Journal of Open
Humanities Data

Accessibility, Discoverability, and Functionality: An Audit of and Recommendations for Digital Language Archives

RESEARCH PAPER

]u[ubiquity press

IRENE YI
AMELIA LAKE
JUHYAE KIM
KASSANDRA HAAKMAN
JEREMIAH JEWELL
SARAH BABINSKI
CLAIRE BOWERN

**Author affiliations can be found in the back matter of this article*

ABSTRACT

While digital archiving has long been standard for linguistics, archives themselves are heterogeneous (Aznar & Seifart 2020), and archived linguistic material is important for researchers and communities, particularly for language reclamation (cf. Baldwin & Olds 2007; Whalen et al. 2016; Hinton 2003, 2018; Kung et al. 2020). The format and usability of scholarly archival collections is shaped by the functions of the management practices at the stewarding institution, making an appreciation of the range of access services provided by such institutions relevant to the evaluation of individual collections.

Here we report on a review of 41 digital language archives. Three factors are examined: 1) **accessibility**, including metadata and site navigation; 2) **discoverability**, or searchability and internal navigation; and 3) **functionality**, the overall ease of data retrieval and use. We recognize that the decisions made by both stewards and depositors can greatly impact the accessibility of archived materials; to that end, we present **recommendations** for how archives might increase the utility of their holdings for their users. We emphasize that our intention is not to dissuade linguists from using archives because of these issues, and we recognize the tremendous amount of work that goes into the upkeep of digital infrastructure, often with very limited institutional support. Implementing such recommendations at an institutional level can establish a fairer peer-review process of archival collections. By delineating precisely what standards fall under the archive management level and what procedures individual depositors are responsible for, the roles of "archivist" and "depositor" become clearer.

CORRESPONDING AUTHOR:

Irene Yi

Linguistics Department, Yale University, New Haven, CT, US

irene.yi@yale.edu

KEYWORDS:
language archives; documentation; accessibility; discoverability; functionality; linguistics; endangered languages; metadata

TO CITE THIS ARTICLE:
Yi, I., Lake, A., Kim, J., Haakman, K., Jewell, J., Babinski, S., & Bowern, C. (2022). Accessibility, Discoverability, and Functionality: An Audit of and Recommendations for Digital Language Archives. *Journal of Open Humanities Data*, 8: 10, pp. 1–19. DOI: *https://doi.org/10.5334/johd.59*

1 INTRODUCTION

It is estimated that 32% of living languages are currently in some state of loss (Simons & Lewis 2013:10); some estimates place the figure at closer to 50% (Campbell et al. 2013; Campbell & Belew 2018). Documentation of endangered languages is vital for preserving them (Berez 2013), whether for study, language reclamation, preventing the irreversible loss of intangible cultural heritage, or any other reason that language is used.

Digital archiving has been standard for linguistics for at least 15 years, but the extent to which this material can be accessed and used for research, education, and activism varies (cf. Evans & Sasse 2004; Kaplan & Lemov 2019; Paterson 2021). Language archives utilize a number of different content management systems and do not provide uniform functionality (Aznar & Seifart 2020). Some language reclamation projects have found success working exclusively with archival sources (Hinton 2003; ANA 2006; Baldwin & Olds 2007; Whalen et al. 2016, 2018);[1] such work is a partnership-at-a-distance between the institutions that store and curate the materials, the researchers who deposit them, and the users of the materials. Archival materials are decontextualized (Schwartz & Dobrin 2016; Gaby & Woods 2020:e273; Dobrin & Schwartz 2021) from their original utterance, and depositors and archives both can do much to ensure that language collections are as robust and as useful as possible. In this paper, we report on the results of a review of language archives, with a concentration on sites and organizations with substantial holdings of digital data in and about endangered languages. We discuss the accessibility, discoverability, and functionality of archival resources, focusing on features of web portals and the special needs of linguistic collections. That is, the focus is around the needs that depositors and language resource users have, and how such needs are or are not met by current practices at the stewardship institutions that manage archives. Finally, we provide recommendations and suggest changes to the access services provided by stewardship institutions. It is our hope that these recommendations will serve as foundation for future guidelines in the creation, curation, and maintenance of web portals, the gateways to language resources at "language archives".

1.1 DIGITAL LANGUAGE ARCHIVES

A "language archive" is defined here as a repository of language data (broadly construed), such as audio and/or video recordings, transcriptions, and translations, whether in physical or digital format, created with the purpose of preserving and disseminating those materials (Kung et al. 2020; Burke et al. 2021; Austin 2021; Paterson 2021). There is substantial variation among repositories that contain linguistic data (cf. Vann 2006)—in scope, functionality, infrastructure, the number of languages or regions covered, and the extent to which they function as research tools or simply data repositories, to name just a few. We follow Austin (2021) in considering the role of archives to be appraising materials (that is, collecting selectively based on a stated goal), preserving those materials, "mak[ing] known their existence", and facilitating their appropriate distribution. For our purposes, we include sites that appear to have these aims. We treat the *process of archiving* as one in which someone places language resources in one of these repositories, as opposed to interacting with an archive or linguistic data in other ways. For this reason, we exclude from our definition of "archives" sites such as OLAC, which do not collect materials themselves but rather act as a directory for other archives. Throughout this paper, we refer to "items", "collections", and "archives", where items are the linguistic materials that are deposited; they are grouped into "collections", and those collections are housed by archives. Archives are repositories that are owned and managed by people, who are employed by institutions. Thus to talk about access to archives we need to think about the web portals, the choices of individuals, their employee obligations at their institution(s), the infrastructure that underlies the repository and its data services, among other topics.

While the advancement of technology has allowed linguists to digitize corpora that were once only available in physical media, digitization and online archiving have problems of their own. The long-term accessibility of digital material is dependent on the continuing availability of

[1] We recognize that language revitalization and reclamation are complex topics, far beyond the scope of what we can cover in this paper. Archived language data is inevitably an incomplete portrayal of languages and their communities.

compatible hardware and software. Necessary equipment may become obsolete and/or fall out of production (e.g., computers are no longer produced with built-in optical drives, making it difficult to access information stored on CDs). Storage media has a limited lifespan. Software, too, can rapidly become obsolete. While linguists have mostly heeded Bird and Simon's (2003) call to use open source software wherever possible, documentation projects can become tied to a specific software platform and version (cf. Bird & Simons 2003). Such issues affect both depositors and archives: while depositors should ensure they archive materials in the most endurable formats possible, digital archives are also subject to these constraints, such as the lifespan of servers and backups.

While the Internet has greatly improved the availability of research materials to far-flung audiences, it is far from an equalizer. Access to a reliable Internet connection is not universal, particularly in remote communities where broadband has yet to be fully implemented. For example, Wasson et al. (2016) note that Dinjii Zhuh K'yaa (Gwich'in Language Archive and Language Revitalization Center), a community archive of Gwich'in people in Fort Yukon, Alaska, is mostly accessed and available only by physically accessing the center where the archive is located because internet access is uneven within the language community. By some estimates, roughly 40% of the global population are not Internet users. Even in the United States, 21 million people lack access to broadband Internet; the FCC believes this figure "radically overstates" the number of people who have reliable connections (Sonnemaker 2020).

Despite the decades-long prevalence of digital archiving in the field, no two archives are alike, some having features that are tailored to the languages of focus. The Digital Archive of Scottish Gaelic,[2] for example, offers a search feature to filter for lenited words, an option especially useful for researchers working with Goidelic languages. But because archives are so decentralized, there is currently no set of protocols or standards for digital language preservation (Aznar & Seifart 2020).[3]

1.2 RATIONALE FOR THE CURRENT PROJECT

Previous research (among others, Bird & Simons 2003; Vann 2006; Berez 2013; Sullivant 2020; Burke et al. 2021) has discussed aspects of linguistic archiving, including the importance of metadata, consistent approaches to creating language materials, and the current state of language archiving. The current paper covers a wider scope of contemporary language collections, as well as contributing to the discussion of how to improve the archival practice in order to help communities and researchers more easily use and reuse these archives. A review of digital archival practices, such as the decisions made in designing websites and displaying content, will provide insight into how archives are and are not meeting the needs of end users—and the steps they can take to rectify these issues.

This paper reports on the results of a review conducted of online digital archives in June-August, 2021. The "audit" was conducted with the aim of investigating the utility of online archives and their accessibility for retrieval of materials. Concomitantly, we investigated a sample of individual collections in a subset of archives for ease of completing certain standard investigations, such as testing whether or not materials could be easily aligned using the Montreal Forced Aligner (a process widely used in the creation of corpus materials for phonological research; McAuliffe et al. 2017). This paper reports on findings that relate specifically to archives; a companion paper (Babinski et al. 2022) details the phonological/typological findings. The remainder of this paper describes the methods (Section 2), results of the archive audit (Section 3), and conclusions (Section 4), focusing on topics ranging from accessibility and discoverability to actual functionality (i.e., use and reuse of archival materials). At the end of each subsection in Section 3, we present suggestions for changes in practice. In engaging with these questions and making suggestions for changes in practice, we do not wish to downplay the efforts and skills of professional archivists, or dissuade researchers from depositing their materials in these archives. We recognize that there are innumerable tradeoffs in all aspects of language

[2] *https://dasg.ac.uk/en*.

