Expressing Language Resource Metadata as Linked Data

The Case of the Open Language Archives Community

Simons, Gary F.; Bird, Steven

Published in:
Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences

DOI:
10.7551/mitpress/10990.001.0001

Published: 01/01/2019

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Simons, G. F., & Bird, S. (2019). Expressing Language Resource Metadata as Linked Data: The Case of the Open Language Archives Community. In A. Pareja-Lora, M. Blume, B. C. Lust, & C. Chiarcos (Eds.), *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences* (1 ed., pp. 117-130). The MIT Press. https://doi.org/10.7551/mitpress/10990.001.0001

7 Expressing Language Resource Metadata as Linked Data: The Case of the Open Language Archives Community

Gary F. Simons and Steven Bird

Introduction

The Open Language Archives Community (OLAC) is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources.[1] The library is virtual because OLAC does not hold any of the resources itself, rather it aggregates a union catalog of all the resources held by the participating institutions. A major achievement of the community has been to develop standards for expressing and exchanging the metadata records that describe the holdings of an archive. Since its founding in 2000, the OLAC virtual library has grown to include over 300,000 language resources housed in 60 participating archives.[2] Because all the participating archives describe their resources using a common format and shared vocabularies, OLAC is able to promote discovery of these resources through faceted search across the collections of all 60 archives.[3]

The OLAC metadata standard prescribes an interchange format that uses a community-specific XML markup schema. In the meantime, Linked Data has emerged as a common data representation that allows information from disparate communities to be linked into an interoperating universal Web of Data. This chapter explores the application of Linked Data to the problem of describing language resources in the context of OLAC. The first section sets the baseline by describing the OLAC metadata standard. The next section discusses Linked Data and how the existing OLAC standards and infrastructure measure up against the rules of Linked Data. The third section then describes how we have implemented the conversion of OLAC metadata records into resources within the Linked Data framework. Finally, the fourth section considers the impact on the OLAC infrastructure, including both changes that have already been implemented in order to bring the resources of OLAC's participating archives into the Linguistic Linked Open Data (LLOD) cloud (Chiarcos et al. 2013), as well as the potential of embracing Linked Data as the basis for a revised OLAC metadata standard.

The OLAC Metadata Standard

OLAC has created an infrastructure for the discovery and sharing of language resources (Simons and Bird 2003, 2008d). The infrastructure is built on three foundational standards: *OLAC Process* (Simons and Bird 2006), which defines the governance and standards process; *OLAC Metadata* (Simons and Bird 2008a), which defines the XML format used for the exchange of metadata records; and *OLAC Repositories* (Simons and Bird 2008b), which defines the requirements for implementing a metadata repository that can be harvested by an aggregator using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[4]

The OLAC metadata scheme (Bird and Simons 2004) is based on Dublin Core, which is a standard originally developed within the library community to address the cataloging of web resources. At its core, Dublin Core has 15 basic elements for describing a resource: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type. To support greater precision in resource descriptions, this basic set has been developed into an enriched set of metadata terms (DCMI 2012) that can be used to further qualify these elements. The qualifications are of two kinds—refinements that provide more specific meanings for the elements themselves and encoding schemes (including controlled vocabularies) that provide for standardized ways of representing the values of the elements.

The OLAC metadata format is defined by a community-specific XML schema that follows the published guidelines for representing qualified Dublin Core in XML (Powell and Johnston 2003). In addition to supporting the encoding schemes defined by the Dublin Core Metadata Initiative, those guidelines provide a mechanism for further extension via the incorporation of application-specific encoding schemes. The OLAC community has used its standards process to define five metadata extensions (Bird and Simons 2003, Simons and Bird 2008c) that use controlled vocabularies specific to language resources:

- Subject Language, for identifying with precision (using a code from the ISO 639 standard)[5,6] which language a resource is about
- Linguistic Type, for classifying the structure of a resource as primary text, lexicon, or language description
- Linguistic Field, for specifying a relevant subfield of linguistics
- Discourse Type, for indicating the linguistic genre of the material.
- Role, for documenting the parts played by specific individuals and institutions in creating a resource

