An OLAC Perspective on Services: The Forgotten Language Resources

Hugh Paterson III[*]
Collaborative Scholar
University of Oregon & University of North Texas
USA
i@hp3.me

Abstract

From the perspective of ethno-linguistic minority communities and those who work with them, finding pedagogical, descriptive, and primary documentation resources about their culture and language have never been easy. This has primarily been due to the lack of language-specific indexing of resources in libraries and archives. The Open Language Archive Community (OLAC) created a metadata profile and aggregator built upon Open Archive Initiative (OAI) protocol and Dublin Core Metadata Standard (DCMS) to address the lack of diversity in indexing language and culture resources. The aggregator supports discovery of resources with language as an entry point into the discovery process. This application profile is used by over sixty data providers. In this paper I address the lack of inclusion of services as identified by the DCMIType "Service" within the aggregated records. I discuss services from a typological perspective and explore how services can themselves be language resources. I propose that indexing services, specifically language resource services, is one way to increase diversity in records while connecting language communities with digital resources they can use in their language.

Keywords: DCMIType; services; language resources

1 Introduction

Discovery of and access to language resources continues to be a critical issue in the success of language and linguistic scholarship. Broadly the issue of metadata and language resource discovery has been well covered in the academic literature. Examples include: OLAC (Bird et al., 2001), IMDI (Broeder & Wittenburg, 2006), CMDI (Broeder et al., 2012), CLARIN/VLO (Van Uytvanck et al., 2012; Van Uytvanck et al., 2010), META-SHARE (among others, Gavrilidou et al., 2012). However, the specific class of resources identifiable with the DCMIType vocabulary term "Services" is absent in the OLAC literature — it seems to be forgotten and unused. Of the listed schemas, the OLAC application profile is most closely aligned with Dublin Core. The OLAC application profile relies on both OAI-PMH and Dublin Core metadata (Bird & Simons, 2001, 2003, 2004; Simons & Bird, 2003) to provide a high-level description of language resources. Originally OLAC was conceived as being used by language archives, e.g., (Michailovsky et al., 2011—LACITO), but has subsequently been embraced by libraries (Hirt et al. 2009—Graduate Institute of Applied Linguistics; Kleiber et al. 2018—University of Hawaii), database creators (Greenhill et al., 2008—Austronesian Basic Vocabulary Database), digital repositories, and even technologists creating digital portfolios presenting the contents of Curriculum Vitae (Paterson, 2021a). That is, both institutional and personal stakeholders use OLAC to make others aware of language resources which may be broadly available or may otherwise require direct communication with the resource stewards. Therefore, the OLAC aggregator would seem to be one place to look to see if anyone has previously considered services as language resources. I suggest that there is a distinct category of language resources which are services that ought to be indexed like other kinds of language resources. This stands in addition to services which may not be distinct language resources. Broadly the resource description literature is vague on how to define services distinctly from other kinds of entities identified in the DCMIType vocabulary such as "Software" or "Interactive Resources". Finding concrete examples of bibliographic records illustrating the three way distinction is a challenge. The absence of concrete description examples is likely due to different conceptual models and how these models treat what Dublin

[*] The author has a MA in linguistics and has worked with SIL International's Language & Culture Archives for several years. He writes for broad audiences in which readers may not be subject specialists. Currently he seeks to collaborate with institutions which desire to make their language resources more discoverable. https://hughandbecky.us/Hugh-CV

20th International Conference on Dublin Core and Metadata Applications 2022 (DC 2022)

https://doi.org/10.23106/dcmi.953120677

Core via the DCMIType vocabulary considers "Services". I suggest that service-typed language resources are distinct from business services and other kinds of language resources. I present a method for conceptualizing services as language resources. I then present three use-cases and two modeled service records using the OLAC application profile (Open Archives Initiative, Dublin Core, and OLAC Vocabularies). I then discuss the utility of advertising services via OLAC and some limitations of the OLAC application profile.

My focus here is on how *services* are described (or not described) in OLAC. In an effort to provide broader applications of this discussion I make connections to other frameworks for describing language resources as well as other aggregation projects. OLAC uses the DCMIType vocabulary to describe the types of items referenced in bibliographic records. As illustrated in Figure 1, *services*, *language resources*, and *their indexing* is a larger topic than how such issues are handled in OLAC. Not all records aggregated via the OLAC aggregator are in fact identified as being about or a primary resource for a specific language (only about sixty-four percent use the `<dc:language>` tag). This suggests that the scope of content discoverable via the OLAC aggregator includes more than narrowly defined "language resources". This is not a problem for the Open Language Archives Community because the community defines a language resource broadly stating: "A language resource is any kind of DATA, TOOL or ADVICE (see the founding vision statement, “The Seven Pillars of Open Language Archiving: A Vision Statement”) pertaining to the documentation, description or development of a human language."[2] For example, using the OLAC definition of a language resource, advice about language documentation methodologies may actually not contain any language description or language content as subject material and yet qualify as a *language resource*.

There are several description models in use by language resource stewarding communities. Not all use the DCMIType vocabulary. In the *digital libraries space*, which is composed of projects which aggregate records from different providers, even those which use the Dublin Core Metadata Elements[3] differ in their conceptual models and the terminology referring to types of entities in the models. For example, neither the *National Science Digital Library*[4] nor *Europeana*,[5] use the term *services* as they refer to data providers or metadata handling processes. Before discussing OLAC-centric specifics, it is helpful to generally describe the field of language resource indexing, where language resources are catalogued on the basis of the language(s) to which they pertain.

1.1 Language Resource Indexers

There are several tools which help users find software tools, research reports, scholarly outputs, and language data in and about specific languages. Tools, research reports, scholarly outputs, and language data are broadly known as *language resources*, though the exact definition of what qualifies as a language resource varies by community of practice. For example, one OLAC contributor, the European Language Resource Association (ELRA, 2015) states that “the term *Language Resource* refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users.”[6] It goes on to state that “examples of Language Resources are written and spoken corpora, computational lexica, terminology databases, speech collection, etc.” Like the more broadly scoped OLAC definition discussed previously, the ELRA definition is not exclusive of services, neither does it overtly mention services. In my review of other major language resource aggregators, I find that entities for which the DCMIType “Service” definition would apply are often categorically lumped with software. This is in part understandable as software is often the most tangible part of a service. However, the DCMIType vocabulary suggests that software and services should be distinct types as they can be very different kinds of “things”.

