

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



Open Language Archives Community

www.language-archives.org

- ▶ 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

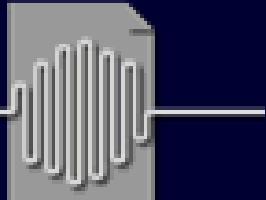
(> 500 items; see [complete list](#))

- ▶ Aboriginal Studies Electronic Data Archive
- ▶ Alaska Native Language Archive
- ▶ C'ek'aedi Hwnax Ahtna Regional Archive
- ▶ California Language Archive
- ▶ Collections 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:
 - <http://www.language-archives.org/documents.html>
- ▶ 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



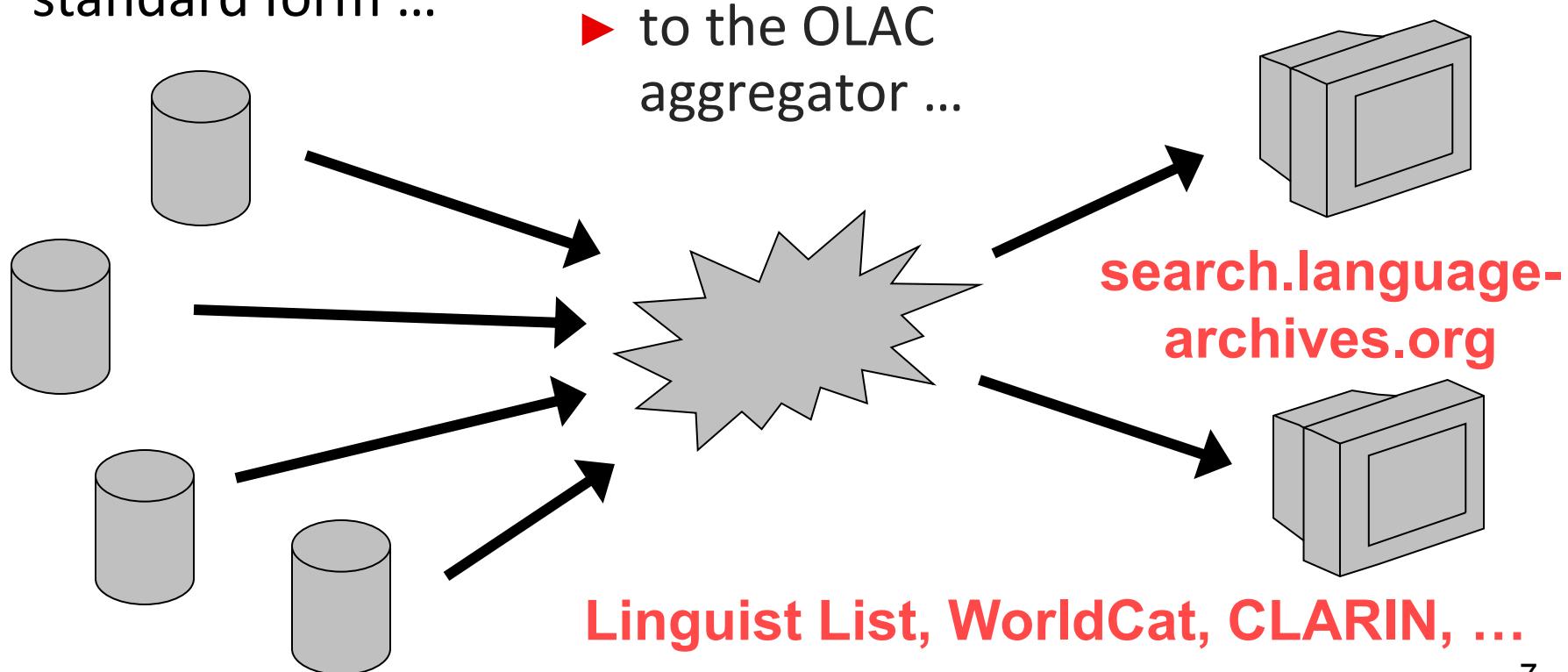
A sample metadata record

```
<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:identifier xsi:type="dcterms:URI">http://www.lapsyd.ddl.ish-lyon.cnrs.fr/lapsyd/index.php?data=view&code=692</dc:identifier>
  <dc:type xsi:type="dcterms:DCMType">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.