[3] The challenges around digital archives are not unique, as issues such as the longevity of software and hardware, internet accessibility, and the like, are common across many digital media repositories. However, because of the complexity of language archive collections, their many filetypes, heterogeneity of construction (and resulting metadata), to name just a few, they are probably a good illustration of a very broad array of challenges.

documentation and archiving, and that any safeguarding is preferable to none. However, we also consider it appropriate to evaluate the extent to which archival practices—that is, those practices that are primarily controlled by archives and their management—are serving the aims of those using archives. To this end, we are not yet at the stage where we can present a full set of recommendations for archival practice. Rather, we raise the issues we have found across archives so that those in the field, including archivists, can consider them in future archive development and management.

Region-focused archives, such as the Alaska Native Language Archive (ANLA)[4] and the Survey of California and Other Indian Languages/California Language Archive (CLA),[5] draw an audience of language communities who access materials for the purposes of cultural, historical, and language learning. It is believed that the usage of language archives by Indigenous communities is underestimated (cf. Austin 2011; Holton 2012; Woodbury 2014), as a single representative may bring resources back to a community that are then more widely disseminated and used by many more individuals. In discussing issues with language archives, we wish to emphasize that roadblocks created by archives will also greatly affect language communities, and, to best suit the needs of their audiences, it may be critical for archives to be accessible and interpretable to users without specialized linguistic training or extensive technical knowledge. Holton (2012) and Woodbury (2014) discuss the different audiences and users of language archives, drawing particular attention to the fact that non-Indigenous linguists are not the only audiences of archives, and that both linguist and non-linguist members of Indigenous communities are using archives (e.g., the DOBES[6] Archive, the Archive of Indigenous Languages of Latin America [AILLA],[7] and ANLA)[8] for community-oriented purposes like language revitalization. In discussing the role of archived collections in promotion or hiring, therefore, it is also important to recognize that academics are not the only users of this material.

Additionally, implementing such recommendations at an archive level (i.e., having inter-archive standards maintained by those who manage archives) can help establish a fairer peer-review process of archival collections. By delineating precisely what standards fall under the archive management level and what procedures individual depositors are responsible for, the roles of "archivist" and "depositor" become clearer. Thus, in reviewing depositors' archival collections, we avoid evaluating the individual for aspects of archiving which are outside their control. Having standardization on the side of archives will create more equitable standards by which individuals are reviewed.

2 METHODS

An archive review was conducted between June and August, 2021, by the authors of this paper. Our audit focused on archival usability as a whole, as well as two aspects of collections: files suitable for phonetic and phonological analysis, and textual archives/archives not exclusively maintained for linguistic research. The general archive audit included 41 archives (as listed in the Supplementary Materials).[9] The archive list was compiled from OLAC's list of participating archives[10] as well as Digital Endangered Languages and Musics Archives Network (DELAMAN)

[4] *https://www.uaf.edu/anla/*.

[5] *https://cla.berkeley.edu/*.

[6] *https://archive.mpi.nl/tla/*.

[7] *http://ailla.utexas.org*.

[8] *https://www.uaf.edu/anla/*.

[9] The supplement is available from *https://osf.io/daksh/*. Anonymous reviewers of this submission had differing opinions on the extent to which this choice of archives was appropriate. One reviewer suggested that the sample should be expanded, while another felt that it was too broad, including too many archives of different types (and that it was inappropriate to generalize across archives with very different levels of institutional support and access to funding). It was unclear from our survey how many of the archives in the OLAC and DELAMAN lists are actively maintained, what their support is, and how they backup and preserve their holdings. This is itself an important issue which should be investigated further. For our purposes, rather than restrict the focus to archives that are clearly actively maintained, we preferred to cast a wider net and examine as many digital archives as possible (with caveats further discussed in Section 3.1.3 below).

[10] *http://www.language-archives.org/archives*.

members and associate members.[11] For this reason, the archives examined are heavily skewed towards English-language based collections, though (as discussed in Section 3.1.2 below) we actively attempted to address this bias (unfortunately without much success).[12]

We compiled general information on metalanguages, search and retrieval functions, corpus structure, access condition options, and types of materials archived. Prior to the audit, we created a questionnaire that probed various aspects of archives and collections that could prove problematic in linguistic research. This questionnaire was used to systematically document information regarding the archives content, accessibility restrictions, search functions, metadata, download options, and file manipulation necessary for analysis (see Babinski et al. in prep for a larger summary of findings). Members of the team examined archives individually; the results were discussed as a group and CB & IY spot-checked data coded by other authors. We found a very high degree of inter-rater consistency, with the exception of problems arising from web browsers and access to sites which were blocked from Yale's campus internet.[13] We focus on the following points in this paper:

- Accessibility
    - Which language(s) must a user know in order to navigate sites and collections?
    - Is the site accessible to users of screen readers?
    - Are there aspects of the site design that impede or promote accessibility?
- Restrictions
    - How available is material in collections?
    - If restrictions are placed on access, what is needed to access collections?
    - What types of controls are in place, and for what reason?
- Finding information
    - How easy is it to find information on the site?
- File manipulation
    - How usable are the collection materials? Are there aspects of the site and archive design that promote or impede the use of materials?

Another set of possible metrics are the FAIR principles.[14] FAIR data is *findable, accessible, interoperable,* and *reusable*. Our points overlap with FAIR in a number of respects, but the FAIR framework was unsuitable for our evaluation for two reasons. Firstly, the findability criterion focuses exclusively on metadata structure, whereas we consider issues of "findability" to be much broader, as discussed further below. Secondly, the FAIR criteria mostly apply to collections, rather than to the overall structure of the archive *qua* repository.

There are a range of reasons why an archive might have a particular property, ranging from constraints introduced by the Content Management System (CMS), to decisions made in light of the amount of funding or staffing, to philosophical decisions about the appropriate structure of an archive. Therefore, rather than focus on the particular properties of individual archives, we instead focus on implications of current design for what end users can accomplish. We do list selected examples to illustrate and explain findings, however. While our findings are therefore perhaps not fully *reproducible* (cf. Berez-Kroeker et al. 2018), we have endeavored to make the findings *replicable* by including information in the supplementary materials. This represents a snapshot of archival issues as of August, 2021, which will no doubt evolve as sites are updated.

[11] *https://www.delaman.org/members/*.

[12] While there are other archives (such as Kielipankki, the Language Bank of Finland; *https://www.kielipankki.fi*), restricting the sample to OLAC/DELAMAN archives provided some principle for inclusion in the survey. We acknowledge that it is unclear at this point how representative or comprehensive this list is. Organizations differ considerably in the extent to which they focus on preservation or access to files, or serving as research resources or content delivery platforms, making a clear definition of "language archive" difficult. There is, to our knowledge, no global list of language archives. The closest are the DELAMAN and OLAC compilations.

[13] We were unable to diagnose why some sites loaded and others did not, based on IP addresses. We noted issues when they arose, since, if they arose during this sampling process, they will likely arise for other users too. A reviewer asked why we do not exhaustively list, enumerate, and quantify all points made in this paper. We argue that doing so would give rise to misleading precision. As discussed in Footnotes 9 and 12, it is impossible to know how representative this sample is. It would therefore be misleading to draw detailed conclusions about small differences in prevalence. Instead, we concentrate on reporting common trends in this set of data. This allows us to evaluate recurring issues among commonly used language archives without being unduly focused on small differences.

[14] *https://www.go-fair.org/fair-principles/*.

There were some points which we wished to investigate but were unable to include. The extent of institutional support may be a critical component of an archive's longevity, but such information was typically unavailable. Other points relating to archival infrastructure, such as long-term plans, backup procedures, storage procedures, type of content management system, and staffing will also have a major impact on what the archive can deliver. Because we are evaluating archives in terms of their usefulness to end users and not in terms of institutional and financial barriers they must overcome, we do not consider these points in our analysis, though we recognize that archives vary greatly in this dimension.

3 RESULTS

3.1 ACCESSIBILITY

The contents of an archive are only useful as far as they are findable and accessible. Accessibility can be impacted by a number of factors, both on the user end and through archive design choices. "Web accessibility" is generally understood to refer to compatibility with assistive technology. We discuss accessibility in this narrow sense in Section 3.1.3. However, we also discuss registration and account-creation requirements and procedures, display language, and site navigation. These are also points which may facilitate or impede a user's access to the archive contents.

3.1.1 Accounts and registration

The majority of language archives we surveyed have materials that are available for download for free and with minimal registration requirements. Many archives appropriately have access restrictions for collections to respect the wishes of language communities and researchers (Nathan 2010). Five archives, including the Repository and Workspace for Austroasiatic Intangible Heritage (RWAAI),[15] required registration to access any materials at all, including the catalog; other archives had a public-facing catalog, even if registration was required for download. Four archives, including ELAR and the DOBES Archive, had multiple tiers of access, where some tiers required registration and/or permission of the depositor for listening or download, while other tiers were unrestricted. One archive, the CLA, has materials that are closed-access, in that they are not available online and must be accessed in person. Two archives, Kaipuleohone[16] and LIA Sápmi (Sami Speech Corpus),[17] restrict all or most of their contents specifically to academic institutions and institution-affiliated researchers, a limitation that may exclude members of language communities. Others restrict only parts of their materials to those affiliated with academic institutions. The CHILDES Data Repository[18] includes password-protected collections available only to faculty members, and the CLARIN Slovenian Repository[19] requires that users access certain materials restricted for "academic use" through their institutional emails. While there are good reasons why collections may be not freely available, some of the convoluted, unclear, or heavily outdated procedures for requesting permission could be fixed. Archives that do not streamline the permission forms or provide unclear contact information could be updated. For example, account registrations requiring manual approvals, or emailing specific individuals, should be automated.

