Where Have all the Collections Gone? Analysis of OLAC Data Contributors' use of DCMIType 'Collection'

HUGH J. PATERSON III
Collaborative Researcher

Abstract: Language materials, as commonly conceptualized by academics, are resources which specifically exhibit or provide evidence of a naturally spoken language. The modern area of academic practice known as *language documentation* has its roots in anthropological linguistics but maintains a strong adherence to ideals which call for the archiving of source materials. The purpose for archiving is to benefit the many stakeholders involved in language development activities. Language archives, hosting language resources, have by and large adopted Dublin Core as a metadata standard along with the additional metadata terms of the Open Language Archive Community (OLAC) application profile as described in Bird and Simons (2001; 2003). This study is a first look at how the DCMIType "Collection" is used across aggregated records from language archives. This study finds that current practices of arrangement and description at language resource preservation institutions participating in OLAC do not currently follow archival best practices in arrangement and description as described in frameworks like *Describing Archives: A Content Standard* including honoring principles like *respect des fonds*. This has multiple impacts including consequences in web-based navigation and discoverability.[1]

Introduction

The ability to group items together for description, management, discovery, and presentation has been an indisputable necessity within archival practice. However, the classification, discovery process, description, management, and presentation process for these groupings of items has been the object of many discussions within the scholarly discourse of information science. In this study I report on how *Open Language Archive Community*[2] (OLAC) metadata contributors are and are not reporting on groupings of items according to current capabilities of the OLAC application profile. I do this by investigating the use of the DCMIType vocabulary[3] value "Collection" [4] as used to refine the Dublin Core *type* element.[5] Within Dublin Core, this is the most appropriate method to indicate that a record is about a group of items. The Open Language Archive Community was formed to assist newly established preservation organizations which focus on digital and digitized language resources. It is a domain-based collaboration of scholars with initial funding by the NSF. The collaboration formed a metadata standard (Bird & Simons 2001) and an exchange protocol based on OAI (Simons & Bird 2003) and Dublin Core (Bird & Simons 2003). The community of scholars also created and deployed an OAI harvester and aggregator (viewer) for parties who opted to share metadata about language resources via the OLAC metadata conventions.

[1] I would like to thank Matthew Lee who helped me craft XSLT code for isolating the target metadata attributes used in the analysis.

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

[3] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-7

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

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

Published by Society of American Archivists, May 2022.

Broadly, the term *language archives* may be conceived of as any **archival unit**[6] comprised of or containing language resources such as the *Manchu-Language Archives of the Qing Dynasty* (Elliott 2001) or an institution which stewards language resources, e.g., the *British Library*, the *UCLA Ethnomusicology Archive*,[7] Indiana University's *Archives of the Languages of the World*,[8] or the many tribal museums and libraries as discussed in Roy et al. (2011). More recently in the literature (Simons & Bird 2003; Hughes 2004; Wasson et al. 2016; Holton 2012; Seyfeddinipur et al. 2019; Yi et al. 2022) the term **language archives** has been used to refer to a class of stewardship institutions which have more recently been established specifically out of response to global paradigm shifts within the context of academic field-linguistics, language documentation, and linguistic anthropology.[9] The three greatest influences on paradigm shifts impacting field-linguistics in the last thirty years are: (1) the introduction of digital tools and workflows within the discipline, (2) the general acknowledgment of the moral duties (or fiduciary duties as language information stewards) to assist language communities in reaching their language use goals,[10] and (3) the perception that new institutions and archival units established and managed by linguists can better meet the needs of field practitioners independent of long-established resource stewardship organizations.

Language resource preservation and management in some senses is in its infancy with several dedicated digital language resource stewardship organizations being established in the 1990s and 2000s. Some examples of these newly established digital language resource stewards include the Open Language Archive Community members: *The Pangloss Collection of the Collection de Corpus Oraux Numériques* (Pangloss) which was established in 1994 (Thieberger & Jacobson 2010),[11] *SIL International's Language & Culture Archives* (L&CA) which was established in 1999 (Nordmoe 2018),[12] *The Pacific and Regional Archive for Digital Sources in Endangered Cultures* (PARADISEC) which was established in 2003 (Barwick 2003),[13] *Endangered Languages Archive* (ELAR) which was established in late 2004 (Munro & Nathan 2005),[14] and *Kaipuleohone* which was established in 2008 (Albarillo & Thieberger 2009).[15] Even though there has been a response in the scholarly community of linguistics to embrace the use of repositories and preservation institutions, the task of managing the life-cycle of language resources developed through scholar-engaged, community-oriented collaborations[16] has often fallen out of the

[6] **Archival unit** is a term of art from *Describing Archives: A Content Standard* (DACS) (Society of American Archivists 2013:§2.3.19). It is a generic term for a "set of resources". DACS also identifies *Papers*, *Collection*, and *Records* as terms of art with specific implications. They are all types of archival units. Innovative linguists have also created their own terminology for archival units. The three terms which come to mind are *assemblages*, *language documentation corpus*, and *corpus*. The nuances of the variation in these terms is discussed in Paterson (2021b:footnote 8).

[7] https://schoolofmusic.ucla.edu/facilities/ethnomusicology-archive

[8] https://media.dlib.indiana.edu/collections/2f75rd115

[9] Authors (such as Bird and Simons 2021) use the term **language archives** to reference all data providers to the *Open Language Archive Community* aggregator adding further ambiguous nuance to the term *language archives*. Not all data providers are in fact preservation institutions, a function and concept that the term *archive* has come to represent.

[10] For remarks on the latter, see Krauss' contribution in Hale et al. (1992).

[11] https://pangloss.cnrs.fr

[12] https://www.sil.org/resources/language-culture-archives

[13] https://www.paradisec.org.au

[14] https://www.elararchive.org

[15] https://scholarspace.manoa.hawaii.edu/handle/10125/4250

[16] *Language resources* is a term which can have variable meaning depending on its usage context. However, the kinds of materials preserved by OLAC participating institutions generally fall into the broad categories of *language description resources*, *language documentation resources*, and *language development resources*. These kinds of resources include: audio recordings and video recordings (each with their transcriptions); software; publications; field notes; photographs; databases; and datasets. Materials may be published or unpublished.

hands of academic linguists and onto the plates of stewardship organizations, which have been set up to, and usually exclusively to, steward language resources.[17]

Archival units support at least three essential functions in the service life-cycle of language resources, and therefore should be present at all stewardship institutions.[18] First, *archival units are important to contextual description of artifacts*. For example, institutionally adopted models of description (as described in DACS and other resources) often prescribe the kinds of things which need to be described at each layer of groupings, i.e., multilevel descriptions. These layers of description are then supported by software, tools, and encoding formats such as Encoded Archival Description (EAD).[19] EAD has a telescoping model where a general description applies to the whole archival unit and then for each component grouping in the archival unit, additional specific description elements are used. Progressive layers of description are available to describe individual items of merit.

Second, *archival units are important to the faithful management of resources.* Exercising stewardship responsibilities is an important part of trust building for cultural heritage institutions. Subdivisions of archival units may incur separate curation and preservation activities, have different provenances, or allow for different viewing audiences, i.e., access permissions. Each of these is a resource management activity at the stewardship institution. Sometimes the conflicting fiduciary loyalties that stewardship institutions face require them to subdivide archival units beyond the designs of the original arranger or collector. These management criteria point to the increased importance that archival units and their subunits have to stewardship institutions. If a stewardship institution removes subdivisions to fit their predefined organizational structure, then it raises interesting questions about how they conceive of their fiduciary duties given that the structure of complex resources is recognized by scholars such as Bartlett (1992) and Haworth (2001) as part of its provenance and part of the original creative work.

