Expressing language resource metadata as Linked Data: A potential agenda for the Open Language Archives Community

Gary F. Simons

SIL International

Co-coordinator,
Open Language Archives Community

Workshop on Linguistic Linked Open Data (LLOD)

LSA Summer Institute, Chicago, 25-26 July 2015

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 library of >225,000 items from 57 archives

Largest participants by number of items

The Language Archive's IMDI portal	Netherlands	94,755
SIL Language and Culture Archives	USA	30,177
California Language Archive	USA	13,965
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	Australia	9,781
COllections de COrpus Oraux Numeriques (COCOON)	France	8,850
Graduate Institute of Applied Linguistics Library	USA	8,176
Glottolog 2.5	Germany	7,817
Ethnologue: Languages of the World	USA	7,480
World Atlas of Language Structures (WALS) Online RefDB	Germany	7,155
The Rosetta Project: A Library of Human Language	USA	6,571
A Digital Archive of Research Papers in Comp. Linguistics	USA	3,280
World Atlas of Language Structures (WALS) Online	Germany	2,622
The LINGUIST List Language Resources	USA	2,563
Living Archive of Aboriginal Languages	Australia	2,462

Enter Linked Data

- When OLAC began, purpose-specific XML markup was the best common practice for metadata interchange
- In the meantime, Linked Data has emerged as a means for linking purpose-specific datasets into an interoperating universal Web of Data
    - Linked Data is picking up momentum in the metadata community
    - *E.g.*, in BIBFRAME the Library of Congress is working on a Linked Data successor to the MARC format
- What would it look like for OLAC to adopt Linked Data?