While the majority of archives are entirely free for use, and we did not encounter any archives[20] requiring payment of fees to access collections during our audit, we acknowledge that for some researchers, particularly those who do not have institutional membership with archives or sufficient funding, the cost of accessing an archive may be prohibitive. The Linguistic Data

[15] *https://projekt.ht.lu.se/rwaai*.

[16] *http://ling.hawaii.edu/kaipuleohone-language-archive/*.

[17] *https://tekstlab.uio.no/glossa2/saami*.

[18] *https://childes.talkbank.org/access/*.

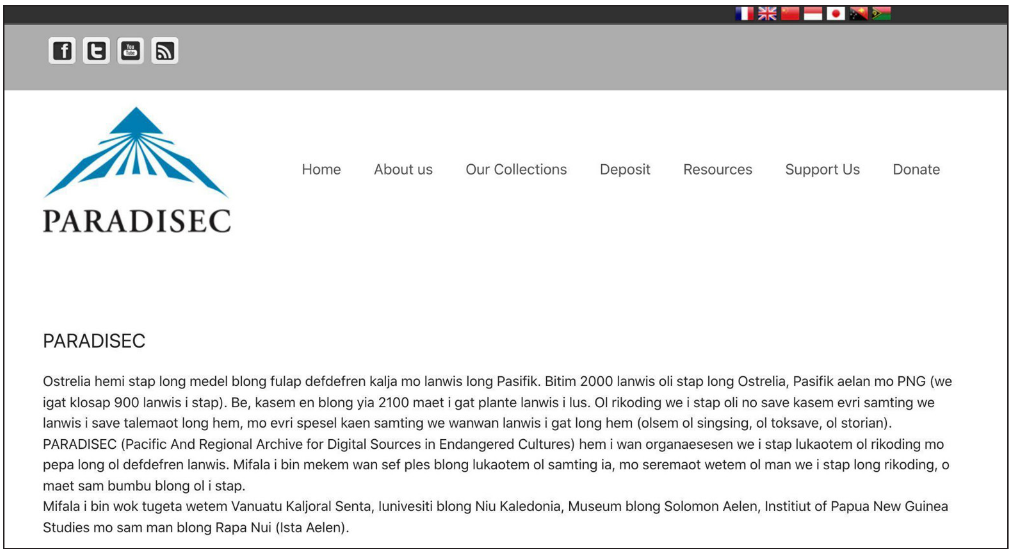
[19] *http://www.clarin.si/info/about*.

[20] We did not include predominantly physical archives that also have digital materials. This excluded archives such as AIATSIS (*https://mura.aiatsis.gov.au*), which requires the purchase of physical media for accessing digital collection items.

Consortium (LDC)[21] and the European Language Resources Association (ELRA)[22] are two examples of archives requiring payment in return for access to materials; these fees can range into tens of thousands of dollars (Vann 2006). Endangered language archives have tended towards a model where the archive is supported through institutional or grant funds, with costs supplemented by fees from depositors (similar to "gold" open access models for academic publication). Some archives have recently requested that depositors include archiving charges in research grant applications. Clearly the funding model for ongoing support for endangered language archives needs to be investigated in more detail.

3.1.2 Display language

The language(s) used to display metadata and to navigate the site may also limit accessibility of the materials. Bias towards English-language users and lack of built-in site translations disadvantages researchers whose primary language is not English and may prevent community members from accessing documentation of their own languages or other languages which they regularly use.[23] While the arrival of digital media devices and technologies can facilitate the creation of a "social network of digital exchange" of cultural heritage for Indigenous communities (Mansfield 2014:66), unavailability of these resources in endangered languages further entrenches generational and educational divides in language communities where acquisition of literacy, particularly in English, is not widespread. A number of linguists and Indigenous community members have expressed concern that "the majority of digital resources available to Indigenous users are in English, even though English is not a first language for many" (Carew et al. 2015:310). Only 14 of the archives we examined provide more than one language interface, and not all of these had fully functional language options. We point to PARADISEC[24] (see ***Figure 1***) as an example of an archive providing information through 7 regional languages (though unfortunately not on the mobile site).



PARADISEC

Ostrelia hemi stap long medel blong fulap defdefren kalja mo lanwis long Pasifik. Bitim 2000 lanwis oli stap long Ostrelia, Pasifik aelan mo PNG (we igat klosap 900 lanwis i stap). Be, kasem en blong yia 2100 maet i gat plante lanwis i lus. Ol rikoding we i stap oli no save kasem evri samting we lanwis i save talemaot long hem, mo evri spesel kaen samting we wanwan lanwis i gat long hem (olsem ol singsing, ol toksave, ol storian).
PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures) hem i wan organaesesen we i stap lukaotem ol rikoding mo pepa long ol defdefren lanwis. Mifala i bin mekem wan sef ples blong lukaotem ol samting ia, mo seremaot wetem ol man we i stap long rikoding, o maet sam bumbu blong ol i stap.
Mifala i bin wok tugeta wetem Vanuatu Kaljoral Senta, Iunivesiti blong Niu Kaledonia, Museum blong Solomon Aelen, Institiut of Papua New Guinea Studies mo sam man blong Rapa Nui (Ista Aelen).

Figure 1 PARADISEC's informational language options (top right corner of banner).

Archives that focused on languages of a particular region often provided interfaces relevant to their users. For example, AILLA has interface options in Spanish, and PanGloss is fully implemented in both English and French.[25] ELAR and CLA have interfaces only in English,

[21] *https://www.ldc.upenn.edu*.

[22] *http://www.elra.info/en*.

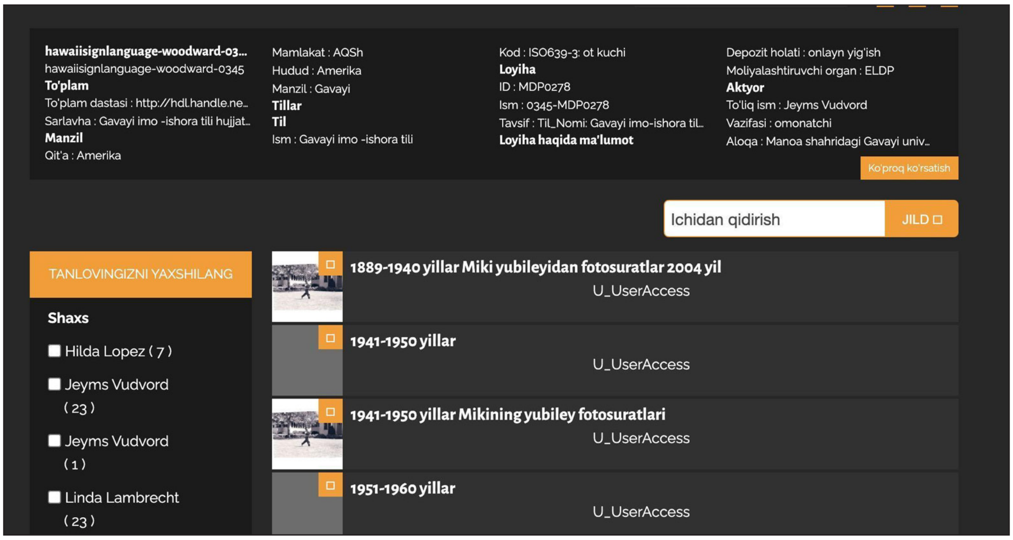
[23] The finding that there is a lack of archives with a primary interface in other languages may be, in part, due to our own biases as all English-dominant researchers in the USA. However, we made a substantial effort to search out archives written in other languages (e.g., Spanish, Russian, French), but they were largely difficult to search for because of Internet search engine rankings, which return results based on language and geographic region. This should be noted as an issue for linguistics that leads to a substantial reduction in findability of materials, though one beyond the control of individuals.

[24] *https://www.paradisec.org.au*.

[25] However, the translations caused issues with file matching. Where .mp3 files were labeled in French but the transcripts were auto-generated and downloaded by the site, and given English filenames.

though at the collection level, ELAR allows filenames and metadata to be in other languages and scripts, which helps users if they know of the collection.

We applied Google Translate to the exclusively English archives (testing languages such as Korean, Uzbek, Kyrgyz, and French). Translations were inconsistent, incomplete, and sometimes misleading. Some localizations translated only parts of the site text, leaving others, such as an embedded map and filenames, in English (see ***Figures 2, 3,*** and ***4***). Therefore using Google Translate as a workaround for untranslated sites is not a straightforward alternative.



hawaiisignlanguage-woodward-03...
hawaiisignlanguage-woodward-0345
To'plam
To'plam dastasi : http://hdl.handle.ne...
Sarlavha : Gavayi imo -ishora tili hujjat...
Manzil
Qit'a : Amerika

Mamlakat : AQSh
Hudud : Amerika
Manzil : Gavayi
Tillar
Til
Ism : Gavayi imo -ishora tili

Kod : ISO639-3: ot kuchi
Loyiha
ID : MDP0278
Ism : 0345-MDP0278
Tavsif : Til_Nomi: Gavayi imo-ishora til...
Loyiha haqida ma'lumot

Depozit holati : onlayn yig'ish
Moliyalashtiruvchi organ : ELDP
Aktyor
To'liq ism : Jeyms Vudvord
Vazifasi : omonatchi
Aloqa : Manoa shahridagi Gavayi univ...