Finally, *archival units are important for discoverability purposes.* Finding and browsing resources by archival unit is a common way to navigate a large set of artifacts or records in many long-established stewardship institutions. Archival units and the group/item relationship is an important part of delivering contextual information to artifact discoverers. This is noted by Sullivant (2020:522) as he makes the case for what ought to be included in a well-formed description of archival units containing language resources. He states:

> Unless they [users] are following a direct link from somewhere else, users discover materials in a digital collection through browsing and searching. … Browsing can be difficult or time consuming in a collection with many subparts, especially if they are labeled in ways that do not indicate their contents.

In multi-institutional aggregator such as OLAC, the function of browsing by archival unit is especially important. The ability to drill down through an archival unit's constituents is immensely important in the browsing process of discovering resources. As a metadata-driven web-portal, OLAC web views depend on the presence of Dublin Core metadata elements and their refining terms. Archival units (*collections*, *papers*, *records*) as well as *aggregate works*, and *records for volume/issues of serials* are all semantically

[17] This arrangement between academic scholars, language communities, and archives raises interesting questions about to whom fiduciary loyalties are to be directed, to the scholars, or to the ethno-linguistic communities? For discussion see Brown (1998). If the primary loyalty is to the communities: who is a valid legal, moral, or authoritative representative of that community? See Patrick (2008) for a discussion of various definitions of "speech community". For the impact of the concept of speech or language community on language resource stewardship see Seyfeddinipur et al. (2019) and Nathan (2013).

[18] Wickett et al. (2014) point out additional functions of archival units. In their work they reference archival units as *collections*.

[19] https://www.loc.gov/ead

equivalent to the concept of the DCMIType "Collection". The importance of the DCMIType term "Collection" is to make the group/item distinction between a set of records clear.

Dublin Core elements may make use of different externally managed vocabularies, e.g., *ISO 639-3* to refine the *language* element, or *Library of Congress Subject Headings* to refine the *subject* element. The only vocabulary contained within the Dublin Core standard is the DCMIType vocabulary. In this study I looked at the presence and absence of the DCMIType "Collection" within the records aggregated to OLAC. A baseline investigation provides insights into how language resources might be more usefully interlinked. For example, one practical application of linking between records relates to collections, i.e., archival units, and published descriptions about their contents. Sullivant in his well-received 2020 paper *Archival description for language documentation collections* presents a framework to guide the writing of collection descriptions with the anticipation that they will be published as standalone articles in serials rather than included or integrated with archival unit contents. A good number of these papers are also indexed in OLAC. How then should these stand-alone works be indexed in OLAC? And further, how should they be related to archival unit materials? A baseline of usage for the DCMIType "Collection" can reveal current and past practice within the set of OLAC providers. Since the DCMIType vocabulary is present in the Dublin Core standard, it is available to all implementers of OLAC without any additional modifications to standards or application profiles.

The term "Collection" is unambiguously defined within the context of Dublin Core as follows:

Definition: *An aggregation of resources*. **Comment**: *A collection is described as a group; its parts may also be separately described.*

Even with a simple definition, there are several possible applications of the term "Collection" to records of entities represented in the OLAC application profile. First, is the distinction between a managed or curated archival unit and a dynamic aggregation of references to items such as one finds in search results for a query using a particular term. Pragmatically, it is often not useful to create archival records for dynamic sets of references. The second distinction would be the difference between an archival unit and an aggregate work.[20] Aggregate works are creative works with multiple parts. Examples may include an edited volume with individual works (chapters), an encyclopedia with multiple volumes, or a curriculum which is comprised of audio, video, and textual components. Language resource preservation institutions have both managed archival units and aggregate works. Therefore, it is expected that some ambiguity will be present in archival records accessed via the OLAC aggregator.

Burke and Zavalina (2020a) characterize the metadata practices at these organizations saying: "the field of linguistics is not up to date on best practices in information organization...". More recently, Burke et al. (2022) point out that many language resource stewardship organizations, but not all to the same degree, rely heavily on submitter-provided descriptions and metadata. Reports of variation, even within the records of a single language resource stewarding institution, are in-line with reports from other community and scholar driven repositories which rely on user-contributed metadata (especially those without metadata field-specific data-validation or the use of authority control records).

Preservation organizations stewarding language resource materials, and presenting them via OLAC are challenged to not only present the group/item dimension but also how it correlates with the dimension of records following the *Functional Requirements for Bibliographic Records* (FRBR) model (Tillett 2001, 2004; Riva, Le Bœuf & Žumer 2017). That is, records may represent the following entities: *Work*,

[20] An indirect way of using Dublin Core to distinguish between aggregate works and archival units is if archival unit records conformed to an application profile such as the *Dublin Core Collection Description Application Profile* https://www.dublincore.org/specifications/dublin-core/collection-description/collection-application-profile.

Expression, *Manifestation*, or *Item*. This second dimension invoking the FRBR entity model is motivated by Dublin Core's *1:1 Principle* as discussed in the Dublin Core guidelines in section 1.2.[21] Currently it is not clear that any OLAC data contributor is providing records with a faithful implementation of the Dublin Core 1:1 Principle. This is not unexpected as Park and Childress (2009:737) address the 1:1 Principle in the results of their survey reporting on the semantics of Dublin Core by interviewing professionals implementing Dublin Core. Their report indicates variation in the application of the 1:1 Principle; they summarize the situation saying, "there is a lack of studies focusing on how the one-to-one principle is reflected in the metadata creation process among cataloguing and metadata professionals as well as in actual DC metadata item records." The interview example they publish agrees with the sentiment expressed by Miller (2010) suggesting that curators, database administrators, or others responsible for the creation of records are not familiar with abstract models to which metadata usage is assumed to conform. Urban (2010) shows in his study that sixty-five percent of the repositories surveyed did not honor the 1:1 Principle, leaving records in a confusing state. The same kinds of record keeping behaviors discussed by Urban (2010) were found in OLAC data providers and reported on in my work (Paterson 2021b:§4.6.3; 159–163).[22]

The two relevant points related to not honoring the 1:1 Principle in record management are: (1) it makes navigating resources, their permutations, enhancements, and related resources a challenge for stakeholders within both the OLAC aggregator view and the web-portals of many of OLAC's participating data providers (where users are expected to acquire and interact with stewarded web resources);[23] (2) groups of related records along the FRBR dimension do not constitute a DCMIType "Collection".

Even though many of these newly established stewardship institutions with digital holdings have less than a thirty-year history, they collectively hold a rich repertoire of resources demonstrating the unique linguistic cultural heritage of many minority people groups around the world. My research takes a look at the practice of reporting archival unit records (commonly referred to as collection records) to a common record aggregator intended to increase the discoverability of resources across institutions. No one has yet reported on how newly established language resource stewardship institutions are sharing archival unit level records or if the contents of archival units have metadata which can be used to indicate inclusion in fonds or sub-fonds. For metadata-driven web experiences, the inclusion of archival unit level metadata via a distinct record can have significant impact on the usability of aggregation tools. Archival unit level records and metadata are important for several reasons including the identification of component members of the archival unit, the independent provenance of those component members, and the management practices applied to the whole archival unit.

[21] https://www.dublincore.org/specifications/dublin-core/usageguide/#whatis

[22] For additional discussion of the application of the 1:1 Principle within the cultural history preservation context see Wijesundara & Sugimoto (2018) and Sugimoto et al. (2018).

