i Library
- PHOIBLE Online
- Research Papers in Computational Linguistics
- Rosetta Project Library of Human Language
- SIL Language and Culture Archives
- TransNewGuinea.org
- WALS Online, Germany
How do we get data?

► Participating archives contribute the metadata on their archive holdings using standard formats that have been defined by the community. They are at:

► Including
  - OLAC Metadata — XML format of metadata records
  - OLAC Repositories — Protocol for metadata harvesting and the requirements on conformant repositories
  - OLAC Metadata Usage Guidelines — Explains the available metadata elements and how to use them
A sample metadata record

<olac:olac>
  <dc:title>LAPSyD Online page for Cape Verde Creole, Santiago dialect</dc:title>
  <dc:description>This resource contains information about phonological inventories, tones, stress and syllabic structures</dc:description>
  <dcterms:modified xsi:type="dcterms:W3CDTF">2012-05-17</dcterms:modified>
  <dc:type xsi:type="dcterms:DCMIType">Dataset</dc:type>
  <dc:format xsi:type="dcterms:IMT">text/html</dc:format>
  <dc:publisher xsi:type="dcterms:URI">www.lapsyd.ddl.ish-lyon.cnrs.fr</dc:publisher>
  <dcterms:license>http://creativecommons.org/licenses/by-nc-nd/3.0/</dcterms:license>
  <dc:contributor xsi:type="olac:role" olac:code="author">Maddieson, Ian</dc:contributor>
  <dc:subject xsi:type="olac:linguistic-field" olac:code="phonology"/>
  <dc:subject xsi:type="olac:linguistic-field" olac:code="typology"/>
  <dc:type xsi:type="olac:linguistic-type" olac:code="language_description"/>
  <dc:language xsi:type="olac:language" olac:code="eng"/>
  <dc:subject xsi:type="olac:language" olac:code="kea">Cape Verde Creole, Santiago dialect</dc:subject>
</olac:olac>
An overview

The 60 archives submit catalogs in a standard form...

which supplies information to search services.

► to the OLAC aggregator...

search.language-archives.org

Linguist List, WorldCat, CLARIN, ...
How do researchers access the metadata?

- Via Google search (or any web search engine) since OLAC exposes everything as pages that crawlers can access.
- Via our faceted search engine which exploits the controlled vocabularies to give search with complete recall and precision.
- Via links from language-related sites like Ethnologue.
- Via services like WorldCat, CLARIN, Linguist List which use OAI-PMH to harvest the metadata from OLACA.
- By consuming the raw XML or RDF/XML directly from OLAC.
Use any ISO 639-3 code at end of URL
Today: 77 total resources indexed to [bbb]
From: PARADESIC, SIL; plus Crubadan, Ethnologue, GIAL, Glottolog, Rosetta, TransNewGuinea, U Hawaii Library, WALS
### Sample catalog record

**OLAC Record**

oai:paradisec.org.au:TD1-P028

#### Metadata

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>Barai stories</td>
</tr>
<tr>
<td><strong>Access Rights</strong></td>
<td>standard, as per PDSC Access form</td>
</tr>
<tr>
<td><strong>Bibliographic Citation</strong></td>
<td>Dutton, Tom (recorder), Silolo (speaker); Barai stories, WAV/MP3 <a href="http://paradisec.org.au/repository/TD1/P028">http://paradisec.org.au/repository/TD1/P028</a> 2009-11-07</td>
</tr>
<tr>
<td><strong>Contributor (recorder)</strong></td>
<td>Dutton, Tom</td>
</tr>
<tr>
<td><strong>Contributor (speaker)</strong></td>
<td>Silolo</td>
</tr>
<tr>
<td><strong>Coverage (Box)</strong></td>
<td>northlimit=-9.06936; southlimit=-9.61342; westlimit=147.61114; eastlimit=148.20572</td>
</tr>
<tr>
<td><strong>Coverage (ISO3166)</strong></td>
<td>PG</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>Digitised: yes; Media: BASF LGS 35 RtoR 270m; Audio notes: Side A only -- A) 3 3/4ips. Some flutter. -- B) BLANK -- A) -- 0.00 BARAI - Origin story in Police Motu... -- 1.48 Story number 2 by the same informant... -- 3.35 Cowboy song... -- 5.27 BARAI - Origin story in Police Motu... -- 7.20 BARAI songs -- 7.54 2nd one... -- 8.16 And a church song in Motu... -- 9.50 ... five times... -- 11.30 This is a crying song for funerals... -- 12.17 end -- B) -- BLANK;</td>
</tr>
<tr>
<td><strong>Identifier</strong></td>
<td>TD1-P028</td>
</tr>
<tr>
<td><strong>Identifier (URI)</strong></td>
<td><a href="http://paradisec.org.au/repository/TD1/P028">http://paradisec.org.au/repository/TD1/P028</a></td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>Barai, Hiri Motu</td>
</tr>
<tr>
<td><strong>Language (ISO639)</strong></td>
<td>bbb</td>
</tr>
</tbody>
</table>

