

A Short Reflection on the Long History of Language Archives

Hugh Paterson III

i@hp3.me

The Institute for Advanced Study

Princeton, New Jersey, USA

ABSTRACT

This reflection presents a historical narrative surrounding the term *language archives*. Three elements are highlighted: (1) the history of the use of the term, (2) the sociological connections between individual voices central to the modern development of language archives, and (3) the evolution of metadata schemas across institutions in newly established language archives. By highlighting these points I contrast a narrative developed within language documentation [51, 73] and the Open Language Archives Community [9, 10]—that language archives are a new innovation. I center a substantial number of obscure digital publications and primary archival resources. This historical view and the links to original source materials are helpful to new information professionals and language scholars.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; • **Applied computing** → **Digital libraries and archives**; **Document metadata**; • **Information systems** → **Digital libraries and archives**.

KEYWORDS

Language Archives, Scholastic History, Metadata Schemas

1 INTRODUCTION

The concept of *language archives* has received significant attention in recent years, as evidenced by conferences in: (1) Australia via *Sustainable Data from Digital Fieldwork: From creation to archive and back* in 2006, *Sustainable data from digital research: Humanities perspectives on digital scholarship* in 2011,¹ and *PARADISEC@100* in 2021,² (2) Europe broadly via the Common Language Resources and Technology Infrastructure’s (CLARIN) *Annual Conference* (2012-2026),³ (3) Germany via the *Language Documentation and Archiving Conference* (2022, 2024, 2026),⁴ and (4) the United States via the *JCDL International Workshop on Digital Language Archives* (2021, 2023, 2025).⁵

Other significant venues often blend language documentation focused research and the preservation of the created information resources, e.g., the *International Conference on Language Documentation and Conservation* (2009-2027),⁶ and several diversely focused and irregularly held satellite workshops at the European Language Resource Association (e.g., 2002, 2006, 2008, 2012, 2019).⁷ The nature of these conferences and workshops vary. That is, some are

much more focused on language research, community engagement methodologies, and research impacts, while others are focused more narrowly on information resource management and information architecture at the level of resource management. Additionally, scholars engaging in the broader discourse of language resource stewardship come from no less than three different scholarly disciplines:

- (1) The people-centric—Field & Anthropological Linguistics
- (2) The resource- or institution-centric—Library & Archival Science
- (3) The computer-centric—Computational Linguistics & Information Management.

These three areas of scholarship each conceptualize language resources differently, and therefore, emphasize different arrangement and stewardship practices.

While the worldwide discussion is well underway, it is useful for individuals new to the conversation to have a historical narrative. For the new information professional, there are at least three basic questions which, if well answered, can help one better grasp the status and trajectory of the global conversation. Crucial to the conversation around language archives are the history, use, and concepts behind the term *Language Archive*. That is, *What is a language archive?* A second question is, *What are the sociological connections between individual voices central to the development of global discussions?* And yet a third question of particular interest to the information professional is, *What were the evolutionary trajectories of metadata schemas across institutions?* Some answers to these questions are provided in the following sections.

2 TERMINOLOGICAL AND CONCEPTUAL DEFINITIONS

2.1 Language Archives

The term-of-art *Language Archive* is an idiom. That is, it cannot be understood from the individual meanings of its elements. To really provide clarity on *What is a language archive?*, context is needed. The following sections provide context for defining *language archives* along three dimensions: (1) institution versus collection, (2) the usage of the term over time, and (3) contrasting language archives with language resources.

2.1.1 Institution vs. Collection. *Language archives* as a term carries unfortunate ambiguity due to the difference between *an Archives*—A type of memory institution; and *an archive*—A set of artifacts grouped based on specific criteria, one of which is often a common origin. According to Caswell [19] *an Archives* and *an archive* represent distinct concepts in archival praxis. Furthermore, as explained by Paterson [85] the large-A—*an Archives*—is one of a variety of types memory institutions which are best distinguished by business model rather than institutional title. The

¹<https://paradisec.org.au/2011Conf.html>

²<https://sites.google.com/view/paradisec-100/home>

³<https://www.clarin.eu/content/clarin-annual-conference>

⁴<https://langdoc.org>

⁵<https://digital.library.unt.edu/explore/collections/LANGARC>

⁶<https://ling.lll.hawaii.edu/sites/icldc>

⁷<http://www.elda.org/en/lrec/proceedings>; <https://www.elra.info/elra-events/lrec>

business model is often driven by or chosen alongside of the institutional mandate and purpose. It impacts managerial decisions such as metadata schema choice and resource life-cycle actions. On the small-*a* side of the oft-confused distinction are a different set of categorical distinctions. *Describing Archives: A Content Standard*—DACS (SAA 2013) recommends a three way distinction: *records*, *papers*, and *collection*. These are useful for framing aggregations of resources reflective of named entities (individuals or groups) such as the *Personal papers of John Doe* or *Records of the XYZ project* (p. xvii, Principle 3, §2.3.19). However, there are also other terms for arrangement structures within information science such as *fonds*, *series*, *archival collection*, *aggregates*, and *record set*. Each term-of-art carries specific implications related to provenance, and the inter-relationship of component artifacts to each other.⁸

Therefore, in practice and recent literature, things referred to as *language archives* (or even *corpora*) may be a set of resources that are more accurately matched to concepts familiar to archivists as *collections*, i.e., *papers* or *records generated by a project*. These may be small-*a* archives depending on their internal compositions and provenance. In other situations, entities referenced as *language archives* may be *special libraries*,⁹ *departments within larger institutions*, or *standalone institutions* which engage in the stewardship of language artifacts. These are all memory institutions, but not necessarily large-*A* Archives based on their mandates and purpose for existence.

2.1.2 Language Archives Through Time. The relatively recent attention given to language archives as described in the introduction may lead one to think that the concept of language archives is a new concept. And with limited exceptions [e.g., 3] the following narrative is often repeated in recent literature [e.g., 22, 24, 51, 54, 98]. The narrative generally starts the story around 1991, and goes something like this: Krauss [48] gives a motivational speech in 1991 at the annual meeting of the largest linguistics society challenging the status quo in linguistic and language research praxis, while also highlighting language loss around the world. Krauss argued for more tangible community-centric outputs as part of research practice. By 1998, Himmelmann [52] sets the stage for a method of conducting language research which incorporates the use of stewardship institutions for the benefit of minority language communities. Both events are significant in the history and formation of the *language documentation movement in the digital era*. However, this plot line simplification jumps over a great deal of language archival activity prior to 1990, which has largely, but not completely, gone unnoticed [3, 32, 33, 63] (within scholarly communication centering language archives).

While the *Language Documentation Movement* pulls from several scholarly disciplines—resource stewardship being one of them—*Language Archives*, and language resource stewardship has a much

longer history. Figure 1 is provided to illustrate the relative frequency of the terms: *Language Archive(s)*, *Linguistic Archive(s)*, and *Folklore Archive*. The data is provided via the English Corpus via the Google N-Gram viewer. The increased scholarly attention discussed in the introduction is clearly evident in the chart. However, evidence presented via Figure 1 also suggests that the concept of a *Language Archive* pre-dates the digital modern era. In fact, *Language Archives* as a term can be dated back as early as the 1900s. However, if we think about the function of language archives as stewarding language resources, the same functions have been conducted under different nomenclature. Many of the oldest discussions surrounding third-party stewardship of language resources also involves the term *folklore archive(s)*, which Figure 1 indicates is also frequently used throughout the 20th century.

While Figure 1 shows English term usage, terms for these kinds of collections and their stewardship institutions also start to appear in the scholarly literature in other languages. Other commonly used English terms were *speech archive*, or *sound archive*, which are both also concepts found in German scholarship (*Spracharchiv*, *Lautarchiv*; German also has *Lautbibliothek*, ‘sound library’). In German scholarship *Spracharchiv* as a collection of evidentiary resources for regional variants is reported to date back to at least 1772 [50]. Audio recordings of French date back to 1860 [40].¹⁰ It is natural then to anticipate there might be language related archival collections in France, and indeed there are [among others see: 17, 18, 35, 62, 65, 74]. Some examples include: the *CNRS - Musée de l’Homme*¹¹ and the *Phonothèque Nationale* at the *Bibliothèque Nationale*. French also had its own terms such as *archives linguistiques* (for language and linguistic archives) and *Phonothèque* (sound archives). *Archives linguistiques* is used for a broad range of evidentiary modalities related to language, especially the evidentiary record created in the process of crafting dialectical or regional dictionaries.

Through time what has been popularly understood as a *language archive* has varied. For example, in 1954, Wells [110, p. 101] conceptualizes an archive as a collection of artifacts in the written modality indicating that it includes: “written transcriptions of dictated material, and other written records”.¹² This is largely consistent with how the notion of the Boasian-Trilogy in minority-language focused philology and linguistics was constructed. Franz Boas advocated a research approach that sought to describe languages via the creation of a dictionary, a grammar, and a collection of transcribed and glossed texts [see among others: 39, 44, 47, 97]. The understood interactive carrier is paper.

