From community-specific XML markup to Linked Data and an abstract application profile: A possible path for the future of OLAC

Gary F. Simons
SIL International

Co-coordinator,
Open Language Archives Community

OLAC / DELAMAN Workshop, Austin, TX, 11 April 2016

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 237,000 items from 58 archives

Is it time for an update?

- When OLAC was established, we followed the best practice of the time
  - Creating a community-specific XML metadata format
- In the intervening years, new best practices have emerged
  - Representing metadata as Linked Data
  - Expressing a community standards as a Metadata Application Profile
- Is it time for a significant update?

Overview

1. How OLAC metadata works now
2. Enter Linked Data
   - What is it and why might we want it?
3. Expressing OLAC metadata as Linked Data
   - Progress to-date and some open issues
4. Looking to the future
   - Or, Are we ready for OLAC 2.0?

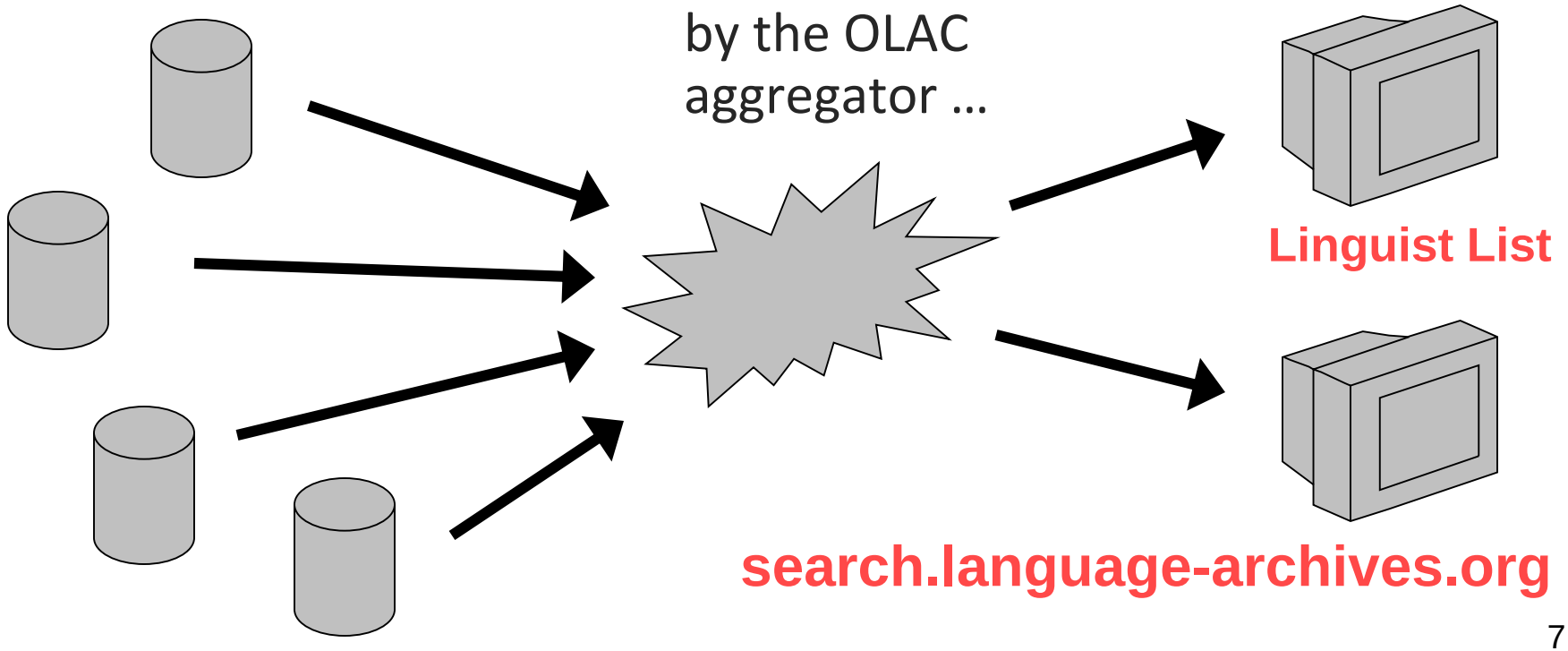
1. How OLAC metadata works now

Standards for interoperation

- The community has defined standards for the encoding and exchange of language resource metadata to permit discovery and sharing. 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

OLAC infrastructure

- The 58 archives publish catalogs in a standard XML form …
- to be harvested by the OLAC aggregator …
- which supplies information to search services.