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



Via Google search

Google

barai grammar

Search

[Advanced Search](#)

Web [Show options...](#)

Results 1 - 10 of about 12,000 for bar

[Resources in and about the Barai language](#) - 11:02am

Barai grammar highlights. Olson, Michael Leon. 1975. Canberra Australia : Australian National University. oai:gial.edu:28481; **Barai** sentence structure and ...
www.language-archives.org/language/bbb - [Cached](#) - [Similar](#) -

Use any ISO 639-3 code at end of URL

[OLAC Record: oai:paradisec.org.au:TD1-P034](#)

Barai Grammar con't from tape P29 -- 3. More **Barai** vocabulary. -- Side 2 -- 4. Orokaiva (Hamara Village) Lexico Stats. -- 5. Kwena (Managalasi) -- 6. ...
www.language-archives.org/.../oai:paradisec.org.au:TD1-P034 - [Cached](#) - [Similar](#) -

[Show more results from www.language-archives.org](#)

[SIL Bibliography: Barai grammar highlights](#)

"**Barai grammar** highlights." In T. E. Dutton (ed.), Studies in languages of central and south-east Papua, 471-512. Pacific Linguistics C, 29. ...

www.ethnologue.com/show_work.asp?id=11852 - [Cached](#) - [Similar](#) -
by ML Olson - 1975 - [Cited by 6](#) - [Related articles](#) - [All 2 versions](#)

Ethnologue report for language code: bbb



Resources in and about the Barai language

ISO 639-3: [bar](#)

The combined catalog of all OLAC participants contains the following:

- Primary texts
- Lexical resources
- Language descriptions
- Other resources about the language
- Other resources in the language

Other known names and dialect names: Birarie, Muguan

Primary texts

1. Barai stories. Biliilo (speaker); Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P028](#)

Lexical resources

1. Barai, Kokila, Manubara. Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P027](#)
2. Barai (Dorobisori) Lexico. Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P028](#)
3. Barai Lexicostats. Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P030](#)
4. Lexico Stats - Barai (Doe). Taboro, Kwale, Ikega. Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P031](#)
5. Barai - Vocabulary and Grammar. Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P034](#)
6. Kanga, Karukuru, Namanedza, Emo. Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P036](#)

Language descriptions

1. Barai grammar highlights. Olson, Michael Leon. 1975. Canberra, Australia : Australian National University. [calolai.edu.28481](#)
2. Barai sentence structure and embedding. Olson, Michael L. 1973. Language Data, Asian-Pacific Series, 3. Santa Ana, CA: Summer Institute of Linguistics. viii, 141 p. [calill.org.11849](#)
3. Barai grammar/highlights. Olson, Michael L. 1975. In T. E. Dutton (ed.), Studies in languages of central and south-east Papua, 471-512. Pacific Linguistics C, 29. Canberra: Australian National University. [calill.org.11850](#)