Wells’ (and others’) emphasis on written modalities as the basis for a language archive was not universally accepted. Representing a different approach to ‘evidence’, Zwirner collected audio recordings for the linguistic study of German dialects from as early as 1932 [69, p. 142; 102, 119].¹³ Zwirner was preceded

⁸The work of Geoffrey Yeo [115–118] is especially enlightening with regard to the terms of art used within archival science. Other scholars such as Curral et al. [29], Muller et al. [78], and Gilliland et al. [46, p. 170] also provide clarity and historical perspective on the terms of art and the concepts they represent.

⁹Special libraries function for immediate knowledge retrieval and knowledge-as-service capacity without explicit mandate for resource preservation beyond what the current information clients require. Special libraries have historically been understood to be physical collections and places, but in a digital world *information centers*, *shared network drives*, and *wikis* may serve a similar purpose.

¹⁰<https://www.firstsounds.org/sounds/scott.php>

¹¹<https://archives.crem-cnrs.fr>

¹²Wells’ list of artifact types, all in a written modality, excludes the source audio materials. In 2009, David Nathan, former archivist at a major language archive, is quoted in the New York Times as saying: “For most linguists, audio is just an inconvenience on the way to transcription” [82].

¹³Zwirner’s collections are now housed at the *Archiv für Gesprochenes Deutsch*, a research center for stewarding corpora of spoken German. <https://agd.ids-mannheim.de/index.shtml>

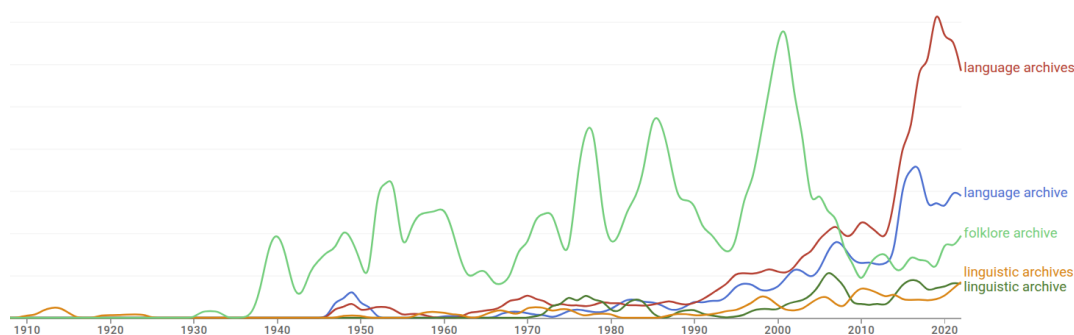


Figure 1: Google N-Gram of the relative frequency of the terms: *Language Archive(s)*, *Linguistic Archive(s)*, and *Folklore Archive*.

by pioneers like Jesse Walter Fewkes (1890) who worked in the United States and used anthropological approaches under folklore studies [42, 43].¹⁴ The strong connection between folklore archives and language archives is exemplified in publications that reference or describe the *Memorial University of Newfoundland Folklore and Language Archive* (MUNFLA) [e.g., 112, p. 316].¹⁵ MUNFLA is the largest folklore archive in Canada. In contrast to folklore archives and general sound collections which may contain music, political speeches, radio programs, etc., Zwirner’s sound collection is variously referred to as a *language archive* in literature relevant to the time [e.g., 111].

By the 1950s it was clear to observant linguists such as Voegelin [106] that audio collections created by linguists and in service of linguistic research—especially in the United States—were in need of preservation.¹⁶ In 1950 Voegelin [106] laments the current situation that there is no language archive that rivals those set up for music in Berlin or Indiana. In 1953, Voegelin convened a four-day meeting concurrent with the Linguistic Society of America’s Summer Institute [107] to discuss the state of the field. His primary concern was scholars studying languages who were not working with stewardship institutions towards the preservation of audio materials—a stark contrast to scholars like Zwirner! It is during this time frame that we see the establishment of several audio-focused resource stewardship centers.¹⁷ For example, *The California Language Archive* was established in 1953,¹⁸ initially as *The Survey of California and Other Indian Languages* in the Department of Linguistics at the University of California, Berkeley. *The UCLA Ethnomusicology Archive* was established in 1961,¹⁹ and the *Archives of Traditional Music*²⁰ had recently moved to Indiana University

where the LSA Summer School was hosted and Voegelin was faculty.

Two of the three afore mentioned stewardship institutions have the term *Music* in their names. While Voegelin is advocating for stewardship solutions and practices specifically among linguists, a look at the history of audio recording will show that interest in the recording of minority languages—even if music is included—has been there since almost the invention of the technology [among others: 4, 5, 64, 68, 105, 114]. Seeger, whose career included at different times being the director of the *Archives of Traditional Music* and *The UCLA Ethnomusicology Archive*, wrote in 1986 that ethnomusicology archives are just as relevant for linguists and folklorist [96]. Therefore, as we consider the concept of a *Language Archive*, we should also consider why one might draw lines between ethnomusicology, folklore, and linguistics.

So far, the discussion here has been exemplified through contrasting textual and audio modes. However, this is not to exclude materials primarily characterized via the visual mode inclusive of illustrations, cartographic depictions, photographs, portraits, and video as materials relevant to language archives—linguists and language scholars use and create artifacts with these modes too. In fact, these modes of evidence can and often do belong in language centric collections. However, a carrier or modality-based perspective is only one dimension of archives and “suitable content” for inclusion in language resource collections.

There is a different dimension related to the way a stewardship institution is situated relative to the public it serves. That is, the concept of a *language library* may operate under the nomenclature *language archive*. Fundamentally the difference between a *language library* and a *language archive* is the purpose of a collection and the way that the act of collecting benefits (some segments of) society. While the act of collecting, especially of language resources in minority languages, has often been associated with colonialism and practices related to the inequitable distribution of power. It is not universally true. There are examples where collections have been established with *the* or one of the express purposes (at the time) being to facilitate language learning by native speakers such as in: Germany [103], the United States [106, fnt. 4], and Australia [13]. Other examples include the collecting of language resources for language learning by non-native speakers

¹⁴Original materials produced by Jesse Walter Fewkes are stewarded at the US Library of Congress: <https://www.loc.gov/collections/ancestral-voices>

¹⁵<https://www.mun.ca/munfla>

¹⁶Voegelin is also known for a classification and index of languages of the world [108]. This work was published with a notice that its copyright would revert to the public domain after 20 years. Voegelin’s work became the basis of the MultiTree project managed by the Linguist list [20, 70]. MultiTree in turn became the seed data for the Glottolog.

¹⁷The dates provided here are those advertised as their official founding dates. However, often the case was that collecting has started several decades earlier. It is also frequently the case that these Archives had undergone several name changes or moves prior to their current instantiation.

¹⁸<https://cla.berkeley.edu>

¹⁹<https://schoolofmusic.ucla.edu/facilities/ethnomusicology-archive>

²⁰<https://libraries.indiana.edu/archives-traditional-music>

[36]. This social function is more in line with the prototypical mission of libraries than with the prototypical mission of archives.²¹ In fact these libraries, in some cases may be most accurately portrayed as “special libraries” [i.e., 79, 99]. For example, Delaney [36] discusses *language libraries* in the context of the US Government intelligence operations and language libraries at teaching schools that function as a service to the teaching capacity. However, special libraries and special collections also exist on a continuum. For example, the University of Hawai‘i has a *Slavic Collection* [89] and *The Hawaiian and Pacific Collections* [66]. However, *The Hawaiian and Pacific Collections* are housed in closed stacks and functions in a much more restricted context suggesting a management approach more like that of a special library.

Since the late 1990s the term *language archives* has almost exclusively been used in conjunction with digital interaction with language resources via digital repositories. This limitation to digital is an unhelpful constraint when considering the long history of multimedia tools and carriers used to record global ethno-linguistic heritage. The merger of concepts also conflates digital exhibits and research platforms with digital platforms serving the traditional “library” function, let alone those platforms serving as a repository within the management framework of an Archives.