The following is a sample metadata record in the XML format prescribed by the *OLAC Metadata* standard as it has been published by the Lyon-Albuquerque Phonological Systems Database, or LAPSyD. The described resource provides information on the phonological inventory, syllable structures, and prosodic patterns of the Cape Verde Creole language. The example below shows the complete metadata record as it is returned in a GetRecord request of the OAI-PMH:

```
<oai:record xmlns:oai=http://www.openarchives.org/OAI/2.0/
      xmlns:olac=http://www.language-archives.org/OLAC/1.1/
      xmlns:dc=“http://purl.org/dc/elements/1.1/”
      xmlns:dcterms=“http://purl.org/dc/terms/”
      xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”>
        <oai:header>
          <oai:identifier>oai:www.lapsyd.ddl.ish-lyon.cnrs
          .fr:src692</oai:identifier>
          <oai:datestamp>2009-10-07</oai:datestamp>
        </oai:header>
        <oai:metadata>
          <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:publisher xsi:type=“dcterms:URI”>www.lapsyd.ddl.ish
          -lyon.cnrs.fr
          </dc:publisher>
          <dcterms:license xsi:type=“dcterms:URI”>http://
          creativecommons.org/licenses/by-nc-nd/3.0/
          </dcterms:license >
          <dc:type xsi:type=“dcterms:DCMIType”>Dataset</dc:type>
          <dc:format xsi:type=“dcterms:IMT”>text/html</dc:format>
          <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:metadata>
</oai:record>
```

In the example, we can see the basic features of OLAC metadata. Metadata elements come from the 15 elements of the basic dc namespace, plus the additional refinements from the dcterms namespace. The xsi:type attribute is used to declare the encoding scheme that is used to express a value precisely. When the encoded value comes from a controlled vocabulary that is enumerated in one of the OLAC recommendations listed above, the olac:code attribute is used to encode the value. In that case, the element content can optionally be used to express the denotation more specifically. For instance, the final element in the example above illustrates using an ISO 639-3 code to identify the language and adding a note to say more specifically that the resource pertains to a particular dialect.

Enter Linked Data

When OLAC began, developing purpose-specific XML markup for information interchange was a best current practice. In the intervening years, Linked Data (Berners-Lee 2006; Bizer, Heath, and Berners-Lee 2009) has emerged from the Semantic Web[7] activity of the World Wide Web Consortium as a strategy for linking disparate purpose-specific datasets into a single interoperating global Web of Data. The impetus for reframing OLAC metadata in terms of Linked Data has come from two directions. The first is the general trajectory of the Dublin Core Metadata Initiative and the wider library community. Librarians are recognizing that Linked Data represents an opportunity for libraries to integrate their information resources with the wider web (see, for instance, Byrne and Goddard 2010). Whereas Dublin Core was initially conceived as a simple record format, a new best practice has emerged in which an abstract model[8] is used in defining application profiles[9] that provide semantic interoperability with other applications within the Linked Data framework (Baker 2012). There is perhaps no stronger evidence for a major trend toward Linked Data in cataloging than the BIBFRAME[10] initiative at the Library of Congress, which is building on the Linked Data model to develop a replacement for the

MARC standard (Miller et al. 2012). Players in the OAI-PMH world are also working with Linked Data (Haslhofer and Schandl 2008, 2010; Davison et al. 2013).

The second impetus has come from the application of the Linked Data framework to the linking of linguistic data and metadata (Chiarcos, Nordhoff, and Hellmann 2012). With the emergence of a Linguistic Linked Open Data cloud (Chiarcos et al. 2013), OLAC as a major source of linguistic metadata has been notable by its absence. The work described herein has therefore sought to rectify this gap by bringing OLAC into the cloud of Linked Data.