Commercial discovery tools, focused mostly on scholarly publications, include: Clarivate’s *Linguistics & Language Behavior Abstracts* (Adlington, 2018), *Brill’s Linguistic Bibliography* (Shparberg, 2015), Springer’s *Linguistic Abstracts Online* (Shparberg, 2022). The *Glottolog* (Hammarström, 2015; Nordhoff & Hammarström, 2012) is an open, scholar-driven, alternative to these commercial offerings. Outside of the scope of scholarly publications three major discovery portals exist: OLAC (a global scholar-driven com-

[2] http://www.language-archives.org/documents/faq.html

[3] https://www.dublincore.org/specifications/dublin-core/dcmi-terms

[4] https://nsdl.oercommons.org

[5] https://www.europeana.eu/en

[6] Most stewards of language resources would consider materials which are not machine readable to be valid language resources, e.g., a printed book, or hand written document.

https://doi.org/10.23106/dcmi.953120677

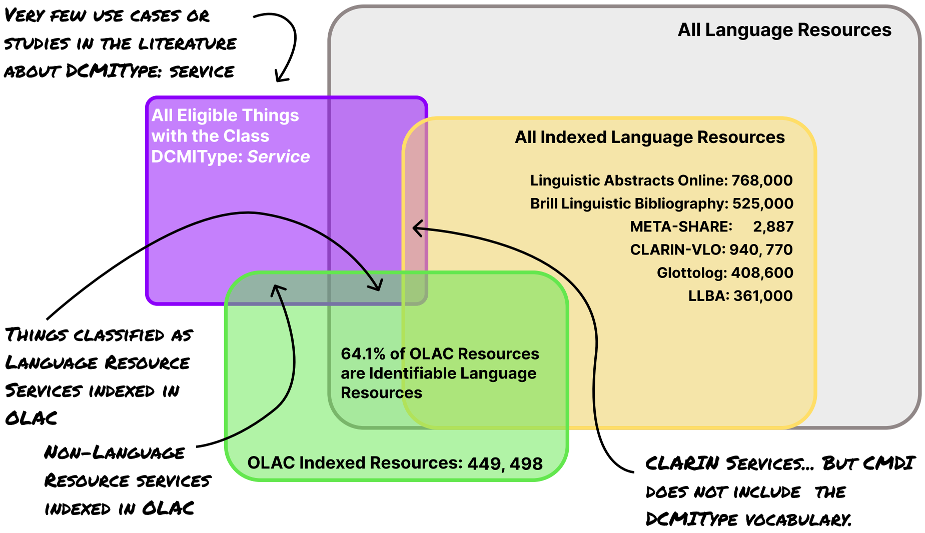
VERY FEW USE CASES OR STUDIES IN THE LITERATURE ABOUT DCMITYPE: SERVICE
All Language Resources
All Eligible Things with the Class DCMIType: *Service*
All Indexed Language Resources
Linguistic Abstracts Online: 768,000
Brill Linguistic Bibliography: 525,000
META-SHARE: 2,887
CLARIN-VLO: 940, 770
Glottolog: 408,600
LLBA: 361,000
THINGS CLASSIFIED AS LANGUAGE RESOURCE SERVICES INDEXED IN OLAC
64.1% of OLAC Resources are Identifiable Language Resources
NON-LANGUAGE RESOURCE SERVICES INDEXED IN OLAC
OLAC Indexed Resources: 449, 498
CLARIN SERVICES... BUT CMDI DOES NOT INCLUDE THE DCMITYPE VOCABULARY.

Figure 1: Euler diagram showing the overlap of generic services, OLAC indexed resources, indexed language resources, and un-indexed language resources. Quantities of indexed language resources are sourced from respective websites as of May 2022.

munity),[7] CLARIN/VLO (an European Union funded community focused on European institutions),[8] and META-SHARE (a member driven community with mostly European institutional members).[9] These portals, representing three different networks of institutions supporting scholarship, are each driven by metadata about language resources, inclusive of software, corpora, and archival collections. Each network facilitates resource discovery through its own metadata schema. For example, OLAC has extended Dublin Core via a set of vocabularies further qualifying DC elements (Bird & Simons, 2003, 2004; Bird et al., 2001), the VLO and CLARIN infrastructure is powered by application profiles which conform the CMDI framework (which is the evolutionary result of IMDI), while META-SHARE (among others, Federmann et al., 2012; Gavrilidou et al., 2012; Piperidis, 2012) maintains its own metadata standard and application profile (Desipri et al., 2015).[10] The META-SHARE application profile is more constrained than CMDI, which can be highly tailored for linguistic variables. The semantics of metadata elements in each of the different schemas are not always compatible in a one-to-one relationship across the different schemas. For example, META-SHARE lumps services and tools together in a common category titled *technologies*, whereas OLAC via the DCMIType vocabulary splits these into "Software" and "Services". Additionally, the description model invoked by the schemas are also not always compatible. That is, the parts of a complex resource may be described to different degrees under different schemas and some properties in one schema may not have an equivalence in the other schemas.

The topic of the description of linguistic or language services in aggregators is only tangentially treated in the literature (Ide et al., 2014; Odijk, 2016, 2019). A brief review of the CMDI documentation (CMDI Taskforce, 2016), the metadata profile which drives the CLARIN/VLO portal suggests that CLARIN does have a way to reference web-search tools, and they label these interfaces *services*.[11] Odijk (2019) has written about accessing software via CLARIN discovery tools. This seems to be very similar to the kind of use-case for which the DCMIType "Service" was intended. Albeit, the term "software" chosen by Odijk (or perhaps the CLARIN team) seem to miss the point that these "software" units are acting as "services". Odijk (2016) makes it clear that some software in the CLARIN system does act on data in the way that services would be expected to act on data. As a tentative statement, it appears that for CMDI users, in some cases referenced "things" are categorized differently (software vs. service) than if those same "things" were to be categorized using Dublin Core's DCMIType vocabulary definitions.

[7] http://search.language-archives.org

[8] https://vlo.clarin.eu

[9] http://www.meta-share.org

[10] http://www.meta-share.org/knowledgebase/overviewOfTheMetadataModel

[11] See also the YouTube video "CLARIN Services in the European Open Science Cloud (EOSC)" https://youtu.be/YvZ9Y_uyr7M.

https://doi.org/10.23106/dcmi.953120677

1.2 Dublin Core & The DCMIType Vocabulary

Even though current use-case discussions of DCMS often cast it as a key element in the linked data or RDF universe, two use-cases remain outside of RDF uses: 1) embedding Dublin Core in HTML, e.g., GoogleScholar indexing; and 2) OAI aggregation using Dublin Core, such as the *Digital Public Library of America*,[12] and the *Directory of Open Access Journals*.[13] It is under this second use-case that the OLAC community aggregator falls. The DCMS contains fifteen element properties. The DCMIType vocabulary refines the `<dc: type>` element with one of twelve terms. "Services" is one of those terms.

Various DCMS components (elements and refinements) have been discussed regarding their percentage of use, syntax, and institutional semantics applied to the component (among others see: Hutt & Riley, 2005; Kurtz, 2010; Park & Childress, 2009; Phelps, 2012; Shreeves et al., 2005; Toves & Hickey, 2014; Ward, 2002). A subset of studies specifically report on the use of the DCMIType vocabulary (Balatsoukas et al., 2018; Park & Tosaka, 2010; Paterson, 2022; Ward, 2003, 2004), but even the few reports which look at some aspect of the DCMIType vocabulary are silent on possible occurrences of “Service” in aggregators and collections of Dublin Core records. This paper explores the current and potential use by the OLAC application profile of the DCMIType “Service”.

