January 2021

Language Archive Records: Interoperability Of Referencing Practices And Metadata Models

Hugh J. Paterson Iii

LANGUAGE ARCHIVE RECORDS:
INTEROPERABILITY OF REFERENCING PRACTICES AND METADATA MODELS

by

Hugh Joseph Paterson III
Bachelor of Arts, Southern Illinois University Edwardsville, 2007

A Thesis

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of
Master of Arts

Grand Forks, North Dakota
May
2021

This thesis, submitted by Hugh Joseph Paterson III in partial fulfillment of the requirements for the Degree of Master of Arts from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

_____

J. Albert Bickford, Chair

_____

Douglas M. Fraiser

_____

Robert Fried

This thesis is being submitted by the appointed advisory committee as having met all of the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.

_____

Chris Nelson
Dean of the School of Graduate Studies

_____

Date

PERMISSION

Title Language Archive Records:
Interoperability of Referencing Practices and Metadata Models

Department Linguistics

Degree Master of Arts

In presenting this thesis in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my thesis work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this thesis or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my thesis.

Hugh Joseph Paterson III
4/21/2021

One can’t archive people.
Share life with those around you.
In memory of Stephen Muscarella.
My friend.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful to Jeremy Nordmoe (L&CA), Mandana Seyfeddinipur (ELAR), Nick Thieberger (PARADISEC), and Aurelia Vasile (Pangloss) for constructive comments on earlier versions of this work. I am also grateful for assistance in proof reading from Jedidiah Paterson (abstract and chapter 1), Moriah Paterson (chapters 2, 6 & 7), Monica Paterson (chapters 3 & 5), Catherine MacLeod (abstract, chapters 1, 2, 3 & 5) and Rebecca Paterson (the whole thesis). Andy Black has been immensely helpful in the creation of XLingPaper and his unparalleled support of that product. I am grateful for each bug he has fixed. It has been a pleasure to work with him to create pathways to export bibliographic metadata from Zotero via MODS and import it into XLingPaper. Sebastian Karcher and Brenton Wiernik have helpfully increased my understanding of the Zotero application. I am grateful to Albert Bickford, my advisor, for being flexible and pushing me towards excellence. I also thank my committee members, Robb Fried and Doug Fraiser for inspirational discussions prior to this research and for their attention to detail in the final draft. All errors are, unfortunate, and my own.

ABBREVIATIONS

ACM	Association for Computing Machinery
AILLA	The Archive of the Indigenous Languages of Latin America
APA	American Psychological Association
API	Application Programming Interface
ARK	Archival Resource Key
BCP	Best Current Practice
BOLD	Basic Oral Language Documentation
CD	Compact Disc
CLA	California Language Archive
CMS	Content Management System
CNRS	Centre national de la recherche scientifique
COinS	Context Objects in Spans
CSL	Citation Style Language
DACS	Describing Archives: A Content Standard
DC	Dublin Core
DELAMAN	Digital Endangered Languages and Musics Archives Network
DOI	Digital Object Identifier
ELAR	Endangered Languages Archive
FRBR	Functional Requirements for Bibliographic Records
GA1	Greg Anderson Collection 1
HTML	Hypertext Markup Language
HTML5	HyperText Markup Language 5
IANA	Internet Assigned Numbers Authority

IETF	Internet Engineering Task Force
IFLA	International Federation of Library Associations and Institutions
IMT	Internet Media Type
ISBN	International Standard Book Number
ISO	International Standards Organization
IT	Information Technology
JISC	Joint Information Systems Committee
JSON	JavaScript Object Notation
L&CA	Language and Culture Archives
LACITO	Langues et civilisations à tradition orale
LRM	Library Reference Model
LSA	Linguistic Society of America
MARC	Machine-Readable Cataloging
MIME	Multipurpose Internet Mail Extensions
MODS	Metadata Object Description Schema
NGO	Non-Governmental Organization
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OLAC	Open Language Archives Community
p.c.	Personal Communication
PARADISEC	Pacific and Regional Archive for Digital Sources in Endangered Cultures
RB5	Roger Blench Collection 5
RDF	Resource Description Framework
RFC	Request for Comments
SOAS	School of Oriental and African Studies
TLA	The Language Archive
UHM	University of Hawai‘i at Mānoa
URI	Uniform Resource Identifier

URL	Universal Resource Locator
USD	United States Dollar
W3CDTF	W3 Consortium Date and Time Formats
XHTML	EXtensible HyperText Markup Language
XML	Extensible Markup Language
ZF1	Zygmunt Frajzyngier Collection 1

ABSTRACT

With the rise of the digital language archive and the plethora of referenceable content, a critical question arises: “How easy is it for authors to use existing tools to cite the content they are referencing?” This is especially important as people use archived materials as evidence within published language descriptions.

Archived resource metadata is well discussed in language documentation circles; however, bibliographic metadata and its accessibility are less discussed. Discoverability metadata, a subset of archived resource metadata, serves aggregators like OLAC by declaring a resource exists. In contrast, bibliographic metadata functions within documents by declaring where to find a resource that is known to exist.

In this thesis I look at the interaction between Zotero, an open source reference manager, five different archives (PARADISEC, Pangloss, SIL Language & Culture Archives, ELAR, and Kaipuleohone), and three methods of importing metadata from them into Zotero (DOI import, HTML embedded metadata, and file based import). I report on collection and audio artifact metadata provided by the archive to the author via Zotero’s interfaces: what’s included, what’s missing, and what’s misaligned.

Understanding the processes by which authors collect metadata for the purpose of citation and referencing, what metadata they need, and if it is being provided, facilitates the design of useful interfaces to archives which elevate the value of archives to all groups who interact with them. I propose that interaction design is an additional factor to those presented by Chang (2010) in her well received checklist for evaluating language archives. Interaction design, the technical field concerned with designing how people interact with objects and services, is the design process by which archives manage the interactions they have with those they serve. I specifically argue that interaction design adds value to an

archive's brand, as perceived by the network of archive users, when it facilitates the interaction with bibliographic metadata about artifacts within holdings. This added value speaks to the sustainability of an archive within its sphere of influence. It is increasingly important in the career development of scholars to meet metric-based assessments of their influence in scholarly discussions. Reference counts, including those pointing to the evidentiary record housed in archives, play a significant role in establishing quantitative baseline metrics for scholars.

CHAPTER 1

Background

Within the fields of linguistics and language documentation there has been a favorable response to a larger movement across scholarly disciplines to "prevent the erosion of the base" (Altman 2013). That is, there is a need to secure the evidence on which the discipline depends. This commitment to accessible evidence has led to an embraced use of language archives to house the evidentiary record.[1] However, making evidence accessible is only one part of the picture. It is equally important to link social (including scholarly) uses and observations about the evidence back to where that evidence is preserved—and accessible, that is, through proper referencing. An early discussion of the poor state of referencing can be found in Bird and Simons (2003a:§3.5) where they present the challenges that linguists in a digital world and language documentation practitioners should address with more specificity. The challenge that Bird and Simons laid out was in part answered via the work that went into creating both the *Austin Principles*[2] and a position statement (Berez-Kroeker et al. 2018) which together lay out the imperatives for data referencing and citation.[3,4]

[1] The embraced use of language archives by language documentation practitioners has also been influenced by scholarly activism, in which scholars have generally broadened their understanding of how scholarly pursuits impact language communities.