Shaxs

□ Hilda Lopez (7)
□ Jeyms Vudvord (23)
□ Jeyms Vudvord (1)
□ Linda Lambrecht (23)

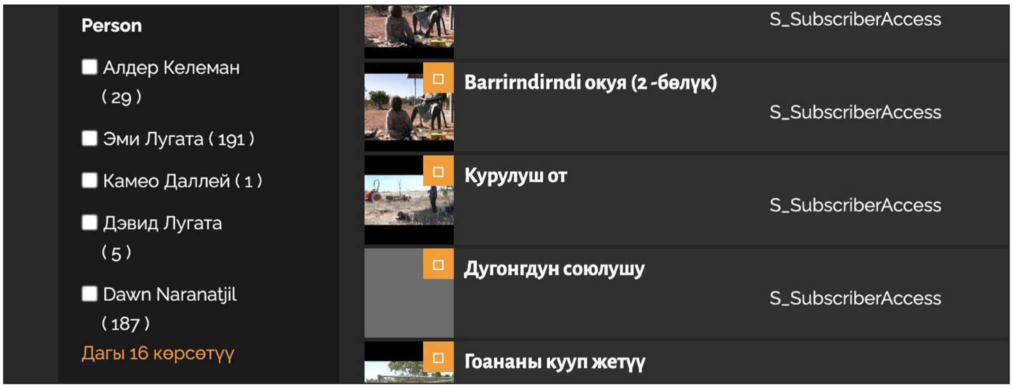
□ **1889-1940 yillar Miki yubileyidan fotosuratlar 2004 yil**
U_UserAccess

□ **1941-1950 yillar**
U_UserAccess

□ **1941-1950 yillar Mikining yubiley fotosuratlari**
U_UserAccess

□ **1951-1960 yillar**
U_UserAccess

Figure 2 ELAR interface in Uzbek with Google Translate overlay. Names on the side are not consistently translated or transliterated: "Hilda Lopez" is not transliterated, but "James Woodward" becomes "Jeyms Vudvord". Selective transliteration breaks links elsewhere in the collection.



Person

□ Алдер Келеман (29)
□ Эми Лугата (191)
□ Камео Даллей (1)
□ Дэвид Лугата (5)
□ Dawn Naranatjil (187)

S_SubscriberAccess

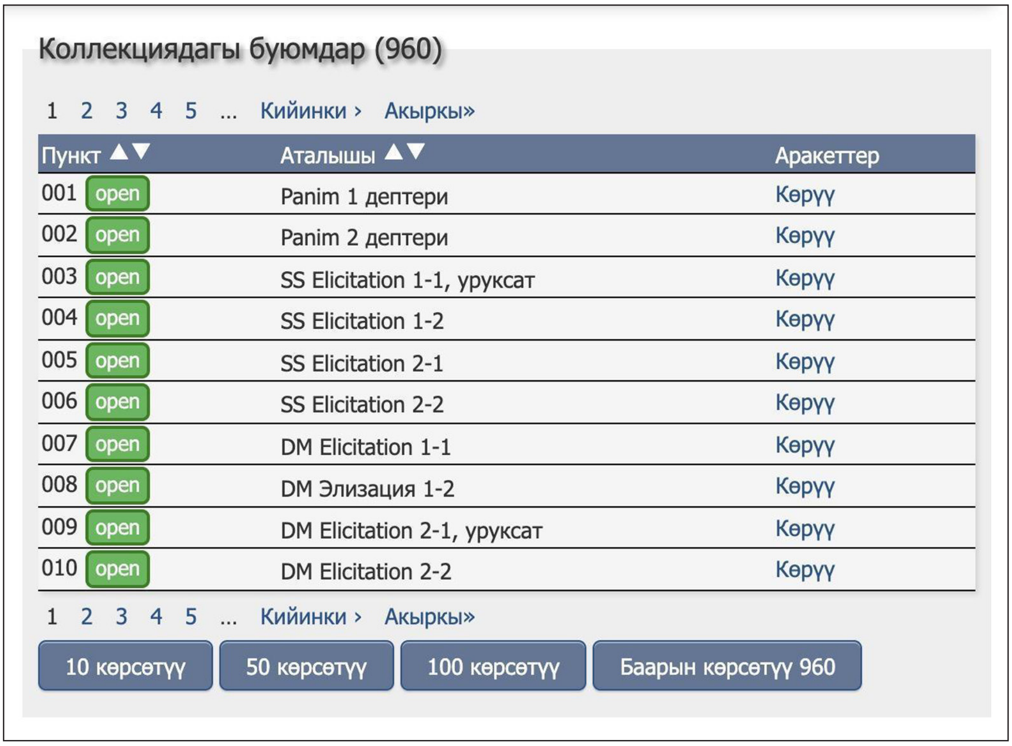
□ **Barrirndirndi окуя (2 -бөлүк)**
S_SubscriberAccess

□ **Курулуш от**
S_SubscriberAccess

□ **Дугонгдун союлушу**
S_SubscriberAccess

□ **Гоананы кууп жетүү**

Figure 3 ELAR interface in Kyrgyz (Google Translate overlay). The "View 16 more" button on the menu (in orange text) no longer works with Google Translate as an overlay.



Коллекциядагы буюмдар (960)

1 2 3 4 5 ... Кийинки › Акыркы»

Пункт ▲▼	Аталышы ▲▼	Аракеттер
001 open	Panim 1 дептери	Көрүү
002 open	Panim 2 дептери	Көрүү
003 open	SS Elicitation 1-1, урукcат	Көрүү
004 open	SS Elicitation 1-2	Көрүү
005 open	SS Elicitation 2-1	Көрүү
006 open	SS Elicitation 2-2	Көрүү
007 open	DM Elicitation 1-1	Көрүү
008 open	DM Элизация 1-2	Көрүү
009 open	DM Elicitation 2-1, урукcат	Көрүү
010 open	DM Elicitation 2-2	Көрүү

1 2 3 4 5 ... Кийинки › Акыркы»

Figure 4 PARADISEC items where the term "Elicitation" is translated into Kyrgyz once, but not other times. Further, green "Open" buttons are not translatable as the text is part of the icon.

It is also worth noting that when site translations are available, options are predominantly languages of European origin. This is especially striking given the scope of archived languages, most of which are indigenous to Africa, Asia, and the Americas. Lack of translations into major regional languages limits the abilities of scholars to use these archives, creating a bias in the demographics of researchers and restricting potential scholarly innovation. For non-Indo-European languages whose structures differ greatly from that of languages like English, French, and Spanish, automatic translation programs like Google Translate and Yandex are especially prone to offering confusing and poor-quality translations. We recognize that this is a much bigger problem than what individual archives can solve. For example, search engines filter out search queries in languages other than the query language (which made it almost impossible for us to search for archives outside the anglosphere internet).[26] However, at the collection level, depositors should be encouraged to provide materials in languages that will be most usable for community members, and the substantial additional time costs for doing so should be recognized explicitly.

3.1.3 Disability accommodations

We acknowledge that disability accommodation remains a critical, and often-overlooked element of archive accessibility, and indeed, the accessibility of any digital material. In regards to the structure of websites and storage of data such as text files, it is essential that Web content – including archives – is presented in a way that is accessible for visually-impaired researchers.[27]

It is generally agreed as a principle of accessible Web design not to make different elements of a site distinguishable only by their color (Campbell 2018). In order to assess color blindness accessibility, we put each archive through a filter[28] mimicking how each site would look to users with 3 of the most common kinds of color blindness. Sites were subjectively reviewed by a member of the team who is colorblind. The archives we surveyed, largely, performed well in this regard. The main issue raised by our survey was the low contrast between font and background colors, which may compromise readability for users with certain kinds of color blindness and other visual impairments; it may also inconvenience users with certain color and brightness settings on their computers and browsers. The websites for SIL's International Language & Culture Archives and the Rosetta Project revealed such problems.

3.1.4 Recommendations

The restrictions archives place on access to language data are there for a reason; however, it is important that these restrictions do not place too much of a burden on researchers and language communities looking to access their contents. Therefore, we suggest archives streamline the process of requesting access permission. More specifically, we recommend that request forms be built into the site itself, with additional capacities for automated password retrieval. This is especially critical for those archives (such as ELAR) whose code is not built for long-term accessibility, as passwords cannot be reset and the application permission form is built on Google Surveys. For archives not already implemented in multiple languages, we strongly suggest expanding the scope of display languages offered, especially those languages which may be relevant to language communities and local researchers. Furthermore, we recommend against applications that must be physically posted to the archive, given their inefficiency and potential to disadvantage researchers in areas underserved by the postal system.[29] Following the principles of accessible Web design will make great strides in overcoming barriers for researchers who require assistive technology. Even in the absence of laws like the American Disabilities Act (1990) (or varying legal

[26] *https://developers.google.com/search/blog/2010/03/working-with-multilingual-websites* provides some information about how Google determines relevance for multilingual sites; this includes the domain name suffix and IP address of the server, as well as language identification for monolingual web pages. It does not include HTML language attribute tags of georeferencing in HTML.