2 Services

2.1 Definitions and Types of Services

In defining *service* for an audience of linguists and language archivists, Aristar-Dry and Simons (2006) point to the business function of services suggesting that services solve problems that individual organizations and resource users cannot solve on their own, that services use automation tools to solve these problems, and that services are provided by organizations. Phipps et al. (2006) in writing to librarians and information professionals describe nine different kinds of services which support better resource discovery. These include: *Harvesting*, *Primary metadata generation*, *Metadata augmentation*, *Transformation* (safe and collection-specific), *Equivalence*, *Crosswalking* (schema and vocabulary), *Archiving/Persistence*, *Annotation*, *Metadata improvement and rating*. A tenth one is *Aggregation or Display* which they do not overtly mention. The DCMIType definition for “Service” is: “A system that provides one or more functions”.[14] However, what constitutes a function is open to interpretation. The DCMIType definition leaves a broad possible range of uses of the term *service*. To help narrow possible ambiguities DCMS suggests that implementers follow the one-to-one principle in DCMS. When applied, the DCMIType definition of “Service” could only apply to something that does not also fit another DCMIType definition. The definition provided by Aristar-Dry and Simons fits within the DCMIType definition of “Service” as do the ten types of services previously listed. Common across all the presented definitions are their digital nature (though there seems to be no reason digital should be required) and their tendency to facilitate repeatable tasks.

To address possible ambiguities between an interactive resource and a service in digital contexts, I suggest that a service has two conditions. First, it receives input data and computes on the data, and then outputs some or all of data in some new structure, format. Essentially changing the expression or the manifestation of the input. Second, it allows for multiple inputs to the defined workflow. In contrast to the presented criteria for services, interactive resources act on a single set of input data and provide some sort of enhanced understanding on the basis of that single input data. The data operating in an interactive resource may be sub-divided or visualized by the interactive resource for user engagement, but it is always the same data for a given expression of the interactive resource.

Often the semantics of a DCMS element finds its final definition and utility from how the community of users adopts it and implements the element. DCMIType “Service” has not received as much attention in the literature relative to other portions of the DCMS. There appears to be only two uses of the DCMIType “Service” discussed across five papers.[15] Apps (2004, 2005, 2006) presents the use-case where collections and their associated services are sought to be described within the context of the JISC project Information Environment

[12] https://dp.la

[13] https://doaj.org

[14] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dcmitype/Service

[15] Review of documentation and description of metadata schemes developed for cultural heritage both within and outside of linguistic contexts, e.g., Europeana Data Model, Center for Intercultural Documentation-Conceptual Reference Model, Digital Public Library of America, National Science Digital Library, etc., has yet to reveal an application profile which makes an explicit equivalence statement to DCMIType “Service”, or even require the use of DCMIType “Service”.

https://doi.org/10.23106/dcmi.953120677

Service Registry. In the work reported on by Apps, collections may be accessed programmatically via OAI, web or SOAP, or other protocol. They sought to advertise collection specific access services in a central registry. The second case is presented by Roberts (2001, 2002). It consists of records for services of the New Zealand government. No discussion exists addressing usage or semantics of the DCMIType "Service" in the broader context of library science. Additionally, in Apps' case, to meet the project requirements, the project created an application profile which supported specific service attributes which were not directly indicated via DCMS. That is, further reuse of the DCMIType "Service" within Apps' context only makes sense if future projects adopt and reuse their entire metadata application profile including metadata outside of the DCMS. The specific project requirements make future adoption of the solution in new projects unlikely.

It seems that in the design of DCMS the initial purpose for "Service" was to do exactly what Apps does with it—to indicate records which were for other bibliographic record providers to consume indicating where other resource servers were located (e.g., see: Guenther, 1999, 2000). Before DCMS was created, Z39.50 servers made library catalogs available for digital discovery (Lynch, 1991, 1997).[16] These servers were considered services. The general use-case of 'bibliographic record proliferation' was overtaken in popularity by OAI-PMH[17] and OAI-ORE[18] servers/services. OAI servers frequently serve records composed in DCMS. Therefore it makes sense that records about servers would also be "distributable", implying that the drafters of DCMS had these kinds of use-cases in mind when they added the term to the DCMIType vocabulary.

2.2 Existing Records of Language Resource "Services"

Because, the OLAC aggregator dataset is one place to look to see if anyone has previously considered services as language resources I extracted the declared DCMITypes used by OLAC data providers (as of January 1st 2022) and present these in Table 1.[19] The DCMIType "Service" is not used in any OLAC records.

Most OLAC data providers are institutional archives that generally focus on the cataloging of artifacts; services do not easily fit into acid-free boxes or other containers which may be used by archives. Services cease to be services and are more accurately described as "Software" at the point of archival. However, records present in aggregators like those at OLAC, are not bound to point to or describe only archived language resources—they can point to any language resource.

2.3 The nature of services and language resources

As Phipps et al. point out, services come in different types. Organizations may provide digital access services, or a more limited records search/discovery service, akin to the Z39.50 services of libraries, but an access service is not a language resource—it is a gateway to general records about general resources (some of which may be language resources). Organizations such as language archives may (hopefully) also provide preservation services. In contrast to the access and preservation services of archives, OLAC is a service which provides aggregation of records about language resources, but the records are not (in the general sense) language resources. For these reasons, it is not entirely surprising that no language archives have published a language resource record with the DCMIType "Service" via OLAC. This does not preclude others from advertising services via OLAC.

Additionally, an important distinction exists between a *language resource as a service* and a *business function (service) offered in a language*. For example, a hospital may offer medical care and offer patients the opportunity to engage with hospital staff via their language of choice: Spanish, English, Chinese, etc. In this sense, the hospital is offering healthcare as a service via a language, but the service is not a language resource—it is a business function. In a similar sense then, an institutional repository may offer access to

[16] https://www.loc.gov/z3950/agency

[17] https://www.openarchives.org/pmh

[18] https://www.openarchives.org/ore

[19] These numbers are dynamic and will reflect changes by data providers. Source: http://dla.library.upenn.edu/dla/olac/browse.html?browse=dcmi_type_facet&. However, a stable dataset of OLAC records from 2021 is available (Paterson, 2021b).

[20] Best practice among the DCMS and OAI data exchanges is to follow the casing of terms as they appear in the DCMS standard. Some OLAC providers do not provide records with a DCMIType specification, other providers list multiple DCMITypes on a record as OLAC documentation indicates that DCMIType is optional, and repeatable. However, best practice within the DCMS community is to only provide one DCMIType per record. If multiple are needed, then it is best to create multiple records and link the records with a relation element. OLAC reports the item types as the data providers provide them. Spelling and capitalization are not changed. Validation with the DCMIType vocabulary is also not required.

https://doi.org/10.23106/dcmi.953120677

Table 1: Distribution of DCMITypes across OLAC data providers

DCMIType[20]	Usage Count
Collection	781
Dataset	5620
Event	6
Image	1383
image	63
InteractiveResource	12
Moving Image	1
MovingImage	61581
movingimage	4
PhysicalObject	4
Software	507
Sound	131647
sound	4
StillImage	7391
Text	194287
text	65

language resources (files from a language documentation project). The repository is acting as a gateway service to language resources, but is itself not a language resource. Aristar-Dry and Simons (2006) at the close of the NSF funded E-MELD project forecasted that services were needed to solve the research community's needs for language resource discovery. In the model they put forward, services are not specifically language resources, rather they imagine services akin to the function of Z39.50 in the library space and others outlined by Phipps et al.

