OLAC:
Accessing the World's Language Resources

Steven Bird
CSSE, University of Melbourne
LDC, University of Pennsylvania

Gary Simons
SIL International
Graduate Institute of Applied Linguistics

What is OLAC?

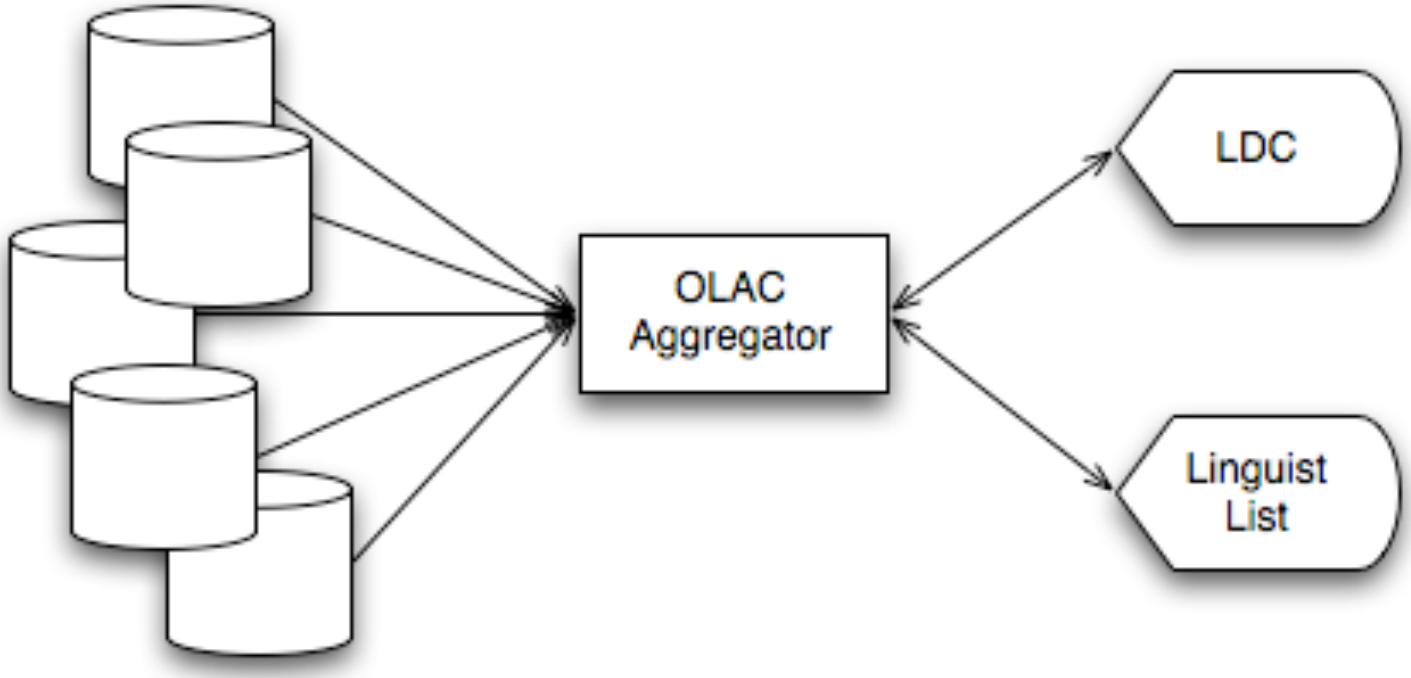
- OLAC is an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources by:
  - Developing consensus on best current practice for the digital archiving of language resources
  - Developing a network of interoperating repositories and services for housing and accessing such resources
- Founded in December 2000
  - now has 34 participating archives
  - hosted at www.language-archives.org

Who is involved in OLAC?

Aboriginal Studies Electronic Data Archive
1. Academia Sinica
2. Alaska Native Language Center
3. Archive of Indigenous Languages of Latin America
4. ATILF_Resources
5. Berkeley Language Center
6. Centre de Ressources pour la Description de l'Oral
7. CHILDES_Data Repository
8. Comparative Corpus of Spoken Portuguese
9. Cornell Language Acquisition Laboratory
10. Dictionnaire Universel Boiste 1812
11. DOBES catalogue (MPI, Nijmegen)
12. Ethnologue: Languages of the World
13. European Language Resources Association
14. Laboratoire Parole et Langage
15. Linguistic Data Consortium Corpus Catalog
16. LINGUIST List Language Resources
17. Natural Language Software Registry
18. Online Database of Interlinear Text (ODIN)
19. Oxford Text Archive
20. PARADISEC
21. Perseus Digital Library
22. Research Papers in Computational Linguistics
23. Rosetta Project 1000 Language Archive
24. SIL Language and Culture Archives
25. Surrey Morphology Group Databases
26. Survey for California and Other Indian Languages
27. TalkBank
28. Tibetan and Himalayan Digital Library
29. TRACTOR
30. Typological Database Project
31. University of Bielefeld Language Archive
32. University of Queensland Flint Archive
33. Virtual Kayardild Archive (Melbourne)