4. WALS Online Resources for Barai. Haspelmath, Martin (editor); Dryer, Matthew S. (editor); Gil, David (editor); Comrie, Bernard (editor). 2008-09-01. Max Planck Digital Library (<http://mpdl.mpg.de/>). [calwals.info/barai.html#des](#)

Other resources about the language

1. Barai sentence structure and embedding. Olson, Michael L. 1973. Language data, Asian-Pacific series ; no. 3. [calolai.edu.24982](#)
2. Barai clause/junctives : toward a functional theory of interclausal relations. Olson, Michael L. 1979. Graduate Institute of Applied Linguistics Library. [calolai.edu.24983](#)
3. Barai grammar/highlights. Olson, Michael. 1975. Studies in Languages of Central and Southeast Papua. [calwals.info.2824](#)
4. Barai clause/junctives: Toward a functional theory of interclausal relations. Olson, Michael L. 1979. WALS Online ReIDB. [calwals.info.4146](#)
5. Barai sentence structure and embedding. Olson, Michael L. 1973. Summer Institute of Linguistics. [calwals.info.4885](#)
6. Barai: a language of Papua New Guinea. Gordon, Raymond G., Jr. (editor). 2005. SIL International (www.sil.org). [calethnologue.com/pbg](#)
7. Barai pre-school report. Evans, Beverley. 1984. Read 19(2): 27-32. [calill.org.18570](#)
8. Results of Barai pre-school reading tests. Evans, Beverley. 1985. Read 20(1): 26-33. [calill.org.18571](#)
9. From beneficiary to facilitator. Olson, Michael L. 1982. Read 17(1): 1-6. [calill.org.15256](#)
10. Adult literacy-follow-up or don't start!. Evans, Peter. 1981. Read 16(1): 27-31. [calill.org.15647](#)
11. Saturation literacy. Evans, Peter. 1981. Read 16(1): 32-34. [calill.org.15843](#)
12. Barai (Northern Province). Evans, Beverley; Evans, Peter. 1980. In Neville Southall, Mary Stringer and Joice Franklin (eds.), Reports of vernacular literacy programmes conducted by the Summer Institute of Linguistics in Papua New Guinea, 44-47. Workpapers in Papua New Guinea Languages, 28. Ukarumpa: Summer Institute of Linguistics. [calill.org.23090](#)
13. Agency cooperation and learning needs in the Oro Province. Olson, Michael L. 1985. Read 20(2): 3-7. [calill.org.20042](#)
14. Wem clothing using a bilum stitch. Evans, Beverley. 1988. Read 23(1): 36-37. [calill.org.21955](#)
15. Reading readiness. Evans, Peter. 1991. Read 26(2): 42-45. [calill.org.31394](#)

Other resources in the language

1. Intelligibility Test Tape B' (Toll Stories). Dutton, Tom (recorder). n.d. Pacific And Regional Archive for Digital Sources In Endangered Cultures (PARADISEC). [calparadisecc.org.au/TP1-P008](#)



Sample catalog record



OLAC Record

[oai:paradisec.org.au:TD1-P028](http://oai.paradisec.org.au:TD1-P028)

Metadata

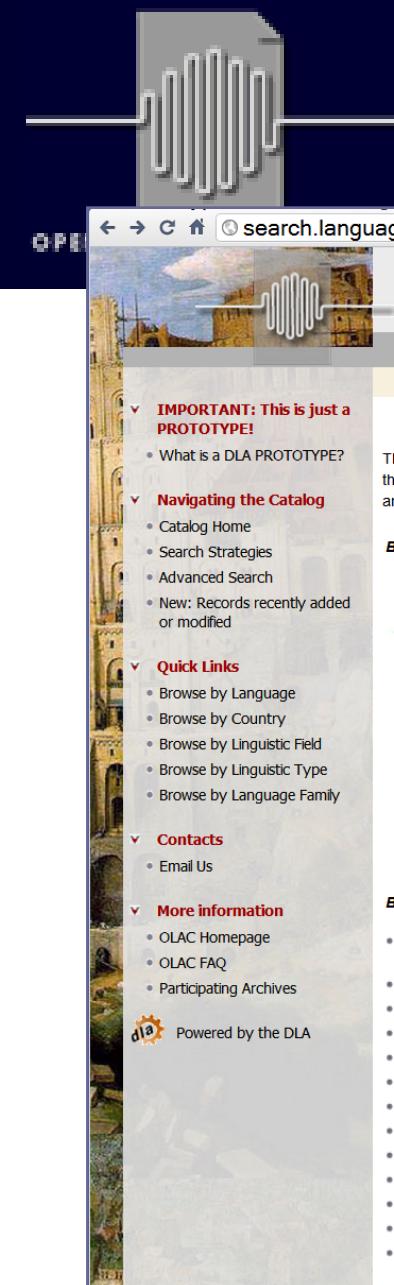
<i>Title:</i>	Barai stories
<i>Access Rights:</i>	standard, as per PDSC Access form
<i>Bibliographic Citation:</i>	Dutton, Tom (recorder), Silolo (speaker) ; Barai stories, WAV/MP3 http://paradisec.org.au/repository/TD1/P028 2009-11-07