Of the ten service types previously mentioned, *Aggregation*, *Harvesting*, and *Metadata rating* are within the functionality provided by OLAC (CLARIN's VLO is an example of aggregation). Given the difference between the number of types of services mentioned by Phipps et al. and those provided by OLAC, it leaves open the possibility that there remains a substantial need for services within the community of users who are looking for language resources.

Finally, just as some services are not language resources, similarly some language resources are not services—even if they are engaged with via the internet. For example, the *Brown Corpus* (Francis & Kučera, 1961) is not a service but is a language resource comprised of several texts or sub-units. In DCMIType terminology, resources composed of sub-resources fall under the definition of "Collection". More recently compiled collections (corpora or otherwise labeled for marketing purposes) often have tools for exploring or exploiting their contents via digital means. Several examples include: the *Digital Archive of Scottish Gaelic*,[21] the *Pahka'anil (Tübatulabal) Text Project*,[22] and *Multi-CAST, the Multilingual Corpus of Annotated Spoken Texts*.[23] Additionally, a slew of resources exist for enabling greater digital access to collections, including among others: *IATH ELAN Text-Sync Tool* (Dobrin & Ross, 2017) and *LingView* (Pride et al., 2020). The referenced language resources in this section certainly have an access service component. But that component is generic and not specifically a language resource. The access service does provide access to language resources, but the ones mentioned in this section are all corpora; DCMIType: "Collection".[24] I suggest that to have a service which is both a language resource and a service, there must be language data in the input and in the output. When language data (or language resources) are merely viewed, this is an indication that the service is an *access service or other kind of infrastructure service* rather than a language resource. Library and infrastructure services likely don't need language input from the end-user to provide that user with access to a language resource in a dynamic way.

[21] https://dasg.ac.uk/en

[22] https://web.csulb.edu/colleges/cla/projects/lingresearch/pahka'anil/index

[23] https://multicast.aspra.uni-bamberg.de

[24] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dcmitype/Collection

3 OLAC records

Examples in the following sections provide code modeled on an OLAC example record. For this reason I present details of OLAC records here. OLAC uses an XML encoded structure of DCMS. This is compliant with OAI harvesting protocols and tools.

OLAC provides the following minimal example (replicated as Example 1 below) for a book, specifically a grammar.[25] The record can be parsed as containing a header statement where links to defined terms are established followed by a description of the resource.

```
<olac:olac
    xmlns:olac="http://www.language-archives.org/OLAC/1.1/"
    xmlns="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.language-archives.org/OLAC/1.1/
    http://www.language-archives.org/OLAC/1.1/olac.xsd">
        <title>A grammar of Kayardild. With comparative notes on Tangkic.</title>
        <creator>Evans, Nicholas D.</creator>
        <subject>Kayardild grammar</subject>
        <subject xsi:type="olac:language" olac:code="gyd">Kayardild</subject>
        <language xsi:type="olac:language" olac:code="en">English</language>
        <description>Kayardild Grammar (ISBN 3110127954)</description>
        <publisher>Berlin - Mouton de Gruyter</publisher>
        <contributor xsi:type="olac:role" olac:code="author">Nicholas Evans</
            contributor>
        <format>hardcover, 837 pages</format>
        <relation>related to ISBN 0646119966</relation>
        <coverage>Australia</coverage>
        <type xsi:type="olac:linguistic-type" olac:code="language_description"/>
        <type xsi:type="dcterms:DCMIType">Text</type>
</olac:olac>
```

Example 1: OLAC example record from the OLAC metadata standard.

Neither OLAC nor OAI prescribe an order to elements, but the example record does present the elements in the following order: Title, Creator, Subject, Subject Language, Language of the work, Description, Publisher, Contributor, Format, Relation, Coverage, Linguistic Type, and DCMIType. The DCMIType value in Example 1 is “Text”.

4 Language Resources as Services

So what might a language resource as a service look like? Do language and linguistic resources as services actually exist? Yes. They exist. But no one has listed them in OLAC yet, nor has anyone discussed them in the context of OAI-PMH or DCMS record sharing. In the following sections, I present three services each with a different need for description. The first is GoogleTranslate (Wu et al., 2016), an automated text-to-text, speech-to-text, and speech-to-speech translation service. The second is RefLex (Segerer, 2016; Segerer & Flavier, 2013), a reference lexicon for the languages of Africa. The third is a speech-to-text (ASR) service tailored to a specific language. Each of these is a product in the business sense, and therefore, a resource. Each deals with language data, and therefore, is a language resource in the OLAC sense. Most importantly, each also qualifies as a service according to the DCMIType definition. They each perform at least one function. I suggest it is helpful for us to think of services as having clear input and output mechanisms which provide or act

[25] The example is found as a linked attachment to the document *OLAC Metadata* (Simons & Bird, 2008, §3). It is hosted at the following link: http://www.language-archives.org/OLAC/1.1/olac.xml. However, the reference example is problematic as the example does not validate in the OLAC validator found at: http://www.language-archives.org/tools/metadata/freestanding.html. A second example which does validate is found on the validator page. Further examples in this paper follow the validating example with one exception. The validating example uses ISO 639-1 language codes and throughout the OLAC documentation these codes are indicated as ISO 639-3 codes. My examples use ISO 639-3 codes. Differences between the two OLAC provided examples are primarily in XML syntax.

https://doi.org/10.23106/dcmi.953120677

upon data, and are actively operational. In this way we maintain a clear distinction between the DCMITypes "Service" and "InteractiveResource".[26] Each of the following examples also meets these additional criteria.

4.1 GoogleTranslate

As an interface, the consumer puts data in and gets data out. They provide data in one language and receive language data in another language. Quality may vary. The service maintains continuity of user experience and access to the features of the service (front end), while the back end (the technical components determining the results) of the service may change. For example, the translation service may use a variety of statistical models to provide "the most appropriate" translation pairs. Or as more recently has been done, they may switch the service to using a Multilingual Neural Machine Translation System. Service providers may keep the training data and translation models inaccessible as they are independent language resources. Ultimately, the translation model and the neural network changes over time, while the service and the service gateway remain fairly stable from the end-user's perspective. The accessible and interactive unit is the service. It is a language resource rather than a library service. So how could this service be advertised via OLAC?

```
<olac:olac
    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">
    <dc:title>GoogleTranslate</dc:title>
    <dcterms:created xsi:type="dcterms:W3CDTF">2006</dcterms:created>
    <dc:creator xsi:type="olac:role" olac:code="author">Google, Inc.</dc:creator>
    <dc:creator xsi:type="olac:role" olac:code="developer">Och, Franz Josef</dc:
        creator>
    <dc:description>Google's free service instantly translates words, phrases,
        and web pages between English and over 100 other languages.</dc:
        description>
    <dc:language xsi:type="olac:language" olac:code="eng"/>
    <dc:subject xsi:type="olac:linguistic-field" olac:code="
        translating_and_interpreting"/>
    <dc:type xsi:type="dcterms:DCMIType">Service</dc:type>
    <dc:publisher>Google</dc:publisher>
    <dc:identifier xsi:type="dcterms:URI">https://translate.google.com</dc:
        identifier>
    <dc:subject xsi:type="olac:language" olac:code="eng">English</dc:language>
    <dc:subject xsi:type="olac:language" olac:code="deu">German</dc:language>
</olac:olac>
```

Example 2: An OLAC record for GoogleTranslate using DCMIType: Service.