[2] https://site.uit.no/linguisticsdatacitation/austinprinciples

[3] In my work, I follow the usage of *citation* and *references* as outlined in the *Chicago Manual of Style* (2017:898 §15.10) and the *Publication manual of the American Psychological Association* (2010:6). That is, I use the term *citation* to mean an in-line pointer to a specific item within a *references section* (a document-internal link). I use the term *references* to refer to the items in a references section which provide the bibliographic metadata necessary for someone to find the cited resource (a document-external link). A references section is also sometimes called a *bibliography section*, though other document sections, such as *works consulted* may also contain references. References may also appear throughout a work depending on the selected style sheet or the purpose of the document as is the case with annotated bibliographies.

[4] I use the term **artifact** to mean any physical or digital object. In this sense, archives hold artifacts. In my work, I define **evidence** as those things used to support an argument. Evidence may be true or fabricated, but evidence speaks to the use of artifacts as logical propositions for belief. When I use the term **data**, I refer specifically to an artifact which can be described with the DCMIType element **dataset**, e.g., tabulated values which are often designed to be read by computer. In this way, it is not uncommon for scientists in the physical sciences to create a laboratory event (e.g., with an electron or proton) and observe that event with instruments that only produce datasets. Taking certain values from these datasets, then, and using them in an argument allows for the datasets (or some part of a dataset) to serve as evidence. In the activities of linguistics,

This was then followed up with work aimed at authors and publishers which offers a practical method to implement data citation, resulting in *The Tromsø Recommendations for Citation of Research Data in Linguistics* (Andreassen et al. 2019).

In order to promote a culture of informative and effective linking between evidence and its usage, still more work needs to be done. In this research I take a first look at how language archives—acting as evidence distributors—provide metadata to authors. I look at the interaction between authors and archives via an open source tool—*Zotero*—which is used to craft citations and references during the authorship process. Zotero[5] was chosen for two reasons: first, a poll conducted in 2015 indicated that sixty percent of respondents use Zotero; and second, Zotero has more import options and a larger ecosystem of plugins and compatible authorship tools than other open source reference managers. That is, Zotero is only one example of technology which sits between artifacts and the creative outputs of scholarly discussion, but there is reason to believe that a larger number of authors use it. Reference managers like Zotero are used in various ways, but three benefits are:

1. the abstraction of bibliographic metadata from the formats required by different style sheets (allowing the reuse of the metadata in different authorship projects with different style sheets),
2. the storage of associated research notes with the bibliographic metadata, and
3. the storage and management of a representation of the referenced work (usually a PDF).

In April of 2015, I sent a poll to an SIL International mailing list to which 143 language researchers were subscribed. Approximately thirty percent of the list members responded

language documentation, and language archiving, linguists generally create primary artifacts of the following kinds: video, audio, and text. These artifacts may then be processed to create secondary artifacts and compiled into datasets, e.g., using tools like PRAAT to extract formant values. However, generally, the original artifact produced in a linguistic investigation is not a dataset nor data—it is the *artifact*. My perspective has certain similarities with Himmelmann's (2012) analysis of linguistic data types. However, I do not use the term *data* with as broad a range as he does.

[5] https://www.zotero.org

to the following question: *Do you use a citation manager and if so which one?* The results are tabulated in Table 1.

Table 1. Breakdown of reference manager poll responses

Citation Manager	Users	% of total
Zotero	26	60.47
Endnote	5	11.63
Nothing	3	6.98
Citavi	1	6.98
Sente	1	2.33
JabRef / Bibtex	1	2.33
Biblioscape	1	2.33
Papers	1	2.33
XLingPaper	1	2.33
RefWorks	1	2.33

In general there were two classes of reference manager software users who responded to the poll: those who use a reference manager currently in ongoing research, and those who have used a reference manager in the past but no-longer are conducting research. Both classes of users were treated as "users" and are presented in the same manner. Several people responded that they are not using any reference manager during their research process. I suspect that many more people who are actively involved in research but do not use a reference manager simply did not respond. I interpret these results to indicate that many people don't use a reference manager at all, but of those who do, there is a significantly larger portion of researchers who use Zotero.[6]

There is a linking cycle between scholarly works, archival artifacts, and authors. The cycle is illustrated in Figure 1 presented in the order of evolutionary history. At point Ⓐ, Bird and Simons (2003a) urge the discipline to find descriptive and articulate reference formats for linking scholarly works to archival artifacts. More recently indicated by point Ⓑ, the authors of the *The Tromsø Recommendations for Citation of Research Data in Linguistics* (Andreassen et al. 2019) propose formatting guidelines for use within scholarly works.

[6] This kind of question should be investigated with a larger audience of language researchers. One possible bias might be generational. I have incidentally noticed that senior language researchers have a tendency to not use reference managers.

This thesis addresses the relationship between the archive and Zotero at point Ⓒ and seeks to give an account of the state of the field and which direction the discipline needs to move. Finally, point Ⓓ indicates that even after bibliographic metadata is in Zotero, Citation Style Language files for specific style sheets need to be updated to accommodate authors as they use tools like Zotero.

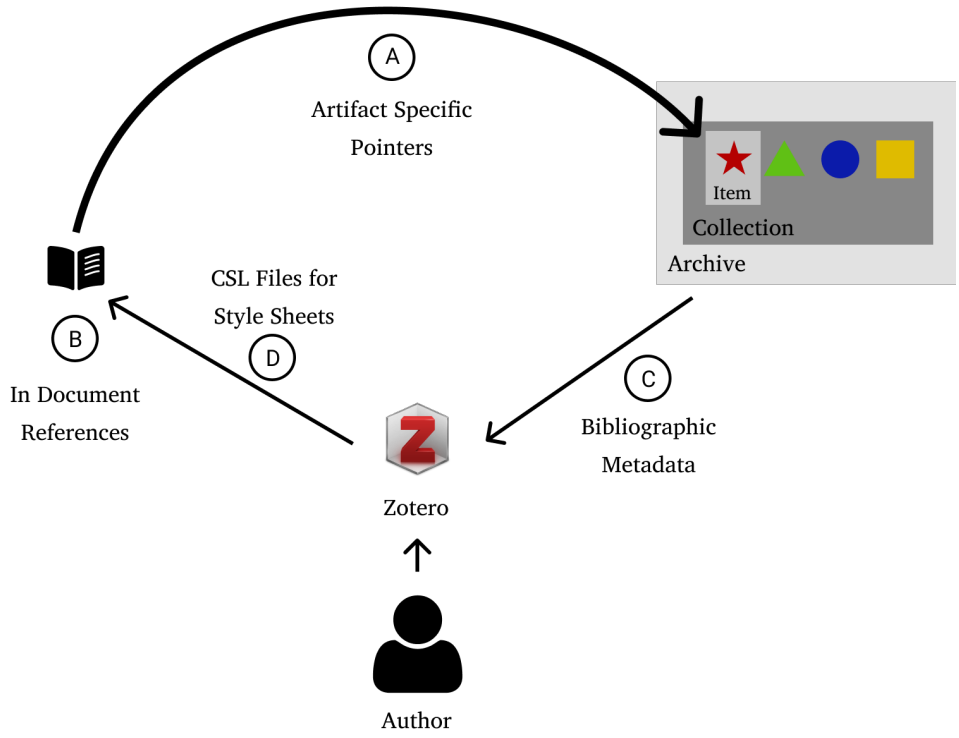

Figure 1. The linking cycle between scholarly works and archives

One could reasonably expect that the minimum bibliographic metadata required to be transferred to Zotero via these interactions would be the sum of metadata required to create references across important citation styles such as *Chicago 17th edition Author-date*,[7] *APA 6th*, *APA 7th*, *Unified style sheet for linguistics*, etc.