[23] In my review of web-portals associated with large language resource preservation institutions done as part of this study and previous work (Paterson 2021b), no institutional web-portal consistently provided an interface option which displayed links to, information about, or commentary on records for versions or the differences between *manifestations*. Popular pre-print repositories such as arxiv.org and bioarxiv.org both support the linking of manifestations. Of the OLAC data providers, SIL's L&CA was the only one which I have noticed to have occasional linking. Sometimes it does have links to or between components of an aggregate work, and sometimes it has links between editions of a work (editions are often considered variations at the *expression* level in the FRBR model). However, these relationships are not consistently applied to all resource records and some manifestations or editions which have records in the institutional catalogue may not have visible records in the public facing web-portal, e.g., the second edition of a work might have a record in the public web-portal, but the record for the first edition only appears in the institutional catalogue with staff-only viewing access.

Considering the broader interdisciplinary scope of general archiving, it is surprising that even less has been published on how the DCMIType "Collection" is used broadly among institutions that publish Dublin Core compliant metadata.[24] Therefore, this paper provides valuable insight to the broad archiving community into the under-discussed usage of the DCMIType "Collection". Equally important, this paper speaks to the degree that language archives have availed themselves of the total descriptive power of Dublin Core.

Literature Review

Literature about the quality of language archive metadata is sparse. Hughes (2004) presents a framework for evaluating a record's completion and the quality of a metadata feed to OLAC on the basis of the number of Dublin Core elements included by a provider. Hughes' work has been accompanied by more recent discussions including Burke and Zavalina (2019; 2020a; and 2020b) who attempt to measure the quality of unrestricted text fields of description elements. Burke and Zavalina (2019; 2020a) established that record composition for the free-text description field is used in various ways across the three OLAC participating archives they evaluated.[25] Some of these free-text description fields indicated that the object to which the record referred was more like a set of items/artifacts than a single item/artifact. However, archival unit records should have a different composition from individual artifact records because they each have distinct scopes. With this in mind, different record types should have distinct evaluation criteria when compared with artifact records. For example, archival unit records should link to the records of items in the archival unit, and thereby support the browsing of associated groups of items within an aggregator (Zavalina 2011). Wasson et al. (2016) and Paterson (2021b) both discuss the quality of language archive metadata pointing to resource management choices at individual archives as contributing causes of record variation. However, neither proposes any specific framework for ongoing quality analysis. Based on the needs of both linguists and language community members, Wasson et al. (2016) note that usability requirements are not met by language-archive records. Unexplored in the literature is how OLAC record providers are utilizing distinct archival unit records. Existing evaluations of OLAC records (Hughes 2004) do not take record types into account. Rather than treating distinct record types (as identified by the use of a vocabulary on the *type* element), Hughes simply counts the number of elements present in a record. This current study seeks to generate discussion about different ways to assess record quality within OLAC and other aggregator-centric communities by suggesting that different types of records might be better-described or under-described on the basis of their type. For example, an audio artifact record may be under-described at a certain quantitative threshold of elements but an archival unit record may be adequately described at the same quantitative threshold. This study does not try to propose any sort of metric for quality of the OLAC feed and reserves that for future work; rather, this study looks at the semantics of a particular DCMIType attribute as it was designed to be used and then investigates its actual use. In the sense that this study investigates a Dublin Core dataset for patterns of metadata usage, it is comparable to studies like Park (2006; 2009), Zavalina (2011), and Park and Childress (2009).

[24] I found no prior literature investigating or describing the use of DCMIType "Collection" across sets of Dublin Core records in multi-institutional aggregations. Wickett et al. (2014) addresses DCMIType "Collection" from a data model perspective but the focus of their work does not include reporting on use. Archivists in the USA speculated that while discussed in training programs, Dublin Core as a metadata standard is rarely used "in the field" at archival institutions. Some archivists have pointed out that some institutions hold archival unit records in one system and artifact records in another system. However, based on investigative work conducted as part of Paterson (2021b) I can say that the OLAC data providers contributing the most records do not operate split records management systems.

[25] In contrast to the current study which looks at archival records after they are passed to OLAC, Burke and Zavalina's work took records directly from the institutional websites.

Stakeholder interest in archival units containing language resources and the use of those language resources is significant. Kipp (2007) describes the use of physical materials in archives within the context of creating minority-language language-learning materials. Bow et al. (2014) describe a project to collect, digitize, and make accessible minority language resources (literacy resources) which were originally only available in print media. Thieberger and Harris (2022) discuss supporting the reuse of materials by ethnolinguistic communities through metadata enrichment via collaborations with community members and the support of geographically disparate museums via redistribution of copies of holdings. Holton (2012) and Woodbury (2014) both discuss the nature of the audiences of collections focused on stewarding language resources. They point out that stakeholders include ethno-linguistic community members to whom the language content in the resources is either part of their cultural past and/or part of their present context. Collections and artifacts are also of interest to archivists, academics (linguists, historians, anthropologists, etc.) and technologists. With specific regards to the role that the OLAC aggregator plays in pointing people to language resources, Bird and Simons (2021) report that OLAC gets over 8,000 record views a month with over 2,000 of those resulting in click-throughs to resource hosting web-portals.

Accessing resources, in general, remains a challenge for these stakeholders and is the impetus for work reported on in Wasson et al. (2016), Wasson et al. (2018), Burke et al. (2022), and Yi et al. (2022). User-centered design (as discussed by Wasson et al.) and considerations of audiences can only improve resource browsing in limited ways. That is, user-centered design methods can only address some of the felt constraints in archive user-interfaces, because of the divergent approaches of the audiences. These methods will never be able to address the desire to sovereignly determine available user-interface actions, nor will they be able to completely address the discontinuity between what user-interfaces offer and how different stakeholders approach the task of discovering language resources, e.g., by language, by contributor, by date, by technical quality or even by collection. No doubt, these more sociological issues, related to how web-engagements support felt needs, have facilitated the growth of the exhibit (collection presentation) software *Mukurtu* (Wiberg 2014; Christen et al. 2017; Ferreira et al. 2021),[26] which is a *Drupal install profile*.[27]

Institutions responding to the need for digital language resource stewardship early on during the initial stages of global paradigm shifts in the field of linguistics and language documentation often adopted the Open Language Archive Community (OLAC) metadata standards. OLAC metadata standards are an implementation of OAI (Simons & Bird 2003) and an extension of Dublin Core (Bird & Simons 2003).[28] The OLAC metadata extension of Dublin Core is an application profile (Hillmann and Phipps 2007) as defined by Heery and Patel (2000). This contrasts with Simons (2016) and Bird and Simons (2021) who indicate that OLAC still needs to develop an application profile.[29] However, this and other research support the idea that OLAC-participating data providers apply different abstract models to the generation of archival records while still using the same metadata elements. For example, Paterson (2021b:§4.6.3; 159–163) exemplifies and discusses how a record at *Kaipuleohone* conflates metadata about multiple manifestations into a single record. This means that the stewardship institution has not adopted a model of description which generates independent records for *Work*, *Expression*, *Manifestation*, and *Item*. This