Renewed interest in direct asset management and the persistence of the evidentiary record in academic scholarship have mutually spurred the establishment of a slew of new language archives. In response to these felt need the following stewardship organizations are some of those established since Krauss’ call to action: *The Language Archive* (est. 1990s),²² *Pangloss* (est. 1994),²³ *The SIL Language & Culture Archives* (est. 1999),²⁴ *The Archive of the Indigenous Languages of Latin America* (est. 2000),²⁵ *Endangered Languages Archive* (est. 2002),²⁶ *Pacific and Regional Archive for Digital Sources in Endangered Cultures* (est. 2003),²⁷ *Leipzig Endangered Languages Archive* (est. 2005),²⁸ *Kaipuleohone* (est. 2008),²⁹ the *Computational Resource for South Asian Languages* (CoRSAL; est. 2016),³⁰ and the *South African Centre for Digital Language Resources* (SADiLaR; est. 2016).³¹

Bird and Simons [7, §2.4] intentionally conflated the ideas of digital exhibits, research platforms, large-*A*, and small-*a* archives in their influential paper. This served to only add to an already broad misunderstanding of terms-of-art in the domains of scholarly communications and stewardship [91, 109], especially around the concept of *archives* and their preservation roles in society. That is, the role of a specific large-*A* an Archives in society is linked to

its mandate and legally established purpose. The stewards are subject to fulfill those legally established goals. This gives stewarding institutions different trajectories for their stewardship efforts.

2.1.3 Language Resources. Defining a language resource can be challenging. Language scholars often have challenges distinguishing the difference between a language resource and an archive—in part because language resources are often aggregates of several independent works.

There is a copious literature which describes the creation methodologies of the language resources frequently created by linguists [among others: 5, 21, 30, 38, 86, 90, 93, 107]. However, the aforementioned Boasian-Trilogy illustrates the complexity nicely. A Boasian-Trilogy (even digitally [49, 72]), is the result of several unique works³² and their derivative works. These works may be the result of the efforts of several people, a long period of time, and may be the result of several distinct projects in which participants have an evolving understanding of the language and culture under investigation. These resources themselves do not constitute a *corpus* which is a specific aggregation of independent works in a query-able arrangement. Rather, these resources often constitute several small-*a* archives, document groups, or fonds depending on their creation context. Although the works themselves are independent works, there are frequently threads of scholarly inquiry which form relationships between the created works.³³ The propensity for language scholars (like Wells) to call groups of language resources *archives* complements what Watson [109] describes as the lack of terminological distinctions maintained by humanities and history scholars related to the term *Archives* (where they mean small-*a*) in contrast to how information professionals use the term (meaning large-*A*).

The distinction between aggregate and non-aggregate resource is followed by a second dimension of complexity—scarcity. This is articulated by Simons [100] in a presentation describing the problem for linguists interested in archiving and defining the scope of cataloging for language archives. First, Simons points out something obvious to anyone who works with undocumented or under-documented languages:

A language resource is any resource that is an input to or an output of language documentation, description, or development.

He continues later, writing,

When there are millions of books in a language, any book could be an input to language description, but declaring every one of them to be a language resource creates information noise that hides the true resources. When there are very few books in a language: We want to flag every single one as a potential input. If we don’t, they’ll be lost in the global Web of Data.

²¹In recent times stewards of collections and whole institutions have come to realize that a greater engagement with the public creates long-term value via social relevance which strengthens the institution making it more sustainable.

²²<https://archive.mpi.nl/tla>

²³<https://pangloss.cnrs.fr/?lang=en>

²⁴<https://www.sil.org/resources/language-culture-archives>

²⁵<https://www.ailla.utexas.org>

²⁶<https://www.elararchive.org>

²⁷<https://www.paradisec.org.au>

²⁸<https://www.eva.mpg.de/linguistics/past-research-resources/resources/leipzig-endangered-languages-archive-lela>

²⁹<https://hdl.handle.net/10125/4250>

³⁰<https://digital.library.unt.edu/explore/collections/CORSAL>

³¹<https://sadilar.org/en>

³²I mean *Work* in the sense of a WEMI *work*. WEMI is discussed later in this section.

³³These threads of scholarly inquiry may be conceived of as subjects by the information professional, but language scholars rarely conceive of them using this term-of-art.

Therefore determining the exact status of an object as a language resource is contextually dependent. In many cases this is dependent on a genre classification.³⁴ However, the conundrum that Simons discusses is where information science via cataloging may possibly make its greatest contribution to language resource description. That is, the International Federation of Library Associations and Institutions—Library Reference Model (IFLA-LRM) [58, 94] use a Work-Expression-Manifestation-Item (WEMI) ontology as a basis for establishing new cataloging records. Significant in this model is the acknowledgment and modeling of aggregates [83, 104]. Recent advances in the conceptualization of WEMI (as OpenWEMI [27, 28])³⁵ has de-buckled WEMI from the IFLA-LRM allowing for its use as a taxonomy within other systems such as Dublin Core.³⁶ By modeling the WEMI entities in cataloging efforts, many of the cataloging challenges surrounding language resources can be eliminated and clarified in citation, referencing, and aggregation.

3 NETWORKS OF LANGUAGE RESOURCE STEWARDS

The rise of the global change involving digital methods of language resource creation, also gave rise to the different networks of actors working both independently and in concert to solve information resource persistence issues. This gives rise to two questions: *Who are these actors?*, and, *what are (or were) the sociological connections between individual voices central to the development of global discussions?* For context, it is important to keep in mind that there are several dimensions which drive scholars and their supporting infrastructure personnel to create networks. Frequently these are: (1) the area of linguistics they are working within, (2) the presence of an adequate institutional or national support network, and, (3) the trust and legal factors that mitigate information resource sharing environments.

At the beginning of the 21st century, several influential voices arose. Among them were Steven Bird (Charles Darwin University), Gary Simons (SIL International), Mark Liberman (University of Pennsylvania), Thierry Declerck (Deutsches Forschungszentrum für Künstliche Intelligenz), Laurent Romary (Institut national de recherche en sciences et technologies du numérique), Khalid Choukri (European Language Resource Association), Nick Thieberger (University of Melbourne), Linda Barwick (University of Sydney), Nancy Ide (Brandeis University), and Helen Aristar-Dry (Linguist's List).³⁷ These scholars worked in North America, Europe, and Australia. They also worked in the subject areas of: (1) Minority Language Scholarship, Language Documentation, and Linguistic Field Work, (2) Dialectology, (3) Lexicography, (4) Text Annotation, and (5) Sign-language Studies.

One of the first notable efforts was the E-MELD workshops whose funding was secured through the work of Helen Aristar-Dry and her husband Anthony Aristar.³⁸ E-MELD (Electronic Metastucture for Endangered Languages Data) was a 5-year (2001-2006) project with a dual objective: (1) to aid in the preservation of endangered languages data and documentation and (2) to aid in the development of the infrastructure necessary for effective collaboration among electronic archives. They were very successful in practically moving scholars within the discipline of linguistics towards a data-persistence and evidence-preserving praxis, in part due to relevant and significant papers given at each workshop [e.g., 1, 95],³⁹ publications about the workshops [e.g., 15, 75], and updates about E-MELD discussions given at other conferences and workshops [e.g., 2, 60].

Helen Aristar-Dry and Anthony Aristar also founded *The Linguist List*,⁴⁰ a nonprofit that supports linguistic communication and has in the past hosted infrastructure projects. Several language archive related mailing lists have been hosted on Linguist List,⁴¹ as well as discussions more broadly on the subject of preserving the evidentiary record in linguistic scholarship.

In Europe, one of the early digital language resource stewardship projects was conducted as part of *DoBeS: Dokumentation Bedrohter Sprachen* (Documentation of Endangered Languages), which was funded with a grant from the *VolkswagenStiftung* (the Volkswagen Foundation) in 2000. This endangered language focused project had an out-sized impact on language documentation praxis and digital language resource stewardship. This funded multi-year project pushed the *ISLE Metadata Initiative* (IMDI)⁴² into the limelight of language resource metadata [among others: 16, 76, 77, 113].⁴³ This initiative sought to solve the descriptive and analytical needs of linguists, administrative needs of stewards, as well as search and discovery interface needs. DoBeS, and more importantly IMDI was able to be as influential as it was due to funding and institutional stability. Stable funding and institutional support is not the norm in the broad scope of language resource creation and stewardship. Thus, in contrast to the institutionally supported work done at DoBeS, the Open Language Archives Community (OLAC)⁴⁴ formed as a loose-knit, global community driven by scholars working in the areas of Minority Language Scholarship, Language Documentation, and Linguistic Field Work. These scholars often lacked institutional support, although many scholars from institutions with significant interests in language resources were participants. OLAC efforts, led by Steven Bird and Gary Simons [6, 8, 11] who secured U.S. National Science Foundation funding in late 2000, focused on a communication protocol utilizing Dublin Core with infrastructure for the aggregation of basic language resource descriptions from catalogs (which in many cases also didn't exist).

³⁴ Add to this that language resources, especially those evidencing endangered languages, have content which is often inaccessible to memory institution staff due to language barriers, thus preventing memory institutions from performing curation and cataloging services otherwise performed in other contexts.

³⁵ <https://www.dublincore.org/specifications/openwemi>

³⁶ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms>

³⁷ In fact there were many more voices across several networks, each with their own sphere of influence and scholarly specialization. As the global discussion has evolved new voices continue to make their mark on the trajectory of language resource creation and stewardship.