[7] The *Chicago Manual of Style* outlines two major ways of presenting citations and references, each with several sub-variations. The two major styles presented are often referred to as *Chicago 17th edition* and *Chicago 17th edition Author-date*. The "Author-date" part of the title refers to the presentation style's format for references where the first element of the reference is the author's name and the second element is the date. Throughout this thesis I refer to the Author-date style.

This thesis reviews collection records and audio artifacts contained within a collection[8] as the point of analysis across the archives surveyed. As a point of reference and a baseline, the following examples are provided from *APA 6th edition*[9] and *Chicago 17th edition Author-date*[10] in order to demonstrate the kinds of references expected when using these well-respected and well-described style sheets.[11] Particularly note how material types are indicated in the references, how contributors are treated, and how the location of the artifact is specified including the artifact's location within a larger collection.

Below is the APA 6th edition collection reference template (VandenBos 2010:212):

[8] There has been some discussion of what constitutes a **collection** among language documentation practitioners. Johnson (2004:142) equates a collection and a corpus saying: "A ***collection***, or ***corpus***, is the body of documentary materials created by linguists and native speakers in the course of their research." Simons (2008) and Boerger (2011:Footnote 2) advance the idea that the materials generated as part of a language documentation effort are a corpus, not distinguishing between a corpus and a collection. Many other language documentation practitioners (Lüpke 2010, Good 2011, Austin 2013) extend or modify the meaning of "corpus" when referring to a collection; using a phrase like ***language documentation corpus***. In contrast to the coalescence of corpus and collection, Thieberger (2018), in a blog post, credits Jane Simpson with proposing the following terms: **Assemblage** – all material collected, working files, early sources, multiple versions and drafts; **Collection** – the archived material, a subset of the above, but curated with sufficient metadata to allow the user to know what all items are; **Corpus** – a crafted set of texts in the language that can be used for further analysis. For a recent treatment of these ideas see Sullivant (2020) which expounds on Simpon's ideas. I am not convinced that the distinction between assemblage and collection is relevant for a variety of reasons. Most relevant for this thesis is their bearing on the crafting of a bibliographic reference. I suggest that reference formulations do not need to be different between an assemblage and a collection. Lest I leave the definitions distinguishing a corpus and a collection undisputed, let me suggest some pragmatic distinctions. Biber (1993a; 1993b) set up diversity-based criteria for corpus design. One interpretation of Biber's work is that if a body of evidence does not meet the diversity threshold then it does not qualify as a corpus. The diversity dynamic as a goal for a set of materials generated through language documentation efforts has been a long standing tenant of creating an evidence based documentary record for a language (Himmelmann 1998). Biber's diversity arguments have even been suggested as guidelines for guiding elicitation in language documentation (Lüpke 2010). In this sense it is clear that selection is a process which impacts both corpus design and what is collected during fieldwork. However, a second useful criterion for defining a corpus (in contrast to a collection) might be that there is an existing tool set which can be used to exploit the collection of artifacts as a single work. This could lead to an interpretation that a corpus is more like a **dataset** in the Dublin Core sense, while a collection is more like a **collection** in the Dublin Core sense—and indeed we see that Xia et al. (2016) blur the line between database and corpus in their work. Both datasets (of which databases are a type) and collections can be considered examples of aggregate works. I realize that linguists and language materials archivists have applied these labels loosely but I side with Society of American Archivists (2013:rule 2.3.19) in its suggested distinguishing terms for the highest level of an archival unit of materials: "Archival materials are frequently described by devised aggregate terms such as *papers* (for personal materials), *records* (for organizational materials), or *collection* (for topical aggregations)" (see Wahbeh 2009 for an example of application). There are other terms which still need to be fleshed out among language archive practitioners such as: archival **deposit**, **accession**, and what one might call an **artifact and its associated metadata file**.

[9] In this thesis APA formatted references are presented in ***brown boxes***.

[10] In this thesis Chicago formatted references are presented in ***lavender boxes***.

[11] It is not uncommon for websites to offer a "suggested citation" or "suggested reference". Because the formats of these suggestions are generally not either Chicago or APA they are presented in ***blue boxes***.
In some cases, publishers do not provide a template or discussion for a reference pattern. They only provide exemplars. This is the case with Chicago 17, the journal *Language*, and the journal *Linguistic Inquiry*. Specifically in chapter 5, I provide an analysis of reference patterns provided by publishers. I place these analyses in ***yellow boxes***.

Author, A. A. (Year, Month Day). Title of material. [Description of material]. Name of collection (Call number, Box number, File name or number, etc.). Name and location of repository.

Below are examples of referenced collections components APA 6th edition (VandenBos 2010:213):

Frank, L. K. (1935, February 4). [Letter to Robert M. Ogden]. Rockefeller Archive Center (GEB series 1.3, Box 371, Folder 3877), Tarrytown, NY.

Berliner, A. (1959). Notes for a lecture on reminiscences of Wundt and Leipzig. Anna Berliner Memoirs (Box M50). Archives of the History of American Psychology, University of Akron, Akron, OH.

Below is the APA 6th edition audio recording reference template (VandenBos 2010:209, Skutley 2012:25):[12]

Writer, A. A. (copyright year). Title of song [Recorded by B. B. Artist if different from writer]. On *Title of album* [Medium of recording: CD, mp3, record, cassette, etc.]. Retrieved from http://xxxxx (Date of recording if different from song's copyright date)

Below is the Chicago 17th edition collection reference for a whole collection in Author-date format (Harper 2017:§15.54):

[12] I lean on APA descriptions of published audio works for their object-in-a-container nature, e.g., track on an album. Skutley (2012:27 #53) does present an example of a speech, unfortunately, the location of where to access the content is ephemeral.

"King, M. L., Jr. (1963, August 28). *I have a dream* [Audio file]. Retrieved from http://www.americanrhetoric.com/speeches/mlkihaveadream.htm"

APA like Chicago (as mentioned in footnote 13) does not have a straightforward approach to signifying the difference between an object title and a container title for collections and artifacts. Speeches such as *I have a dream*, may have author assigned titles. Collections however seem more frequently to have titles assigned by the institutions which house them. It seems to be that it is in these cases which titles are not italicized, the distinguishing characteristic used in APA. None of the collection titles are italicized in any of the examples provided by VandenBos (2010:§7.08-7.10).

Egmont Manuscripts. Phillipps Collection. University of Georgia Library.

Below is the Chicago 17th edition collection reference for a single item in a collection in Author-date format (Harper 2017:§15.54):

Dinkel, Joseph, n.d. Description of Louis Agassiz written at the request of Elizabeth Cary Agassiz. Agassiz Papers. Houghton Library, Harvard University.

Below are Chicago 17th edition audio recording references for a single item in a collection in Author-date format (Harper 2017:§15.57):[13]

Coolidge, Calvin. [1920?]. "Equal Rights" (speech). In "American Leaders Speak: Recordings from World War I and the 1920 Election, 1918-1920." Library of Congress. Copy of an undated 78 rpm disc, RealAudio and WAV formats, 3:45. http://memory.loc.gov/ammem/nfhtml/.

Holiday, Billie, vocalist. 1958. "I'm a Fool to Want You." By Joel Herron, Frank Sinatra, and Jack Wolf. Recorded February 20,1958, with Ray Ellis. Track 1 on Lady in Satin. Columbia CL 1157, 33x/3 rpm.