OLAC Language Resource Catalog at search.language-archives.org

Search for language resources go

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

Results: *« First • Previous • Next • Last »*

Showing hits **1 - 6** out of **6** Show 50

PHOIBLE Online phonemic inventories for Kabuverdianu
n.a. 2014. Max Planck Institute for Evolutionary Anthropology.

LAPSyD Online page for Cape Verde Creole, Santiago dialect
Maddieson, Ian. 2012. www.lapsyd.ddl.ish-lyon.cnrs.fr.

Glottolog 2.4 Resources for Kabuverdianu
n.a. 2014. Max Planck Institute for Evolutionary Anthropology.

APiCS Online Resources for Cape Verdean Creole of Santiago
n.a. 2013. Max Planck Institute for Evolutionary Anthropology.

APiCS Online Resources for Cape Verdean Creole of Brava
n.a. 2013. Max Planck Institute for Evolutionary Anthropology.

APiCS Online Resources for Cape Verdean Creole of São Vicente
n.a. 2013. Max Planck Institute for Evolutionary Anthropology.

« First • Previous • Next • Last »

Currently Used Filters
- ✓ Subject language: Kabuverdianu
- ✓ Linguistic type: Language description

Sort Results By:

Possible Sorts: all

Narrow Results By:

Archive browse

Linguistic field browse
- Phonetics 1
- Phonology 1
- Typology 1

DCMI type browse
- Text 4
- Dataset 2

Contributor browse

Publisher browse

OLAC Language Resource Catalog

Search for language resources

Navigating the Catalog
- Catalog Home
- Back to Search Results
- 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

LAPSyD Online page for Cape Verde Creole, Santiago dialect

Title:	LAPSyD Online page for Cape Verde Creole, Santiago dialect
Link to the object:	http://www.lapsyd.ddl.ish-lyon.cnrs.fr/lapsyd/index.php?data=view&code=692
Online:	Yes
Archive:	LAPSyD (see archive description)
Contributor:	Maddieson, Ian (author)
Date:	2012-05-17
Publisher:	www.lapsyd.ddl.ish-lyon.cnrs.fr
Description:	This resource contains information about phonological inventories, tones, stress and syllabic structures
Content language:	English
Subject language:	Kabuverdianu
Language	Creoles and pidgins, Portuguese-based

Find Related Information:

- Archive: LAPSyD
- Online: Yes
- Subject language: Kabuverdianu
- Language family: Creoles and pi
- Language family: Creoles and pi Portuguese-based
- Geographic region: Africa
- Linguistic type: Language descri
- Linguistic field: Phonology
- Linguistic field: Typology
- DCMI type: Dataset
- Format: text/html
- Content language: English
- Date: 2000 and later

OLAC metadata standard

- OLAC starts with the 15 basic Dublin Core (*dc:*) elements:
  - Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type
- Use *dcterms:* namespace to add the refined elements of Qualified Dublin Core
  - abstract, accessRights, alternative, audience, available, bibliographic-Citation, conformsTo, created, dateAccepted, dateCopyrighted, dateSubmitted, educationLevel, extent, hasFormat, hasPart, hasVersion, instructionalMethod, isFormatOf, isPartOf, isRefer-encedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, license, mediator, medium, modified, provenance, references, replaces, requires, rightsHolder, spatial, tableOfContents, temporal, valid

Adding precision for values of metadata elements

- Use *xsi:type* attribute to indicate that the element value is from an encoding scheme recognized in *dcterms*
  - *E.g.,* Box, DCMIType, IMT, ISO3166, LCSH, Period, Point, TGN, URI, W3CDTF
- Use *olac:code* attribute when the value is from a vocabulary defined in an OLAC recommendation:
  - Code for Discourse Types [olac:discourse-type]
  - Code for Identifying Languages [olac:language] (ISO 639-3)
  - Code for Linguistic Field [olac:linguistic-field]
  - Code for Linguistic Data Types [olac:linguistic-type]
  - Code for Participant Roles [olac:role]

Metadata as published

```
<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&amp;amp;code=692</dc:identifier>
    <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>
```