What does it take to link into the Web of Data? The Linked Data paradigm is based on four simple rules (Berners-Lee 2006):

1. Use uniform resource identifiers (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 them with useful information using RDF and other Semantic Web standards.
4. Include links to other URIs so that users can discover more things.

These rules serve as the backdrop for the discussion in the next sections, which describe how OLAC resources have been expressed as Linked Data and how those expressions have been incorporated into the OLAC infrastructure.

As the rules indicate, the Linked Data paradigm is built on two foundational standards. The first is the Resource Description Framework (RDF),[11] which is a model for the representation and interchange of data that is semantically interoperable. The second is Uniform Resource Identifiers (URI),[12] which provide a syntax for the creation of globally unique names for things in the world (including concepts). The RDF approach to semantic representation can be summarized as follows. Information is expressed as a set of statements. Each statement is a triple consisting of a subject, a predicate, and an object. The subject is a resource that is named by a URI. The predicate is a URI that names a property. In the case of representing Dublin Core in RDF, the metadata elements (like Title, Date, Creator) become properties. The object may be another resource named by a URI or it may be a literal value. A set of statements forms a directed graph, in which the resources and literals are nodes and the properties are directed arcs from subject to object. The fact that any collection of RDF graphs can be merged into a single, large graph forms the basis for the interoperation across data sources within the Linked Data approach.

Expressing OLAC Metadata as Linked Data

OLAC is a source for information about three kinds of resources: the controlled vocabularies it has developed for language resource description, descriptions of the archives that participate in OLAC, and descriptions of the language resources those archives hold. The next three subsections describe how each of these is expressed as Linked Data. A final subsection considers the issue of personal and organizational names, which is an area in which the current solution is not yet in line with the rules of Linked Data.

Controlled Vocabularies

The *OLAC Metadata Usage Guidelines*[13] specify many best practices in terms of controlled vocabularies that should be used in representing the values of the metadata elements. To comply with the rules of Linked Data, those values need to be represented as URIs. All the controlled vocabularies that are specified as encoding schemes in Dublin Core (such as DCMI Type and MIME Type) already have URIs and RDF descriptions in common use. This includes the ISO 639-1 and ISO 639-2 standards for language identification, which are implemented at the Linked Data Service[14] of the Library of Congress. For instance, the 639-2 code [deu] for German is represented by *http://id.loc.gov/vocabulary/iso639-2/deu*. Work is in progress to implement the entire ISO 639-3 code set in the same way at the LC Linked Data Service; in the meantime, we are using *lexvo.org* URIs—for example, *http://lexvo.org/id/iso639-3/deu*.

The four controlled vocabularies defined by OLAC itself (Linguistic Type, Linguistic Field, Discourse Type, and Role) were not previously implemented in RDF. These have now been implemented as hash namespaces, so that "lexicon" from the Linguistic Type vocabulary is now represented by *http://www.language-archives.org/vocabulary/type#lexicon*. The vocabularies are implemented in RDF by means of the Simple Knowledge Organization System (SKOS).[15] The vocabulary as a whole is first defined as an instance of a concept scheme. For instance, the following is the definition of the Linguistic Type scheme. This RDF sample (as are all the samples that follow) is expressed in the N3[16] syntax. The first line is a complete subject-predicate-object triple in which "a" is shorthand for the property rdf:type. A semicolon indicates that the next line will be another predicate-object pair for the same subject, whereas a comma indicates an additional object for the same subject and predicate:

```
<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 exten-
        sion 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> .
```

Each term in the vocabulary is then defined as a SKOS concept by mapping the definition, examples, and comments from the published vocabulary documentation[17] into the appropriate SKOS properties. Here is the definition of the term “lexicon”:

```
<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 com-
          parative 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." .
```

In the case of the Linguistic Type, Linguistic Field, and Discourse Type vocabularies, the terms are concepts that serve as the values of metadata properties. In the case of the Role vocabulary, the terms of the vocabulary are properties themselves. More specifically, they are refinements of the dc:contributor property. The implementation of those terms adds that declaration.