The remainder of this thesis is organized as follows. In chapter 2 I discuss how bibliographic metadata is imported to Zotero. In chapter 3 I discuss how authors make use

[13] To complicate matters, it would appear that the editors of the 17th edition of the *Chicago Manual of Style* have taken some liberties in the generation of their references. The material referenced actually does have a known date, and has a link to its specific record. The link provided in the *Chicago Manual of Style* directs users to the whole collection, rather to the specific item. For the purposes of this thesis I am going to overlook this discrepancy. The following is how I would craft a reference to this resource in Chicago 17th edition style, which includes the bibliographic content for the formally published audio, the archival collection it appears in, and a link to its online location:

Calvin, Coolidge. 1920. "Equal Rights" (speech). Recorded in June 29, 1920, New York, N.Y., Nation's Forum: 49853 (matrix). Bridgeport, Conn.: Columbia Graphophone Manufacturing Company. In "American Leaders Speak: Recordings from World War I and the 1920 Election, 1918-1920". Copy of 78 rpm disc, RealAudio and WAV formats. Nation's Forum Collection, item 17 [Nation's Forum matrix 49853], Side A. Library of Congress. https://www.loc.gov/item/2016655159.

In the above case, I would not just be referencing the published audio in a general sense but I am referencing the digital audio file made from a particular item in the Library's collection.

One additionally confusing inconsistency in the Chicago 17th edition style is the use of quotes and italics in references with regards to collections. Generally, a whole work such as a book's title is italicized, the title of a portion of that book would appear in quotes. This is not how audio tracks which are part of an album are presented, nor is it how items in a collection are presented. It would appear that this is an inconsistent application of the part-whole distinction which the use of italics and quotes attempts to denote.

of their Zotero database when composing documents. In chapter 4 I present a survey of interoperability of bibliographic metadata between five language archives and Zotero. In chapter 5 I offer some example references and include a brief discussion of archive resources in the journal style sheets for *Language* and *Linguistic Inquiry*. In chapter 6 I discuss the implications of the survey results and some of the underlying reasons why archives struggle to provide value to users via bibliographic metadata exchange. Finally, in chapter 7 I discuss some social approaches for addressing technical challenges.

CHAPTER 2

Importing bibliographic data

Given that Zotero is well reviewed in the academic literature (Trinoskey et al. 2009; Duong 2010; Mueen Ahmed & Al Dhubaib 2011; Idri 2015; Thomson 2016; Ray 2017; Brander et al. 2019), I discuss only those technical details of its use which are directly relevant to this discussion.[14] Broadly, Zotero is a reference management database which allows users to track the artifacts they reference in scholarly works. It facilitates the easy crafting of formatted citations and references in an author's document.

Zotero allows users to import metadata about items they encounter and wish to potentially cite and reference, in three different automated ways.[15] The first is with a unique identifier such as a DOI (as shown at the number 1 in Figure 2) which Zotero uses to fetch data about an item from online databases. This method or tool is sometimes called the "magic wand tool" due to its icon.

[14] Rueda (2016) is a DataCite blogpost presenting a DataCite webinar (Karcher 2016b, Karcher (2016a) in which a Zotero team presents methods for making Zotero CSL craft references for Datasets. Datasets and Collections are similar in how Zotero and CSL currently implement them. Watching the video (Karcher 2016b) will give my reader an introductory perspective on the technical issues I discuss.

[15] Zotero detects other kinds of identifiers (ISBN, PMIDs, and arXiv IDs) as is shown in Figure 2. None of these apply to language archives and therefore are not discussed. There is also the manual input method, which is necessary to use when automated methods provide errant or missing information.

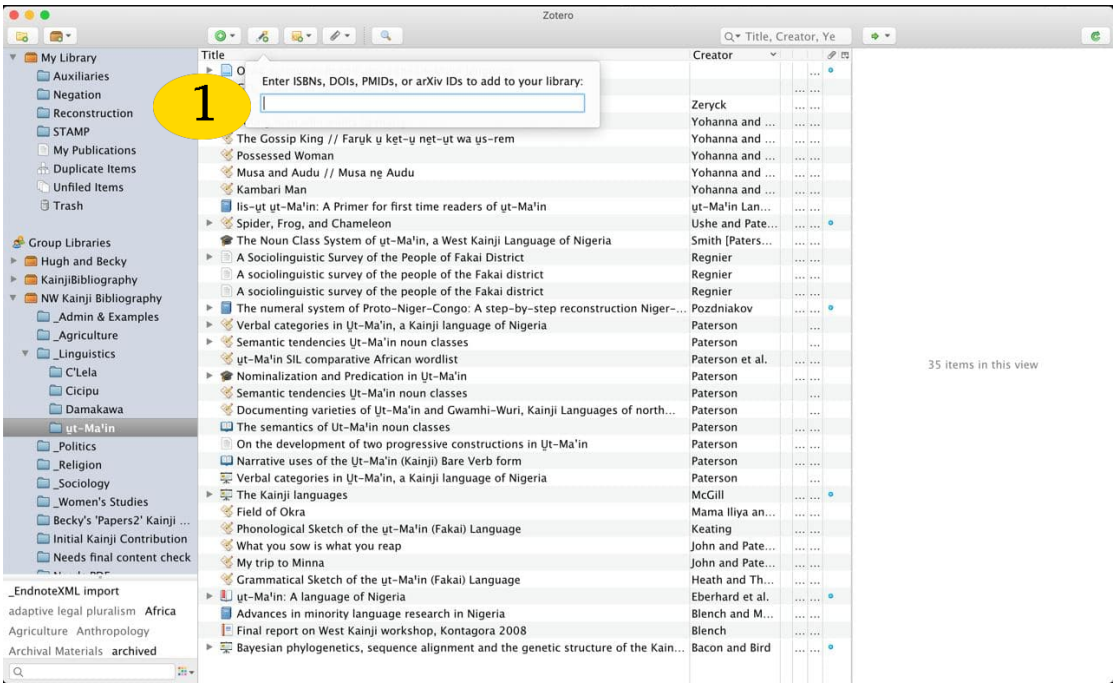

Figure 2. Zotero's ISBN and DOI lookup tool

Second, Zotero allows users to add a browser plugin that can read embedded metadata in HTML pages upon which digital artifacts and records are viewed. Upon selection by the user, Zotero then creates an entry in its database for the desired items. If multiple input methods exist, the connector plugin allows users to select which imput method they wish to use. This is illustrated in Figure 3.

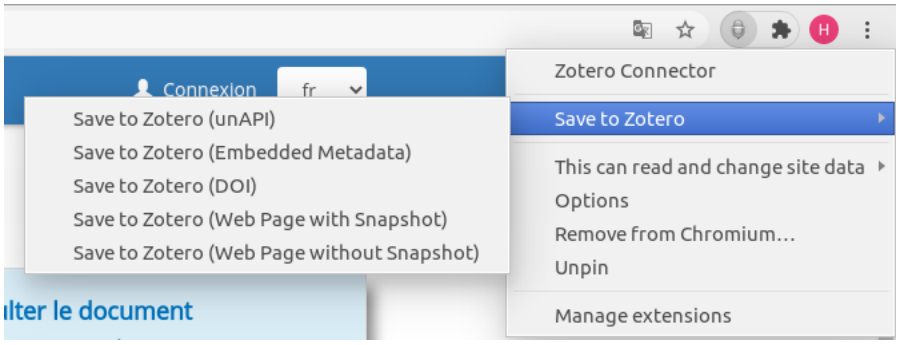

Figure 3. Zotero's browser plugin demonstrating the detection of multiple input sources

Third, Zotero facilitates the import of metadata from several different file types which are used for bibliographic metadata exchange (BibTeX, RIS, Zotero RDF, etc.).