OAI Protocol for Metadata Harvesting

- There are six verbs:
  - GetRecord, Identify, ListIdentifiers, ListMetadataFormats, ListRecords, ListSets
- Harvesting requests are expressed as URLs:
  - *baseURL*?verb=*value&parameters*
- For instance:
  - http://elar.soas.ac.uk/olac?verb=Identify
  - http://elar.soas.ac.uk/olac?verb=GetRecord&identifier=elar.soas.ac.uk
- Response is an XML document

2. Enter Linked Data

Best practice in 2000

- When OLAC began, purpose-specific XML markup was the best common practice for metadata interchange
- Pros
  - Data integrity is automatically verified by a parser
  - Interoperation is possible across all data that conforms to the markup schema
- Cons
  - Interoperation with data that uses a different markup schema is not defined
  - Enriching the markup violates the integrity

An emerging practice

- At the same time, the W3C's Semantic Web activity was pushing ahead with an alternative to purpose-specific XML for information interchange:
  - Developed RDF (Resource Description Framework) as a means to represent information in terms of semantics
  - Each concept represented by a URI
  - Define the formal properties of concepts with RDF Schema and OWL (Web Ontology Language)
  - Information has an abstract and simple graph structure; multiple syntaxes for serializing it

Representing information in RDF

- All information is represented as a set of statements.
- Statement = < subject, predicate, object >
  - The subject is a URI representing a *resource*.
  - The predicate is a URI representing a *property*.
  - The object may be another resource or it may be a *literal* value.
- A set of statements forms a directed graph.
  - *Basis for interoperation: A*ny two RDF graphs can be combined; they will merge if they have any resource URIs in common.

A new best practice

- The Semantic Web, after much initial hype, has receded into the background, but …
- Essentially the same idea, rebranded as Linked Data, is now getting significant traction as
  - a means for linking independently-developed, purpose-specific datasets into an interoperating universal Web of Data
- Even some linguists have gotten on board, *e.g.:*
  - Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.) (2012). *Linked Data in linguistics: Representing and connecting language data and language metadata.* Heidelberg: Springer.

Uptake by the library community

- Librarians are recognizing that Linked Data represents an opportunity for libraries to integrate their information resources with the wider web
  - Dublin Core is now based on an abstract model (implemented in RDF Schema) which is referenced when defining Metadata Application Profiles
    - A MAP identifies all the pre-existing properties and values (as URIs) that will be used in the metadata application
    - Implementers can use whatever serialization they like
  - The BIBFRAME initiative at the Library of Congress is building on the Linked Data model to develop a replacement for the MARC standard

References

- Baker, T. (2012). Libraries, languages of description, and linked data: a Dublin Core perspective. *Library Hi Tech,* 30(1), 116-133.
- Byrne, G., & Goddard, L. (2010). The strongest link: Libraries and linked data. *D-Lib magazine*, *16*(11), 5.
- Miller, E., Ogbuji, U., Mueller, V., & MacDougall, K. (2012). *Bibliographic Framework as a Web of Data: Linked Data model and supporting services.* Washington, DC: Library of Congress. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf
- DC AM: http://dublincore.org/documents/abstract-model/
- DC MAP: http://dublincore.org/documents/profile-guidelines/
- BIBFRAME: http://www.loc.gov/bibframe/

What would it look like for OLAC to adopt Linked Data?

3. Expressing OLAC metadata as Linked Data

The rules of Linked Data

- The four rules as articulated by Tim Berners-Lee:

1. Use URIs to name (identify) things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information using open standards like RDF
4. Include links to other URIs so that they can discover more things

- http://www.w3.org/DesignIssues/LinkedData.html

How does OLAC stack up?