Archive Descriptions

OLAC publishes an index of all participating archives[18] that links to a description of each archive. By virtue of building on the OAI-PMH, every archive has been assigned a unique identifier from the outset, and these are mapped to HTTP URIs to provide a location for the archive description. For instance, the HTTP URI for the LAPSyD archive that is the source of the sample OLAC metadata record given above is *http://www.language-archives.org/archive/www.lapsyd.ddl.ish-lyon.cnrs.fr.* Thus, with respect to archive descriptions, OLAC already complied with the first two rules of Linked Data. But as far as the third rule is concerned, an RDF form of the description was missing.

The OLAC archive description is a mandatory component of an OLAC metadata repository.[19] It was already assigned a namespace and was defined by an XML schema.[20] Providing an RDF rendering of the archive descriptions involved first creating an RDF schema[21] that defines the properties of an OLAC archive description and then implementing an XSLT script that transforms the archive description as harvested from the repository into the RDF equivalent. For example, the following is the RDF description of the LAPSyD archive

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix olac-archive: <http://www.language-archives.org/OLAC/1.1/olac
-archive#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
<http://www.language-archives.org/archive/www.lapsyd.ddl.ish-lyon.cnrs
.fr> a rdfs:Resource ;
        dc:title "Lyon-Albuquerque Phonological Systems Database
        (LAPSyD)" ;
        olac-archive:archiveURL <http://www.lapsyd.ddl.ish-lyon.cnrs
        .fr/lapsyd/> ;
        dc:contributor "Flavier, Sébastien (Developer)",
          "Maddieson, Ian (Creator)",
          "Marsico, Egidio (Editor)",
          "Pellegrino, François (Editor)" ;
        olac-archive:institution "CNRS and University of New Mex-
        ico" ;
        olac-archive:shortLocation "Lyon, FRANCE" ;
        olac-archive:synopsis "This OAI/OLAC metadata repository
        gives a metadata record for every language entry in the
        Lyon-Albuquerque Phonological Systems Database (LAPSyD)
        database. LAPSyD is a searchable database which provides
        phonological information (inventories, syllable structure
        and prosodic patterns) on a wide sample of the world's
        languages." ;
```

```
          olac-archive:access "Each language entry described in this
          repository is a public Web page that may be accessed with-
          out restriction. Reuse of material on the site is subject to
          the Terms of Use shown on the LAPSyD site." .
```

Language Resource Descriptions

Similarly for language resource descriptions, each language resource has always been identified by an HTTP URI, but an RDF form of the description was missing. Another XSLT script has been implemented to transform the OAI-PMH GetRecord response into an RDF equivalent. For instance, this process outputs the sample OLAC metadata record given above as the following RDF statements:

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@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
#>.

<http://www.language-archives.org/item/oai:www.lapsyd.ddl.ish-lyon.cnrs
.fr:src692>
          a rdfs:Resource ;
          dc:publisher <http://www.language-archives.org/archive/www
          .lapsyd.ddl.ish-lyon.cnrs.fr> ;
          dc:title "LAPSyD Online page for Cape Verde Creole, San-
          tiago dialect" ;
          dc:description "This resource contains information about
          phonological inventories, tones, stress and syllabic struc-
          tures" ;
          dcterms:modified "2012-05-17"^^dcterms:W3CDTF ;
          dc:identifier <http://www.lapsyd.ddl.ish-lyon.cnrs.fr/lapsyd
          /index.php?data=view&code=692> ;
          dc:publisher <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> ;
          olac-role:author "Maddieson, Ian" ;
```

```
dc:subject olac-field:phonology, olac-field:typology ;
dc:type olac-type:language _ description ;
dc:language <http://lexvo.org/id/iso639-3/eng> ;
dc:subject <http://lexvo.org/id/iso639-3/kea>,
  "Note for [kea]: Cape Verde Creole, Santiago dialect" .
```