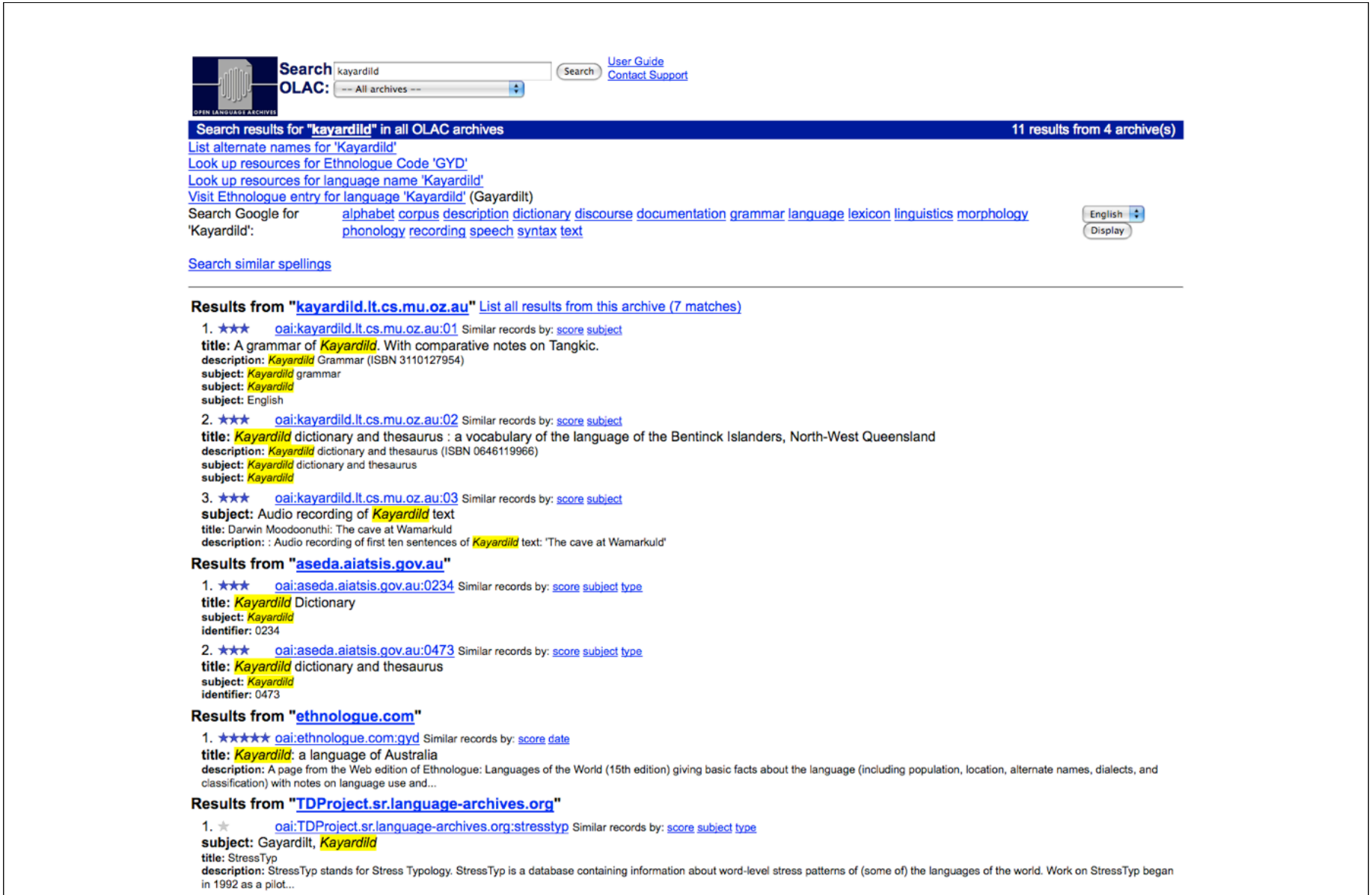
How does OLAC work?