³⁸ <https://web.archive.org/web/20120430232404/http://emeld.org/index.cfm>

³⁹ <https://web.archive.org/web/20111203195400/http://emeld.org/events/index.cfm>

⁴⁰ <https://linguistlist.org>

⁴¹ <https://listserv.linguistlist.org/pipermail/olac-general> and <https://listserv.linguistlist.org/pipermail/olac-implementers>. In the Internet Archive as: <https://web.archive.org/web/20170203144945/http://listserv.linguistlist.org/pipermail/olac-general> and <https://web.archive.org/web/20190220112210/https://listserv.linguistlist.org/pipermail/olac-implementers>.

⁴² <https://archive.mpi.nl/forums/t/indi-metadata-information/2933>

⁴³ ISLE stand for *International Standard for Language Engineering*.

⁴⁴ <http://search.language-archives.org>

Some scholars in the USA, such as Heidi Johnson, felt that IMDI offered significant advantages over other perceived options [59, 61]. A popular misconception at the time was that OLAC and IMDI were two options trying to accomplish the same ends. However, OLAC metadata never tried to provide answers for institutions related to administrative metadata, rather it was focused on the bare minimum for aggregated search and discovery—riding on the back of Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)⁴⁵ and Dublin Core. This has left some scholars to reject OLAC as a useful tool in the creation of collections of resources. However, it should only highlight that the management of language resources is a complex endeavor requiring many administrative decisions which are dependent upon: (1) institutional mandates, (2) legal jurisdictional requirements, and (3) the business model implemented to service the language resource users.

By 2010 work on IMDI had been superseded with work on Component Metadata Infrastructure (CMDI) and infrastructure stewarding the output of the DoBeS projects had been moved to The Max Planck Institute for Psycholinguistics (MPI) in Nijmegen. The Language Archive⁴⁶ also then affiliated with CLARIN.

The *Common Language Resources and Technology Infrastructure* (CLARIN) is an European Commission funded trans-national non-profit organization focused on facilitating Research.⁴⁷ Within the EU legal framework this class of organization is known as an European Research Infrastructure Consortium (ERIC). CLARIN was founded in 2012 to provide infrastructure support for language resource stewardship and reuse. While CLARIN is primarily focused on supporting scholars in Europe, its impact more broadly should not be underestimated. For example, the South African Centre for Digital Language Resources is a CLARIN member. Currently there are no United States based stewardship institutions which are members of CLARIN.

DARIAH,⁴⁸ founded in 2014, is also an ERIC but in contrast to CLARIN, it is an EU-centric support and infrastructure institution for digital arts and humanities. Since a significant number of digital language resources are created through humanities projects, DARIAH becomes a key infrastructure partner for these scholarly endeavors. Linguistic research using existing texts are often better supported through DARIAH than CLARIN but every research project has its own unique support needs.⁴⁹

META-SHARE is another member driven community with mostly European institutional members, launched in 2013.⁵⁰ META-SHARE [among others: 41, 45, 87, 88] maintains its own metadata standard and application profile [37].⁵¹ The META-SHARE application profile is more constrained than CMDI, which can be highly tailored for linguistic variables. There is some evidence that as of 2024 some META-SHARE members, such as Finland, are moving to

CLARIN-based infrastructure.⁵² Therefore, the relevance of META-SHARE may be waning.

DELANMAN,⁵³ the *Digital Endangered Languages and Musics Archives Network* is an association of nineteen language resource stewards. Often the institutions have significant collections of minority language resources, and whose institutional leadership have had a history with the E-MELD and OLAC communities.

Yet another network that emerged among those who use the term *language resources* was one involving scholars who use linked data in their language-centric scholarship [e.g., 23, 25, 84]. *Linked Data* is a way of implementing a database schema such that the schema can be used in different locations, mixed with other schemas, and where the pre-split schemas may each evolve along unique trajectories. For example, two different research institutions might start with the same data structure but decide to augment it independently. While data annotated with those augmentations will no longer be coalesced the same, the portions of the data structure in common use prior to the split will remain. Additionally, if each independent research party used a third party's data structure, then those would be coalesced the same. In this sense, linked data is a distributed database technology. Bosque-Gil et al. [12], in 2018, point out that even then there were over 120 different linked data schemas created for and used in linguistic research. There is a significant difference between how linked data technologies are used by librarians and archivists, linguists, and other text focused humanists. The difference is roughly this: linked data in stewardship contexts is focused on WEMI entity description (catalog access points); however, in linguistics and the digital humanities, [e.g., 34] the content of the WEMI entity is annotated much like an annotated edition. These annotations may range from what information professionals in libraries traditionally call *catalog access points* [e.g., 53, 67, 92] to individual Unicode characters. The expected use case is quite different between those who approach linked data from a resource-stewardship context and those who create annotated editions. Those in a stewardship context generally are concerned with resource level search and discovery, while those who create annotated editions (with the tools of linked data) are generally looking for aggregates of resources from which they can then produce derivative resources. Thus the tools and the technology of language resource management in communities using linked data are very different from other scholars.

Akin to, and partially overlapping with the linked data community in terms of annotation interests, are those scholars who use the Text Encoding Initiative (TEI).⁵⁴ These scholars develop and share XML-based processing tools to create digital editions, annotated editions of texts [e.g., 14] or encode dictionaries [e.g., 31, 71]. TEI is not a singular metadata and annotation schema but rather a framework from which scholars can choose various levels of detail suitable to their project requirements. Thus some “dialectical” variation may occur between the outputs of different TEI-using projects. Nancy Ide [56, 57], one of the founders of TEI, writes extensively on its history.

Even with these previously listed networks of scholars and institutions sharing overlapping interests, un-networked collections

⁴⁵<https://www.openarchives.org/pmh>

⁴⁶<https://archive.mpi.nl/tla>

⁴⁷<https://www.clarin.eu>

⁴⁸<https://www.dariah.eu>

⁴⁹How individual stewardship institutions relate to national and trans-national institutions in the European context is outside of the scope of the current paper. However, these relationships are of interest to many US based scholars and information professionals who work with language resources.

⁵⁰<http://www.meta-share.org>

⁵¹<http://www.meta-share.org/knowledgebase/overviewOfTheMetadataModel>

⁵²<https://www.kielipankki.fi/tools/metashare>

⁵³<https://www.delaman.org>

⁵⁴<https://tei-c.org>

and institutions which steward (often minority and under-described), language resources are extensive. For example, there are over 3,000 collections (containing multiple resources) where language, often a minority language, is a key feature of the collection indexed in Proquest's *Archive Finder*⁵⁵ service. These catalog records do not appear in the OLAC aggregator or the Glottolog,⁵⁶ and therefore estimates on the description status of languages that base their approximations on OLAC and Glottolog indexes are artificially minimizing quantities and increasing the long-tail of under-described languages [e.g., 55].

4 SOCIO-HISTORICAL EVOLUTION OF METADATA SCHEMAS

Since many of the metadata schemas used to describe language resources are not part of the set commonly taught in programs matriculating librarians and archivists, understanding them and their diversity is important. The last of the previously mentioned questions then is *how did we get here?* Or *What were the evolutionary trajectories of metadata schemas across institutions?* To better understand the historical development, Figure 2 presents several of the metadata schemas relevant to archives (broadly), language archives, and scholarly communications. The historical development broadly follows this: The OLAC community developed commensurate with the Dublin Core and the OAI-PMH communities. The IMDI community became part of the CMDI community. CMDI is, in a way, like linked data. That is, there is a large pool of metadata "fields" which are used to create a collection level application profiles. In this way CMDI is much more complex than MARC, but it doesn't permit the use metadata fields from outside of the option pool unless they are universally added for all CMDI users. In a sense, this aspect of CMDI is analogous to application profiles in *Resource Description & Access*.

Broadly within the context of digital language archives, the metadata is resource-depositor supplied. This puts administrative pressure on managers to keep metadata schemas lean. However, this comes at a significant cost to discovery for re-usability. Dublin Core, via OAI-PMH is continuing to be used and is required of OpenAIRE⁵⁷ compliant repositories. Dublin Core is also expanding its market presence in many of the unaffiliated language archives as many of the digital asset management software tools have it bundled with them (e.g., DSpace).

5 CONCLUSIONS

The history of language archives is defined by the social history of stewards and the contexts defining the relationships between stewards and content creators. Networks form for different reasons: sometimes it is an ethos—a way of treating the evidentiary record (e.g., Carl Stumpf, Erich Moritz von Hornbostel, and George Herzog, Jaap Kunst, Klaus Wachsmann, etc.), sometimes it is infrastructure (e.g., CLARIN), sometimes it is search and discovery—an idealized service environment (e.g., OLAC), and sometimes it is the materiality of the scholarly work or its production process (e.g., linked data). Open (gratis) business models meet the economical

need that many stewards and minority language using communities share. But *minority* language using communities are not the only language communities, and language resources are needed by *all* communities of language users.