[26] https://mukurtu.org

[27] https://www.drupal.org/docs/7/install/using-an-installation-profile

[28] https://www.dublincore.org/specifications/dublin-core

[29] An alternative interpretation is that OLAC should develop a Dublin Core Application Profile (DCAP) following the *Singapore Framework* as discussed by Žumer et al. (2010), but this is not clear. Simons (2016) does reference DCAP ideas and literature, but Bird and Simons (2021) use the term “general profile”. If it is the case that Bird and Simons are arguing for an FRBR aligned implementation of OLAC metadata, I fully concur. However, my interpretation of the evidence looked at in this study leads me to believe that it is not just the OLAC records which need to become FRBR aligned but the curation practices at the preservation institutions themselves which need to change.

impacts the composition of records (inclusive of relationships between records) when they are transmitted to an aggregator or relied upon for generating browsing experiences. If manifestations are not handled independently and are reduced to a common item-type record for aggregator, then a valid question arises: *Is the same type of conflation happening for archival unit records?* An equally important question follows: *If conflation occurs, does it occur as a result of stewardship management practice or is the conflation a technological anomaly on how the records were exported for aggregation via OLAC?* This research does not address either of these important questions, but rather attempts to articulate a basic understanding of what archival unit (collection) records exist in the OLAC aggregator.[30] However, Miller (2010:150–151) states in regards to record generation that:

> many local database and user interface systems do not have the capacity to adequately link separate records and to display them together in a clear and meaningful way for end users. This becomes a usability issue, and may drive the creation of single records instead of separate records for original and digital versions of a resource.

This suggests that software choices for curation tasks at preservation institutions is a contributing factor to the state of records. In the case of Kaipuleohone , they use DSpace—software originally designed to create FRBR work-level records for text-based materials. In its structure, materials are not by nature hierarchical but rather are held in a flat equivalence structure. Records can be related to each other creating a logical hierarchical structure using Dublin Core relationships. However, these metadata-declared relationships must be added by a curator after initial record creation, and therefore, are costly due to labor costs. Most language archives rely on submitters for curation tasks, but the field of linguistics has yet to find a way to incentivize or acknowledge the academic merit of curation activities (Weber 2021).

The OLAC metadata vocabularies (Bird & Simons 2001) have served two roles. First, they function as a template of sorts for those who are new to language material preservation and stewardship on the kinds of metadata which should be collected about language resources. Second, the metadata vocabularies serve to drive the interactions within the OLAC aggregator. The OLAC aggregator provides a single gateway experience to metadata from sixty-plus participating data providers which all use the OLAC metadata application profile (Bird & Simons 2021). In many ways the OLAC aggregator is like other multi-institutional aggregator such as *Europeana*,[31] the *National Science Digital Library* (NSDL),[32] the *Digital Public Library of America* (DPLA),[33] *Indiana Memory*,[34] the *Online Archive of California*,[35] and the *Ohio Digital Network* (and other DPLA hubs).[36] However, unlike the general and broad scope of cultural heritage topics within these other projects, OLAC participants and the OLAC application profile focus on *language-based* cultural heritage artifacts, description, and discovery.

Research on other Dublin Core data sets, which often are used to drive web-portals functioning as gateways to multi-collection or multi-institution aggregations, has shown that metadata from providers can vary greatly in several different dimensions including: syntax—data format (Hutt & Riley 2005), semantics—how an element's value is used (Zavalina et al. 2009; Palmer et al. 2010; Park & Tosaka 2010; Zavalina 2011), and the number or types of elements per record (Stvilia et al. 2004; Kurtz 2010). Lynch et al. (2020) report that eighty percent of the items in the *Illinois Digital Heritage*

[30] http://dla.library.upenn.edu/dla/olac/index.html and also http://search.language-archives.org.
[31] https://www.europeana.eu/en
[32] https://nsdl.oercommons.org
[33] https://dp.la
[34] https://digital.library.in.gov
[35] https://oac.cdlib.org
[36] https://ohiodigitalnetwork.org

Hub[37] did not use the Dublin Core *type* element. In contrast, as shown in Table 1, eighty-three percent of the OLAC records contain a DCMIType refinement to the Dublin Core *type* element. Discontinuity of metadata element syntax and semantics across providers and missing elements produce friction for end-users who are interested in cross-collection and cross-institution searches. This friction is the mismatch between provided results and end-user expectations. It is well known in the web-design industry that too much user-friction in a tool causes users to devalue and abort the use of that tool (Cooper 2004).

Methodology

In this study I report on records which are navigable via the OLAC aggregator. The nightly record dump from July 18th, 2021, on which this study is based, is persisted via Zenodo (Paterson 2021a). The XML data were then filtered via specific queries using XSLT. These same sorts of queries can be done using the OLAC faceted search features on the OLAC website. However, future investigations should consider the use of XSLT and persist their data sets for two reasons: First, OLAC data is constantly in flux as archives receive new accessions and metadata on existing artifacts and as collections are updated. By persisting the nightly version used, future researchers can work with the same data set to verify claims and build on those observations. Second, the use of XSLT can allow for more complex queries of the data than using the faceted browsing of the OLAC website.[38]

Given the total lack of prior work in the area of DCMIType vocabulary usage among OLAC data providers, I sought to establish a baseline knowledge which could be used to describe what the usage choices of DCMIType "Collection" looks like across the OLAC dataset. In this investigation I sought to answer several questions regarding the OLAC data set, both across the whole data set and specifically with regards to the application of the DCMIType term "Collection".

1. What was the quantity of the total number of records and providers both across the whole data set and with regard to the use of the DCMIType "Collection"?
2. What are the semantics of when DCMIType "Collection" is used? That is, is it applied to archival unit records, records of aggregate works, some other context, or misapplied?
3. Can I estimate how many collections might be in the dataset but not explicitly marked? To do this, a general search was conducted for the term "Collection" across all fields.[39]

Additionally, I relied on my professional knowledge of the semantics embedded in the URL structures at various language archives. Some archives have archival unit identifiers in their URLs; these can be used to detect clusters of records which belong to an implicit archival collection.

In my work I did not look at Dublin Core relation elements. Dublin Core "hasPart" relation term (or its inverse "IsPartOf") can be used to infer that a record is referring to the existence of an archival unit record which does not exist in the aggregator or that an item record is in fact a record of a multi-part set (such as an aggregate work or archival unit). I did not systematically check for networks of records using relation elements across OLAC records. That being said, I do not recall seeing any relation terms across the several hundred records I have manually checked.

Other OAI based record aggregating projects such as NSDL and DPLA are known to represent archival units in the navigation structures of their web-portal aggregators. Comparing OLAC with other OAI

[37] https://idhh.dp.la

[38] Due to security investigations at the host provider, the OLAC website was down for three months in 2021. As far as I know this is the first time the OLAC website was unavailable in the near twenty years of continuous operation.

[39] This was done case insensitively.

aggregations is challenging because allowances need to be made for variations in terminology and abstract models. For example, NSDL calls data providers “Collection” and DPLA calls them “ore:Aggregation”. These providers may actually be better described using the DCMIType “Service”. Within the OLAC context these providers are often called *archives*, though in this paper I have chosen the more neutral term *data provider*. The DPLA abstract model is based off of the Europeana Data Model (Clayphan et al. 2017) and is discussed by Wickett et al. (2014). According to the metadata documentation of other multi-institutional aggregations such as DPLA hubs, it seems that they use the inferred collection method and list items without directly creating records for the collections in which items are a component (Indiana Memory Project 2020; Ohio Digital Network 2021).[40] Actual use of DCMIType “Collection” within DPLA is not clear. The *DPLA Metadata Application Profile* does call for a class for collections using the DCMIType “Collection” (DPLA MAP 2017:15) and as such seems to indicate an overt record may be possible, due to the documentation including a *dcterms:description* field within the class. How would this element become available to the aggregator without an overt record? However, a DPLA contributor, the *Ohio Digital Network* states in their resources on recommended fields: “Due [to] DPLA’s collection development guidelines, they do not want the DCMI type “collection” used.”[41]

Approaches such as those taken in DPLA may become problematic, though, when seeking to implementing the DACS principle of *respect des fonds* (Society of American Archivists 2013:xvi, xviii [Principle 2 & 7]). This becomes obvious when one is seeking to display archival unit descriptions for each node in the arrangement architecture. This is acknowledged in the Europeana Foundation's special report on hierarchical objects (Bardi et al. 2014). Middle layer groupings have no guarantee of existing even if each node in the fond (middle layers) is listed with an “IsPartOf” relationship. The issue is that no description element or information will be accessible to the aggregator because these are inferred and have no description information associated with them.