- Each participating archive and each language resource has always been identified by an HTTP URI
  - http://www.language-archives.org/archive/www.lapsyd.ddl.ish-lyon.cnrs.fr
  - http://www.language-archives.org/item/oai:www.lapsyd.ddl.ish-lyon.cnrs.fr:src692
    - But looking them up needs to yield a description expressed in RDF as well
- All of the metadata elements and vocabularies we use from DC have URIs and descriptions that comply
  - But the vocabularies defined by OLAC need them, too
  - First step: Turn the OLAC vocabularies into Linked Data resources

The Language extension

- The olac:language extension uses codes from ISO 639, parts 1, 2, and 3, *e.g.,* using codes for German:
  - <dc:language xsi:type="olac:language" olac:code="de"/>
  - <dc:language xsi:type="olac:language" olac:code="deu"/>
- For parts 1 and 2, the Library of Congress Linked Data Service already provides the solution at id.loc.gov
  - Part 1, “de” = <http://id.loc.gov/vocabulary/iso639-1/de>
  - Part 2, “deu” = <http://id.loc.gov/vocabulary/iso639-2/deu>
- For part 3, SIL (the RA for the standard) is working with LC to add ISO 639-3 to their Linked Data Service

The other four extensions

- olac:discourse-type, olac:linguistic-field, olac:linguistic-type, and olac:role are vocabularies defined by OLAC
  - <dc:type xsi:type="olac:linguistic-type" olac:code="lexicon"/>
  - OLAC must provide the Linked Data Service for these
- Solution
  - Convert each vocabulary document into an RDF document using the Simple Knowledge Organization System (SKOS)
  - Use a hash namespace to reference the terms, *e.g.*, <http://www.language-archives.org/vocabulary/type#lexicon>
- *N.B.* The RDF samples which follow are in N3 notation

A controlled vocabulary as a SKOS Concept Scheme

<http://www.language-archives.org/vocabulary/type> a skos:ConceptScheme ;

dc:title "OLAC Linguistic Data Type Vocabulary" ;

dc:description "This document specifies the codes, or controlled vocabulary, for the Linguistic Data Type extension of the DCMI Type element. These codes describe the content of a resource from the standpoint of recognized structural types of linguistic information.";

dc:publisher "Open Language Archives Community" ;

dcterms:issued "2006-04-06" ;

rdfs:isDefinedBy <http://www.language-archives.org/REC/type.html>, <http://www.language-archives.org/vocabulary/type.rdf> ;

skos:hasTopConcept <http://www.language-archives.org/vocabulary/type#language_description>, <http://www.language-archives.org/vocabulary/type#lexicon>, <http://www.language-archives.org/vocabulary/type#primary_text> .

A vocabulary term as a SKOS Concept

<http://www.language-archives.org/vocabulary/type#lexicon> a skos:Concept ;

skos:inScheme <http://www.language-archives.org/vocabulary/type> ;

skos:prefLabel "Lexicon" ;

skos:definition "The resource includes a systematic listing of lexical items.";

skos:example "Examples include word lists (including comparative word lists), thesauri, wordnets, framenets, and dictionaries, including specialized dictionaries such as bilingual and multilingual dictionaries, dictionaries of terminology, and dictionaries of proper names. Non-word-based examples include phrasal lexicons and lexicons of intonational tunes.";

skos:scopeNote "Lexicon may be used to describe any resource which includes a systematic listing of lexical items. Each lexical item may, but need not, be accompanied by a definition, a description of the referent (in the case of proper names), or an indication of the item's semantic relationship to other lexical items.".

Expressing an OLAC metadata record in RDF

- In addition to the usual namespace prefixes for dc:, dcterms: rdf:, rdfs:, the example will use:

@prefix olac: <http://www.language-archives.org/OLAC/1.1/> .
@prefix olac-archive: <http://www.language-archives.org/archive/> .
@prefix olac-item: <http://www.language-archives.org/item/>.

@prefix olac-field: <http://www.language-archives.org/vocabulary/field#> .
@prefix olac-role: <http://www.language-archives.org/vocabulary/role#> .
@prefix olac-type: <http://www.language-archives.org/vocabulary/type#> .

Expressing an OLAC metadata record in RDF (2)

- *The item is curated by the named archive:*

```
olac-item:oai:www.lapsyd.ddl.ish-lyon.cnrs.fr:src692
    a rdfs:Resource ;
    olac:curatedBy olac-archive:www.lapsyd.ddl.ish-lyon.cnrs.fr ;
```

- *Basic DC elements with literal values*

```
    dc:title "LAPSyD Online page for Cape Verde Creole, Santiago
      dialect" ;
    dc:description "This resource contains information about
      phonological inventories, tones, stress and syllabic structures" ;
```

- *Literal value in a standard encoding scheme*

```
    dcterms:modified "2012-05-17"^^dcterms:W3CDTF .
```

Expressing an OLAC metadata record in RDF (3)