Zotero does not glean information from a web page the way a human would by reading the displayed text. Since Zotero does not have any natural language processing features to detect metadata, nothing is automatic. However, website developers and publishers can write translators, i.e., small JavaScript web page scrapers, which tell Zotero which HTML fields and which classed HTML elements contain compatible text.[16]

2.1 DOI import

Digital object identifiers (DOI)[17] are frequently used by publishers to identify a specific publication. They are intended to be globally unique and usable in a URL to access either an item or information about an item. There are several companies, known as registration agencies, that issue DOIs to publishers. Crossref[18] and DataCite[19] are two such registration agencies. Crossref focuses on providing service to traditional format publishers (books and journals). In contrast, DataCite focuses on providing services to data repositories and non-traditional publishers, such as institutional repositories. Both registration agencies provide **application programming interfaces** (APIs) which enable software programs to look up information in their databases.[20] A Zotero user may use the DOI feature to query these APIs to import the metadata on record in the Crossref or DataCite database into Zotero. However, from personal experience I find it is the rare case that an API reply from Crossref or DataCite is without error for content published via linguistic publishers and language documentation data repositories. Tracking down the source of these errors is challenging. Possible sources include Zotero's interpretation of the metadata provided to it, or that the publishers are providing errant metadata to the

[16] https://www.zotero.org/support/translators
[17] https://www.doi.org/factsheets.html
[18] https://www.crossref.org
[19] https://datacite.org
[20] As an organization, Crossref focuses on the issuance of DOIs to publishers in the traditional sense, while DataCite focuses on the issuance of DOIs to archives and repertories which have "datasets". In practice, I have seen archives such as SEALang (http://sealang.net/library) assign DOIs from DataCite to published articles (from serial publications). Doing this raises several questions regarding ideological definitions of "published", and "dataset". Does published mean public access or does it mean processed via a formal press and publication venue? Similarly, can one have a dataset composed of published works? From the archive's perspective they are not the publisher of these objects, as that designation belongs to the originating press. What organizations must deal with when choosing appropriate DOIs for their content is the metadata schema available from the organization providing the DOI. DataCite's schema is designed to be applied in contexts where non-traditional print artifacts or non-print artifacts are the objects being defined. Whereas Crossref's metadata schema is oriented towards traditional academic print media.

DOI providers. Sometimes it is clearly the case that publishers and archives are providing errant or insufficient metadata to the DOI API service. In chapter 4, I present the results of attempts to import metadata from several artifacts in language archives which use DOIs.

2.2 Embedded metadata in HTML import

Zotero can, through a plugin, connect to a user's web-browser and read metadata embedded in HTML pages.[21] This can be very useful as users can then click a button in their browser to add an item they discover on the web to their Zotero database. Zotero attempts to read a variety of kinds of metadata embedded in a website to fill in a Zotero record via JavaScript translators. Zotero is responsive to several "dialects" of embedded metadata includ HTML meta tags such as Dublin Core tags,[22] EPrints tags,[23] or Highwire Press tags.[24,25] It is also able to detect a type of semantic markup using HTML span elements called *ContextObjects in Spans* (COinS).[26] Each set of tags has a different metadata schema associated with it. These schema were each designed to describe various kinds of materials and therefore have different scopes of coverage, resulting in different limitations when describing artifacts. For instance, COinS was designed specifically for books and journal articles to the exclusion of other kinds of materials, while Dublin Core does not overtly specify the difference between a masters thesis and a doctoral dissertation (see discussion in Allinson 2008). So, best practice recommendations for embedding metadata in HTML pages suggest using multiple strategies for full coverage. Bibliographic metadata

[21] https://www.zotero.org/support/dev/exposing_metadata

[22] For Dublin Core elements, and Dublin Core terms HTML tags consult defined tags in the WHATWG Wiki with the namespaces `dc.` and `dcterms.` : https://wiki.whatwg.org/wiki/MetaExtensions. Google Scholar indexing also uses this same namespace for their tags, but has added some Google Scholar specific tags which are not true Dublin Core elements or terms. Consult: https://scholar.google.com/intl/en/scholar/inclusion.html#indexing

[23] There are several vocabularies which make up the EPrints tag set. These can be found in the *Scholarly Works Application Profile* at: http://purl.org/eprint/terms. There are four other EPrints vocabularies which are also used in addition to the terms. These are the *Access Rights Vocabulary*, the *Entity Type Vocabulary*, the *Status Vocabulary*, and the *Type Vocabulary*. These are accessible at the wiki for JISC Digital Repositories Programme's Repositories Research Team: http://www.ukoln.ac.uk/repositories/digirep/index/Category:ScholarlyWorksApplicationProfile

[24] There is no Internet based publication of "The Highwire Press Tags". However, some tags are mentioned in the `citation_` namespace within the WHATWG Wiki. Other Highwire Press tags are mentioned in Google Scholar's documentation for indexing. See footnote 22 for links. I have noticed that several publishers hijack the `citation_` namespace and create custom tags. So, the exact set may be a bit amorphous.

[25] The exact file Zotero uses to convert these tags can be read in the application's Github repository: https://github.com/zotero/translators/blob/master/Embedded%20Metadata.js

[26] https://web.archive.org/web/20151106024244/http://ocoins.info

encoded throughly and presented well can save users many keystrokes. This makes the task of collecting bibliographic metadata and the resulting task of attribution in publications much easier. The embedding of metadata in HTML can also improve the visibility of content in search result rankings (Arlitsch & O'Brien 2012, 2013) for both general search and special search engines like *Google Scholar*.[27] However, negotiating web standards, metadata standards, and how those standards impact tools like Zotero can be a specialist's job.[28] In this thesis I look at how HTML import works on webpages presenting archived artifacts.

Zotero also has the capability of detecting through the HTML metadata that an unAPI server is running (Chudnov et al. 2006; Chudnov & England 2008). The unAPI protocol[29] is not well known outside of specialist circles but is implemented in a variety of open-source library catalog applications. In this context it allows software developers to identify a location where machines can read the metadata, while not directly embedding it in the HTML pages.

2.3 File based import

The final way to import metadata to Zotero is from a variety of metadata file types. This includes most of the popular types like EndNote XML, RIS, BibTeX, and MODS (*Metadata Object Description Schema*). A full list, along with a picture tutorial, is available in the Zotero documentation.[30] It should also be noted that Zotero can be extended to import more file types in two ways: the first is via a plug-in framework, and the second is through a JavaScript file called a translator. Either of these methods could be used to increase the import options to Zotero, but I limit this discussion to capabilities that are already present in Zotero.

[27] Google Scholar Tags: https://scholar.google.com/intl/en/scholar/inclusion.html#indexing

[28] When crafting academic websites, my go-to resources have included: Verrelli (2018), Johnston (2010), and the HTML5 WHATWG Wiki (2020). The ongoing issue is that HTML5 and XHTML approach embedding differently from a standards point of view. Description via Schema.org JSON looks to be a promising replacement technology for embedded metadata challenges, and works for general search like Google, Bing, and Yandex, but is unproven in specialized search like *Google Scholar*, or for capture by Zotero. Fenner et al. (2019) presents some ideas on Schema.org implementation in scholarly contexts. Note, however, that Schema.org is a broad scoped metadata schema much like Dublin Core. There are contexts in which the schema does not align cleanly with academic and preservationist needs.