1. Archives submit catalogs in a standard format
2. Central index is updated every 8 hours
3. Accessed via search services



How does OLAC Metadata relate to Dublin Core?

- OLAC has extended Dublin Core metadata by providing additional descriptors tailored to language resources:
  - Subject language
  - Linguistic type
  - Linguistic field
  - Discourse type
  - Role



Searching for “kayardild” using the OLAC service at LDC

What is the current coverage of OLAC?

	All archives	**Excluding *Ethnologue***
items	36,161	28,892
online items	21,579 (60%)	14,310 (50%)
ISO 639-3 coverage	7,334	3556

OLAC Coverage in Relation to Language Size

Population range	*Languages*	*In OLAC*		*Items*
10,000,000 or more	83	82	99%	3,341
1,000,000 to 9,999,999	264	223	84%	1,431
100,000 to 999,999	892	575	64%	2,607
1,000 to 99,999	3,746	1,797	48%	9,012
100 to 999	1,071	392	37%	2,305
1 to 99	548	271	49%	832
Unknown	308	86	28%	307
All living languages	*6,912*	*3,426*	*50%*	*19,835*
Extinct languages	602	130	22%	315

Online Resources in Relation to Language Size

Population range	*Languages*	*In OLAC*		*Items*
10,000,000 or more	83	75	90%	426
1,000,000 to 9,999,999	264	174	66%	312
100,000 to 999,999	892	342	38%	501
1,000 to 99,999	3,746	929	25%	1,167
100 to 999	1,071	140	13%	194
1 to 99	548	95	17%	147
Unknown	308	27	9%	33
All living languages	*6,912*	*1,782*	*26%*	*2,780*
Extinct languages	602	34	6%	53

Current Issues

- Three shortcomings:
  - metadata quality
  - participation
  - digitisation projects
- Broader Challenges
  - searching for language resources on the web
  - library catalogues and the deep web
  - discovering OLAC

NSF Project

- improve access to resources in language archives:
  - All OLAC repositories should have up-to-date catalogs that contain metadata conforming to best practice.
  - All major language archives should be participating in OLAC.
  - All OLAC repositories should conform to current best practices for the long-term curation of their holdings.
- improve access to language resources on the web:
  - Low-density language materials identified by linguistic web mining should be reliably categorized with OLAC vocabularies.
  - Language resources held in libraries and digital repositories should be indexed in OLAC through services that crosswalk and enrich existing catalog records.
  - Web search engines should index all OLAC records, so that users who discover language resources using a conventional web search quickly find OLAC records and are drawn to the OLAC site for more precise searching.

Conclusion

- OLAC has a functioning infrastructure that allows our community to index and discover endangered language documentation
- We need to encourage every institution with endangered language resources to participate so that the catalog can be complete
- We could enhance standards for resource description to support metrics for summarising the extent of documentation for a language

OLAC Website: Documents

OLAC: Open Language Archives Community

HOME | DOCUMENTS | ABOUT | ARCHIVES
NEWS | ORGANIZATION | TOOLS | SERVICES

SEARCH THIS SITE:

Participating Archives

OLAC has 37 participating archives. This page provides a summary; for full details on any archive follow the link on the right.

- Register an archive
- Machine readable list of registered archives
- Metrics report on all archives
- Comparative archive metrics