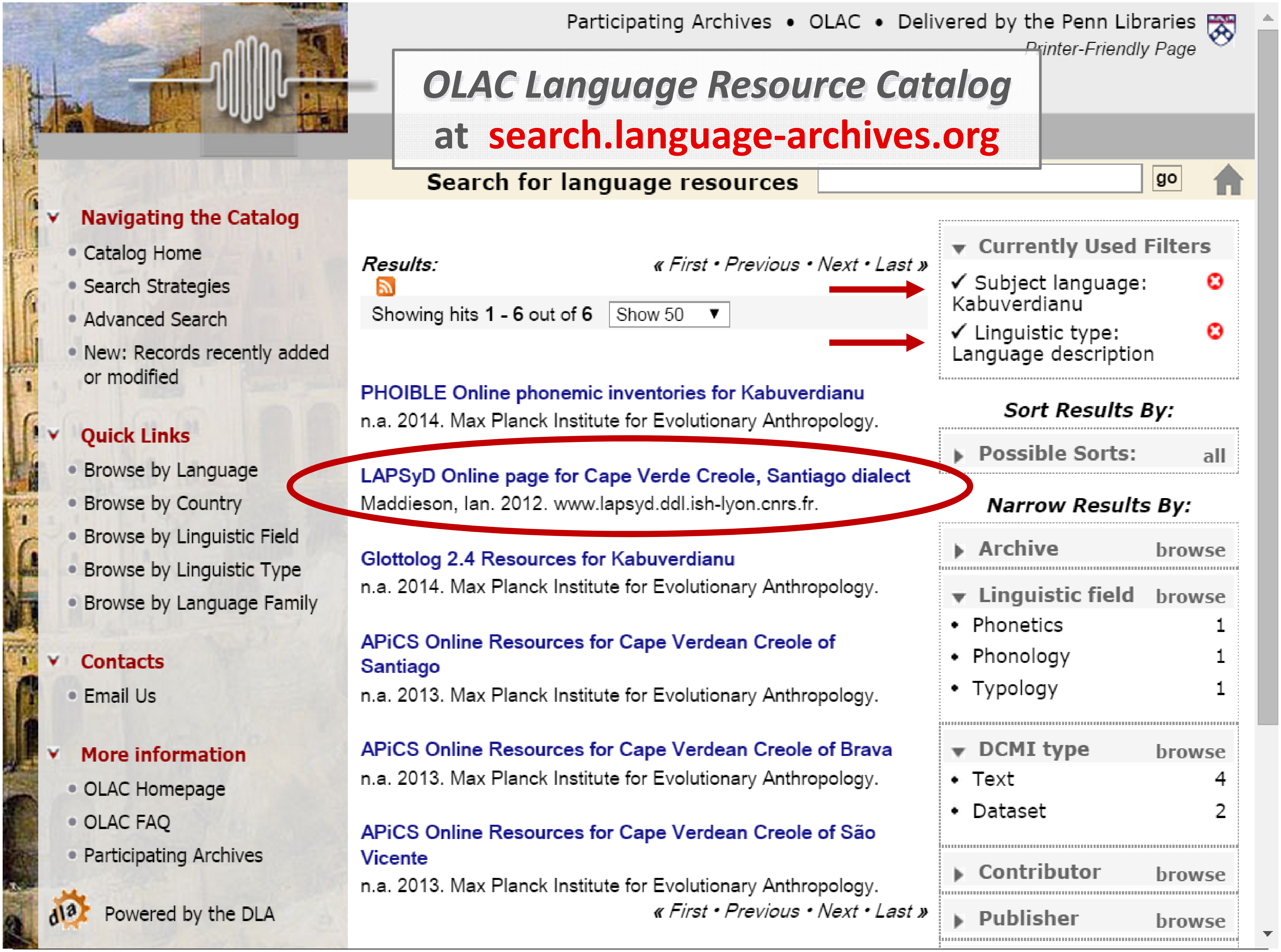
Overview

1. Introduce the OLAC metadata standard
2. Identify known problems with the current OLAC metadata from the standpoint of representing it as Linked Data
    - Then propose solutions for addressing them
3. Discuss possible strategies for incorporating Linked Data into the OLAC infrastructure

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 Language Resource Catalog at search.language-archives.org