[29] https://web.archive.org/web/20140331060734/http://unapi.info

[30] https://www.zotero.org/support/adding_items_to_zotero

CHAPTER 3

Using bibliographic data

In the context of the linking cycle described in chapter 1 and illustrated by Figure 1, this chapter is focused on getting data out of Zotero in useful ways. This is illustrated in Figure 1 by point Ⓓ. Zotero has two main export methods which are relevant to this part of the cycle. That is, there are two main ways to get bibliographic metadata out of Zotero. The first method uses a metadata file which can be read by other applications. The second method creates a reference or citation which is realized inside of an author's document. To implement the second method, Zotero uses the Citation Style Language (CSL)[31] to encapsulate a publisher's style sheet requirements and make them operational in an author's document.

3.1 Export files

Much like how Zotero imports a variety of metadata file types from other bibliographic software, it can natively export to a variety of metadata file types. The most complete (as in lossless data) is the Zotero RDF file which is in an XML format. Plug-ins can increase the export options.

3.2 Citation Style Language based export

When authors use a Zotero database to cite and reference the evidentiary record during their writing process, they depend on CSL descriptions as the primary means to properly format citations and references. Most authors do not think of Citation Style Language as a Zotero "export" because the data is inserted directly into their documents in the reference and citation formats they desire. However, of the ways to extract data from Zotero,

[31] https://citationstyles.org

exporting references and citations in a desired CSL format directly into an authored document is likely the most important to an average user. Furthermore it is the export which completes the linking cycle illustrated in Figure 1.

Citation Style Language is a meta-language for encoding the series of typographical choices necessary for crafting citations and references, e.g., should page numbers include all digits or only different digits (102–105 vs. 102–5) or the sorting of references by title (some styles exclude articles while others include them). The specific details of the style sheet's preferences are encoded in XML. The technical specification for the Citation Style Language is independent from Zotero, but many of the Zotero developers are also contributors to the CSL project.[32] CSL is relevant because some of Zotero's limitations are actually limitations in CSL. For example, the only contributor roles in Zotero are those which are specified in CSL.[33] As a technology, it has wide support across several citation managers, and therefore also has significant leverage in a very niche market. Both BibTeX (Patashnik 1988, 1998, 2003) and BibLaTeX (Lehman, Kime & Wemheuer 2019) style files (`.bst`), which work within the context of the TeX typesetting system, are comparable to CSL in the sense that they encode ways of taking bibliographic data from a database and formatting it into a reference or citation. However, CSL styles have a much broader range of influence, interacting with authorship tools including: *Microsoft Word*, *Apple Pages*, *LibreOffice*, *Open Office*, *TeX*, *Pandoc*, *R Markdown*, *GoogleDocs* and a variety of web content management systems like *Drupal* and *WordPress*.

As of March 2021 there are more than 125 different style sheets in the CSL repository that are relevant to the field of linguistics and neighboring disciplines.[34] Style sheets and journals are not in a one-to-one correspondence. For example, several journals may use the same style sheet, or a single style sheet may have several sub-varieties like the one

[32] CSL 1.0.1 is used in Zotero at the time of writing. In July 2020 the specification was advanced to 1.0.2. The version increment addressed many of the limitations encountered in 1.0.1. The 1.0.2 specification can be found at: https://github.com/citation-style-language/documentation/blob/master/specification.rst. An overview of changes made available by the CSL team is available as a GoogleDoc: https://docs.google.com/document/d/1wY1cOOamDYYh8VNW7h_uleqieBDGOa_LYsRiVdQy1RI/edit#

[33] This is important for those who advocate the use of the *The Tromsø Recommendations for Citation of Research Data in Linguistics* because it allows for all datacite roles, OLAC roles, and CRediT roles to be used in the contributor slot.

[34] Including subjects such as: General Linguistics, Phonetics, Phonology, Syntax, Language Pedagogy, Language Policy, Writing Systems Researcher, Ethno- Studies such as Ethnomusicology, and Ethnic Studies.

for the Modern Language Association. Within these 125 styles, some styles are used to support publishing in specific languages including: French, German, English, Spanish, Italian, and Polish. Some of the reference styles which are included in the CSL repository used by Zotero are listed in Table 2.

Table 2. Sample of style sheets with a CSL file in the Citation Style Language Repository

Style sheet	**Publishing House**
Generic Style Rules for Linguistics	Language Science Press
Unified style sheet for linguistics	LSA
Unified Stylesheet for Linguistics (de Gruyter Literature)	De Gruyter
Language	LSA
American Psychological Association (APA) 6th & 7th editions	APA
Chicago 15th, 16th, 17th editions	Chicago University Press
Transactions on Asian Language Information Processing	ACM
Transactions on Speech and Language Processing	ACM
Association for Computational Linguistics - Conference Proceedings	ACM
Lexicography	Springer
Glossa: a journal of general linguistics	Ubiquity Press
International Journal of Lexicography	Oxford University Press
Lingua	Elsevier
Lingua Sinica	Sciendo (De Gruyter)
Linguistics and Education	Elsevier
Multilingual Education	Springer
Natural Language & Linguistic Theory	Springer
Natural Language Semantics	Springer
Quarterly Journal of Speech	Taylor & Francis
Revista de Filología Española	Consejo Superior de Investigaciones Científicas
Russian Linguistics	Springer
Writing Systems Research	Taylor & Francis
Zeitschrift für deutsche Philologie	Erich Schmidt Verlag

3.3 Crafting references

Crafting a reference is a multifaceted act. References set the context for an argument and reflect the truth basis an author holds. References also define the social connections

authors consider important and the relationships authors wish to acknowledge.[35] While social connection and epistemological connection serve as two distinct functions of a reference, the reference itself still needs to be be clear and informative to readers, regardless of its function. The *Publication manual of the American Psychological Association* says it this way:

> As with any reference, the purpose is to direct the reader to the source, despite the fact that only a single copy of the document may be available and the reader may have some difficulty actually seeing a copy. (VandenBos 2010:212)

Therefore, references should be crafted first and foremost with the human reader in mind, describing not only the artifact being referenced, but also where to find it. Some redundancy may be required so that artifacts can be found both manually and digitally. In opposition to informative references, many publication's style sheets opt for short references, valuing the economy of the page over clarity in communication. This perspective had merit in the pre-digital era and certainly still has merit when considering physical media. The social nature of a reference may also stand in contrast to the information nature of a reference, since it reflects the way the authors craft their social identity via their argumentation.

With the rise of bibliometrics, the study of how many authors reference another given work, the social networking of scholarly discourse is more visible. Bibliometrics can be used to assess the impact of a particular scholarly work by counting only authors who are in agreement with the thesis of that work. However, more often, bibliometrics are generalized to count total mentions from works supporting and countering the position. These broader impact statistics are then used in the career advancement of the scholar, e.g., tenure. This relevance is certainly felt in the academic linguistics community. Haspelmath (2014) and Berez-Kroeker et al. (2018) both argue for changes in referencing practices in order to favorably impact (for scholars) the metrics by which scholars are evaluated. Haspelmath (2014) argues for a simplified presentation structure in the reference.

[35] Both Peroni & Shotton (2012) and Hoffmann et al (2016) provide multiple technical and moral reasons authors choose to cite and reference literature.

> The strongest justification for simple rules is that the references should be automatically parsable (e.g. by Google Scholar), and correct and complete author names should be extractable. In the modern age, this is crucial for scientometric and hence career-building purposes. (Haspelmath 2014:footnote 16)

Berez-Kroeker et al. (2018) argues that authors should clearly and overtly cite and reference the data they use.