A Digital Archive of Research Papers in Computational Linguistics Philadelphia, USA	MORE DETAILS	METRICS	N/A	✓
Aboriginal Studies Electronic Data Archive (ASEDA) Australian Institute of Aboriginal and Torres Strait Islander Studies, Canberra, Australia	MORE DETAILS	METRICS	SAMPLE RECORD	✓
Academia Sinica Balanced Corpus of Modern Chinese Academia Sinica, Taipei, Taiwan	MORE DETAILS	METRICS	SAMPLE RECORD	✓
Academia Sinica Formosan Language Archive Academia Sinica, Taipei, Taiwan	MORE DETAILS	METRICS	SAMPLE RECORD	✓
Academia Sinica Tagged Corpus of Early Mandarin Chinese Academia Sinica, Taipei, Taiwan	MORE DETAILS	METRICS	SAMPLE RECORD	✓
Alaska Native Language Center Archive Alaska Native Language Center, Fairbanks, Alaska, USA	MORE DETAILS	METRICS	N/A	✗ 2008-02-08
Archive of the Indigenous Languages of Latin America University of Texas at Austin, Austin, USA	MORE DETAILS	METRICS	SAMPLE RECORD	✓
ATILF Resources ATILF, Nancy, France	MORE DETAILS	METRICS	SAMPLE RECORD	✓
Audio Archive of Linguistic Fieldwork University of California, Berkeley Language Center, Berkeley, CA, USA	MORE DETAILS	METRICS	SAMPLE RECORD	✓
Boiste OLAC, Philippe KREBS, Paris, FRANCE	MORE DETAILS	METRICS	SAMPLE RECORD	✓
Centre de Ressources pour la Description de l'Oral (CRDO) CNRS / Centre de Ressources pour la Description de l'Oral (CRDO), Paris, France	MORE DETAILS	METRICS	SAMPLE RECORD	✗ 2008-09-12
CHILDES Data repository Carnegie Mellon University, Pittsburgh, USA	MORE DETAILS	METRICS	N/A	✓
Comparative Corpus of Spoken Portuguese IEL Unicamp, Campinas - SP - Brasil	MORE DETAILS	METRICS	SAMPLE RECORD	✓