[27] For further information see *https://www.w3.org/WAI/standards-guidelines/wcag/* (WCAG 2018).

[28] *https://www.toptal.com/designers/colorfilter/*.

[29] One assumes that for a digital archive, users who will access the materials are also able to access a digital registration form.

requirements across different countries), it is important and not too difficult to improve what is already there.

3.2 DISCOVERABILITY

Collections need to be discoverable; that is, users should be able to navigate the site to find what they need. Discoverability encompasses both the abilities of users to find archives through search engines or aggregation portals (such as OLAC) and to perform searches within those archives. The former point is essential for the use and reuse of an archive in general, while the latter point sheds important light on the internal organization/description of material.

3.2.1 Search functions and mislabeling

Search functions are vital for navigating large collections, but they can be made frustratingly slow and even useless depending on their available options. Six archives offer a map search function, which allows users to browse collections by location. This function is especially useful for more casual users or for those who are not searching for specific data, but it presents its own challenges. Archives like ELAR and the California Language Archive use Google Earth, and many others use similar platforms. The California Language Archive does not outwardly indicate whether each collection on the map is available online. This makes it initially seem as if there are more resources readily available to users than there actually are. While these issues are a cause for frustration, they are not necessarily debilitating, and the map function tends to be a useful visual aid for users. We point to PanGloss as an example of the map function at its most useful; its map function is easy to navigate, contains information on the title, researchers, and types of resources available for each collection (as well as a link to each collection), and can be filtered by the criteria "with annotation" and "with video". In contrast, AILLA's map function is non-interactive. That said, some location information within PanGloss was inaccurate.

Lack of transparency about the contents of collections was observed in numerous archives. Users may have little information about what a deposit contains before accessing its contents. Researchers often have specific criteria in their search for language materials—for example, linguists looking to perform certain kinds of phonological analysis may have a preference for collections whose video and audio recordings are all fully transcribed, and rule out collections with too few hours of recorded material or those that consist only of written materials. Other criteria may include the specific dialect(s) documented, date of creation, file type, number, age and location of speakers, or specific individuals. Some of these categories can be aggregated automatically, while others require manual labeling. While the lack of some of this information is due to incomplete metadata provided by depositors, we encourage archives to make such information easy to find. ELAR, for example, includes a collection landing page consisting of sections for "summary of the deposit", "groups represented", "language information", "special characteristics", and "deposit content", though the quality and specificity of the information in these descriptions varied greatly between collections. This could be a point of evaluation for individual collections. In most collections, the metadata about the holdings is a file within the general collection. It is not consistently named and where collections have many files it is difficult to find. Archives could assist the retrieval of such information by flagging such metadata files directly or including an explicit link to the metadata file(s) within the collection overview.

Most archive portals include search bars, but these have varying degrees of usability. One important feature is a filter function, especially for larger archives. All but seven of the archives we investigated have some kind of search filter function. Some filtering options include language, speaker, depositor, file type, topic, and country, among others. However, the availability and usability of the filter function was inconsistent. ELAR's search filter options vary by collection, while The African Language Materials Archive,[30] Digital Himalaya,[31] and AILLA all lack a search

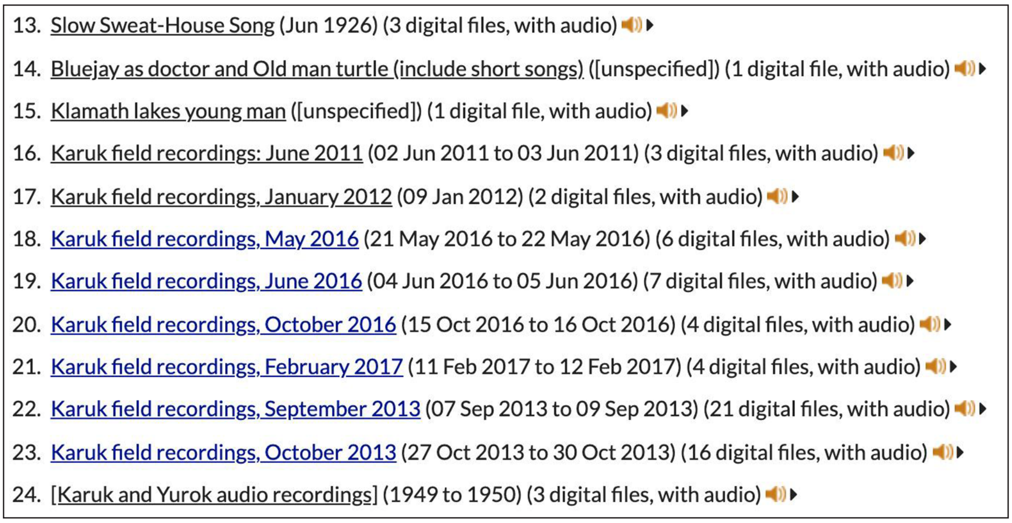
[30] *http://alma.matrix.msu.edu*.

[31] *http://www.digitalhimalaya.com*.

filter function entirely, making large collections more difficult to search. PARADISEC had a flexible search and filtering interface, at the item or collection level.

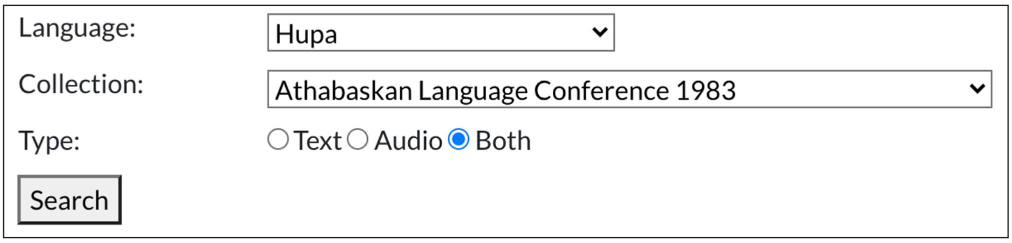
A useful search feature available in some archives is the ability to search within collections. This feature is especially useful, almost necessary for archives that contain large collections. However, despite its importance, we only found the feature in four of the archives we examined. Such free-text search increases finding options for collections extensively, allowing more refined searches than filters alone. For example, a filter may exist to restrict files to .xml, but a test search makes it easier to distinguish between Flextext transcripts, ELAN transcripts, and .xml-format metadata. These are all .xml format files but have very different functions. At the collection level, searches were hampered by missing metadata, incorrect tags, case sensitive searches and inconsistent metadata (e.g., searches returning either *audio* or *Audio* as filetype), empty folders, and broken URLs within collections. Correcting these small but time-consuming errors would improve intra-archive searches.

Two of the most useful search filter categories are media type and file type (see ***Figures 5*** and ***6***). Many researchers using these digital archives can only use files of a specific media type (e.g., videos or sound recordings), or, in cases where they plan to use certain software in their research, certain file types (e.g., .pdf or .wav files). File type and media type filters greatly reduce the time a researcher must spend browsing files to find what they need. Despite this importance, only five of the archives we looked at offer the option to filter by file type, and one of these archives, The Language Commons,[32] returns files that aren't bundled (with related files of different file types) when the file type filter is employed, causing users to miss potentially useful materials. Similarly, only four of the archives we looked at offer the option to filter by media type. Even fewer allowed users to filter by specific file extensions (such as .mp3 or .wav), and, when offered, the archive often displayed results with mislabeled extensions (.xml for .eaf, for example).



13. Slow Sweat-House Song (Jun 1926) (3 digital files, with audio) 🔊▶
14. Bluejay as doctor and Old man turtle (include short songs) ([unspecified]) (1 digital file, with audio) 🔊▶
15. Klamath lakes young man ([unspecified]) (1 digital file, with audio) 🔊▶
16. Karuk field recordings: June 2011 (02 Jun 2011 to 03 Jun 2011) (3 digital files, with audio) 🔊▶
17. Karuk field recordings, January 2012 (09 Jan 2012) (2 digital files, with audio) 🔊▶
18. Karuk field recordings, May 2016 (21 May 2016 to 22 May 2016) (6 digital files, with audio) 🔊▶
19. Karuk field recordings, June 2016 (04 Jun 2016 to 05 Jun 2016) (7 digital files, with audio) 🔊▶
20. Karuk field recordings, October 2016 (15 Oct 2016 to 16 Oct 2016) (4 digital files, with audio) 🔊▶
21. Karuk field recordings, February 2017 (11 Feb 2017 to 12 Feb 2017) (4 digital files, with audio) 🔊▶
22. Karuk field recordings, September 2013 (07 Sep 2013 to 09 Sep 2013) (21 digital files, with audio) 🔊▶
23. Karuk field recordings, October 2013 (27 Oct 2013 to 30 Oct 2013) (16 digital files, with audio) 🔊▶
24. [Karuk and Yurok audio recordings] (1949 to 1950) (3 digital files, with audio) 🔊▶

Figure 5 CLA materials with media type specified next to item name.