Note that in the first dc:publisher statement, the LAPSyD archive (as described in the RDF snippet in the preceding subsection) is declared to be the publisher of the metadata record. This is an application of the fourth rule of Linked Data in which the objects of the RDF statements should link to other URIs so that users can discover more things. The use of OLAC-specific vocabularies is seen beginning with the olac:author property, which comes from the OLAC Role vocabulary. In the next two statements, the property values come from the OLAC Field and OLAC Type vocabularies, respectively. A final feature of note is in the final statement that describes the subject language of the resource. In the OLAC metadata standard, first the subject language is identified by a code from ISO 639-3 as the value of the olac:code attribute and then free text may be added in the element content to give greater detail. This is translated into two RDF statements, one with an HTTP URI as the value and the other with a literal string as the value. In generating the latter, which is a comment for human consumption, the conversion process prepends "Note for [kea]:" to identify which ISO 639-3 language the comment is about.

The Problem of Personal Names

Having implemented the conversions described above, OLAC is now expressing language resource metadata as Linked Data. There is one respect, however, in which the results still fall short of the spirit of Linked Data in that they fail to comply with the fourth rule of Linked Data: "Include links to other URIs so that users can discover more things." The problem area is the use of literal strings to represent the names of persons who are contributors to the language resource. In the specific case of the resource description above, the user should be able to follow a URI to find out who "Maddieson, Ian" is.

For the practice of Linked Data across a general audience, the URI of a person's article in the English Wikipedia is a popular source of URIs for persons. Even better for Linked Data purposes is the corresponding URI from DBpedia,[22] which maps each Wikipedia article into an RDF resource. Within the library cataloging world, the gold standard is to use an identifier from a national library's authority file—as, for instance, the Library of Congress Name Authority File.[23] In this particular case, Ian Maddieson is a sufficiently eminent linguist that he can actually be found in both, though that will not be the case for the vast majority of people who contribute to language resources. An existing single source of URIs for over 34,000 persons across the field of linguistics is the Linguist List Directory of Linguists,[24] though these URIs are not ideal for use in Linked Data because they are not "Cool URIs."[25] Another source that provides even more URIs, but that lacks uniformity, is personal or professional home page URIs. The academic world has recognized the need to develop a standardized way of uniquely identifying those who have made contributions to the academic literature. In 2012 an open, nonprofit, community-based effort named ORCID (Open Researcher and Contributor ID)[26] was launched. In just four years, its registry has grown to include over 2.5 million unique researcher identifiers.

All the following are thus HTTP URIs that could be used to identify this particular author in a Linked Data context (though note that only dbpedia.org, id.loc.gov, and orcid .org comply with all four rules of Linked Data):

- https://en.wikipedia.org/wiki/Ian_Maddieson
- http://dbpedia.org/resource/Ian_Maddieson
- http://id.loc.gov/authorities/names/n84089547
- http://linguistlist.org/people/personal/get-personal-page2.cfm?PersonID=695
- http://www.unm.edu/~ianm/index.html
- http://linguistics.berkeley.edu/person/23
- http://orcid.org/0000-0002-0775-0555

At present the *OLAC Metadata Usage Guidelines*[27] recommend only that a contributor be identified "by means of a name in a form that is ready for sorting within an alphabetical index." Yet the OLAC infrastructure has no means of enforcing this guideline or even of ensuring that each contributor metadata element names only one contributor. As a result, in spite of providing a faceted search service[28] that offers interoperable search on 14 facets that have uniform metadata values across the community of archives, contributor is not one of those facets. This is an area in which the community will need to tighten its metadata guidelines and practices if it intends to support the identification of contributors both in Linked Data and in faceted search.

Incorporating Linked Data into the OLAC Infrastructure

OLAC has taken the first steps of incorporating Linked Data into its infrastructure. The new RDF vocabularies described earlier are in place, as are the RDF transformations for archive descriptions and language resource descriptions. The URIs for all these resources are configured following W3C best practices to support HTTP content negotiation[29] so that they return an HTML document by default, but return an RDF/XML document when the header of the HTTP request specifically asks for the application/rdf+xml MIME type. To contribute[30] to the cloud of Linguistic Linked Open Data (Chiarcos et al. 2013), the nightly metadata harvest creates a gzipped dump[31] of the RDF/XML rendering of every metadata record in the OLAC catalog, and that dataset has been registered at the DataHub[32] of the Open Knowledge Foundation.