In Example 2, I provide the code for what an OLAC record might look like for GoogleTranslate. For brevity I have not included all 109 languages this record should contain. Also for brevity, the example record for GoogleTranslate does not contain a `<dcterms:medium>` element nor does it contain a `<dcterms:format>` element. The `<dcterms:format>` element would be useful to indicate which of GoogleTranslate's available channels is being referenced: API, HTML (web), iOS app, Android app. Given DCMS's one-to-one record principle, each channel should be listed independently and then related with `<dc:relation>` (Hillmann, 2005, §1.2; Urban, 2010).[27] Encoding *relation*, *medium*, and *format* in examples is not demonstrated due to space constraints.

4.2 RefLex

RefLex is a service hosted by CNRS-LLACAN which enables users to load wordlists (or work from existing wordlists) to create and track cognate sets (and the hypotheses correspondence sets these associations represent) across wordlists in different languages. It is a value added service for engaging with language artifacts. It is an

[26] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dcmitype/InteractiveResource

[27] https://www.dublincore.org/resources/glossary/one-to-one_principle

https://doi.org/10.23106/dcmi.953120677

interface to collections of language resources. Users can sub-divide those resources, choose certain artifacts to create custom subsets, then test hypotheses, and create their own relationships between lexical evidence. In this digital environment, RefLex hosts language resources which are compiled from source documents, such as from published wordlists, dictionaries, or archived language documentation materials. However, RefLex also allows its users to create internal reconstruction hypotheses and broad cross-language hypotheses using the historical-comparative method. These "reconstructed" forms are also language resources and are not always available elsewhere. RefLex is therefore more than a gateway to or repository of language resources (though, in a sense it is this as well, but it does not accept preservation responsibility as an archive would). It is a service which allows a user to interact with language resources and create derivative resources—fulfilling input/output requirements common to services.

In Example 3, I provide an OLAC record for RefLex using the DCMIType "Service".

```
<olac:olac
    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">
    <dc:title>RefLex</dc:title>
    <dcterms:created xsi:type="dcterms:W3CDTF">2011</dcterms:created>
    <dc:creator xsi:type="olac:role" olac:code="author">Segerer, Guillaume</dc:
        creator>
    <dc:creator xsi:type="olac:role" olac:code="developer">Flavier, Sbastien</dc:
        creator>
    <dc:description>The RefLex Project aims at providing the scientific community
         with a Reference Lexicon of the languages of Africa, as well as various
        tools to exploit it.</dc:description>
    <dc:description>Le projet RefLex a pour objectif ...</dc:description>
    <dcterms:abstract>The aim of the RefLex project...</dcterms:abstract>
    <dc:subject xsi:type="olac:linguistic-field" olac:code="
        historical_linguistics"/>
    <dc:subject xsi:type="olac:linguistic-field" olac:code="phonology"/>
    <dc:subject xsi:type="dcterms:LCSH">Historical linguistics</dc:subject>
    <dc:subject xsi:type="dcterms:LCSH">Lexicography</dc:subject>
    <dc:subject>Cognate Sets</dc:subject>
    <dc:type xsi:type="olac:linguistic-type" olac:code="lexicon"/>
    <dc:type xsi:type="dcterms:DCMIType">Service</dc:type>
    <dc:publisher>Le laboratoire Langage, Langues et Cultures d'Afrique (CNRS-
        LLACAN)</dc:publisher>
    <dc:identifier xsi:type="dcterms:URI">http://reflex.cnrs.fr</dc:identifier>
    <dc:language xsi:type="olac:language" olac:code="eng">English</dc:language>
    <dc:language xsi:type="olac:language" olac:code="fra">French</dc:language>
    <dc:coverage xsi:type="dcterms:TGN">Africa</dc:coverage>
    <dcterms:spatial>Africa</dcterms:spatial>
</olac:olac>
```

Example 3: An OLAC record for RefLex using DCMIType: Service.

4.3 ASR Service

A third kind of service which is becoming more common in academic linguistic work is the use of automatic speech recognition software. In these types of software, a consumer provides audio data as input and receives transcribed data, segmented data, or a sentiment analysis (a type of annotation) in return as an output. The consumer may provide a language specific model, or the model may be provided as part of the service. Popular packages in academic work include *Montreal Forced Aligner* (McAuliffe et al., 2017), *Persephone* (Michaud et al., 2018), and *Kaldi* (Povey et al., 2011).[28] The *Montreal Forced Aligner*[29] and *SANTLR*[30] (Li et al., 2020) are two such ASR systems. Neither in their default settings is set up for a specific language, but both can be

[28] https://kaldi-asr.org

[29] https://montreal-forced-aligner.readthedocs.io

[30] https://www.dictate.app

https://doi.org/10.23106/dcmi.953120677

tuned for use with a specific language. Commercially, services like IBM's Watson provide sentiment analysis as an output. Commercial services are also tuned for specific languages. Due to space constraints I don't model an ASR service in an OLAC record here. But we can clearly see that it is feasible to deploy ASR services for specific languages, which could be added to OLAC as language resources. Unlike other language resource services, ASR services are tuned. They are not just tuned in their input (oral language) but tuned in their output which can include script. Therefore, indicating the script (or orthography version) a service is compliant with becomes important to an accurate bibliographic record.

5 Discussion

In working through the creation of these records, there are three areas of possible change to the OLAC application profile and associated services which would make the search experience more amenable to end-users of the OLAC aggregator. These three areas include: *Metadata Quality Assessment*, *Negotiating Subject Language*, and *Indicating Scripts & Metadata Language*.

5.1 Metadata Quality Assessment

The OLAC validator (a distinct service from the OLAC aggregator, but used by submitters to the aggregator) contains a pre-submission indicator of metadata quality. Post-submission, an archive is evaluated on the basis of the depth and breadth of their descriptive elements in records (Hughes, 2004).[31] In the case of the GoogleTranslate example, adding the DCMS elements *Relation*, *Medium*, and *Format*, would bring OLAC's metadata record quality evaluation to 9 on the 10 point scale used. OLAC's metadata indicator is flat, in the sense that it does not vary its analysis of a record's descriptive quality by assessing the type of record being parsed. Using the DCMIType indicator to determine the type of record being parsed would present a more useful tool to technologists using the validator, e.g., records for services or collections do not need all the same DCMS elements that other kinds of records should contain. Tailoring the metadata evaluation process for provided data to include record profiles that include variation by DCMIType would provide a more intricate view into the kinds of useful or problematic metadata providers are offering to OLAC. For example, a record with a DCMIType of "Service" may not have a legitimate *OLAC Linguistic Data Type Vocabulary*[32] value. In the current OLAC evaluation system, this prevents records like GoogleTranslate's from receiving a 10 out of 10 rating. This kind of penalty should be addressed by tailoring record quality assessment by DCMIType.