Language: Hupa
Collection: Athabaskan Language Conference 1983
Type: ○ Text ○ Audio ● Both

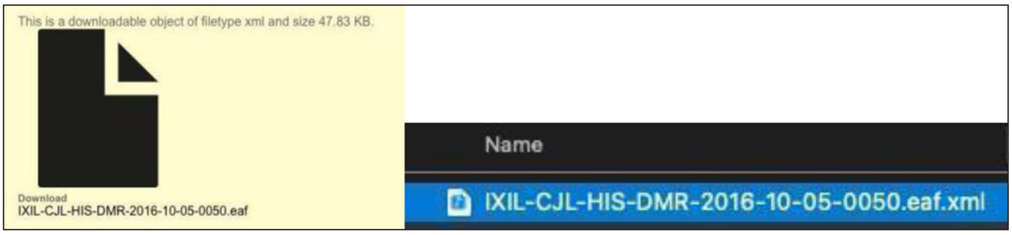
Figure 6 ANLA materials searchable/filterable by media type.

Mislabeled file types are another issue we encountered. ELAR and AILLA, for example, rename ELAN[33] .eaf files and FLEx[34] .flextext files as .xml (see ***Figure 7***). While these are underlyingly XML files and alternate extensions are visible upon downloading the files, one needs to know how

[32] *https://archive.org/details/LanguageCommons?tab=about*.

[33] *https://archive.mpi.nl/tla/elan*.

[34] *https://software.sil.org/fieldworks*.


This is a downloadable object of filetype xml and size 47.83 KB.

Download
IXIL-CJL-HIS-DMR-2016-10-05-0050.eaf

IXIL-CJL-HIS-DMR-2016-10-05-0050.eaf.xml

Figure 7 On the left is a screenshot of a sample AILLA document, in .eaf format (displayed as such). However, when the file is downloaded (at right), the file receives a .eaf.xml extension, which must be manually removed before it is readable by ELAN.

to change the file extensions in order to open the files with the appropriate applications. It is also difficult to differentiate ELAN audio transcripts from FLEx dictionary or interlinearized texts, which are both listed as .xml files but have different underlying data structures.

3.2.2 Metadata

We noted considerable inconsistency in what type of metadata was available, across archives and collections. It is easy for relevant files to be lost in a search because they do not have the type of metadata used in the search.

Another issue we discovered was the use of different layers of metadata. In many cases, important metadata was hidden inside the folders of a collection, making it difficult for a user to find the specific information they need. AILLA, for example, has three layers of metadata: one layer is for the whole collection, another layer is found within each individual folder within each collection, and the final layer is attached to the individual files themselves. Such layering, combined with the frequent gaps in available metadata, makes it extremely difficult to find desired information and reduces the accuracy of the search function.

Sullivant (2020) provides detailed recommendations for collection metadata, breaking down these recommendations into categories based on importance. We point to The California Language Archive and PARADISEC as two archives that do a good job of including "first priority collection metadata". Finally, it is important to note that, while many of the archives we examined do include the most important information in their collection metadata, almost none include the information in Sullivant's next two tiers of recommended metadata. While archives are reliant on the metadata provided by depositors, this only reinforces the points made by Sullivant (2020) and Burke and Zavalina (2020) that metadata is crucial to the usability of a collection. The DACS[35] standards may also be useful for both depositors and archives to introduce and maintain consistency.

3.2.3 Site maintenance

Other issues impeded discoverability, with archives being incompatible with specific browsers, requiring defunct software, or failing to load entirely. This occasionally varied depending on the individual user in ways we were unable to solve. For example, three of the team members found that the APS Digital Library[36] would not open for them unless they accessed it via Yale University's VPN, while the remaining team members could access the site with no difficulties from off campus, all using recent versions of Chrome on MacOS 11.6 or Windows 10.

Six of the 41 archives gave web access errors or were unreachable.[37] Some, such as ALORA,[38] could only be accessed with the Wayback Machine.[39] While these workarounds do allow users to access materials, users who are unfamiliar with the Wayback Machine would be deterred from retrieving relevant information. Moreover, the Wayback machine may provide access to the catalog, but not the files in the collection itself. Links provided within archives often faced

[35] *https://github.com/saa-ts-dacs/dacs*.

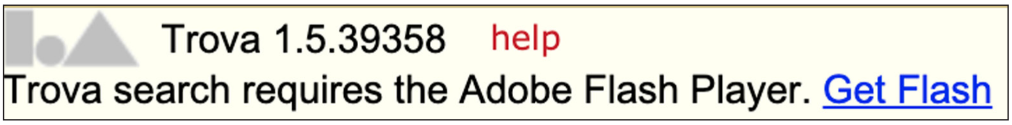
[36] *https://diglib.amphilsoc.org*.

[37] Academia Sinica English corpora (*http://www.ling.sinica.edu.tw/en/announcements/Resources*); ALORA (*https://alora.cerdotola.com*); Multimodal Learning Corpus Exchange (*http://mulce.org*); Standing Rock Sioux Tribe Language and Culture Institute (*http://wooyake.org*); American Philosophical Society Digital Library (*https://diglib.amphilsoc.org*); World Oral Literature Project (*http://www.oralliterature.org*).

[38] *https://alora.cerdotola.com*.

[39] *https://web.archive.org/web/20190208220853/https://alora.cerdotola.com*.

the same issues, defeating the purpose of being an archive that safeguards data.[40] Furthermore, at least two archives[41] still required some use of Adobe Flash Player (see ***Figure 8***), which was phased out by December 2020.



Trova 1.5.39358 help

Trova search requires the Adobe Flash Player. Get Flash

Figure 8 Adobe Flash Player required to access materials on The Repository and Workspace for Austroasiatic Intangible Heritage.

Many archives contained broken links, though this differed in extent and severity. The problems related to both site-internal and -external links, and problems arose due to internal site reconfigurations (such as those of the British Library's Endangered Archives Collections)[42] as well as the removal of individual pages. It would be ideal for archives to not rely on external links, but when necessary, regularly checking for outdated links is crucial. Broken links not only hinder the usability of archival materials from an end-user perspective, they also hinder the discoverability of such webpages. Search engines penalize broken links[43] in search results, thus making archive sites with such links less findable. As language resources are entrusted in archive sites' stewardship, it is important that they remain discoverable by those who wish to access these language materials.

Corrado and Sandy (2017) draw attention to the lifecycle of a project, as defined by the Life Cycle Information for E-Literature.[44] They argue that "institutional commitment...ensuring that enough financial resources are available to sustain the initiative" is necessary for digital preservation to be successful (Corrado & Sandy 2017:11). In order for stewardship organizations to faithfully fulfill their fiduciary duties as language resource stewards, website maintenance must receive ongoing support to keep up with rapidly-changing software and security compatibility requirements.

3.2.4 Recommendations

Offering more detailed descriptions of a collection's contents, specifically, media types (video, audio, text, etc.), completion state of any transcriptions or translations, and number of hours of recorded material, would help researchers evaluate the utility of a collection for a particular purpose, and give community members a sense of what is in the collection. Allowing searches by file type would allow researchers to further refine their queries and determine the usability of a given collection for their research purposes; we also recommend that archives correctly label file types and remove filetype capitalization dependencies on searching.[45] We also suggest that archives make it clear to depositors what types of information are indexed for searching, and how researchers can structure their collections to make them usable. To make archives more easily discoverable, we recommend archive managers use the Sitemaps[46] protocol set to provide site-internal content information to search engines. Finally, we suggest that depositors consider how they use links to external sites in their deposits, archiving copies where appropriate (or pointing links to the Internet Archive). We suggest that archives regularly

[40] Collections within the Endangered Archives Programme, British Museum (*https://eap.bl.uk/project/EAP347*); Online Database of Interlinear Text (*https://odin.linguistlist.org*); ELAR (*https://www.elararchive.org/dk0611*).

[41] The Repository and Workspace for Austroasiatic Intangible Heritage (*https://projekt.ht.lu.se/rwaai*); Yami Corpus (*http://yamiproject.cs.pu.edu.tw/yami/en_index_flash.htm*).

[42] *https://eap.bl.uk*. For approximately 8 months, every collection-level link from the main site catalog was broken. However, as of December 15, 2021, this has been fixed.

[43] See *https://devrix.com/tutorial/crucial-google-penalties/* for more about search engine penalties.

[44] *http://www.life.ac.uk/glossary*.

[45] To be clear, the issue we are discussing here is where a search returns both *Audio* and *audio* filetypes (for example) and treats them as distinct filetypes. This is a claim about variable capitalization in standardized vocabularies, not a point about case sensitivity in searches more generally.

[46] *https://www.sitemaps.org*. We thank the anonymous reviewer who pointed us to Sitemaps.

check for link breaks (e.g., by using automated checking tools that generate reports, such as the Broken Link Checker plugin),[47] particularly to archive-internal pages.

3.3 FUNCTIONALITY

The primary function of an archive is to store and safeguard materials, so it is essential for both the process of depositing and retrieving data to be straightforward; after all, material is safeguarded *for a purpose*, not simply to have an unused record of languages. This section discusses the functionality of data retrieval and use. Section 3.3.1 focuses on the structure and content of various archives, as well as issues surrounding downloads. Section 3.3.2 lists our recommendations for archive functionality.