Link to the resource at PARADISEC
Via our faceted search engine

http://search.language-archives.org
Barai

A language of Papua New Guinea

<table>
<thead>
<tr>
<th>ISO 639-3</th>
<th>bbb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Central province: Rigo district, Rigo Inland RLLG, west of Moni river; Oro province: Afore district, Managalas plateau; Itokama, Madokoro, Naokanane, and Umuate villages (Birarle dialect).</td>
</tr>
<tr>
<td>Language Maps</td>
<td>Papua New Guinea, Map 16</td>
</tr>
<tr>
<td>Language Status</td>
<td>5 (Developing).</td>
</tr>
<tr>
<td>Classification</td>
<td>Trans-New Guinea, Southeast Papuan, Koiarian, Baraic</td>
</tr>
<tr>
<td>Dialects</td>
<td>Birarle, Muguani. Lexical similarity: 50% with Ese [mcq].</td>
</tr>
<tr>
<td>Typology</td>
<td>SOV.</td>
</tr>
<tr>
<td>Language Resources</td>
<td>OLAC resources in and about Barai</td>
</tr>
<tr>
<td>Writing</td>
<td>Latin script [l att]</td>
</tr>
<tr>
<td>Other Comments</td>
<td>Christian, traditional religion.</td>
</tr>
</tbody>
</table>

PLACE IN LANGUAGE CLOUD

Click to enlarge with explanation

ETHNOLOGUE PRODUCTS

Languages of Papua New Guinea, Map 16

An Ethnologue Language Map

$24.55

Add to cart

Browse all products

JOIN THE CONVERSATION

Be notified whenever someone posts Feedback about the language, country or region you are interested in.

Follow Barai
Santa Barbara Corpus of Spoken American English Part I

Author: John W. Du Bois; Chafe, Wallace; Meyer, Charles; Thompson, Sandra

Publisher: Linguistic Data Consortium https://www.ldc.upenn.edu 2000

Edition/Format: Archival material : English

Summary: *Introduction* The Santa Barbara Corpus of Spoken American English is based on hundreds of recordings of natural speech from all over the United States, representing a wide variety of people of different regional origins, ages, occupations, and ethnic and social backgrounds. It reflects many ways that people use language in their lives: conversation, gossip, arguments, on-the-job talk, card games, city council meetings, sales pitches, classroom lectures, political speeches, bedtime stories, sermons, weddings, and more. *Data* The three CD-ROM volumes in Part I contain
Ways of consuming OLAC metadata

► Full or incremental harvest at OLACA (via OAI-PMH)
  ▪ http://www.language-archives.org/cgi-bin/olaca3.pl

► RDF/XML of any metadata record is available by HTTP content negotiation (Accept: application/rdf+xml)
  ▪ E.g., http://www.language-archives.org/item/oai:paradisec.org.au:AA1-001

► Nightly gzipped dumps of the entire metadata catalog
  ▪ OLAC XML: http://www.language-archives.org/xmldump/ListRecords.xml.gz
  ▪ RDF/XML: http://www.language-archives.org/static/olac-datahub.rdf.gz
Increasing coverage

- There are significant collections not yet participating, both archives and special collections within libraries
  - We have observed that implementing a data provider for our idiosyncratic metadata format is too high a bar
- Some archives don’t yet expose the actual resources
  - They expose only a landing page per language, and not the individual corpora or resources
- Linguists need to be able to report resources they discover in places that would never join OLAC
Many archives need to improve metadata quality so as to improve the discoverability of their holdings

- 24 out of 60 archives score below 70% on our metric

Huge gaps in our Linguistic Data Type vocabulary

- Current set of 3 values covers 60% of resources; we are lacking type labels relevant to the rest

Subcommunities could make it relevant for themselves

- E.g., `<dc:type>Sociolinguistic corpus</dc:type>

- E.g., for ELAN: `<dc:format>text/x-eaf+xml</dc:format>`
Increasing sustainability

- We have a sustainability problem at the level of participating archives keeping up with change
  - Today, 20 archives show as failing to harvest
  - An overlapping set of 21 have not updated their catalog within the last 5 years

- We have a sustainability problem at the level of our central infrastructure
  - It is showing its age (> 15 years)
  - Depends on volunteerism and contributions
A deeper issue

► OLAC’s metadata format plus infrastructure is an idiosyncratic solution developed and maintained within the linguistics community

▪ But our community is not particularly well-equipped to implement and manage information systems.

► A more robust solution would be to steer OLAC and the cataloging of language resources into the library and information systems mainstream.
Envisioned way forward

► We are monitoring trends in the library community
  ▪ From standardized markup formats (like XML schemas) to Linked Data (RDF) and Metadata Application Profiles
  ▪ We’ve mapped our metadata to Linked Data and envision a Language Resource Type vocab to anchor a profile

► An ideal future
  ▪ We would move from having an idiosyncratic community-specific infrastructure to a mainstream infrastructure that interoperates with the global Web of Data
  ▪ We would influence mainstream cataloging practices to embrace ISO 639-3 and a Language Resource Type vocab
Conclusion

► OLAC has a functioning infrastructure that allows our community to index and discover language resources
  ▪ See OLAC Implementers' FAQ to learn how to join

► But we are being held back by having an idiosyncratic infrastructure
  ▪ A more promising future would be to move into the mainstream infrastructure of the digital library community