Results

The results are presented in four sections: A general overview of the whole data set in this section, a categorization of overtly indicated collections, a comparison of data providers to the OLAC data set and their use of the DCMIType “Collection”, and an estimation of implicit (unidentified) collections within the OLAC data set.

A summary of the OLAC data set as persisted (Paterson 2021a) is summarized in Table 1. Per the data set, the Open Language Archives Community (OLAC) aggregator compiles 443,217 records from sixty-five providers. Participating archives each provide Dublin Core metadata via an OAI feed. Only 369,520 of the records (83.3 percent) include a DCMIType indicator.

Only eight data providers (12 percent) even use the “Collection” DCMIType. Across the OLAC records, 850 (0.19 percent) use the DCMIType “Collection”. By using the DCMIType “Collection” and relating artifact records with archival unit records via the Dublin Core “hasPart” property, more about the original context of the archival unit is transferred from the host institution to the OLAC aggregator. When properly displayed, this can lead to increased utility in browsing environments.

[40] https://ohiodigitalnetwork.org/elements/collection

[41] https://ohiodigitalnetwork.org/contributors/getting-started/recommended-fields

65	Data providers share data catalogues of language and culture resources via a central aggregator.
443,217	Records are shared via OAI-PMH using the Dublin Core, Dublin Core Terms & OLAC name spaces.
369,520	Records include a DCMIType in their record.
8	Providers use DCMIType “Collection”.[42]
850	OLAC records use the DCMIType “Collection”.

Table 1. *OLAC Record Dataset Summary.*

Categorization of OLAC Collections

In this section I report on the semantics of the use of the DCMIType vocabulary term “Collection”. As expected, there are three categories for which overtly marked collection records may fall. This is due to how the DCMIType “Collection” term is defined and the 1:1 Principle of application within Dublin Core. The results are reported in Table 2. Of the eight data providers which employ the DCMIType vocabulary term “Collection”, one data provider misapplied the term, applying it to items which were not specifically either archival units or aggregate works.[43] Five data providers apply the DCMIType “Collection” to archival unit level records. In some cases, these records are for corpora, a specific type of aggregate work. There is some disagreement among language documentation practitioners on the exact nature of a corpus; i.e., is it a single item or is it an aggregate work?[44] Two data providers apply the DCMIType “Collection” to aggregate works. Of the two data providers who applied the DCMIType “Collection” to records, one is a library; the other is an archive. In the case of the archive PARADISEC, the application was within a very limited number of archival collections, suggesting a unique application of the DCMIType term “Collection” rather than a broad policy-based application across archival collections.

[42] In Paterson (2021c), I stated “seven providers”. One of the eight providers misuses the metadata term, leaving seven valid providers. This is clarified in Table 2.

[43] The Graduate Institute of Applied Linguistics (GIAL) is the data provider which misapplied the DCMIType “Collection”. The GIAL data feed is a single run, static feed converted from library MARC records. The code for the conversion is online in the OLAC Github repository and a report of the work is available in Hirt et al. (2009).

[44] For further discussion see Paterson (2021b:footnote 8).

Institution	Reported Collections	Archival Units	Aggregate Works	Mislabeled
The Sociolinguistic Archive and Analysis Project (SLAAP)	36	☑		
Speech and Language Data Repository (SLDR/ORTOLANG)	23	☑		
Bavarian Archive for Speech Signals (BAS)	53	☑		
Multimodal Learning and Teaching Corpora Exchange	49	☑		
COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO)	163	☑		
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	99		☑	
Pacific Collection at the University of Hawai'i at Mānoa Hamilton Library	272		☑	
Graduate Institute of Applied Linguistics Library	155			☑

Table 2. ***OLAC Collections: Categorization of Reported Collections.***

Data Providers not Contributing Collection Records

Only twelve percent of the OLAC data providers use the DCMIType “Collection” vocabulary term. But quantifying provider usage *or lack of usage* gives insight on how provider-based practices of applying the DCMIType “Collection” vocabulary term impact the whole OLAC dataset. For example, if only a few data providers use the DCMIType “Collection” vocabulary term but their records comprise a significant quantity of total available OLAC records, then coverage may be assumed to be fairly good, and non-use may be assumed to be an issue across smaller data providers. Therefore the number of total records and providers was investigated. Table 3 summarizes the rank of OLAC data providers by the quantity of their record contributions. Three data providers provide over sixty-five percent of all records, while the top twenty-five contributors provide over ninety-eight percent of all records. Only four providers among these top twenty-five provide DCMIType “Collection” records. The top three data providers never make use of the DCMIType “Collection”. Positions number four and five are data providers who do mark items with “Collection”. One of these is PARADISEC, which from the data, appears to have inconsistent application of the DCMIType “Collection”. This suggests that major language resource preservation institutions are not providing archival unit records of any kind to OLAC. No doubt this impacts how end users navigate records in the OLAC aggregator.

Institution	**Number of Records**	**Percentage of OLAC Records**	**OLAC Contributor Rank** ***Based on Number of Records***
The Language Archive	149,763	33.79%	1
Endangered Languages Archive	93,687	21.14%	2
SIL Language and Culture Archives	49,494	11.17%	3
California Language Archive	14,959	3.37%	6
Lund University Humanities Lab corpusserver	12,266	2.77%	7

Table 3. ***Five Largest Data Providers not Contributing DCMIType "Collection" Records.***

Estimating missing Archival Unit Records in OLAC

Given the large number of records which are unmarked for DCMIType "Collection" and that the largest record providers either do not provide archival unit records or do so inconsistently, a follow-on question arises. *How many implicit archival units or aggregate works are detectable in the data and are going unreported?* Estimating implicit sets of records is hard because groupings can happen for various reasons. Within language preservation institutions, archival units may be topical for the area of research (phonetic, phonological, oral history, etc.), they may be periodic by research endeavor (such as the set of resources created by a certain field trip or community-researcher collaboration), or they may center around a single scholar or ethno-linguistic community.

I limited my investigation of missing records to only data providers which did not provide a record with the DCMIType "Collection". So providers like PARADISEC and GIAL are excluded. The assumption was made that since these data providers have included DCMIType "Collection" in some of their data that all data is covered by those DCMIType "Collection" records. Of course this may or may not be true and perhaps records from these providers ought to be further investigated. In this task I used two methods to detect clusters of resource records which are suspect to qualify as DCMIType "Collection". The results are listed in Table 4. First, I used a full-text-search on the term "Collection" across metadata element fields. Second, I used patterns in URLs to detect clusters of records. Therefore, I hold these numbers loosely but estimate that there are over 7,816 records which should have the DCMIType "Collection" applied. Even if many of these records contain a statement like "part of X Collection" and therefore show up in searches for the term *collection*, many of these records also represent bundles or aggregate works. Additionally, I estimate there to be 1,086 "missing" archival unit records which would provide additional descriptions to groups of item records. For example, *The Rosetta Project* contributes fifty-nine item level records from the *Alan Lomax Collection*, but there is no archival unit record to describe the nature of the collection.