OLAC Website: Participating Archives

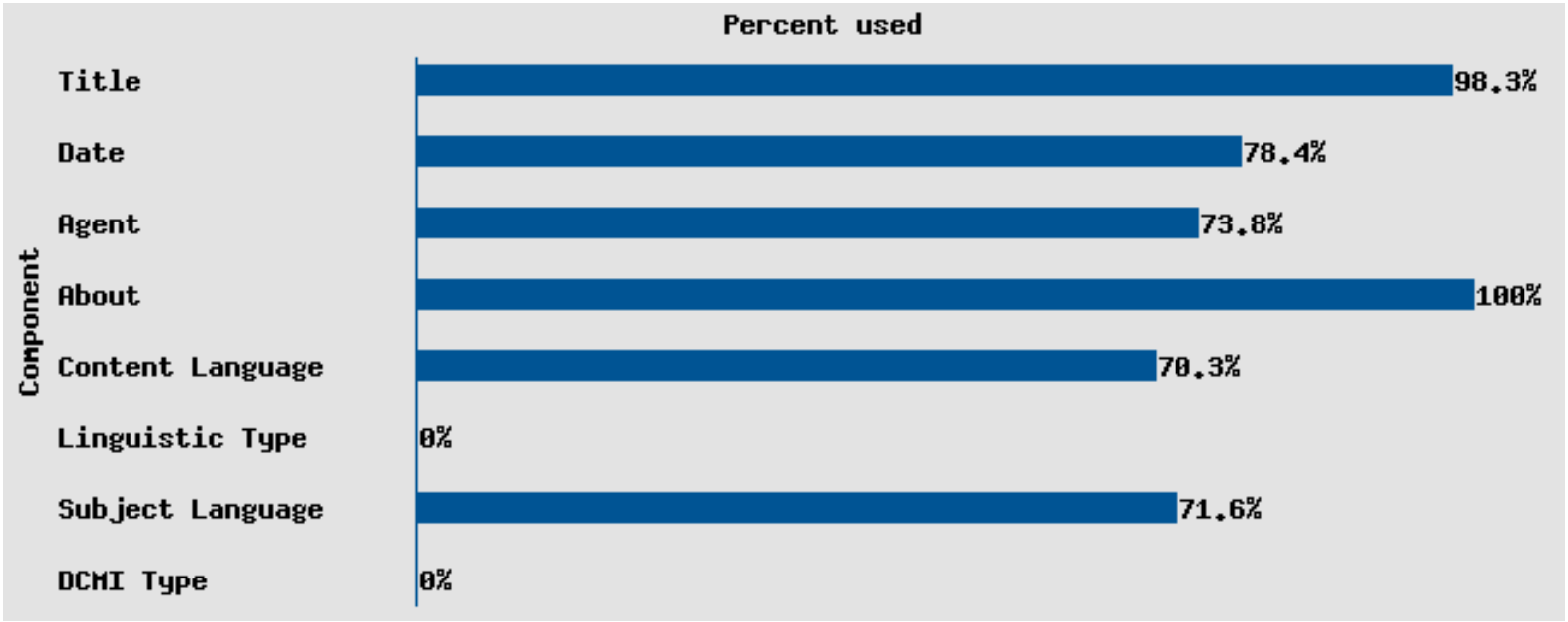
Archive Details

Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

Size:	5702
Repository Name:	Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)
Institution:	A consortium made up of the University of Melbourne, University of Sydney, Australian National University and the University of New Engl
ArchiveURL:	http://paradisec.org.au
Curator:	Nick Thieberger
Location:	Project Manager based at the Department of Linguistics and Applied Linguistics, University of Melbourne, Victoria 3010, Australia
Short Location:	Melbourne, Sydney, Canberra, Armidale, Australia
Synopsis:	PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures) offers a facility for digital conservation and access ethnographic materials from the Pacific region, defined broadly to include Oceania and East and South east Asia. Only some 4908 items are currently digitised. The non-digitised items are part of an assessment of the scope of work that needs to be digitised. They also make material discoverable.
Access:	The current focus of PARADISEC is securing endangered materials. Access to the datastore is by password and is currently only availabl following URL: http://www.paradisec.org.au/repository
Administrator:	nicholas.thieberger@paradisec.org.au
Base URL:	http://azoulay.arts.usyd.edu.au/paradisec/PDSC-DR.php
Repository ID:	paradisec.org.au
OAI Version:	2.0
OLAC MS Version(s):	1.0
Records in Archive:	http://www.language-archives.org/archive_records/paradisec.org.au
Explore:	Visit archive with the Repository Explorer
Last Harvested:	2008-06-25
Reports:	Archive Metrics and Integrity Checks

Name	Value
Number of Resources	5702
Number of Resources Online	0
Distinct Languages	607
Distinct Linguistic Subfields	0
Distinct Linguistic Types	0
Distinct DCMI Types	0
Average Elements Per Record	15
Average Encoding Schemes Per Record	3.9
Average Metadata Quality Score	6
Last Updated	2008-09-17
Known Integrity Problems	0
Overall Rating	★★★

Archive Metrics: PARADISEC


Percent used
Component
Title
98.3%
Date
78.4%
Agent
73.8%
About
100%
Content Language
70.3%
Linguistic Type
0%
Subject Language
71.6%
DCMI Type
0%

```
<olac:olac>
   <dc:contributor xsi:type="olac:role" olac:code="recorder">Thieberger, Nick</dc:contributor>
   <dc:contributor xsi:type="olac:role" olac:code="speaker">Alban, Sailas</dc:contributor>
   <dc:coverage>VU</dc:coverage>
   <dc:coverage xsi:type="dcterms:Box">northlimit=-13.71; southlimit=-20.25; westlimit=166.52; eastlimit=169.89</dc:coverage>
   <dc:date>19950816</dc:date>
   <dc:description>Language as given: South Efate, Bislama</dc:description>
   <dc:description>(Citable as: Thieberger, Nick (recorder), Alban, Sailas (speaker) 1995; DG5-PH0, WAV/MP3
paradisec.org.au/repository/NT1-001 2008-08-18)</dc:description>
   <dc:format>
            Digitised: yes;
            Media: audiocassette;

         </dc:format>
   <dc:identifier>NT1-001</dc:identifier>
   <dc:language xsi:type="olac:language" olac:code="bis"/>
   <dc:language xsi:type="olac:language" olac:code="erk"/>
   <dcterms:hasPart>NT1-001-001A.mp3</dcterms:hasPart>
   <dcterms:hasPart>NT1-001-001A.wav</dcterms:hasPart>
   <dcterms:hasPart>NT1-001-001B.mp3</dcterms:hasPart>
   <dcterms:hasPart>NT1-001-001B.wav</dcterms:hasPart>
   <dcterms:accessRights>standard, as per PDSC Access form</dcterms:accessRights>
   <dc:subject xsi:type="olac:language" olac:code="erk"/>
   <dc:title>DG5-PH0</dc:title>
   <dc:type>primary_text</dc:type>
</olac:olac>
```

XML Format of OLAC Record

Contributor (recorder)	Thieberger, Nick
Contributor (speaker)	Alban, Sailas
Coverage	VU
Coverage (Box)	northlimit=-13.71; southlimit=-20.25; westlimit=166.52; eastlimit=169.89
Date	19950816
Description	Language as given: South Efate, Bislama
Description	(Citable as: Thieberger, Nick (recorder), Alban, Sailas (speaker) 1995; DG5-PH0, WAV/MP3 paradisec.org.au/repository/NT1-001 2008-08-18)
Format	Digitised: yes; Media: audiocassette;
Identifier	NT1-001
Language (ISO639-3)	Bislama [bis]
Language (ISO639-3)	South Efate [erk]
Has Part	NT1-001-001A.mp3
Has Part	NT1-001-001A.wav
Has Part	NT1-001-001B.mp3
Has Part	NT1-001-001B.wav
Access Rights	standard, as per PDSC Access form
Subject (ISO639-3)	South Efate [erk]
Title	DG5-PH0
Type	primary_text