Looking to the future, the OLAC metadata standard has not changed appreciably since version 1.0 was adopted in 2003. In light of the trend toward Linked Data in the wider metadata community, now may be a fitting time to develop a version 2.0 update that brings OLAC into line with Linked Data as well as other current best practices. Doing so would encourage the participating archives to create metadata that better interoperates with the global Web of Data. The open-endedness of the Linked Data approach would further allow archives to create even richer metadata by augmenting their resource descriptions with properties from any RDF vocabulary. Perhaps the greatest advantage would be the long-term benefit for the sustainability of the OLAC vision that could accrue from entering into the mainstream of library practices. However, there is a downside: Developing OLAC 2.0 would have a substantial cost in terms of requiring participating archives to reimplement their OLAC repositories.

One way forward would be to adopt a hybrid approach. The OLAC harvester could support both OLAC 1.1 and 2.0. All 2.0 metadata would be back translated into 1.1 format so that all existing services continue to work. By the same token, all 1.1 metadata would be forward translated into 2.0 format and fed into an RDF aggregator that could capture all the added richness of 2.0 metadata. OLAC could then begin to develop new services that take full advantage of the Linked Data paradigm, including offering semantic search over the OLAC catalog by providing an endpoint for SPARQL (the query language for RDF).[33]

Conclusion

Given the core values of the OLAC process, one of which is that decisions be made by consensus and that the greatest voice is given to those who are implementing the standards, updating the OLAC metadata standard to a new version based on Linked Data is not a step that can be taken lightly. Moving to OLAC 2.0 would be a major effort requiring the participating archives around the world both to agree and to reimplement. Still, the time is surely ripe for OLAC to consider such an update to its standards and infrastructure, particularly in light of the potential for a future in which its language resource descriptions could interoperate seamlessly with the wider library cataloging community—and even more broadly with the global Web of Data.

Notes

1. http://www.language-archives.org/.
2. http://www.language-archives.org/archives.
3. http://search.language-archives.org.
4. https://www.openarchives.org/pmh/.
5. http://www.loc.gov/standards/iso639-2/.
6. http://www.sil.org/iso639-3/.

7. http://www.w3.org/standards/semanticweb/.

8. http://dublincore.org/documents/abstract-model/.

9. http://dublincore.org/documents/profile-guidelines/.

10. http://www.loc.gov/bibframe/.

11. http://www.w3.org/RDF/.

12. http://www.w3.org/Addressing/.

13. http://www.language-archives.org/NOTE/usage.html.

14. http://id.loc.gov/.

15. http://www.w3.org/2004/02/skos/.

16. http://www.w3.org/TeamSubmission/n3/.

17. http://www.language-archives.org/REC/type.html.

18. http://www.language-archives.org/archives.

19. http://www.language-archives.org/OLAC/repositories.html#OLAC%20archive%20description.

20. http://www.language-archives.org/OLAC/1.1/olac-archive.xsd.

21. http://www.language-archives.org/OLAC/1.1/olac-archive.rdf.

22. http://wiki.dbpedia.org/.

23. http://id.loc.gov/authorities/names.html.

24. http://linguistlist.org/people/personal/.

25. http://www.w3.org/TR/cooluris/.

26. http://orcid.org/.

27. http://www.language-archives.org/NOTE/usage.html.

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

29. http://www.w3.org/TR/swbp-vocab-pub/.

30. http://wiki.okfn.org/Working_Groups/Linguistics/How_to_contribute.

31. http://www.language-archives.org/static/olac-datahub.rdf.gz.

32. https://datahub.io/dataset/olac.

33. https://www.w3.org/TR/sparql11-overview/.