Institution	Records which Should Be Associated with an Archival Unit Record	
The Rosetta Project: A Long Now Foundation Library of Human Language	59	Alan Lomax Collection records
Endangered Languages Archive	4,005	Search for “Collection” in records
The Language Archive	1,467	Search for “Collection” in records
Lund University Humanities Lab corpusserver	2,285	Search for “Collection” in records

Table 4. ***Suspect and Implicit Archival Units in OLAC.***

Findings

These data support claims (Wasson et al. 2016; Wasson et al. 2018; Burke et al. 2022) that: (1) User interfaces to web-portals presenting language resources present user friction. Here I show that the OLAC web-views also suffer from user friction because whole/part, collection/component relationships are not effectively communicated to OLAC web-users. (2) Record descriptions are not consistent (archival unit description information is compressed into artifact records). (3) Language resource preservation organizations are not using consistent frameworks for archival unit description such as *Describing Archives: A Content Standard* (Society of American Archivists 2013) which calls for cascading levels of detail applied to items found in archival units.

Overall we should expect a greater number of records with the DCMIType “Collection” within the OLAC aggregator. These archival unit records should link to the records for the constituent parts of the archival unit. We should anticipate that there is a broad range of content for which the DCMIType “Collection” is appropriate. “Collection” records would include records for archival units, records of periodicals (but not their articles, unless they were aggregate works), and aggregate works. This leads me to ask: *Where have all the collections gone?* It would appear that either management practices at language archives (1) prevent language archives from grouping, managing, and documenting digital assets as archival units, or (2) unidentified barriers prevent these institutions from adding archival unit records to their OLAC compliant metadata feeds so that they surface in the OLAC aggregator.

Conclusion

Current practices of arrangement and description at language resource preservation institutions participating in OLAC do not currently follow archival practices in arrangement and description as described in frameworks like DACS, including honoring principles like *respect des fonds*. This has multiple impacts including consequences in web-based navigation and discoverability. Implementing best practice guidelines at language resource preservation institutions and communicating archival unit level records and the relationships between archival units and records representing constituent artifact would increase the general understanding of holdings related to the linguistic cultural heritage of ethno-linguistic minorities around the world. Creating a template for the expression of collection records from metadata elements within the OLAC application profile is the next step in facilitating a common understanding of holdings.

Abbreviations

DACS = Describing Archives: A Content Standard

DCAP = Dublin Core Application Profile

DPLA = Digital Public Library of America

EAD = Encoded Archival Description

ELAR = Endangered Languages Archive

FRBR = Functional Requirements for Bibliographic Records

GIAL = Graduate Institute of Applied Linguistics

ISO = International Standards Organization

L&CA = Language and Culture Archives

MARC = Machine-Readable Cataloging

NSDL = National Science Digital Library

NSF = National Science Foundation

OAI = Open Archives Initiative

OAI-PMH = Open Archives Initiative Protocol for Metadata Harvesting

OLAC = Open Language Archives Community

PARADISEC = Pacific and Regional Archive for Digital Sources in Endangered Cultures

URL = Universal Resource Locator

XML = Extensible Markup Language

XSLT = Extensible Stylesheet Language Transformations

Resources

Albarillo, Emily E., and Nicholas Thieberger. 2009. “Kaipuleohone, the University of Hawai‘i’s Ethnographic Archive.” *Language Documentation & Conservation* 3 (1): 154–81.

Bardi, Alessia, Allison Kupietzky, Antoine Isaac, Dan Matei, Donald Weber, Kerstin Arnold, Maria Luisa Martinez Conde, et al. 2014. “Recommendations for the Representation of Hierarchical Objects in Europeana.” Tech. Netherlands: Europeana Foundation. https://pro.europeana.eu/project/hierarchical-objects.

Bartlett, Nancy. 1992. “Respect Des Fonds: The Origins of the Modern Archival Principle of Provenance.” *Primary Sources & Original Works* 1 (1–2): 107–15. doi:10.1300/J269V01N01_07.

Barwick, Linda. 2003. “Planning for PARADISEC: The Pacific And Regional Archive for Digital Sources in Endangered Cultures.” Presentation Paper presented at the Ozeculture conference, Brisbane Powerhouse, 31 July, Brisbane, Australia. https://web.archive.org/web/20080815073730/http://www.acn.net.au/conference3/barwick/barwick.pdf.

Bird, Steven, and Gary Simons. 2003. “Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources.” *Computers and the Humanities* 37 (4): 375–88. doi:10.1023/A:1025720518994.

———. 2021. “Towards an Agenda for Open Language Archiving.” In *Proceedings of the International Workshop on Digital Language Archives: LangArc 2021*, edited by Zavalina Oksana and Shobhana Lakshmi Chelliah, 25–28. Denton, Texas: University of North Texas. doi:10.12794/langarc1851171.

Bird, Steven, and Gary F. Simons. 2001. “The OLAC Metadata Set and Controlled Vocabularies.” In *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*, edited by Thierry DeClerck, Steven Krauwer, and Mike Rosner, 7–18. Université de Toulouse, France: EACL-ACL; elsnet. https://www.aclweb.org/anthology/W01-1506.

Bow, Catherine, Michael J. Christie, and Brian Devlin. 2014. “Developing a Living Archive of Aboriginal Languages.” *Language Documentation and Conservation* 8: 345–60.

Brown, Michael F. 1998. “Can Culture Be Copyrighted?” *Current Anthropology* 39 (2): 193–222. doi:10.1086/204721.

Burke, Mary, and Oksana Zavalina. 2019. “Exploration of Information Organization in Language Archives.” *Proceedings of the Association for Information Science and Technology* 56 (1): 364–67. doi:10.1002/pra2.30.

Burke, Mary, and Oksana L. Zavalina. 2020a. “Identifying Challenges for Information Organization in Language Archives: Preliminary Findings.” In *Sustainable Digital Communities: Proceedings of the 15th International Conference, IConference 2020, Böras, Sweden, March 23–26, 2020*, edited by Anneli Sundqvist, Gerd Berget, Jan Nolin, and Kjell Ivar Skjerdingstad, 622–29. Lecture Notes in Computer Science 12051. Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-030-43687-2_52.

———. 2020b. “Descriptive Richness of Free-text Metadata: A Comparative Analysis of Three Language Archives.” *Proceedings of the Association for Information Science and Technology* 57 (1). doi:10.1002/pra2.429.

Burke, Mary, Oksana L. Zavalina, Shobhana Lakshmi Chelliah, and Mark E. Phillips. 2022. “User Needs in Language Archives: Findings from Interviews with Language Archive Managers, Depositors, and End-Users.” *Language Documentation & Conservation* 16: 1–24.

Christen, Kimberly, Alex Merrill, and Michael Wynne. 2017. “A Community of Relations: Mukurtu Hubs and Spokes.” *D-Lib Magazine* 23 (5/6). doi:10.1045/may2017-christen.

Clayphan, Robina, Valentine Charles, and Michael Wynne. 2017. *Europeana Data Model – Mapping Guidelines*. 2.4. Netherlands: Europeana Foundation. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf.

Cooper, Alan. 2004. *The Inmates Are Running the Asylum*. Indianapolis, IN: Sams.