> Unfortunately most linguists do not know how to go about advocating that "data work" be given the same kind of attribution as "analysis work" in hiring, tenure and promotion cases. (Berez-Kroeker et al. 2018:11)

More subtle is the issue of who should be privileged with "attributions" and which "works" a publisher should allow to appear in a references section. These more subtle differences are rarely articulated overtly in a publisher's style sheet,[36] rather they are manifested by the examples they do or do not provide for authors to consult. Those who craft CSL files rely on publisher-provided examples to create complete representations of what publishers will accept. One can not simply add a missing item type (such as an audio recording, or a video recording) to a publisher's style sheet by adding these items to a CSL file.[37] For example in the *The Generic Style Rules For Linguistics*, Haspelmath (2014:10) acknowledges six item types (journal article, book, article in edited book, thesis, published conference papers, unpublished materials) which could theoretically be used in a references section,[38] but suggests that only four are dependable options to be included

[36] In the course of reviewing over 40 journal style sheets, I only encountered one, *Natural Language and Linguistic Theory*, which overtly stated that unpublished materials should only be mentioned in the text body, and not included in the references section.

[37] This pressure has been felt broadly across the sciences with some communities rallying around a new type of article called the **data paper**—a paper dedicated to the description of a dataset and published in a journal series so that dataset users can reference the data paper when they encounter publishing constraints. Something similar is also found in some communities who create software. People often publish an introductory paper presenting software. The idea is that future academic users of these resource types would then reference the associated paper. This strategy has some inherent problems in that both software and data sets can have an evolutionary nature, whereas papers when published are generally static.

[38] Haspelmath (2014:8) seems to assume that authored works should have two sections. A references section, and then a sources section, perhaps with extensive usage of footnotes. It states: "When the source is not a bibliographical reference, but is the name of a text or corpus (perhaps unpublished), as in (10), the source is given in square brackets and the article must contain a special section at the end where more information about the sources is given. (When the source indication is unique and quite long, it may of course alternatively be given in a footnote, e.g. when it is a long URL.)"

in the citation of academic works (specifically in the domain of linguistics). The two disprefered item types are published conference papers and unpublished materials:

> Other kinds of publications should be treated like one of these to the extent that this is possible. For example, published conference papers can be treated like articles in edited volumes or like journal articles. Unpublished papers can be treated like journal articles, with information about the location given as a nonstandard part.
>
> In unpublished conference papers, the conference is treated as a nonstandard part in parentheses (but such unpublished papers should only be cited from recent conferences, if it can be expected that the material will eventually be published)... (Haspelmath 2014:11)

The Generic Style Rules For Linguistics with its four acknowledged item types is very influential within linguistic publishing as evidenced by the list of publishers in Table 3 who incorporate it into their style sheet.[39]

[39] There are nearly 1700 linguistic journals among the approximately 40,000 journals and serials tracked by MIAR (Rodríguez-Gairín et al. 2011). http://miar.ub.edu/lista/CAMPO/--TElOR8Ocw41TVElDQQ==. Needless to say, neither the list in Table 3 or Table 4 is exhaustive.

Table 3. Publishing ventures using or incorporating *The Generic Style Rules For Linguistics*

Publishing Venture	**Type**
ELPublishing	Press
Language Science Press[40]	Press
Australian Journal of Linguistics	Journal
Cahiers de Linguistique Asie Orientale	Journal
Cuadernos de Lingüística de El Colegio de México	Journal
Die Welt der Slaven	Journal
Finnisch-Ugrische Forschungen	Journal
Glossa: a journal of general linguistics	Journal
International Journal of Eurasian Linguistics	Journal
Journal of the Southeast Asian Linguistics Society	Journal
LANGUAGE AND LINGUISTICS	Journal
Language Documentation & Conservation	Journal
Language in Africa	Journal
Linguistica Atlantica	Journal
d'*Onoma*	Journal
Mandenkan	Journal
Minpaku Sign Language Studies	Journal
Stellenbosch Papers in Linguistics Plus	Journal
SKY Journal of Linguistics	Journal
CANIL Style Guide: A Guide for Formatting Term Papers to Be Submitted in Linguistic Courses Offered at the Canada Institute of Linguistics	Educational Program
Department of Linguistics at the University of Konstanz: Guidelines for writing an academic paper	Educational Program
Department of General Linguistics University of Bamberg: Style Guide for term papers and final theses in linguistics (v1.4)	Educational Program
Linguistics Program at the University of North Dakota, Grand Forks	Educational Program

A second influential style sheet in linguistic publishing is the *Unified style sheet for linguistics* (Salmons 2007). A list of known users of this style sheet is listed in Table 4.

[40] The Language Science Press guidelines for authors says that they follow *The Generic Style Rules For Linguistics* but their Microsoft Word Template file says to format references following the *Unified style sheet for linguistics*.

Table 4. Publishing ventures using the *Unified style sheet for linguistics*.

Publishing Venture	**Type**
De Gruyter Mouton	Press
Italian Journal of Linguistics	Journal
Journal of African Languages and Linguistics	Journal
Language and Linguistics in Melanesia	Journal
Linguistica Atlantica	Journal
Linguistics of the Tibeto-Burman Area	Journal
Journal of English Linguistics	Journal
Journal of Linguistics	Journal
Journal of Linguistic Geography	Journal
Semantics and Pragmatics	Journal
Studia Neophilologica	Journal
Studies in Language	Journal
North American Conference on Chinese Linguistics	Conference
Catholic University of Eichstätt-Ingolstadt: English Linguistics	Educational Program

The *Unified style sheet for linguistics* also does not contain a formulation for an unpublished work in its examples. No guidance is provided for the referencing of audio materials, video materials, or archival materials which are part of a collection.

When publishing style sheets such as *Unified style sheet for linguistics* or *The Generic Style Rules For Linguistics*[41] lack instructions for citing certain types of items in the evidentiary record, authors are less likely to cite and reference the evidentiary record. I interpret this as a contributing factor to the apparent absence of referencing of primary sources in linguistic works as discussed by Gawne et al. (2017), Gawne, Berez-Kroeker, Andreassen & Okura (2017), Berez-Kroeker et al. (2017), and Gawne & Berez-Kroeker (2018) where the authors show that grammar authors and journal article authors rarely reference the archival copies of the evidence they present.

In addition to the Zotero import challenges that are the focus of this thesis, the publishing requirements and the availability of CSL files for processing bibliographic metadata have an impact on the extent to which archived materials are referenced; and therefore,

[41] *The Generic Style Rules For Linguistics* builds upon the *Unified style sheet for linguistics* and therefore is typographically similar in many respects. Due to their typographical differences I treat them as separate style sheets.

publishing requirements represent a separate challenge in the completion of the linking cycle between artifacts, authors, and publications. The availability of reference formats for collections, and components of collections could be added to CSL files, which implement the style guides for academic publishing venues, if publishers would put these types of references in their style guides. I acknowledge the capability of the CSL technology as an influential component in the relationship between authors and published literature. I have indicated this relationship with point Ⓓ in Figure 1. However, the implementation of reference patterns remains a sociological and philosophical issue stemming from how editors wish to have authors articulate the evidentiary record.

CHAPTER 4

Archives

The following archives are included in the analysis presented in this chapter:

- The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC),[42]
- The Pangloss Collection of the Collection de Corpus Oraux Numériques (Pangloss),[43]
- SIL Language & Culture Archives (L&CA),[44]
- Endangered Languages Archive (ELAR),[45] and
- Kaipuleohone.[46]

There was no specific reason to look at these archives to the exclusion of other archives and special collections, other than I am personally familiar with them in my linguistic research. Together they represent a variety of perspectives on the discipline of archiving. One presents work conducted within a national research lab; one is run by an international NGO and presents work by its staff; and the others are associated with universities presenting a variety of evidence collected or created by scholars and associates. One is centered in France, one in England, one in Australia, and two in the United States. Each has different primary sets of contributors, different hierarchical organization, and different aims or aspirations for user engagement with their holdings. Each of them also has chosen a different technology infrastructure to manage interactions with people interested in their holdings. So, in fact, they represent a fair sampling of the diversity which exists across the "industry" of language artifact archiving.