Engaging with the long history of language resource stewardship allows us to understand how stewardship contexts and the profession have evolved. Today's professional choices in the name of activism, decolonization, and liberation, do not necessarily negate, vilify, or diminish the significance of past scholarly works. There is a growing body of literature re-interpreting or adding historical context to archived language resources. For example, the scholarly discussion around specific recordings made by Wachsmann in Uganda provides descriptive insight where one might be tempted to otherwise press for newly advocated interpretations [26, 80, 81].

By embracing the long history of language resource stewardship, it brings more stewards together in a momentum not experienced by those stewards in the DELAMAN network—or any of the previously mentioned networks. The long term funding of CLARIN has created a context where institutional managers and stewards find long term value for persistence, but issues such as broad resource re-usability and source community engagement may be lower priorities.

Understanding the long history of *Language Archives* through the lens of related professional concepts (and terminology) helps us contextualize the need for search and discovery of language resources—including language focused scholarly literature. Search and discovery tools focused on resource discovery within the niche impacted by language resource management is still an unsolved challenge for the business model success of many stewardship institutions. That is, there are a significant number of stewardship institutions which can benefit from an open economic model that centers collections and resources for those using language and linguistic concepts as information access gateways. There is no mainstream approach or filter on mainstream technologies that can fix the "missing" catalog access points, because those approaches assume that cataloging is being done for these resources in a way that is relevant to those searching by language and linguistic concepts.

Considering then the challenges afflicting search and discovery, it is equally important to maintain clear distinctions between *an Archives*, *an archive*, and a *language resource*. This holds for the information professional and the language scholar.

ACKNOWLEDGMENTS

I am indebted to Paul Unger and Dr. Rebecca Paterson for many proof reading comments. I am also grateful to an anonymous reviewer who clarified some inter-agency relationships among European based institutions.

REFERENCES

- [1] Anthony Aristar. 2001. *E-MELD: Electronic Metastructure for Endangered Languages Data*. Application Proposal Narrative for NSF Award 0094934. National Science Foundation, Washington, D.C. <https://ui.adsabs.harvard.edu/abs/2001nsf....0094934A/abstract> ; https://www.academia.edu/86348556/E_MELD_Electronic_Metastructure_for_Endangered_Languages_Data.

⁵⁵<https://www.proquest.com/archivefinder/index>

⁵⁶<https://glottolog.org>

⁵⁷<https://www.openaire.eu>

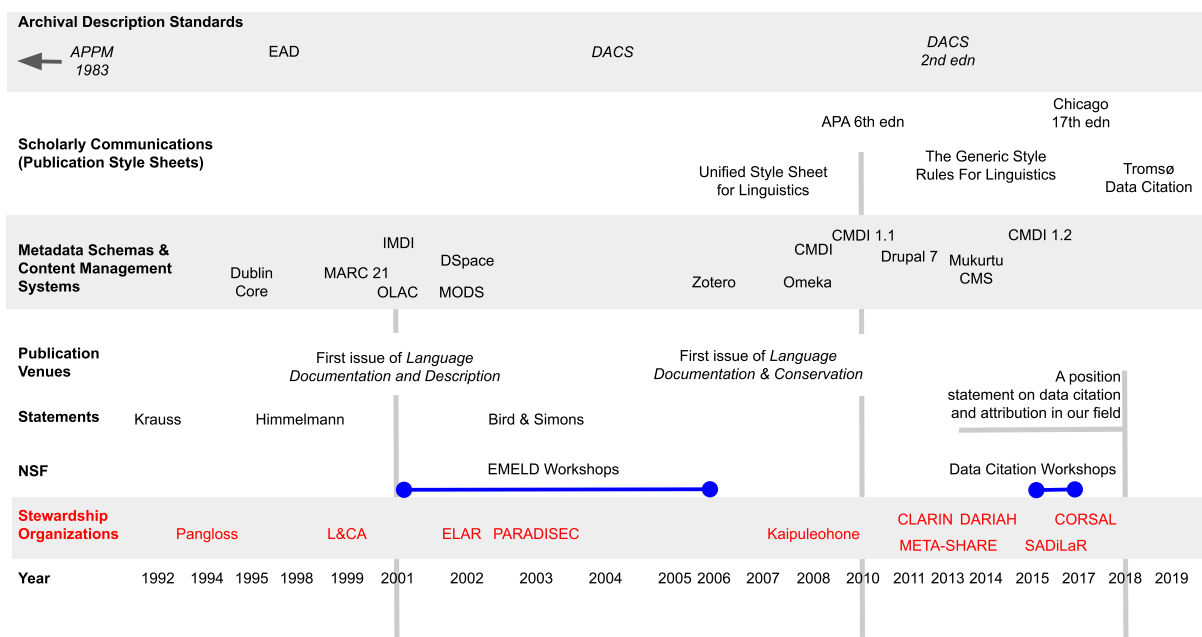


Figure 2: Timeline indicating the starting points of major archival information standards, the founding of several language archives, publishing technology, and key events in the field of language documentation.

- [2] Helen Aristar-Dry. 2002. E-MELD: Overview and Update. In *Proceedings of the International Workshop on Resources and Tools in Field Linguistics, a Workshop at the Third International Conference on Language Resources and Evaluation (LREC'02)* (Las Palmas, Spain), Peter Wittenburg (Ed.). LERC / European Language Resource Association (ELRA), Las Palmas, Canary Islands, Spain, 8 pages. <https://www.mpi.nl/lrec/2002/papers/lrec-pap-03-helen-LREC-DOBES-Workshop.pdf>; <https://www.mpi.nl/lrec/2002/program.html>.
- [3] Andrea L Berez-Kroeker and Ryan E Henke. 2018. Language Archiving. In *The Oxford Handbook of Endangered Languages*, Kenneth L Reh and Lyle Campbell (Eds.). Oxford University Press, Oxford, UK, 347–368.
- [4] Anna Maria Busse Berger. 2013. Spreading the Gospel of *Singbewegung*: An Ethnomusicologist Missionary in Tanganyika of the 1930s. *Journal of the American Musicological Society* 66, 2 (2013), 475–522. <https://doi.org/10.1525/jams.2013.66.2.475>
- [5] Jack Berry and Earl W. Stevick. 1969. Oral Data Collecting and Linguistics in Africa. *Journal of the Folklore Institute* 6, 2/3 (1969), 93–110. <https://doi.org/10.2307/3814171>
- [6] Steven Bird and Gary F. Simons. 2003. Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. *Computers and the Humanities* 37, 4 (2003), 375–388. <https://doi.org/10.1023/A:1025720518994>
- [7] Steven Bird and Gary F. Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79, 3 (2003), 557–582. <https://doi.org/10.1353/lan.2003.0149>
- [8] Steven Bird and Gary F. Simons. 2004. Building an Open Language Archives Community on the DC Foundation. In *Metadata in Practice*, Diane Hillmann and Elaine L. Westbrooks (Eds.). American Library Association, Chicago, 203–222.
- [9] Steven Bird and Gary F. Simons. 2021. Towards an Agenda for Open Language Archiving. In *Proceedings of the International Workshop on Digital Language Archives: LangArc 2021*, Oksana Zavalina and Shobhana Lakshmi Chelliah (Eds.). University of North Texas, Denton, Texas, 25–28. <https://doi.org/10.12794/langarc1851171>
- [10] Steven Bird and Gary F. Simons. 2022. The Open Language Archives Community: A 20-Year Update. *The Electronic Library* 40, 5 (2022), 507–524. <https://doi.org/10.1108/EL-08-2022-0192>
- [11] Steven Bird, Gary F. Simons, and Huang Chu-Ren. 2001. The Open Language Archives Community and Asian Language Resources. In *Proceedings of the Workshop on Language Resources in Asia, 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, Tokyo, November 27–30, 2001. The Association for Natural Language Processing (Japan), Tokyo, Japan, 31–38. <https://doi.org/10.48550/arXiv.cs/0110014> Proceedings archived in the *Internet Archive* at: <https://web.archive.org/web/20060709083754/https://www.afnlp.org/nlprs2001>.
- [12] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda, and A. Gómez-Pérez. 2018. Models to Represent Linguistic Linked Data. *Natural Language Engineering* 24, 6 (2018), 811–859. <https://doi.org/10.1017/S1351324918000347>
- [13] Catherine Bow, Michael J. Christie, and Brian Devlin. 2014. Developing a Living Archive of Aboriginal Languages. *Language Documentation & Conservation* 8 (2014), 345–360. <http://hdl.handle.net/10125/24612>
- [14] Jack Bowers. 2020. *Language Documentation and Standards in Digital Humanities: TEI and the Documentation of Mixtepec-Mixtec*. PhD. Dissertation. École Pratique des Hautes Études, Paris, France. <https://theses.hal.science/tel-03131936>
- [15] Jessica Boynton, Steven Moran, Helen Aristar-Dry, and Anthony Rodrigues Aristar. 2010. Using the E-MELD School of Best Practices to Create Lasting Digital Documentation. In *Language Documentation: Practice and Values*, Lenore A. Grenoble and N. Louanna Furbie (Eds.). John Benjamins Publishing Company, Amsterdam, The Netherlands, 133–146. <https://doi.org/10.1075/z.158.14boy>
- [16] Daan Broeder and Peter Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies* 1, 2 (2006), 119–132. <https://doi.org/10.1504/IJMSO.2006.011008>
- [17] Marie-France Calas. 1990. Les Archives Sonores En France: La Phonothèque Nationale. *Fontes Artis Musicae* 37, 3 (1990), 266–272. <https://www.jstor.org/stable/23507904>
- [18] Marie-France Calas. 1992. Le Département De La Phonothèque et de L'audiovisuel À La Bibliothèque Nationale. *Langue Française* 93 (1992), 120–124. <https://doi.org/10.3406/lfr.1992.5814>
- [19] Michelle L. Caswell. 2016. “The Archive” Is Not an Archives: On Acknowledging the Intellectual Contributions of Archival Studies”. *Reconstruction* 16, 1 (2016), <https://escholarship.org/uc/item/7bn4v1fk>.
- [20] Damir Cavar and Malgorzata Cavar. 2014. Visualization of Language Relations and Families: MultiTree. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland, 698–701. <https://aclanthology.org/L14-1019>