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

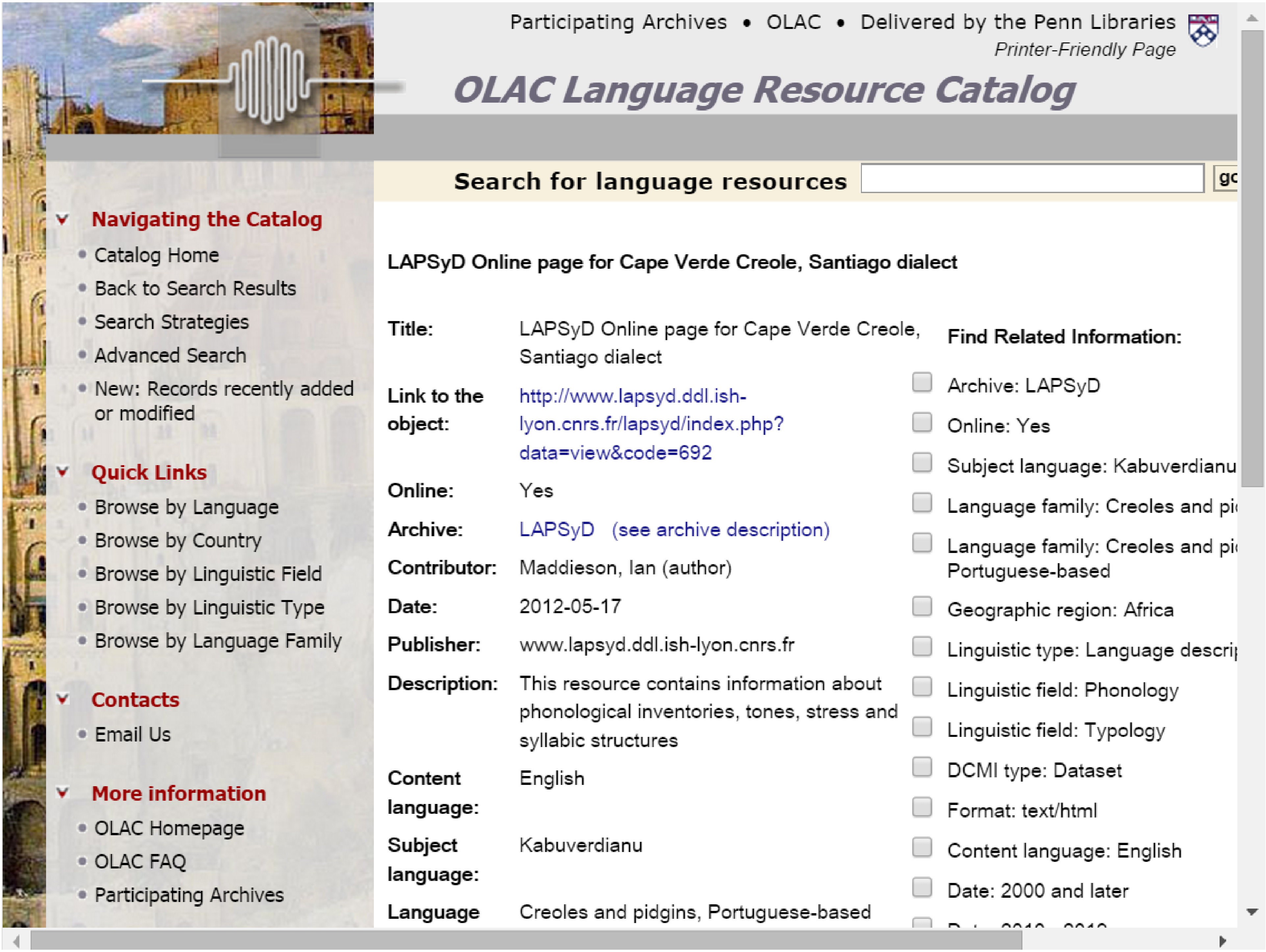
Phonetics	1
Phonology	1
Typology	1

DCMI type browse

Text	4
Dataset	2

Contributor browse

Publisher browse


Participating Archives • OLAC • Delivered by the Penn Libraries
Printer-Friendly Page
OLAC Language Resource Catalog
Search for language resources
go
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

Metadata as published

```xml
<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;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>
```

OLAC metadata standard

- ► OLAC starts with the 15 basic Dublin Core elements:
  - ▪ Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type
- ► Uses *dcterms:* namespace to add the refined elements of Qualified Dublin Core
- ► Uses *xsi:type* attribute to add precision with element values from encoding schemes recognized in *dcterms*
- ► And *olac:code* for extensions specific to our community:
  - ▪ Language Identification (ISO 639-3), Linguistic Data Type, Linguistic Field, Participant Role, Discourse Type

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 is already 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 does not yet yield a description expressed in RDF
- All of the metadata elements and vocabularies we use from DC have URIs and descriptions that comply
    - But the OLAC extensions do not
    - 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 the 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 and use a hash namespace to reference the terms
  - *E.g.,* “lexicon” becomes <http://www.language-archives.org/vocabulary/type#lexicon>

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 case: 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

Linked Data and the OLAC infrastructure

- OLAC standard has not changed appreciably since version 1.0 was adopted 12 years old
  - It may be time for a version 2.0 update to bring OLAC into line with Linked Data and other current best practices
- Advantages
  - Increased interoperation as strings become Cool URIs
  - Archives could create even richer metadata by augmenting it with any RDF vocabulary
  - We would move from being a self-contained community to an interoperating part of the global Web of Data

A less ambitious approach

- ► Developing OLAC 2.0 would have a substantial cost of requiring participating archives to re-implement
- ► A less costly approach would be:
  - Maintain the current metadata standard
  - Operate an automated service at the aggregator which transforms the entire catalog to Linked Data and provides it in a number of formats
  - Add content negotiation to return HTML vs RDF
  - Offer a SPARQL endpoint as a new service to provide semantic search

A compromise solution

- A middle ground would be a hybrid approach
  - The harvester would support both OLAC 1.1 and 2.0
  - Back translate new 2.0 repositories into 1.1 format so that all existing services continue to work
  - Forward translate 1.1 repositories into 2.0 format and begin to operate an RDF aggregator that can capture all the added richness of 2.0 metadata
  - Gradually develop new services that take full advantage of the Linked Data paradigm

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 as we consider the potential of language resource information taking its place within the interoperating global Web of Data