- *Properties where the value is a URL representing a concept:*

olac-item:oai:www.lapsyd.ddl.ish-lyon.cnrs.fr:src692
dc:publisher <http://www.lapsyd.ddl.ish-lyon.cnrs.fr> ;
dcterms:license <http://creativecommons.org/licenses/by-nc-nd/3.0/> ;
dc:type <http://purl.org/dc/dcmitype/Dataset> ;
dc:format <http://purl.org/NET/mediatypes/text/html> ;
dc:language <http://id.loc.gov/vocabulary/iso639-3/eng> ;
dc:subject <http://id.loc.gov/vocabulary/iso639-3/kea> ;
dc:subject olac-field:phonology, olac-field:typology ;
dc:type olac-type:language_description .

A problem remains: Representing contributors

- *The contributor statement:*

  <dc:contributor xsi:type="olac:role" olac:code="author">
  Maddieson, Ian</dc:contributor>

- *Is translated into:*

  olac-role:author "Maddieson, Ian" ;

- But this does not follow the rules of Linked Data
  - Ian Maddieson is a "thing" in the world that should be identified by means of a URL. We need a standard.
  - Perhaps Linguist List Directory of Linguists. But not "Cool URIs"

    http://linguistlist.org/people/personal/get-personal-page2.cfm?PersonID=695

Some possible URLs for our sample contributor

- These URLs do not follow all 4 rules of Linked Data:
  - https://en.wikipedia.org/wiki/Ian_Maddieson
  - http://linguistlist.org/people/personal/get-personal-page2.cfm?PersonID=695
  - http://www.unm.edu/~ianm/index.html
  - http://linguistics.berkeley.edu/person/23
- But the following do!
  - http://dbpedia.org/resource/Ian_Maddieson
  - http://id.loc.gov/authorities/names/n84089547
  - http://orcid.org/0000-0002-0775-0555

A current gap in our search infrastructure

- Current *OLAC Metadata Usage Guidelines* :
  - Identify a contributor "by means of a name in a form that is ready for sorting within an alphabetical index."
- But we have no way to enforce this guideline or to ensure that each Contributor element names only one contributor
- Thus, in spite of providing interoperable search over 14 facets that have uniform metadata values across the community of archives, contributor is not one of them.
- Are we ready to tighten our metadata guidelines and practices in order to support the identification of contributors in Linked Data and in faceted search?

4. Looking to the future

Linked Data in the OLAC infrastructure

- Today we are minimally participating in the Web of Data
  - Each archive description and archived item has a cool URI that returns an RDF description (that is auto-generated from the OLAC 1.1 metadata record)
  - There is a gzipped nightly dump of all the RDF data
  - To join the cloud of Linguistic Linked Open Data, this dataset needs to be registered at the DataHub of the Open Knowledge Foundation
- We could do more
  - Run our own RDF database and mount a SPARQL endpoint to give semantic search over the whole OLAC catalog

Long-term vision

- What is our long-term vision for OLAC?
  - Continue to operate a community-specific technical infrastructure? Or
  - Merge into the mainstream of the digital library infrastructure?
- Could we be so successful at integrating with the mainstream that they provide the basic infrastructure?
  - Using ISO 639-3 becomes the norm when needed
  - Identifying language resource type becomes the norm
  - OLAC could pivot from infrastructure to advocacy

Are we ready for OLAC 2.0?

- OLAC standard has not changed appreciably since version 1.0 was adopted in 2003 (version 1.1 in 2008)
  - It may be time for a version 2.0 update to move OLAC from having a community-specific XML metadata format to being an RDF-based Metadata Application Profile
- Key indicators
  - Do our participating archives want to take advantage of the RDF's expressive power of to create richer metadata?
  - Would key holdouts be ready to participate?
  - Are there partners in the library community who are anxious to help us integrate with the mainstream?

Conclusion

- Given the core values of the OLAC Process that decisions are made by consensus and that the greatest voice is given to those who are implementing the standards
  - Moving to OLAC 2.0 would be a huge effort requiring archives around the world to both agree and re-implement
- But the time is ripe for OLAC to consider a major update to its standards and infrastructure, especially considering
  - The potential of language resource information taking its place within the interoperating global Web of Data
  - The long-term sustainability that could result from entering into the mainstream of library practices