- [21] Wallace L. Chafe (Ed.). 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex Publication Corp, Norwood, New Jersey. <http://catdir.loc.gov/catdir/enhancements/fy1511/80120122-t.html>
- [22] Shobhana L. Chelliah. 2021. *Why Language Documentation Matters*. Springer International Publishing, Cham, Switzerland. <https://doi.org/10.1007/978-3-030-66190-8>
- [23] Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P. McCrae. 2013. Linguistic Linked Open Data (LLOD). Introduction and Overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data* (Pisa, Italy), Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P. McCrae (Eds.). Association for Computational Linguistics, Kerrville, Texas, i – xi. <https://aclanthology.org/W13-5501>
- [24] G. Tucker Childs. 2020. Language Documentation and Revitalization. In *The Cambridge Introduction to Applied Linguistics*, Susan Conrad, Alissa Hartig, and Lynn Santelmann (Eds.). Cambridge University Press, Cambridge, UK, 204–218. <https://doi.org/10.1017/9781108658089.019>
- [25] Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing, Cham, Switzerland. <https://doi.org/10.1007/978-3-030-30225-2>
- [26] Peter Cooke. 2015. A Response to Sylvia Nannyonga-Tamusuza and Andrew N. Weintraub's "The Audible Future: Reimagining the Role of Sound Archives and Sound Repatriation in Uganda". *Ethnomusicology* 59, 3 (2015), 475–479. <https://doi.org/10.5406/ethnomusicology.59.3.0475>
- [27] Karen Coyle. 2022. Works, Expressions, Manifestations, Items: An Ontology. *The Code4Lib Journal* 53 (2022), Article: 16491. <https://journal.code4lib.org/articles/16491>
- [28] Karen Coyle. 2023. openWEMI. In *Proceedings of the 2022 International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, Dublin, Ohio, 7 pages. <https://doi.org/10.23106/DCMI.953115290>
- [29] James Curral, Michael Moss, and Susan Stuart. 2004. What Is a Collection? *Archivaria* 58 (2004), 131–146. <https://www.archivaria.ca/index.php/archivaria/article/view/12480>
- [30] Philip D. Curtin. 1968. Field Techniques for Collecting and Processing Oral Data. *The Journal of African History* 9, 3 (1968), 367–385.
- [31] Ewa Czaykowska-Higgins, Martin D. Holmes, and Sarah M. Kell. 2014. Using TEI for an Endangered Language Lexical Resource: The Nxa?amxcin Database-Dictionary Project. *Language Documentation & Conservation* 8 (2014), 1–37. <http://hdl.handle.net/10125/4604>
- [32] Tjeerd de Graaf. 2009. The Use of Historical Documents and Sound Recordings for the Study and Safeguarding of Endangered Languages. In *FELXIII: Endangered Languages and History*, Hakim Elnazarov and Nicholas Ostler (Eds.). Endangered Language Foundation, Khorog, Tadjikistan, 27–32.
- [33] Tjeerd de Graaf. 2011. Endangered Languages and the Use of Sound Archives. *International Journal of Asia-Pacific Studies* 7, 1 (2011), 27–46. http://ijaps.usm.my/?page_id=477
- [34] Thierry Declerck, Piroska Lendvai, Karlheinz Mörth, Gerhard Budin, and Tamás Váradi. 2012. Towards Linked Language Data for Digital Humanities. In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann (Eds.). Springer, Berlin, Heidelberg, 109–116. https://doi.org/10.1007/978-3-642-28249-2_11
- [35] Roger Decollogne. 1976. Les archives sonores et la Phonothèque nationale. *La Gazette des archives* 92, 1 (1976), 21–27. <https://doi.org/10.3406/gazar.1976.2479>
- [36] Robert Finly Delaney. 1951. The Administration of Intelligence Archives. *College and Research Libraries* 12, 3 (1951), 213–219;232. https://doi.org/10.5860/crl_12_03_213
- [37] Elina Desipri, Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Francesca Frontini, Monica Monachini, Valérie Mapelli, Gil Francopoulo, and Thierry Declerck. 2015. Documentation and User Manual of the META-SHARE Metadata Model. META-SHARE, n.p. http://www.meta-share.org/assets/pdf/META-SHARE_Documentation_v3.1.pdf
- [38] David King Dunaway. 1987. Field Recording Oral History. *The Oral History Review* 15, 1 (1987), 21–42.
- [39] Patience L. Epps, Anthony K. Webster, and Anthony C. Woodbury. 2017. A Holistic Humanities of Speaking: Franz Boas and the Continuing Centrality of Texts. *International Journal of American Linguistics* 83, 1 (2017), 41–78. <https://doi.org/10.1086/689547>
- [40] Patrick Feaster. 2010. Édouard-Léon Scott de Martinville: An Annotated Discography. *ARSC Journal* 41, 1 (2010), 43–82.
- [41] Christian Federmann, Ioanna Giannopoulou, Christian Girardi, Olivier Hamon, Dimitris Mavroeidis, Salvatore Minutoli, and Marc Schröder. 2012. META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari, Khalid Choukri, Thierry DeClerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 1090–1097. <https://aclanthology.org/L12-1483>
- [42] J. Walter Fewkes. 1890. A Contribution to Passamaquoddy Folk-Lore. *The Journal of American Folklore* 3, 11 (Oct. 1890), 257–280. <https://doi.org/10.2307/534065>
- [43] Jesse Walter Fewkes. 1890. On the Use of the Phonograph in the Study of the Languages of American Indians. *Science* 15, 378 (1890), 267–269. <https://doi.org/10.1126/science.ns-15.378.267.b>
- [44] William A. Foley. 1999. Review of *The Korowai of Irian Jaya*. *Language in Society* 28, 3 (1999), 470–472. <https://doi.org/10.1017/S0047404599303067>
- [45] Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Monica Monachini, Francesca Frontini, Thierry Declerck, Victoria Arranz, and Valerie Mapelli. 2012. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari, Khalid Choukri, Thierry DeClerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 1090–1097. <https://aclanthology.org/L12-1647>
- [46] Anne J. Gilliland, Sue McKemmish, and Andrew J. Lau. 2016. *Research in the Archival Multiverse*. Monash University Publishing, Clayton, Victoria, Australia. https://doi.org/10.26530/oaopen_628143
- [47] Jeff Good. 2018. Reflections on the Scope of Language Documentation. In *Reflections on Language Documentation 20 Years after Himmelmann* 1998. Number 15 in Language Documentation & Conservation Special Publications. University of Hawai'i Press, Honolulu, Hawai'i, 13–21. <http://hdl.handle.net/10125/24804>
- [48] Ken Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. Endangered Languages. *Language* 68, 1 (1992), 1–42. <https://doi.org/10.2307/416368>
- [49] Patrick James Hall. 2022. *Participatory Design in Digital Language Documentation: A Web Platform Approach*. PhD. Dissertation. University of California, Santa Barbara, Santa Barbara, California. <https://www.proquest.com/openview/db943d69bd368295036b639ce12d06dd/1>
- [50] Gottlob David Hartmann. 1890. *Von Und Aus Schwaben: Ein Lebensbild Aus Der Sturm- Und Drangzeit*. Number 7 in Geschichte, Biographie, Litteratur Von Wilhelm Lang. Verlag von W. Kohlhammer, Stuttgart, Germany.
- [51] Ryan Henke and Andrea L. Berez-Kroeker. 2016. A Brief History of Archiving in Language Documentation, with an Annotated Bibliography. *Language Documentation & Conservation* 10 (2016), 411–457. <http://hdl.handle.net/10125/24714>
- [52] Nikolaus P. Himmelmann. 1998. Documentary and Descriptive Linguistics. *Linguistics* 36, 1 (1998), 161–195. <https://doi.org/10.1515/ling.1998.36.1.161>
- [53] Birger Hjørland and Lykke Kylesbech Nielsen. 2001. Subject Access Points in Electronic Retrieval. *Annual Review of Information Science and Technology* 35 (2001), 249–298. <https://babel.hathitrust.org/cgi/pt?id=mdp.39015052101956&seq=269>
- [54] Gary Holton. 2012. Language Archives: They're Not Just for Linguists Any More. In *Potentials of Language Documentation: Methods, Analyses, and Utilization*, Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (Eds.). University of Hawai'i Press, Honolulu, Hawai'i, 111–117. <http://scholarspace.manoa.hawaii.edu/handle/10125/4523>
- [55] Kavon Hooshair. 2015. A Quantitative Analysis of Linguistic Metadata. <http://hdl.handle.net/10125/25299> Paper Presented at the 4th International Conference on Language Documentation and Conservation (ICLDC 4), February 26th–March 1st, Honolulu, HI. The University of Hawai'i at Mānoa. <http://hdl.handle.net/10125/25299..>
- [56] Nancy Ide. 2003. Fifteen (and a Half) Years of the TEI: A Retrospective and Prospective. In *Presenations at the third annual TEI Members Meeting, 7-8 November 2003*. TEI, Université de Nancy, Nancy, France., 35 pages. <https://www.tei-c.org/Vault/MembersMeetings/2003-info/ide.pdf>
- [57] Nancy Ide and Jean Véronis (Eds.). 1995. *Text Encoding Initiative*. Springer Netherlands, Dordrecht. <https://doi.org/10.1007/978-94-011-0325-1>
- [58] IFLA Study Group on the Functional Requirements for Bibliographic Records. 1998 [2009]. *Functional Requirements for Bibliographic Records: Final Report* (2 ed.). K.G. Saur, Munich, Germany. <https://www.ifla.org/VII/s13/frbr/> <https://www.ifla.org/VII/s13/frbr>
- [59] Heidi Johnson. 2002. The Archive of the Indigenous Languages of Latin America: Goals and Visions. In *Proceedings of the International Workshop on Resources and Tools in Field Linguistics, a Workshop at the Third International Conference on Language Resources and Evaluation*. European Language Resources Association, Las Palmas, Spain, 4 pages. <https://www.mpi.nl/lrec/2002/program.html>
- [60] Heidi Johnson. 2003. EMELD Working Group on Resource Archiving [Presentation Slides]. In *Language Digitization Project, Conference 2003: Digitizing and Annotating Texts and Field Recordings*. E-MELD, East Lansing, Michigan,