Display Format for OLAC Record

Component	+	-	Comments
Title	1	0	
Date	1	0	
Agent	1	0	
About	1	0	
Depth	1	0	
Content Language	1	0	
Subject Language	1	0	
OLAC Type	0	1	Add a dc:type element that uses the OLAC linguistic-type encoding scheme to identify the type of the resource from a linguistic point of view.
DCMI Type	0	1	Add a dc:type element that uses the DCMIType encoding scheme to identify the generic type of the resource.
Precision	0.67	0.33	For the full score, make use of at least one more encoding scheme in addition to the ones counted explicitly in other components of the score. For instance, • use dcterms:W3CDTF on dc:date (or its refinements) • use dcterms:IMT on dc:format • use dcterms:Box or dcterms:Point or dcterms:TGN on dcterms:spatial
Quality score	7.67		

Archive	Overall Rating	Number of Resources	Number of Resources Online	Distinct Languages	Distinct Linguistic Subfields	Distinct Linguistic Types	Distinct DCMI Types	Average Elements Per Record	Average Encoding Schemes Per Record	Average Metadata Quality Score	Last Updated	Known Integrity Problems
Audio Archive of Linguistic Fieldwork	★★★★★	111	0	80	0	2	2	68.0	16.6	9.9	2008-08-28	0
Ethnologue: Languages of the World	★★★★	7296	7296	7296	0	0	1	11.2	8.2	8.5	2008-07-21	0
UQ Flint Archive	★★★★	1	0	1	0	1	2	22.0	9.0	8.3	2002-12-14	0
Surrey Morphology Group Databases	★★★★	2	0	114	0	0	1	82.5	73.5	7.3	2002-12-14	0
Centre de Ressources pour la Description de l'Oral (CRDO)	★★★	1948	1948	64	10	3	3	24.4	16.5	8.9	2008-08-29	36
Academia Sinica Balanced Corpus of Modern Chinese	★★★	1	1	2	0	0	1	64.0	17.0	8.3	2002-12-14	1
Academia Sinica Formosan Language Archive	★★★	1	1	3	0	0	3	51.0	16.0	8.3	2002-12-14	2
Academia Sinica Tagged Corpus of Early Mandarin Chinese	★★★	1	1	2	0	0	1	63.0	22.0	8.3	2002-12-14	1
DOBES catalogue - OLAC 1.0	★★★	11	11	15	0	1	3	16.5	8.4	8.3	2006-03-02	11
CHILDES Data repository	★★★	239	238	44	1	1	0	16.4	8.1	8.0	2004-08-09	200
U Bielefeld Language Archive	★★★	13	7	5	3	2	0	15.8	5.2	7.1	2003-10-31	2
Tibetan and Himalayan Digital Library	★★★	1	1	5	0	0	0	13.0	5.0	6.8	2002-12-14	0
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	★★★	5702	0	607	0	0	0	15.0	3.9	6.0	2008-09-17	0
Magoria Books' Carib and Romani Archive	★★★	2	1	1	2	0	0	10.5	3.0	5.6	2008-03-12	0
Cornell Language Acquisition Laboratory, CLAL	★★★	5	3	3	0	1	0	20.0	2.8	5.3	2006-05-27	0
Oxford Text Archive	★★★	1264	1264	4	0	0	0	20.0	2.0	5.2	2003-10-23	0
The LDC Corpus Catalog	★★	417	0	131	1	2	0	20.0	3.7	6.8	2008-09-16	112
The Typological Database Project	★★	8	2	353	0	0	2	122.8	94.8	6.7	2002-12-14	1
The Natural Language Software Registry	★★	69	69	16	0	0	0	15.0	1.5	5.6	2002-12-14	8
SIL Language and Culture Archives	★★	7875	380	1750	23	3	0	8.8	2.2	5.3	2005-06-08	1
ODIN - The Online Database of Interlinear Text	★★	717	717	711	0	0	0	7.0	1.0	5.0	2007-06-27	0

Comparative Archive Metrics