DPLA MAP Working Group. 2017. *Introduction to the DPLA Metadata Model*. 5th ed. Boston, MA: Digital Public Library of America. https://pro.dp.la/hubs/metadata-application-profile.

Elliott, Mark C. 2001. “The Manchu-Language Archives of the Qing Dynasty and the Origins of the Palace Memorial System.” *Late Imperial China* 22 (1): 1–70. doi:10.1353/late.2001.0002.

Ferreira, Vera, Leonore Lukschy, Buachut Watyam, Siripen Ungsitipoonpor, and Mandana Seyfeddinipur. 2021. “A Website Is a Website Is a Website: Why Trusted Repositories Are Needed More Than Ever.” In *Proceedings of the International Workshop on Digital Language Archives: LangArc 2021*, edited by Oksana L. Zavalina and Shobhana Lakshmi Chelliah, 1–4. Denton, Texas: University of North Texas. doi:10.12794/langarc1851176.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. “Endangered Languages.” *Language* 68 (1): 1–42. doi:10.2307/416368.

Haworth, Kent M. 2001. “Archival Description: Content and Context in Search of Structure.” *Journal of Internet Cataloging* 4 (3–4): 7–26. doi:10.1300/J141v04n03_02.

Heery, Rachel, and Manjula Patel. 2000. “Application Profiles: Mixing and Matching Metadata Schemas.” *Ariadne* 25. http://www.ariadne.ac.uk/issue/25/app-profiles/.

Hillmann, Diane I., and Jon Phipps. 2007. “Application Profiles: Exposing and Enforcing Metadata Quality.” In *DC-2007--Singapore Proceedings*, edited by Stuart A. Sutton, Abdus Sattar Chaudhry, and Christopher Khoo, 52–62. Singapore; Dublin, Ohio, USA: Dublin Core Metadata Initiative & National Library Board Singapore. https://dcpapers.dublincore.org/pubs/article/view/866.

Hirt, Christopher, Gary Simons, and Joan Spanne. 2009. “Building a MARC-to-OLAC Crosswalk: Repurposing Library Catalog Data for the Language Resources Community.” In *Proceedings of the 2009 Joint International Conference on Digital Libraries - JCDL ’09*, 393–94. Austin, Texas: ACM Press. doi:10.1145/1555400.1555479.

Holton, Gary. 2012. “Language Archives: They’re Not Just for Linguists Any More.” In *Potentials of Language Documentation: Methods, Analyses, and Utilization*, edited by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, 111–17. Language Documentation & Conservation Special Publication 3. Honolulu, Hawai’i: University of Hawai’i Press. http://scholarspace.manoa.hawaii.edu/handle/10125/4523.

Hughes, Baden. 2004. “Metadata Quality Evaluation: Experience from the Open Language Archives Community.” In *Digital Libraries: International Collaboration and Cross-Fertilization*, edited by Zhaoneng Chen, Hsinchun Chen, Qihao Miao, Yuxi Fu, Edward Fox, and Ee-peng Lim, 320–29. Lecture Notes in Computer Science 3334. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-30544-6_34.

Hutt, Arwen, and Jenn Riley. 2005. “Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data Providers of Cultural Heritage Materials.” In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL ’05*, 262–70. Denver, CO, USA: ACM Press. doi:10.1145/1065385.1065447.

Indiana Memory Project. 2020. *Dublin Core Metadata Guide—Indiana Memory Project*. Indianapolis, Indiana: Indiana University Perdue University Indianapolis. https://www.in.gov/library/files/IndianaMemoryMetadata2020.pdf.

Kipp, Darrell. 2007. “Swimming in Words.” *Cultural Survival Quarterly* 31 (2): 36–43.

Kurtz, Mary. 2010. “Dublin Core, DSpace, and a Brief Analysis of Three University Repositories.” *Information Technology and Libraries* 29 (1): 40–46. doi:10.6017/ital.v29i1.3157.

Lynch, Joshua D., Jessica Gibson, and Myung-Ja Han. 2020. “Analyzing and Normalizing Type Metadata for a Large Aggregated Digital Library.” *The Code4Lib Journal* 47 (February). https://journal.code4lib.org/articles/14995.

Miller, Steven J. 2010. “The One-To-One Principle: Challenges in Current Practice.” In *DCMI ’10: Proceedings of the 2010 International Conference on Dublin Core and Metadata Applications*, 150–64. Pittsburgh, Pennsylvania: Dublin Core Metadata Initiative. https://dcpapers.dublincore.org/pubs/article/view/1043.html.

Munro, Robert, and David Nathan. 2005. “Introducing the ELAR Information System Architecture.” Presentation Paper presented at the DELAMAN III, University of Texas at Austin 21-22 November. http://www.robertmunro.com/research/munro05elar.pdf.

Nathan, David. 2013. “Access and Accessibility at ELAR, a Social Networking Archive for Endangered Languages Documentation.” In *Oral Literature in the Digital Age*, edited by Mark Turin, Claire Wheeler, and Eleanor Wilkinson, 21–41. World Oral Literature Series 2. Cambridge, England: Open Book Publishers. doi:10.111647/OBP.0032.03.

Nordmoe, Jeremy. 2018. “SIL International Language and Culture Archives.” Presentation Slides presented at the Planning Workshop on Data Archives and Languages of the Americas, February 8th-9th, Philadelphia, PA. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/nordmoe.pdf.

Ohio Digital Network. 2021. *Ohio Digital Network Metadata Application Profile*. 1.6. Ohio DPLA Project. http://ohiodigitalnetwork.org/wp-content/uploads/metadata-application-profile-v1-6.pdf.

Palmer, Carole L., Oksana L. Zavalina, and Katrina Fenlon. 2010. “Beyond Size and Search: Building Contextual Mass in Digital Aggregations for Scholarly Use.” *Proceedings of the American Society for Information Science and Technology* 47 (1): 1–10. doi:10.1002/meet.14504701213.

Park, Jung-ran. 2006. “Semantic Interoperability and Metadata Quality: An Analysis of Metadata Item Records of Digital Image Collections.” *Knowledge Organization* 31 (1): 20–34.

Park, Jung-Ran. 2009. “Metadata Quality in Digital Repositories: A Survey of the Current State of the Art.” *Cataloging & Classification Quarterly* 47 (3–4): 213–28. doi:10.1080/01639370902737240.

Park, Jung-ran, and Eric Childress. 2009. “Dublin Core Metadata Semantics: An Analysis of the Perspectives of Information Professionals.” *Journal of Information Science* 35 (6): 727–39. doi:10.1177/0165551509337871.

Park, Jung-ran, and Yuji Tosaka. 2010. “Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability.” *Information Technology and Libraries* 29 (3): 104–16. doi:10.6017/ital.v29i3.3136.

Paterson III, Hugh J. 2021a. “Language Archive Records: Interoperability of Referencing Practices and Metadata Models.” M.A. Thesis, Grand Forks, North Dakota: University of North Dakota. Theses and Dissertations. 3937. University of North Dakota Scholarly Commons. https://commons.und.edu/theses/3937/.

Paterson III, Hugh J. 2021b. “Where Have All the Collections Gone?” Poster presented at the 15th Annual Society of American Archivists Research Forum, Hosted Virtually. https://www2.archivists.org/am2021/research-forum-2021/agenda#posters.

———. 2021c. “OLAC Nightly Data Dump (XML) from 18 July 2021.” OLAC: Open Language Archives Community. Zenodo. doi:10.5281/ZENODO.5112131.