<i>Contributor (recorder):</i>	Dutton, Tom
<i>Contributor (speaker):</i>	Silolo
<i>Coverage (Box):</i>	northlimit=-9.06936; southlimit=-9.61342; westlimit=147.61114; eastlimit=148.20572
<i>Coverage (ISO3166):</i>	PG
<i>Description:</i>	1. Barai - Origin Story in Police Motu. Gori in Police Motu -- 2. Cowboy song at Karekadobu (050-075) -- 3. Barai - Origin Story in Police Motu by Silolo -- 4. Barai Songs (104-150). Language as given: Barai
<i>Format:</i>	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;
<i>Identifier:</i>	TD1-P028
<i>Identifier (URI):</i>	http://paradisec.org.au/repository/TD1/P028
<i>Language:</i>	Barai
	Hiri Motu
<i>Language (ISO639):</i>	bbb

Link to the resource
at PARADISEC



Via our faceted search engine

<http://search.language-archives.org>



OPEN

search.language-archives.org/index.html

Participating Archives • OLAC • Delivered by the Penn Libraries
Printer-Friendly Page

Prototype: OLAC Language Resource Catalog

Search for language resources go

IMPORTANT: This is just a PROTOTYPE!

- What is a DLA PROTOTYPE?

Navigating the Catalog

- Catalog Home
- Search Strategies
- Advanced Search
- New: Records recently added or modified

Quick Links

- Browse by Language
- Browse by Country
- Browse by Linguistic Field
- Browse by Linguistic Type
- Browse by Language Family

Contacts

- Email Us

More information

- OLAC Homepage
- OLAC FAQ
- Participating Archives

Powered by the DLA

This catalog, developed by the **Open Language Archives Community (OLAC)**, provides access to a wealth of information about thousands of languages, including details of text collections, audio recordings, dictionaries, and software, sourced from dozens of digital and traditional archives.

Browse the OLAC records by Geographic region or by Language:

- English (3509)
- Spanish (2748)
- Yuracare (1269)
- Beaver (1044)
- French (1003)
- Ixõõ (845)
- Bora (748)
- Ocaina (678)
- Chhintange (673)
- Kriol (578)
- Portuguese (575)
- Nepali (559)
- Motlav (493)
- Trumai (470)
- Achinese (435)
- Saliba (429)
- Occitan (post 1500) (426)
- Afrikaans (408)
- South West Bay (405)
- Bahinemo (401)

[View more...](#)

Browse the OLAC records by Archive:

- A Digital Archive of Research Papers in Computational Linguistics (3280)
- ATILF Resources (9)
- Aboriginal Studies Electronic Data Archive (ASEDA) (707)
- Academia Sinica Collections (3)
- African Language Materials Archive (53)
- Alaska Native Language Center Archive (24)
- Archive of the Indigenous Languages of Latin America (100)
- Audio Archive of Linguistic Fieldwork (199)
- Boiste (1)
- CHILDES Data repository (275)
- CRDO archive (161)
- Central Institute of Indian Languages: Publications (345)
- Centre de Ressources pour la Description de l'Oral (CRDO) (3752)
- Magoria Books' Carib and Romani Archive (2)
- Multimodal Learning and teaching Corpora Exchange (25)
- ODIN - The Online Database of Interlinear Text (710)