3.3.1 Site content, structure, and downloads

The available content and structures of archive sites posed the first issue with functionality. We note that some of the following concerns are affected by the choice of CMS of individual archives. We attempted to track CMS use across archives, such as whether the archive used a common CMS such as Mukurtu,[48] DSpace,[49] or a bespoke platform. However, information about the CMSs underlying the archives in our audit was not easily accessible; an overwhelming majority of archives had no publicly available information at all about their CMS. Half of the archives mentioned the institutions that supported the development of the archive, or external servers where related language corpora were hosted, but the information about infrastructure was not available for enough archives for us to track it. We acknowledge, however, that site structure and content capabilities are closely linked to choice of CMS.

The sites examined here vary extensively in their holdings and scope. Some sites labeled as archives only hosted one or two resources (Magoria Books Carib & Romani Archive),[50] which sometimes required purchase, while others hosted none at all (Multimodal Learning Corpus Exchange).[51] Others, such as the SIL International Language & Culture Archives,[52] appeared to function more as directories with both links to external resources and hosted materials. They were not "archives" in the sense of storing and safeguarding materials. This is in contrast to archives such as the ELAR archive, which has full hosting and offers (per *R3data.org*) more than 462,048 results.[53] The most prevalent issue impeding archives' functionality was the lack of a bulk download option. The vast majority (34/41) had no bulk download option for either text or audio/video. Two[54] had bulk download options for text files only, and five[55] provided download links for zip files containing all or a selection of the files in the corpus. Requiring users to download files individually not only results in loss of time, but also renders some collections (e.g., those with 15,000 audio files) virtually inaccessible because of the sheer number of clicks, ranging from 1 to 7 per file, required to download their contents. Further, when individual downloads are the only option, users would benefit from knowing exactly how many files are in each collection, allowing them to assess their own storage capacity before attempting to download a corpus.

Another concern that results from downloading files individually is the loss of arrangement of items within a collection. For example, nested files lose their relationships to each other and must be manually re-sorted when downloaded onto a drive. This is assuming that the archive

[47] *https://www.outlookstudios.com/tools-to-find-broken-links-on-your-website/#Broken-Link-Checker*.

[48] *https://mukurtu.org*.

[49] *https://duraspace.org/dspace*.

[50] *http://archive.magoriabooks.com*.

[51] *http://lrl-diffusion.univ-bpclermont.fr/mulce2/accesCorpus/accesCorpusMulce.php*.

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

[53] *https://www.re3data.org/repository/r3d100013583*.

[54] LIA Sápmi - Sami Speech Corpus (*http://tekstlab.uio.no/LIA/samisk/index.html*); CHILDES Data Repository (*https://childes.talkbank.org/access*).

[55] DOBES The Language Archive (*https://archive.mpi.nl/tla*); The Language Commons (*https://archive.org/details/LanguageCommons?tab=about*); Slovenian language resource repository (*http://www.clarin.si/info/about/*); Eurac Research CLARIN Centre (*https://clarin.eurac.edu/index.html*); Open Resources and Tools for Language (ORTOLANG) (*https://www.ortolang.fr*).

site has not already collapsed structures that existed when researchers originally deposited their files. When this happens, crucial information can be lost for collections that depend on file structure to match transcripts and metadata files to audio and video files (for further discussion of arrangement, see Patterson 2021: §6.3.1).

We do recognize that there are non-trivial issues concerning bandwidth, web server traffic, and validation of large files that limit download capabilities and may require additional funding to resolve. Still, since these issues directly affect archives' functionality, they should be addressed sooner rather than later. Even if downloads must be done individually, solutions such as putting all of a collection's download links on a single page (as opposed to requiring users to enter into individual folders to download) exist. We draw attention to the DOBES Archive for providing an effortless method of downloading files in bulk. Their "basket" system allows users to select and bundle individual files or entire collections, then after an amount of time proportional to the number of files they have requested, a link to a zip file is emailed directly to them.

Other issues surrounding downloads included non-functioning download buttons or downloads that resulted in unreadable data. AILLA's download links are blocked by Chrome and Firefox browsers due to security settings, and could only be accessed by changing web browsers. The Hindu-Kush Areal Typology,[56] while not strictly an archive, had a bulk download option for wordlists. However, users had to ensure that they were properly opening the UTF-8 encoded CSV file in order to read the data without broken text. While workarounds like these exist, they may deter users with less familiarity with technology from using such archives effectively.

3.3.2 Recommendations

Firstly, and perhaps most critically, we suggest adding the option to download files in bulk, including an option for the entire corpus and for each folder in it, while preserving the original arrangement configuration. We recognize that this may be a complex request, given how file storage may work for the archive, but it is a necessary part of making files accessible. A 15,000-item collection with no bulk download option is neither accessible nor realistically usable. Furthermore, we suggest archives either allow depositors to preserve the original file structure of their collections upon deposit, or develop tools to help them better structure collections once archived in-site, for example through tags. It is vital that archives provide layout guides and naming conventions for depositors, so that users may quickly locate corresponding files and recreate file structures in the event that they are lost, and care should be taken when depositing collections to make sure that vital information about metadata and collection structure is not lost.

4 CONCLUSIONS

Digital archives, even when poorly maintained, may offer protection to language data that may otherwise have been lost, forgotten, or destroyed. We recognize that decisions made by both archives and depositors can greatly impact the accessibility of archived materials. We further recognize that there are tradeoffs in the creation of archives and some decisions that were made long ago continue to affect our methods, procedures, and choices. The power that both archivists and depositors have over these materials conveys a responsibility to ensure that materials will be able to be used and reused into the future. To that end, these findings and recommendations can help set procedural standards that greatly help those who access archives. We recognize that additional resources are necessary for this to succeed.

One incentive for depositors to increase the usability of their collection is for that work to be included in evaluations for promotion. By setting out how archives vary, and how that variation can affect the utility of collections and the user experience, we provide clarification to the scope of possible review. Individuals should not be evaluated for aspects of archiving which are outside their control; and if archives are to feature in hiring and/or promotion reviews, they may need to provide more explicit information about the scope and limitations of their services.

[56] *https://hindukush.clld.org/*.

APPENDIX

Information about the archive review:

- Archive name: the name of the archive
- Site link: the url of the web portal for the archive
- Metalanguage(s): the primary language which is used to deliver the records and to navigate the site
- Broken links: a qualitative assessment of the number of broken links encountered
- Types of materials available: a broad description of the filetypes available for download from the web portal
- Access restrictions: the types of access restrictions found across the site (or as described in the archive meta-information).
- Search function: information about how searches can be conducted on the site and the types of materials returned
- Filter by: discussion of how search results may be filtered.
- Bulk download: whether collection items must be downloaded individually (e.g. using the "save as" command through a web browser) or whether there are options for downloading multiple files at once.
- Number of clicks to download: how many steps does it take between a collection item's information and being able to download it.
- Metadata location: where metadata for a collection is accessed

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplementary Files 1.** Archive Audit Spreadsheet. Summarizes findings and comments. DOI: *https://doi.org/10.5334/johd.59.s1*

ACKNOWLEDGEMENTS

We are grateful to the Fieldwork Reading Group at Yale University for their valuable feedback and discussion throughout this project that have led to considerable improvements. We would also like to thank the audiences of The 7th International Conference on Language Documentation & Conservation (ICLDC2021), PARADISEC@100, and the UC Berkeley Language Revitalization Working Group for their insightful and useful feedback.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

IY: Data curation, formal analysis, investigation, validation, writing—original draft, writing—review & editing

AL: Data curation, formal analysis, investigation, validation, writing—original draft, writing—review & editing

JK: Conceptualization, data curation, formal analysis, investigation, methodology, validation, writing—review & editing

KH: Data curation, formal analysis, investigation, validation, writing—original draft, writing—review & editing

JJ: Data curation, formal analysis, investigation, validation, writing—review & editing

SB: Conceptualization, data curation, formal analysis, investigation, methodology, validation, writing—review & editing

CB: Conceptualization, data curation, formal analysis, investigation, methodology, validation, writing—review & editing

AUTHOR AFFILIATIONS

Irene Yi *orcid.org/0000-0001-9255-4235*
Linguistics Department, Yale University, New Haven, CT, US

Amelia Lake
Linguistics Department, Yale University, New Haven, CT, US

Juhyae Kim
Linguistics Department, Cornell University, Ithaca, NY, US

Kassandra Haakman
Linguistics Department, Yale University, New Haven, CT, US

Jeremiah Jewell
Linguistics Department, Yale University, New Haven, CT, US

Sarah Babinski *orcid.org/0000-0001-7764-5876*
Linguistics Department, Yale University, New Haven, CT, US

Claire Bowern *orcid.org/0000-0002-9512-4393*
Linguistics Department, Yale University, New Haven, CT, US

REFERENCES

Administration for Native Americans (ANA). (2006). Native language preservation: A reference guide for establishing archives and repositories. *http://www.aihec.org/our-stories/docs/NativeLanguagePreservationReferenceGuide.pdf*

Americans With Disabilities Act of 1990, Pub. L. No. 101–336, 104 Stat. 328 (1990).

Austin, P. (2011). "Who uses digital language archives?" *PARADISEC blog*. *https://www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/* (last accessed 27 September 2021).

Austin, P. (2021). "Corpora and archiving in language documentation, description, and revitalization." *Presented at FieldLing Seminar 2021*. Paris. *http://www.peterkaustin.com/docs/teaching/2021-09-09_FieldLing.pdf*

Aznar, J., & **Seifart, F.** (2020). RefCo: An initiative to develop a set of quality criteria for fieldwork corpora. *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, 95–101. *https://hal.archives-ouvertes.fr/hal-03066031/file/lift.pdf#page=100* (last accessed 27 January 2022).