5.2 Negotiating Subject Language

The OLAC application profile provides the opportunity to use `<dc:subject>` to specify the language a resource is about using ISO 639-3.[33] This relationship between the resource and a language is in contrast with `<dc:language>` which is used to indicate the language a resource is in. For example, a linguistic grammar is usually written in one language about another. The OLAC application profile allows both languages to be captured in the metadata. Services like GoogleTranslate may not have an immediately clear subject-language/language distinction. Especially when user interfaces are reduced to icons and empty text fields on web page forms. However, if we abstract the distinction in the OLAC application profile a bit from the book-like grammar example, we can ask "what language is this service engaged with?" (input) and "what language are we presented with?" (output). Languages of the structure of the service are clearly within scope of `<dc:language>` (languages in the User Interface). I suggest that language content outside of the structure of the service is best described with `<dc:subject>` (output). This leaves an open question concerning the best way to indicate the language of input to the service. In the case of GoogleTranslate, all the input languages are also output languages, but this is not a universal attribute of services. It seems for now, considering the designed limitations of the OLAC application profile, that input languages should also be indicated as subject languages, i.e., using `<dc:subject>`. In the case of RefLex, we would include French and English via `<dc:language>` tags for its user interface, while indicating about 1000 languages via `<dc:subject>` tags.

[31] http://www.language-archives.org/NOTE/metrics.html

[32] http://www.language-archives.org/REC/type.html

[33] https://iso639-3.sil.org

https://doi.org/10.23106/dcmi.953120677

5.3 Indicating Scripts & Metadata Language

For all language resources, including services, indicating the language used within fields of the bibliographic record and the scripts used within both the bibliographic record and the actual resource is important. Both use-cases are addressable by using two different techniques within existing structures of XML and Dublin Core. However, within the OLAC framework it is not currently possible to indicate in which script a record's metadata is written (or search for resources by script of bibliographic record) due to the limits of the application profile, validator, and aggregator. For example, XML[34] allows one to indicate via the `lang` tag that a work is written in *en-GB* or *en-US*, or in traditional Chinese characters *zh-TW* (as used in Taiwan) vs. modern characters *zh-CN* (as used in Mainland China). The XML `lang` tag relies on BCP-47 codes (Philips & Davis, 2009) which include script tags from ISO 15924.[35] Implantation of the XML `lang` is application specific and applications may restrict values to some subset of BCP-47. The OLAC validator does not currently allow for `lang` attributes on elements, e.g., titles, abstracts, descriptions.[36]

In addition to the XML `lang` tag's use of BCP-47, the `<dc:language>` can use BCP-47 tags indicating script when syntax encoding scheme from IETF-RFC5646 is indicated.[37] BCP-47 currently points to RFC5646 as the best practice.[38] For OLAC participants, portions of RFC5646 are going to be familiar, such as its use of ISO 639-3 which the OLAC application profile uses to qualify the `<dc:language>` and `<dc:subject>` elements. However, RFC5646 does not always allow for the use of ISO 639-3 as the IANA *Language Subtag Registry* needs to be consulted for grandfathered tags.[39]

Expanding the OLAC validator and the OLAC application profile to include the XML `lang` tag for metadata and the BCP-47 tag for the `<dc:language>` and `<dc:subject>` elements would give a way for OLAC aggregator end-users to search and filter by scripts. The Dublin Core comment for `<dc:language>` already mentions that using BCP-47 is a best practice.[40]

6 Conclusion

Language Resource Services such as those presented in this paper create derivative resources via transformations on the input data. Often those transformations are optimized for specific languages. In this way, they are similar to Phipps et al. (2006) two service types: *Transformation* (safe and collection-specific), and *Crosswalking* (schema and vocabulary). While they were specifically considering the needs of libraries engaged in digital activities and specifically addressing "metadata" curation needs, "metadata" is just another kind of data. Therefore, their categories apply equally well to language data and thereby also language resources.

I have shown that language resources may be described using "Service" when applying the DCMIType vocabulary. The utility of the "Service" type and its utilization in the OLAC aggregator allows for the rapid discovery of language resource services. A greater inclusion of service-typed language resources within OLAC can lead to a greater understanding of digital vitality in small and minority languages. These types of records provide an quantitative perspective on the opportunity that researchers and communities have to use their languages in digital contexts. Additionally, clearly distinguishing services along the ten part typology presented in section 2.1 allows greater insight and the setting of appropriate expectations as scholars investigate the nature of accessible language resources.

Finally, the broader Dublin Core community of users benefits from clear examples illustrating DCMITypes. Within the linguistics community there exists a plethora of resources to illustrate the three way distinction between the DCMIType values: "InteractiveResources", Software", and "Service". Catalogers and indexers benefit from the additional discussion drawing out the distinguishing features of these three under-discussed DCMITypes.

[34] XML 1.1: https://www.w3.org/TR/2006/REC-xml11-20060816/#sec-lang-tag XML 1.0: https://www.w3.org/TR/xml/#sec-lang-tag

[35] https://www.unicode.org/iso15924

[36] The OLAC Metadata Usage Guidelines does show examples with the XML `lang` tag in on the language element (see also discussion in section 2). However, the usage example uses ISO 639-3 codes rather than BCP-47 codes. http://www.language-archives.org/NOTE/usage.html#language

[37] http://tools.ietf.org/html/rfc5646

[38] BCP-47 is a static reference while RFC5646 is a specification. The static reference may through time point to different specifications.

[39] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry

[40] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/elements/1.1/language

https://doi.org/10.23106/dcmi.953120677

References

Adlington, J. (2018). Linguistics and language behavior abstracts (LLBA). *The Charleston Advisor*, *19*(4), 24–27. https://doi.org/10.5260/chara.19.4.24

Apps, A. (2004). A registry of collections and their services: From metadata to implementation. *DC-2004–Shanghai Proceedings*. Retrieved June 8, 2021, from https://dcpapers.dublincore.org/pubs/article/view/771

Apps, A. (2005). A middleware registry for the discovery of collections and services. *NCeSS2005 : The First International Conference on e-Social Science, Manchester (UK), 22-24 June 2005*. Retrieved June 8, 2021, from http://eprints.rclis.org/6977

Apps, A. (2006). Disseminating service registry records. In M. Dobreva & B. Martens (Eds.), *ELPUB2006: Proceedings of the tenth international conference on electronic publishing - digital spectrum: Integrating technology and culture, bansko, bulgaria, 14-16 june 2006* (pp. 37–47). FOI-COMMERCE Sofia. Retrieved April 16, 2021, from http://eprints.rclis.org/7733

Aristar-Dry, H., & Simons, G. (2006). *Good, better, and best practice: The experience of the e-MELD project in developing and evaluating digital archiving recommendations* (Presentation Slides) [Paper presented at the Meeting of the Deutsche Gesellschaft für Sprachwissenschaft]. https://scholars.sil.org/sites/scholars/files/gary_f_simons/presentation/dgfs_2006.pdf

Balatsoukas, P., Rousidis, D., & Garoufallou, E. (2018). A method for examining metadata quality in open research datasets using the OAI-PMH and SQL queries: The case of the dublin core 'subject' element and suggestions for user-centred metadata annotation design. *International Journal of Metadata, Semantics and Ontologies*, *13*(1), 1–8. https://doi.org/10.1504/IJMSO.2018.096444

Bird, S., & Simons, G. F. (2001). The OLAC metadata set and controlled vocabularies. In T. DeClerck, S. Krauwer, & M. Rosner (Eds.), *Proceedings of ACL/EACL workshop on sharing tools and resources for research and education* (pp. 7–18). EACL-ACL; elsnet. https://www.aclweb.org/anthology/W01-1506

Bird, S., & Simons, G. F. (2003). Extending dublin core metadata to support the description and discovery of language resources. *Computers and the Humanities*, *37*(4), 375–388. https://doi.org/10.1023/A:1025720518994

Bird, S., & Simons, G. F. (2004). Building an open language archives community on the DC foundation. In D. Hillmann & E. L. Westbrooks (Eds.), *Metadata in practice*. American Library Association.