- Oxford Text Archive (1264)
- Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (6755)
- Perseus Digital Library (1451)
- SIL Language and Culture Archives (22377)
- Surrey Morphology Group Databases (2)
- Survey of California and Other Indian Languages (2448)
- TALKBANK Data repository (182)
- The LDC Corpus Catalog (486)
- The LINGUIST List Language Resources (2039)
- The Natural Language Software Registry (69)
- The Rosetta Project: A Long Now Foundation Library of Human Languages (6574)

Sort by:

Possible Sorts: all

- Title [a-z][z-a]
- Id [a-z][z-a]
- Date [a-z][z-a]

Browse by:

Archive browse

- SIL Language and Culture Archives 22377
- IMDI to OAI bridge 14883
- Graduate Institute of Applied Linguistics Library 8176
- Ethnologue: Languages of the World 7413
- [view more...](#)

Online browse

- Yes 49386
- No 44554

Subject language browse

- English 3509
- Spanish 2748
- Yuracare 1269
- Beaver 1044
- [view more...](#)

Language family browse

- Austronesian 11969
- Malayo-Polynesian 10730
- South American Indian 8926
- Indo-European 8299
- [view more...](#)

Geographic region browse

- Americas 21252
- Asia 16047
- Pacific 14954
- [view more...](#)

ECCD-On....zip

eccc-tk-f....sef

Ethnolog....htm

Show all downloads... x

12



Barai

 [Print](#)[LANGUAGE](#)[MAP](#)[FEEDBACK](#)

A language of Papua New Guinea

ISO 639-3bbb**Population**

800 (2003 SIL).

Location

Central province: Rigo district, Rigo Inland RLLG, west of Moni river; Oro province: Afore district, Managalas plateau; Itokama, Madokoro, Naokanane, and Umuate villages (Birarie dialect).

Language Maps[Papua New Guinea, Map 16](#)**Language Status**

5 (Developing).

Classification[Trans-New Guinea](#), [Southeast Papuan](#), [Koiarian](#), [Baraic](#)**Dialects**Birarie, Muguani. Lexical similarity: 50% with Ese [[mcq](#)].**Typology**

SOV.

Language Development

Literacy rate in L1: 50%. Literacy rate in L2: 50%–60%. NT: 1994–2006.

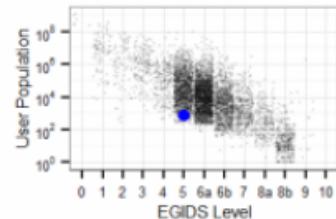
Language Resources[OLAC resources in and about Barai](#)**Writing**

Latin script [Latn]

Other Comments

Christian, traditional religion.

PLACE IN LANGUAGE CLOUD

[Click to enlarge with explanation](#)

ETHNOLOGUE PRODUCTS

[Languages of Papua New Guinea, Map 16](#)

An Ethnologue Language Map
\$24.95

[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](#)



Harvested via OAI-PMH from OLAC Aggregator



[Advanced Search](#) | [Find a Library](#)

[**<< Return to Search Results**](#)

[Cite/Export](#) [Print](#) [Email](#)

[Add to list](#) [Add tags](#) [Write a review](#)

Rate this item:



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



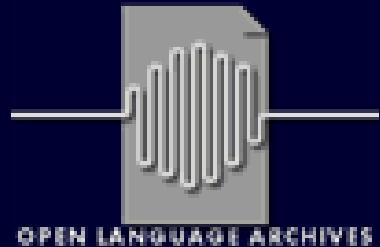
Increasing relevance

- ▶ 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



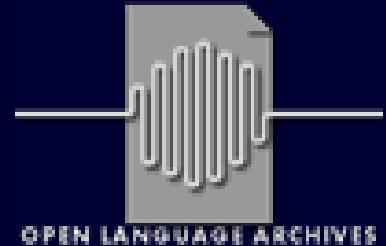
A deeper issue

- ▶ 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