- 24 pages. https://web.archive.org/web/20190903185349/https://islandora-ailla.lib.utexas.edu/sites/default/files/documents/archive_bp.pdf
- [61] Heidi Johnson and Arienne Dwyer. 2002. Customizing the IMDI Metadata Schema for Endangered Languages. In *Proceedings of the International Workshop on Resources and Tools in Field Linguistics, a Workshop at the Third International Conference on Language Resources and Evaluation (LREC'02)*, Peter Wittenburg (Ed.), LERC / European Language Resource Association (ELRA), Las Palmas, Canary Islands, Spain, 4 pages. <https://www.mpi.nl/lrec/2002/papers/lrec-pap-05-JohnsonDwyer.pdf>
- [62] Aude Julien da Cruz Lima. 2011. The CNRS - Musée de l'homme Audio Archives: A Short Introduction. *IASA journal* 36 (2011), 33–37. <https://www.iasa-web.org/iasa-journal-no-36-january-2011>
- [63] Judith Kaplan and Rebecca Lemov. 2019. Archiving Endangerment, Endangered Archives: Journeys through the Sound Archives of Americanist Anthropology and Linguistics, 1911–2016. *Technology and Culture* 60, 2 (2019), S161–S187. <https://doi.org/10.1353/tech.2019.0067>
- [64] Roshanak Kheshti. 2015. The Female Sound Collector and Her Talking Machine. In *Modernity's Ear*. NYU Press, New York, USA, 15–38. <https://doi.org/10.18574/nyu/9781479867011.003.0002>
- [65] Stéphanie Khoury and Joséphine Simonnot. 2014. Applications and Implications of Digital Audio Databases for the Field of Ethnomusicology: A Discussion of the CNRS — Musée de l'Homme Sound Archives. *First Monday* 19, 10 (2014), n.p. <https://doi.org/10.5210/fm.v19i10.5554>
- [66] Eleanor Kleiber, Andrea L. Berez-Kroeker, Michael Chopey, Danielle Yarbrough, and Ryan Shelby. 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 (2018), 110–122. <https://doi.org/10.1353/cp.2018.0005>
- [67] Gunnar Knutson. 1986. Does the Catalog Record Make a Difference? Access Points and Book Use. *College & Research Libraries* 47, 5 (1986), 460–469. https://doi.org/10.5860/crl.47_05_460
- [68] Olivia Landry. 2023. Colonial Listening and the Making of a Sound Archive. In *A Decolonizing Ear*. University of Toronto Press, Toronto, Ontario, Canada, 28–56. <https://www.jstor.org/stable/10.3138/j.ctv31nzk9x.6>
- [69] Werner F. Leopold. 1959. The Decline of German Dialects. *WORD* 15, 1 (1959), 130–153. <https://doi.org/10.1080/00437956.1959.11659689>
- [70] LinguistList. 2023. linguistlist/multitree: MultiTree: A digital library of language relationships. Archived at Zenodo [Dataset comprising the MultiTree project, v0.1] (2023). <https://doi.org/10.5281/zenodo.10006641>
- [71] Megan Lukaniec and Martin Holmes. 2023. Modern Wendat Lexicography: Using XML to Reflect the Grammar and Lexicon of an Iroquoian Language. *Dictionaries: Journal of the Dictionary Society of North America* 44, 2 (2023), 75–105. <https://doi.org/10.1353/dic.2023.a915066>
- [72] Florian Matter. 2021. Combining Description and Documentation: A Digital Boasian Trilogy. In *Posters of the 7th International Conference on Language Documentation and Conservation (ICLDC)*. University of Hawai'i at Mānoa. <http://hdl.handle.net/10125/74479>
- [73] Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton (Eds.). 2018. *Reflections on Language Documentation 20 Years After Himmelmann 1998*. University of Hawai'i Press, Honolulu, Hawai'i. <http://hdl.handle.net/10125/24800>
- [74] Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and Researching Endangered Languages: The Pangloss Collection. *Language Documentation & Conservation* 8 (2014), 119–135. <http://hdl.handle.net/10125/4621>
- [75] Steven Moran. 2005. The E-MELD School of Best Practices in Digital Language Documentation. *Proceedings from the Annual Meeting of the Chicago Linguistic Society* 41, 2 (2005), 71–80.
- [76] MPI IMDI Team. 2003. *IMDI. Metadata Elements for Session Descriptions* (draft proposal version 3.0.4 ed.). Max Planck Institute, Nijmegen, The Netherlands. http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf Via the Internet Archive: https://web.archive.org/web/20090411054359/http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf
- [77] MPI IMDI Team. 2009. *Metadata Elements for Catalogue Descriptions* (version 3.0.13 ed.). Max Planck Institute, Nijmegen, The Netherlands. http://www.mpi.nl/IMDI/documents/Proposals/IMDI_Catalogue_3.0.0.pdf Via the Internet Archive: http://www.mpi.nl/IMDI/documents/Proposals/IMDI_Catalogue_3.0.0.pdf
- [78] S. Muller, Johan Adriaan Feith, and R. Fruin. 2003. *Manual for the Arrangement and Description of Archives* (second edition ed.). Society of American Archivists, Chicago, IL.
- [79] Tara E. Murray. 2013. What's So Special About Special Libraries? *Journal of Library Administration* 53, 4 (2013), 274–282. <https://doi.org/10.1080/01930826.2013.865395>
- [80] Sylvia Nannyonga-Tamusuza and Andrew N. Weintraub. 2012. The Audible Future: Reimagining the Role of Sound Archives and Sound Repatriation in Uganda. *Ethnomusicology* 56, 2 (2012), 206–233. <https://doi.org/10.5406/ethnomusicology.56.2.0206>
- [81] Sylvia Nannyonga-Tamusuza and Andrew N. Weintraub. 2015. A Response to Peter Cooke's Response. *Ethnomusicology* 59, 3 (2015), 480–482. <https://doi.org/10.5406/ethnomusicology.59.3.0480>
- [82] Chris V. Nicholson. 2009. Linguist's Preservation Kit Has New Digital Tools. *The New York Times Online* edition: Science. <https://www.nytimes.com/2009/07/28/science/28prof.html>
- [83] Edward O'Neill, Maja Žumer, and Jeffrey Mixer. 2015. FRBR Aggregates: Their Types and Frequency in Library Collections. *Library Resources & Technical Services* 59, 3 (2015), 120–129. <https://doi.org/10.5860/Lrts.59n3.120>
- [84] Antonio Pareja-Lora, Maria Blume, Barbara C. Lust, and Christian Chiarcos (Eds.). 2020. *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. The MIT Press, Cambridge, Massachusetts. <https://doi.org/10.7551/mitpress/10990.001.0001>
- [85] Hugh J. Paterson, III. 2023. Diversity and Identity: Categories for OAI Data-Providers in the Open Language Archives Network. *Proceedings from North American Symposium on Knowledge Organization (NASKO)* 9, 1 (2023), 18–31. <https://doi.org/10.7152/nasko.v9i1.16301>
- [86] Lee Pederson. 1974. Tape/Text and Analogues. *American Speech* 49, 1/2 (1974), 5–23. <https://doi.org/10.2307/3087915>
- [87] Stelios Piperidis. 2012. The META-SHARE Metadata Schema for the Description of Language Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari, Khalid Choukri, Thierry DeClerck, Mehmet Ugur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Paris, France, 36–42. <https://aclanthology.org/L12-1647>
- [88] Stelios Piperidis, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi. 2014. META-SHARE: One year after. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland, 1532–1538. <https://aclanthology.org/L14-1610>
- [89] Patricia Polansky. 2004. Slavic Collection Descriptions - University of Hawaii Collection. *The Slavonic and East European Journal* 5, 3 (2004), 131–133. <http://hdl.handle.net/10125/1524>
- [90] Ivan Polunin. 1970. Visual and Sound Recording Apparatus in Ethnographic Fieldwork. *Current Anthropology* 11, 1 (1970), 3–22.
- [91] Kenneth M. Price. 2009. Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name? *Digital Humanities Quarterly* 3, 3 (2009), Article 53. <https://dhq.digitalhumanities.org/vol/3/3/000053/000053.html>
- [92] Pal V. Ral. 1982. The Relationship between Card Catalog Access Points and the Recorded Use of Education Books in a University Library. *College & Research Libraries* 43, 4 (1982), 341–345. https://doi.org/10.5860/crl.43_04_341
- [93] D Will Reiman. 2010. Basic Oral Language Documentation. *Language Documentation & Conservation* 4 (2010), 254–268. <http://hdl.handle.net/10125/4479>
- [94] Pat Riva, Patrick Le Boeuf, and Maja Žumer (Eds.). 2017. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information* (December 2017 ed.). International Federation of Library Associations and Institutions (IFLA), Den Haag, The Netherlands. <https://www.ifla.org/publications/node/11412>
- [95] Thomas Schmidt, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. 2006. Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art* (Michigan State University in East Lansing, Michigan). E-MELD, East Lansing, Michigan, 25 pages. <https://web.archive.org/web/20110514212211/http://www.emeld.org/workshop/2006/proceedings.html>
- [96] Anthony Seeger. 1986. The Role of Sound Archives in Ethnomusicology Today. *Ethnomusicology* 30, 2 (1986), 261–276. <https://doi.org/10.2307/851997>
- [97] Frank Seidel. 2016. Documentary Linguistics: A Language Philology of the 21st Century. *Language Documentation and Description* 13 (2016), 23–63. <https://doi.org/10.25894/ldd154>
- [98] Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language Documentation Twenty-Five Years On. *Language* 94, 4 (2018), e324–e345. <https://muse.jhu.edu/pub/24/article/712110>
- [99] Eva Semertzaki. 2011. *Special Libraries as Knowledge Management Centres*. Chandos, Oxford.
- [100] Gary F. Simons. 2016. From Linguistic Data Type to Language Resource Type: Laying the Groundwork for a Metadata Application Profile. https://scholars.sil.org/sites/scholars/files/gary_f_simons/presentation/simons-language_resource_type_vocabulary.pdf
- [101] Society of American Archivists. 2013. *Describing Archives: A Content Standard* (2nd ed.). Society of American Archivists, Chicago, Illinois. http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf

- [102] Ulf-Michael Stift and Thomas Schmidt. 2014. Mündliche Korpora Am IDS: Vom Deutschen Spracharchiv Zur Datenbank Für Gesprochenes Deutsch. In *Ansichten und Einsichten: 50 Jahre Institut für Deutsche Sprache*. Institut für Deutsche Sprache, Mannheim, Germany, 360–375.
- [103] Ernst Surkamp. 1926. *Die Sprechmaschine Als Hilfsmittel Für Unterricht Und Studium Der Neuern Sprachen* (2 ed.). Sperleng, Stuttgart.
- [104] Barbara B. Tillett. 2009. Definition of Aggregates as Works: Tillett Proposal. <https://www.ifla.org/files/assets/cataloguing/frbrg/aggregates-as-works.pdf>
- [105] Viktoria Tkaczyk. 2023. Early Sound Archiving and the Making of Scientific Resources. *Isis* 114, 2 (2023), 373–379. <https://doi.org/10.1086/724797>
- [106] Charles Frederick Voegelin. 1950. Magnetic Recording of American Indian Languages and the Relationship of This to Other Kinds of Memory. *Proceedings of the American Philosophical Society* 94, 3 (1950), 295–300.
- [107] Charles Frederick Voegelin and Florence Marie Robinett. 1954. Obtaining a Linguistic Sample. *International Journal of American Linguistics* 20, 2 (April 1954), 89–100. <https://doi.org/10.1086/464259>
- [108] Charles Frederick Voegelin and Florence Marie Voegelin. 1977. *Classification and Index of the World's Languages*. Elsevier, New York.
- [109] Brian M. Watson. 2021. Please Stop Calling Things Archives: An Archivist's Plea. *Perspectives on History* 59(1): <https://www.historians.org/research-and-publications/perspectives-on-history/january-2021/please-stop-calling-things-archives-an-archivists-plea>.
- [110] Rulon Wells. 1954. Archiving and Language Typology. *International Journal of American Linguistics* 20, 2 (1954), 101–107. <https://doi.org/10.1086/464260>
- [111] Stanley N. Werbow. 1967. Toward a Taxonomy of Spoken German. In *Report of the Eighteenth Annual Round Table Meeting on Linguistics and Language Studies*, Edward L. Blansitt, Jr. (Ed.). Number 20 in Monograph Series on Languages and Linguistics. Georgetown University Press, Washington, D.C., 151–158. <https://repository.digital.georgetown.edu/handle/10822/555460>
- [112] John David Allison Widdowson. 1968. The Dialect of Fortune Harbour, Newfoundland: A Pronouncing Glossary. *Folia Linguistica* 2, 3–4 (1968), 316–326. <https://doi.org/10.1515/flin.1968.2.3-4.316>
- [113] Peter Wittenburg. 2003. The DOBES Model of Language Documentation. *Language Documentation and Description* 1 (2003), 122–139. <https://doi.org/10.25894/ldd310>
- [114] meLê yamomo and Barbara Titus. 2021. The Persistent Refrain of the Colonial Archival Logic / Colonial Entanglements and Sonic Transgressions: Sounding Out the Jaap Kunst Collection. *The World of Music* 10, 1 (2021), 39–70.
- [115] Geoffrey Yeo. 2009. Custodial History, Provenance, and the Description of Personal Records. *Libraries & the Cultural Record* 44, 1 (2009), 50–64. <https://www.jstor.org/stable/25549536>
- [116] Geoffrey Yeo. 2012. The Conceptual Fonds and the Physical Collection. *Archivaria* 73 (2012), 43–80. <https://archivaria.ca/index.php/archivaria/article/view/13384>
- [117] Geoffrey Yeo. 2015. Contexts, Original Orders, and Item-Level Orientation: Responding Creatively to Users' Needs and Technological Change. *Journal of Archival Organization* 12, 3–4 (2015), 170–185. <https://doi.org/10.1080/15332748.2015.1048626>
- [118] Geoffrey Yeo. 2018. *Records, Information and Data: Exploring the Role of Record Keeping in an Information Culture*. Facet, Cambridge, UK. <https://doi.org/10.29085/9781783302284>
- [119] Eberhard Zwirner. 1983. Fünfzig Jahre Deutsches Spracharchiv. *Zeitschrift für Dialektologie und Linguistik* 50, 1 (1983), 35–43. <https://www.jstor.org/stable/40501783>

Received 20 October 2025; revised 6 December 2025