Bird, S., Simons, G. F., & Chu-Ren, H. (2001). The open language archives community and asian language resources. *Proceedings of the Workshop on Language Resources in Asia, 6th Natural Language Processing Pacific Rim Symposium (NLPRS), Tokyo, November 27-30, 2001*. http://arxiv.org/abs/cs/0110014

Broeder, D., Windhouwer, M., van Uytvanck, D., Trippel, T., & Goosen, T. (2012). CMDI: A component metadata infrastructure. *Proceedings of Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR (A Worshop at LERC 2012)*, 1–4. http://www.lrec-conf.org/proceedings/lrec2012/index.html

Broeder, D., & Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, *1*(2), 119–132. https://doi.org/10.1504/IJMSO.2006.011008

CLARIN ERIC. (2020). *CLARIN services in the European Open Science Cloud (EOSC)* [Released via YouTube: 2020-11-16]. CLARIN ERIC. Retrieved June 1, 2022, from https://youtu.be/YvZ9Y_uyr7M

CMDI Taskforce. (2016). *Component metadata infrastructure (CMDI) component metadata specification – version 1.2* (Metadata Specification CE-2016-0880). CLARIN. Retrieved July 20, 2020, from https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf

Desipri, E., Gavrilidou, M., Labropoulou, P., Piperidis, S., Frontini, F., Monachini, M., Mapelli, V., Francopoulo, G., & Declerck, T. (2015). *Documentation and User Manual of the META-SHARE Metadata Model* (P. Labropoulou, Ed.; v3.1). META-SHARE. http://www.meta-share.org/assets/pdf/META-SHARE_Documentation_v3.1.pdf

Dobrin, L. M., & Ross, D. (2017). The IATH ELAN text-sync tool: A simple system for mobilizing ELAN transcripts on- or off-line. *Language Documentation & Conservation*, *11*, 94–102. http://hdl.handle.net/10125/24726

ELRA. (2015). *What is a language resource?* [Association website]. Retrieved May 26, 2022, from http://www.elra.info/en/about/what-language-resource

Federmann, C., Giannopoulou, I., Girardi, C., Hamon, O., Mavroeidis, D., Minutoli, S., & Schröder, M. (2012). META-SHARE v2: An open network of repositories for language resources including data and tools. In N. Calzolari, K. Choukri, T. DeClerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language*

https://doi.org/10.23106/dcmi.953120677

resources and evaluation (LREC'12) (pp. 3300–3303). European Language Resources Association (ELRA). https://aclanthology.org/L12-1483

Francis, W. N., & Kučera, H. (1961). *Brown corpus of standard american english*. Brown University.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Monachini, M., Frontini, F., Declerck, T., Arranz, V., & Mapelli, V. (2012). The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In N. Calzolari, K. Choukri, T. DeClerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 1090–1097). European Language Resources Association (ELRA). https://aclanthology.org/L12-1593

Greenhill, S. J., Blust, R., & Gray, R. D. (2008). The austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, *4*, EBO.S893. https://doi.org/10.4137/EBO.S893

Guenther, R. (Ed.). (1999). *Dublin Core Resource Types list*. Library of Congress. Retrieved April 11, 2022, from https://www.loc.gov/marc/dc/typelist-19991118.html

Guenther, R. (Ed.). (2000). *Dublin Core Type WG - Dublin Core Subtype Vocabulary*. Library of Congress. Retrieved April 11, 2022, from https://www.loc.gov/marc/dc/subtypes-20000612.html

Hammarström, H. (2015). Glottolog: A free, online, comprehensive bibliography of the world's languages. In E. Kuzmin (Ed.), *Proceedings of the 3rd international conference on linguistic and cultural diversity in cyberspace* (pp. 183–188). UNESCO. Retrieved March 29, 2022, from https://pure.mpg.de/rest/items/item_2354764/component/file_2354763/content

Hillmann, D. I. (2005). *Using Dublin Core*. Dublin Core Metadata Initiative. Retrieved May 31, 2022, from https://www.dublincore.org/specifications/dublin-core/usageguide/#whatis

Hirt, C., Simons, G. F., & Spanne, J. (2009). Building a MARC-to-OLAC crosswalk: Repurposing library catalog data for the language resources community. *Proceedings of the 2009 Joint International Conference on Digital Libraries - JCDL '09*, 393–394. https://doi.org/10.1145/1555400.1555479

Hughes, B. (2004). Metadata quality evaluation: Experience from the open language archives community. In Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, & E.-p. Lim (Eds.), *Digital libraries: International collaboration and cross-fertilization* (pp. 320–329). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30544-6_34

Hutt, A., & Riley, J. (2005). Semantics and syntax of dublin core usage in open archives initiative data providers of cultural heritage materials. *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05*, 262–270. https://doi.org/10.1145/1065385.1065447

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., & Wright, J. (2014). The language application grid. In N. Calzolari, K. Choukri, T. DeClerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (lrec'14)* (pp. 22–30). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/926_Paper.pdf

Kleiber, E., Berez-Kroeker, A. L., Chopey, M., Yarbrough, D., & Shelby, R. (2018). Making pacific languages discoverable: A project to catalog the university of hawai'i at mānoa library pacific collection by indigenous languages. *The Contemporary Pacific*, *30*(1), 110–122. https://doi.org/10.1353/cp.2018.0005

Kurtz, M. (2010). Dublin core, DSpace, and a brief analysis of three university repositories. *Information Technology and Libraries*, *29*(1), 40–46. https://doi.org/10.6017/ital.v29i1.3157

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig, G., Black, A. W., & Metze, F. (2020). Universal phone recognition with a multilingual allophone system. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8249–8253. https://doi.org/10.1109/ICASSP40776.2020.9054362

Lynch, C. A. (1991). The z39.50 information retrieval protocol: An overview and status report. *ACM SIGCOMM Computer Communication Review*, *21*(1), 58–70. https://doi.org/10.1145/116030.116035

Lynch, C. A. (1997). The z39.50 information retrieval standard: Part i: A strategic view of its past, present and future. *D-Lib Magazine*, *April*. Retrieved January 6, 2022, from https://www.dlib.org/dlib/april97/04lynch.html

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Michailovsky, B., Michaud, A., & Guillaume, S. (2011). A simple architecture for the fine-grained documentation of endangered languages: The LACITO multimedia archive. *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, 14–23. https://doi.org/10.1109/ICSDA.2011.6085973

Michaud, A., Adams, O., Cohn, T. A., Neubig, G., & Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit. *Language Documentation & Conservation*, *12*, 393–429. http://hdl.handle.net/10125/24793