Babinski, S., Jewell, J., Kim, J., Haakman, K., Lake, A., Yi, I., & **Bowern, C.** (2022). "How usable are digital collections for endangered languages? A review." *Proceedings of the Linguistic Society of America (PLSA)* 7(1). 5219.

Baldwin, D., & **Olds, J.** (2007). Miami Indian language and cultural research at Miami University. In D. Cobb & L. Fowler (Eds.), *Beyond red power: American Indian politics and activism since 1900*, 280–90. Santa Fe: SAR Press.

Berez, A. L. (2013). The Digital Archiving of Endangered Language Oral Traditions: Kaipuleohone at the University of Hawai'i and C'ek'aedi Hwnax in Alaska. *Oral Tradition*, *28*(2), 261–270. DOI: *https://doi.org/10.1353/ort.2013.0010*

Berez-Kroeker, A., Gawne, L., Kung, S., Kelly, B., Heston, T., Holton, G., Pulsifer, P., Beaver, D., Chelliah, S., Dubinsky, S., Meier, R., Thieberger, N., Rice, K., & **Woodbury, A.** (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics, 56*(1), 1–18. DOI: *https://doi.org/10.1515/ling-2017-0032*

Bird, S., & **Simons, G.** (2003). Seven Dimensions of Portability for Language Documentation and Description. *Language*, *79*(3), 557–582. DOI: *https://doi.org/10.1353/lan.2003.0149*

Burke, M., & **Zavalina, O. L.** (2020). Descriptive richness of freetext metadata: A comparative analysis of three language archives. *Proceedings of the Association for Information Science and Technology*, *57*(1), e429. DOI: *https://doi.org/10.1002/pra2.429*

Burke, M., Zavalina, O. L., Phillips, M. E., & **Chelliah, S.** (2021). Organization of Knowledge and Information in Digital Archives of Language Materials. *Journal of Library Metadata, 20*(4), 185–217. DOI: *https://doi.org/10.1080/19386389.2020.1908651*

Campbell, L., & **Belew, A.** (2018). Introduction: Why catalogue endangered languages? In L. Campbell & A. Belew (Eds.), *Cataloguing the World's Endangered Languages*, 1–14. London: Routledge. DOI: *https://doi.org/10.4324/9781315686028*

Campbell, L., Lee, N. H., Okura, E., Simpson, S., & **Ueki, K.** (2013). New Knowledge: Findings from the *Catalogue of Endangered Languages* ("ELCat"). *3rd International Conference on Language Documentation & Conservation*. *https://scholarspace.manoa.hawaii.edu/bitstream/10125/26145/2/26145.pdf*

Campbell, M. H. (2018). *Accessibility of Archives' Digital Resources for Users with Hearing and Visual Impairments*. Master's Thesis, University of North Carolina at Chapel Hill. *https://doi.org/10.17615/c11t-gs09*

Carew, M., Green, J., Kral, I., Nordlinger, R., & **Singer, R.** (2015). Getting in touch: Language and digital inclusion in Australian Indigenous communities. *Language Documentation & Conservation*, *9*, 307–323. http://hdl.handle.net/11343/57354

Corrado, E. M., & **Sandy, H. M.** (2017). *Digital Preservation for Libraries, Archives, and Museums*. Lanham, MD: Rowman & Littlefield.

Dobrin, L., & **Schwartz, S.** (2021). The social lives of linguistic field materials. *Language Documentation and Description, 21*. http://www.elpublishing.org/docs/1/21/ldd21_01.pdf

Evans, N., & **Sasse, H.-J.** (2004). Searching for meaning in the Library of Babel: field semantics and problems of digital archiving. In L. Barwick, A. Marett, J. Simpson & A. Harris (Eds.), *Researchers, Communities, Institutions, Sound Recordings,* 1–42. Sydney: University of Sydney. http://hdl.handle.net/2123/1509

Gaby, A., & **Woods, L.** (2020). Towards linguistic justice for Indigenous people: A response to Charity Hudley, Mallinson, and Bucholtz. *Language*, *96*(4), e268–e280. DOI: https://doi.org/10.1353/lan.2020.0078

Hinton, L. (2003). How to teach when the teacher isn't fluent. In J. Reyhner, O. V. Trujillo, R. L. Carrasco, & L. Lockard (Eds.), *Nurturing Native Languages*, 79–92. Flagstaff, AZ: Northern Arizona University. https://jan.ucc.nau.edu/~jar/NNL/NNL_6.pdf

Hinton, L. (2018). Approaches to and Strategies for Language Revitalization. In K. L. Rehg & L. Campbell (Eds.), *The Oxford Handbook of Endangered Languages*, 442–465. Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780190610029.013.22

Holton, G. (2012). Language archives: They're not just for linguists any more. In F. Seifart, G. Haig, N. P. Himmelmann, D. Jung, A. Margetts, & P. Trilsbeek (Eds.), *Potentials of Language Documentation: Methods, Analyses, and Utilization*, 111–117. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/4523

Kaplan, J., & **Lemov, R.** (2019). Archiving Endangerment, Endangered Archives: Journeys through the Sound Archives of Americanist Anthropology and Linguistics, 1911–2016. *Technology and Culture 60*(2), S161-S187. DOI: https://doi.org/10.1353/tech.2019.0067

Kung, S. S., Sullivant, R., Pojman, E., & **Niwagaba, A.** (2020). Archiving for the Future: Simple Steps for Archiving Language Documentation Collections. New York, NY: Teach Online with Teachable. https://archivingforthefuture.teachable.com

Mansfield, J. (2014). *Polysynthetic sociolinguistics: The language and culture of Murrinh Patha youth*. PhD dissertation, Australian National University. https://doi.org/10.25911/5d723cd88582b

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & **Sonderegger, M.** (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association*. https://montrealcorpustools.github.io/Montreal-Forced-Aligner/. DOI: https://doi.org/10.21437/Interspeech.2017-1386

Nathan, D. (2010). Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing*, 4(1–2), 111–124. https://doi.org/10/c7ct5f. DOI: https://doi.org/10.3366/ijhac.2011.0011

Paterson, H. J., III (2021) "Where Have All the Collections Gone?" *Poster presented at the 15th Annual Society of American Archivists Research Forum*. https://hughandbecky.us/Hugh-CV/publication/2021-where-have-all-the-collections-gone/Where-have-all-the-Collections-Gone.pdf

Schwartz, S., & **Dobrin, L.** (2016). The cultures of Native North American language documentation and revitalization. *Reviews in Anthropology*, *45*, 88–123. DOI: https://doi.org/10.1080/00938157.2016.1179522

Simons, G. F., & **Lewis, M. P.** (2013). The world's languages in crisis: A 20-year update. In E. Mihas, B. Perley, G. Rei-Doval, & K. Wheatley (Eds.), *Responses to language endangerment. In honor of Mickey Noonan,* 3–19. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/slcs.142.01sim

Sonnemaker, T. (2020). "The Number of Americans without Reliable Internet Access May Be Way Higher than the Government's Estimate—and That Could Cause Major Problems in 2020." https://www.businessinsider.com/americans-lack-of-internet-access-likely-underestimated-by-government-2020-3 (last accessed 20 September 2021).

Sullivant, R. (2020). Archival description for language documentation collections. *Language Documentation & Conservation*, *14*, 520–578. http://hdl.handle.net/10125/24949

Vann, R. E. (2006). Frustrations of the Documentary Linguist: The State of the Art in Digital Language Archiving and the Archive that Wasn't. *Department of Spanish Research*, *1*. Western Michigan University. https://scholarworks.wmich.edu/spanish_research/1

Wasson, C., Holton, G., & **Roth, H. S.** (2016). Bringing User-Centered Design to the Field of Language Archives. *Language Documentation & Conservation*, *10*, 641–681. http://hdl.handle.net/10125/24721

Web Content Accessibility Guidelines (WCAG). (2018). *Web Accessibility Initiative. WCAG 2.1.* https://www.w3.org/WAI/standards-guidelines/wcag/

Whalen, D. H., DiCanio, C., & **Dockum, R.** (2018). Phonetic documentation in the literature: Coverage rates for topics and languages. *The Journal of the Acoustical Society of America, 144*(3), 1936–1936. DOI: *https://doi.org/10.1121/1.5068471*

Whalen, D. H., Moss, M., & **Baldwin, D.** (2016). Healing through language: Positive physical health effects of indigenous language use. *F1000Research*, *5*. DOI: *https://doi.org/10.12688/f1000research.8656.1*

Woodbury, A. C. (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In D. Nathan & P. K. Austin (Eds.), *Language Documentation and Description: Special Issue on Language Documentation and Archiving, 12,* 19–36. London: SOAS. *http://www.elpublishing.org/PID/135*

TO CITE THIS ARTICLE:
Yi, I., Lake, A., Kim, J., Haakman, K., Jewell, J., Babinski, S., & Bowern, C. (2022). Accessibility, Discoverability, and Functionality: An Audit of and Recommendations for Digital Language Archives. *Journal of Open Humanities Data*, 8: 10, pp. 1–19. DOI: *https://doi.org/10.5334/johd.59*

Published: 24 March 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.