Patrick, Peter L. 2008. “The Speech Community.” In *The Handbook of Language Variation and Change*, edited by J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 573–97. Malden, MA: Blackwell Publishing Ltd. doi:10.1002/9780470756591.ch23.

Riva, Pat, Patrick Le Bœuf, and Maja Žumer, eds. 2017. *IFLA Library Reference Model A Conceptual Model for Bibliographic Information*. December 2017. Den Haag, Netherlands: International Federation of Library Associations and Institutions (IFLA). https://www.ifla.org/publications/node/11412.

Roy, Loriene, Anjali Bhasin, and Sarah K. Arriaga, eds. 2011. *Tribal Libraries, Archives, and Museums: Preserving Our Language, Memory, and Lifeways*. Lanham: Scarecrow Press.

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, et al. 2019. “Public Access to Research Data in Language Documentation: Challenges and Possible Strategies.” *Language Documentation & Conservation* 13 (October). University of Hawaii Press: 545–63.

Simons, Gary, and Steven Bird. 2003. “Building an Open Language Archives Community on the OAI Foundation.” *Library Hi Tech* 21 (2): 210–18. doi:10.1108/07378830310479848.

Simons, Gary F. 2016. “From Linguistic Data Type to Language Resource Type: Laying the Groundwork for a Metadata Application Profile.” Presentation Slides presented at the OLAC / DELAMAN Workshop, Austin, TX. https://scholars.sil.org/sites/scholars/files/gary_f_simons/presentation/simons-language_resource_type_vocabulary.pdf.

Society of American Archivists. 2013. *Describing Archives: A Content Standard*. 2nd ed. Chicago, Illinois: Society of American Archivists. http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf.

Stvilia, Besiki, Les Gasser, Michael B Twidale, Sarah L Shreeves, and Tim W Cole. 2004. “Metadata Quality for Federated Collections.” In *Proceedings of the Ninth International Conference on Information Quality (ICIQ-04)*, 111–25. https://www.ideals.illinois.edu/handle/2142/721.

Sugimoto, Shigeo, Senan Kiryakos, Chiranthi Wijesundara, Winda Monika, Tetsuya Mihara, and Mitsuharu Nagamori. 2018. “Metadata Models for Organizing Digital Archives on the Web: Metadata-Centric Projects at Tsukuba and Lessons Learned.” In *DC-2018--The Porto, Portugal Proceedings*, 95–105. DCMI 2018: Proceedings of the International Conference on Dublin Core and Metadata Applications. Dublin, Ohio: Dublin Core Metadata Initiative. https://dcpapers.dublincore.org/pubs/article/view/3968/.

Sullivant, Ryan. 2020. “Archival Description for Language Documentation Collections.” *Language Documentation & Conservation* 14: 520–78.

Thieberger, Nicholas, and Michel Jacobson. 2010. “Sharing Data in Small and Endangered Languages: Cataloging and Metadata, Formats, and Encodings.” In *Language Documentation: Practice and Values*, edited by Lenore A. Grenoble and N. Louanna Furbee, 147–58. 158. Amsterdam; Philadelphia, PA: John Benjamins Publishing Company. doi:10.1075/z.158.15thi.

Thieberger, Nick, and Amanda Harris. 2022. “When Your Data Is My Grandparents Singing. Digitisation and Access for Cultural Records, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).” *Data Science Journal* 21 (9): 1–7. doi:10.5334/dsj-2022-009.

Tillett, Barbara. 2004. *What Is FRBR? — A Conceptual Model for the Bibliographic Universe*. Washington, D.C.: Library of Congress Cataloging Distribution Service. https://www.loc.gov/cds/downloads/FRBR.PDF.

Tillett, Barbara B. 2001. “Bibliographic Relationships.” In *Relationships in the Organization of Knowledge*, edited by Carol A. Bean and Rebecca Green, 19–35. Information Science and Knowledge Management 2. Dordrecht: Springer Netherlands. doi:10.1007/978-94-015-9696-1_2.

Urban, Richard J. 2010. “Principle Violations: Revisiting the Dublin Core 1:1 Principle.” In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, Article: 173, pp.1–2. ASIS&T ’10. Pittsburgh, PA, USA: American Society for Information Science.

Wasson, Christina, Gary Holton, and Heather S. Roth. 2016. “Bringing User-Centered Design to the Field of Language Archives.” *Language Documentation & Conservation* 10: 641–81.

Wasson, Christina, Melanie Medina, Miyoung Chong, Brittany LeMay, Emma Nalin, and Kenneth Saintonge. 2018. “Designing for Diverse User Groups: Case Study of a Language Archive.” *Journal of Business Anthropology* 7 (2): 235–67. doi:10.22439/jba.v7i2.5605.

Weber, Tobias. 2021. “The Curation of Language Data as a Distinct Academic Activity: A Call to Action for Researchers, Educators, Funders, and Policymakers.” *Journal of Open Humanities Data* 7 (28): 1–10. doi:10.5334/johd.51.

Wiberg, Andrew. 2014. “Mukurtu: Information Retrieval System Engineered for Indigenous Individuals and Communities.” Paper presented at the IFLA WLIC 2014 - Libraries, Citizens, Societies: Confluence for Knowledge in Session 118 - Indigenous Matters Special Interest Group, 16-22 August 2014, Lyon, France. http://ifla-test.eprints-hosting.org/id/eprint/922.

Wickett, Karen M., Antoine Isaac, Martin Doerr, Katrina Fenlon, Carlo Meghini, and Carole Palmer. 2014. “Representing Cultural Collections in Digital Aggregation and Exchange Environments.” *D-Lib Magazine* 20 (5/6). doi:10.1045/may2014-wickett.

Wijesundara, Chathurangani, and Shigeo Sugimoto. 2018. “Metadata Model for Organizing Digital Archives of Tangible and Intangible Cultural Heritage, and Linking Cultural Heritage Information in Digital Space.” *Libres* 28 (2): 58–80.

Woodbury, Anthony C. 2014. “Archives and Audiences: Toward Making Endangered Language Documentations People Can Read, Use, Understand, and Admire.” In *Special Issue on Language Documentation and Archiving*, edited by David Nathan and Peter K. Austin, 19–36. Language Documentation and Description 12. London: SOAS. http://www.elpublishing.org/PID/135.

Yi, Irene, Amelia Lake, Juhyae Kim, Kassandra Haakman, Jeremiah Jewell, Sarah Babinski, and Claire Bowern. 2022. "Accessibility, Discoverability, and Functionality: An Audit of and Recommendations for Digital Language Archives." *Journal of Open Humanities Data* 8 (10): 1–19. doi:10.5334/johd.59.

Zavalina, Oksana L. 2011. "Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval." *Journal of Library Metadata* 11 (3–4). Routledge: 104–28. doi:10.1080/19386389.2011.629957.

Zavalina, Oksana L., Carole L. Palmer, Amy S. Jackson, and Myung-Ja Han. 2009. "Evaluating Descriptive Richness in Collection-Level Metadata." *Journal of Library Metadata* 8 (4): 263–92. doi:10.1080/19386380802627109.

Žumer, Maja, Marcia Lei Zeng, and Athena Salaba. 2010. "FRBR: A Generalized Approach to Dublin Core Application Profiles." In *DCMI '10: Proceedings of the 2010 International Conference on Dublin Core and Metadata Applications*, 21–30. Pittsburgh, Pennsylvania: Dublin Core Metadata Initiative. https://dcpapers.dublincore.org/pubs/article/view/1024.