Nordhoff, S., & Hammarström, H. (2012). Glottolog/langdoc: Increasing the visibility of grey literature for low-density languages. *Proceedings of the 8th International Conference on Language Resources and Evaluation [LREC 2012]*, 3289–3294. Retrieved March 29, 2022, from https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_1752643

Odijk, J. (2016). Linguistic research using CLARIN. *Lingua*, *178*(4), 1–4. https://doi.org/10.1016/j.lingua.2016.04.003

Odijk, J. (2019). Discovering software resources in CLARIN. In I. Skadin & M. Eskevich (Eds.), *Selected papers from the CLARIN annual conference 2018, pisa, 8-10 october 2018* (pp. 121–132). LiU Electronic Press. https://ep.liu.se/en/conference-article.aspx?series=&issue=159&Article_No=13

Park, J.-r., & Childress, E. (2009). Dublin core metadata semantics: An analysis of the perspectives of information professionals. *Journal of Information Science*, *35*(6), 727–739. https://doi.org/10.1177/0165551509337871

Park, J.-r., & Tosaka, Y. (2010). Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability. *Information Technology and Libraries*, *29*(3), 104–116. https://doi.org/10.6017/ital.v29i3.3136

Paterson, H. J., III. (2021a). *From CV to OLAC* (Presentation Abstract & Video from *The 7th International Conference on Language Documentation & Conservation (ICLDC)*, 4–7 March 2021). https://hughandbecky.us/Hugh-CV/talk/2021-from-cv-to-olac

Paterson, H. J., III. (2021b). *OLAC nightly data dump (XML) from 18 July 2021*. Zenodo. https://doi.org/10.5281/zenodo.5112131

Paterson, H. J., III. (2022). Where have all the collections gone?: Analysis of OLAC data contributors' use of DCMIType 'collection'. *Proceedings of the 15th Annual Society of American Archivists Research Forum*. https://www2.archivists.org/am2021/research-forum-2021/agenda#peer

Phelps, T. E. (2012). An evaluation of metadata and dublin core use in web-based resources. *Libri*, *62*(4), 326–335. https://doi.org/10.1515/libri-2012-0025

Philips, A., & Davis, M. (2009). *Tags for identifying languages*. Internet Engineering Task Force (IETF) - Network Working Group. https://tools.ietf.org/html/bcp47

Phipps, J., Hillmann, D. I., & Paynter, G. W. (2006). Orchestrating metadata enhancement services: Introducing lenny. *International Journal of Metadata, Semantics and Ontologies*, *1*(3), 189–197. https://doi.org/10.1504/IJMSO.2006.012343

Piperidis, S. (2012). The META-SHARE metadata schema for the description of language resources. In N. Calzolari, K. Choukri, T. DeClerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 36–42). European Language Resources Association (ELRA). https://aclanthology.org/L12-1647

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The kaldi speech recognition toolkit [IEEE 2011 Workshop on Automatic Speech Recognition and Understanding]. *Proceedings of Automatic Speech Recognition and Understanding*. https://infoscience.epfl.ch/record/192584

Pride, K., Tomlin, N., & AnderBois, S. (2020). LingView: A web interface for viewing FLEx and ELAN files. *Language Documentation & Conservation*, *14*, 87–107. Retrieved May 25, 2022, from http://hdl.handle.net/10125/24916

Roberts, J. (2001). Between a rock and a hard place: Dealing with NZGLS development issues. *DC-2001–Tokyo Proceedings*, 278–280. https://dcpapers.dublincore.org/pubs/article/view/683.html

Roberts, J. (2002). Describing services for a metadata-driven portal. *DC-2002–Florence Proceedings*, 165–169. Retrieved January 31, 2022, from https://dcpapers.dublincore.org/pubs/article/view/707

Segerer, G. (2016). RefLex: La reconstruction sans peine. *Faits de Langues*, *47*(1), 201–213. https://doi.org/10.1163/19589514-047-01-900000013

Segerer, G., & Flavier, S. (2013). *The RefLex project: Documenting and exploring lexical resources in africa* (Oral Presentation at *Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures*). http://hdl.handle.net/2123/9854

Shparberg, A. (2015). Linguistic bibliography online. *The Charleston Advisor*, *17*(2), 28–31. https://doi.org/10.5260/chara.17.2.28

Shparberg, A. (2022). Linguistics abstracts online. *The Charleston Advisor*, *23*(3), 32–36. https://doi.org/10.5260/chara.23.3.32

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practices for Federated Collections. In H. A. Thompson (Ed.), *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7-10 2005, Minneapolis, MN* (pp. 223–237). Association of College and Research Libraries. http://hdl.handle.net/2142/145

https://doi.org/10.23106/dcmi.953120677

Simons, G. F., & Bird, S. (2000). *The seven pillars of open language archiving: A vision statement* (Digital Manuscript). http://www.language-archives.org/documents/vision.html

Simons, G. F., & Bird, S. (2003). Building an open language archives community on the OAI foundation. *Library Hi Tech*, *21*(2), 210–218. https://doi.org/10.1108/07378830310479848

Simons, G. F., & Bird, S. (2008). *OLAC Metadata* (2008-05-31) [OLAC Standard]. http://www.language-archives.org/OLAC/metadata.html

Simons, G. F., Bird, S., & Spanne, J. (Eds.). (2008). *OLAC Metadata Usage Guidelines* (2008-07-11). Open Language Archive Community. Retrieved May 31, 2022, from http://www.language-archives.org/NOTE/usage-20080711.html

Toves, J. A., & Hickey, T. B. (2014). Parsing and matching dates in VIAF. *The Code4Lib Journal*, *26*, 9607. Retrieved March 17, 2021, from https://journal.code4lib.org/articles/9607

Urban, R. J. (2010). Principle violations revisiting the Dublin Core 1:1 Principle: Principle Violations Revisiting the Dublin Core 1:1 Principle. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–2. https://doi.org/10.1002/meet.14504701441

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The virtual language observatory. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 1029–1034). European Language Resources Association (ELRA). Retrieved January 9, 2022, from http://www.lrec-conf.org/proceedings/lrec2012/pdf/437_Paper.pdf

Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardellini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)* (pp. 900–903). European Language Resources Association (ELRA). Retrieved January 10, 2022, from https://aclanthology.org/L10-1187

Ward, J. H. (2002). *A quantitative analysis of dublin core metadata element set (DCMES) usage in data providers registered with the open archives initiative (OAI)* (M.S. in Information Science). University of North Carolina at Chapel Hill. Chapel Hill, NC, USA. https://ils.unc.edu/MSpapers/2816.pdf

Ward, J. H. (2003). A quantitative analysis of unqualified dublin core metadata element set usage within data providers registered with the open archives initiative. *2003 Joint Conference on Digital Libraries, 2003. Proceedings.*, 315–317. https://doi.org/10.1109/JCDL.2003.1204883

Ward, J. H. (2004). Unqualified dublin core usage in OAI-PMH data providers. *OCLC Systems & Services: International digital library perspectives*, *20*(1), 40–47. https://doi.org/10.1108/10650750410527322

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., . . . Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation* (Digital Manuscript) [Digital Manuscript, version: 1]. Retrieved January 4, 2022, from http://arxiv.org/abs/1609.08144

https://doi.org/10.23106/dcmi.953120677