References

Baker, Thomas. 2012. "Libraries, Languages of Description, and Linked Data: A Dublin Core Perspective." *Library Hi Tech* 30 (1): 116–133.

Berners-Lee, Timothy. 2006. "Design Issues: Linked Data: World Wide Web Consortium." http://www.w3.org/DesignIssues/LinkedData.html.

Bird, Steven, and Gary F. Simons. 2003. "Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources." *Computers and the Humanities* 37 (4): 375–388.

Bird, Steven, and Gary F. Simons. 2004. "Building an Open Language Archives Community on the DC Foundation." In *Metadata in Practice*, edited by Diane I. Hillmann and Elaine L. Westbrooks, 203–222. Chicago: American Library Association.

Bizer, Christian, Thomas Heath, and Timothy Berners-Lee. 2009. “Linked Data—The Story So Far.” *International Journal on Semantic Web and Information Systems* 5 (3): 1–22, doi:10.4018/jswis.2009081901.

Byrne, Gillian, and Lisa Goddard. 2010. “The Strongest Link: Libraries and Linked Data.” *D-Lib Magazine* 16 (11): 5.

Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann, eds. 2012. *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Heidelberg: Springer.

Chiarcos, Christian, Steven Moran, Pablo N. Mendes, Sebastian Nordhoff, and Robert Littauer. 2013. “Building a Linked Open Data Cloud of Linguistic Resources: Motivations and Developments.” In *The People’s Web Meets NLP: Collaboratively Constructed Language Resources*, edited by I. Gurevych and J. Kim, 315–348. Berlin: Springer. doi:10.1007/978-3-642-35085-6_12.

Davison, Stephen, Yukari Sugiyama, Elizabeth McAulay, and Claudia Horning. 2013. “Enhancing an OAI-PMH Service Using Linked Data: A Report from the Sheet Music Consortium.” *Journal of Library Metadata* 13 (2–3): 141–162. doi:10.1080/19386389.2013.826067.

DCMI. 2012. “DCMI Metadata Terms.” Dublin Core Metadata Initiative. http://dublincore.org/documents/dcmi-terms/.

Haslhofer, Bernhard, and Bernhard Schandl. 2008. “The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data.” In *Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008)*. https://eprints.cs.univie.ac.at/284/

Haslhofer, Bernhard, and Bernhard Schandl. 2010. “Interweaving OAI-PMH Data Sources with the Linked Data Cloud.” *International Journal of Metadata, Semantics and Ontologies* 5 (1): 17–1. doi:10.1504/IJMSO.2010.032648.

Miller, Eric, Uche Ogbuji, Victoria Mueller, and Kathy MacDougall. 2012. *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services*. Washington, D.C.: Library of Congress. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.

Powell, Andrew, and Peter Johnston. 2003. “Guidelines for Implementing Dublin Core in XML.” Dublin Core Metadata Initiative. http://dublincore.org/documents/dc-xml-guidelines/.

Simons, Gary F., and Steven Bird. 2003. “The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources.” *Literary and Linguistic Computing* 18 (2): 117–128.

Simons, Gary F., and Steven Bird. 2006. *OLAC Process*. Open Language Archives Community. http://www.language-archives.org/OLAC/processa.html.

Simons, Gary F., and Steven Bird. 2008a. *OLAC Metadata*. Open Language Archives Community. http://www.language-archives.org/OLAC/metadata.html.

Simons, Gary F., and Steven Bird. 2008b. *OLAC Repositories*. Open Language Archives Community. http://www.language-archives.org/OLAC/repositories.html.

Simons, Gary F., and Steven Bird. 2008c. *Recommended Metadata Extensions*. Open Language Archives Community. http://www.language-archives.org/REC/olac-extensions.html.

Simons, Gary F., and Steven Bird. 2008d. “Toward a Global Infrastructure for the Sustainability of Language Resources.” In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, 20–22 November 2008, Cebu City, Philippines*, edited by R. Roxas, 87–100. Manila: De La Salle University.