[42] https://www.paradisec.org.au
[43] https://pangloss.cnrs.fr
[44] https://www.sil.org/resources/language-culture-archives
[45] https://elar.soas.ac.uk
[46] https://scholarspace.manoa.hawaii.edu/handle/10125/4250

4.1 Archive organization & record management

While there is quite a bit of diversity with regards to the sample of language archives, discussed in this thesis there are also some commonalities across the five archives:

1. They all have strong institutional support (rather than being collections managed at a department or individual level).
2. They are all members of DELAMAN,[47] an association of language archives.
3. They all, with the exception of some records in SIL's Language & Culture Archives, have eschewed best practice for archival collection description (which includes preserving hierarchical structures in which artifact creators organized their creations) in favor of a flatter three tiered structure.
4. They also all participate in the Open Language Archive Community (OLAC)[48] by using or reducing some of their collection holding records to Dublin Core plus OLAC extensions and sending metadata from their holdings to the OLAC catalogue aggregator via an OAI feed.

Commonalities (1) and (2) are fairly straightforward. However, as (3) and (4) deal with archival collection curation in general, I introduce these and related concepts in the rest of Section 4.1 and further explore the implications in Chapter 6.

4.1.1 Tiered structure of archives

I define **archives** as institutional organizations which facilitate the preservation and access to artifacts. Archives do this by organizing artifacts into **collections**. Libraries sometimes also steward or choose to manage sets of artifacts as "special collections" which is an analogous concept to an archives' concept of collections. Collections are also organized internally by curators and catalogers working within an archive or library.

[47] https://www.delaman.org
[48] http://www.language-archives.org

The Society of American Archivists (2013) lays out their standard for the description of collections in *Describing archives: a content standard* (hereafter DACS). While every collection is unique, there are general guiding principles for organization and description of collections. Principle number two from DACS, **Respect des Fonds**, says that items should be kept "in their original order". This is further explained as:

> Inherent in the overarching principle of *respect des fonds* are two sub-principles —provenance and original order. The principle of provenance means that the records that were created, assembled, accumulated, and/or maintained by an organization or individual must be represented together, distinguishable from the records of any other organization or individual. The principle of original order means that the order of the records that was established by the creator should be maintained by physical and/or intellectual means whenever possible to preserve existing relationships between the documents and the evidential value inherent in their order. Together, these principles form the basis of archival arrangement and description. (Society of American Archivists 2013:xvi)

To conduct preservation activities and to facilitate access, archivists must negotiate and reconcile the concepts of collection **arrangement** and **description**.

> *Arrangement* is the intellectual and/or physical processes of organizing documents in accordance with accepted archival principles, as well as the results of these processes. *Description* is the creation of an accurate representation of the archival material by the process of capturing, collating, analyzing, and organizing information that serves to identify archival material and to explain the context and records systems that produced it, as well as the results of these processes. (Society of American Archivists 2013:xvi)

Arrangement and description are built into the third and fourth DACS principles: *Arrangement involves the identification of groupings within the material* and *Description reflects arrangement*. These particular DACS principles are relevant when referencing archival material. If references are to be informative (to where one can access an artifact) and descriptive (to the nature of the artifact), then the reference needs to also contain context about the arrangement and the tiers of the archival collection. For example, a reference to an audio artifact is more informative if it contains the larger structure of which it is a part (e.g.,

group of recordings from the same session) as well as its collection (e.g., entire fieldwork trip).

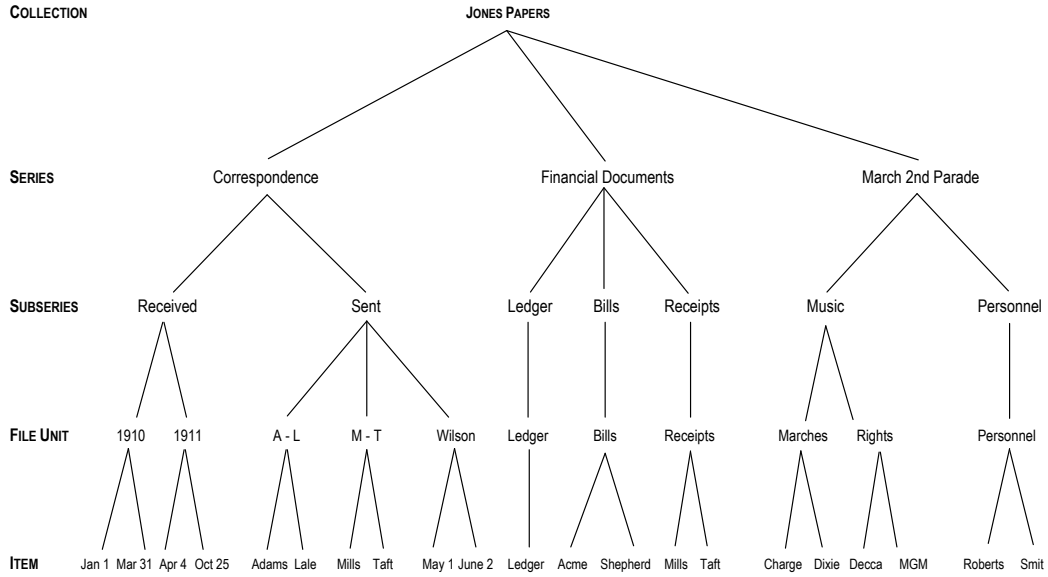

Figure 4. Hodges & McClurkin (2011:3)'s typical archival structures

While the number of tiers a collection should contain is not prescribed by DACS, Hodges & McClurkin (2011), in the *Archives and Manuscripts Processing Manual* from the University of Texas at Arlington's Special Collections Library, demonstrate a multi-tier structure which is a widely held norm across archives. This multi-tier structure, shown in Figure 4, is comprised of levels with the titles: **Collection level**, **Series level**, **Sub-series level**, **File unit level**, **Item level**. Such a system allows for the arrangement and description of series of aggregate works and their components. Aggregate works are a type of collection, which by their provenance (i.e., created together or edited together) in some contexts should be viewed as a whole unit, but in other contexts may be viewed in its component parts. For example, one could reference a whole book using Chicago 17th edition:

> Shopen, Timothy, ed. 2007. Language Typology and Syntactic Description. 2nd ed. Vol. 1. 3 vols. Grammatical Categories and the Lexicon. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511619427.

Or one could reference only a section of that book:

> Dryer, Matthew S. 2007. "Clause Types." In Language Typology and Syntactic Description, edited by Timothy Shopen, 2nd ed., 1:224–75. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511619427.004.

In the same way a scholar may wish to reference an individual archived `.wav` file or the aggregate work that includes a particular `.wav` file along with its transcription or other related recordings.

Linguists frequently use, create, and reference aggregate works. A corpus is a type of aggregate work composed of a balance of materials. An edited volume is another type of aggregate work. Within audio publication, most playlists, CDs, and phonograph records (except the single), e.g., LP 33 ½, are all types of aggregate works. Linguistic fieldwork often creates aggregate works in audio, video, text, and mixed media. For instance, when a recording session runs long the tape may need to be flipped, or the digital audio recorder might need to have its card changed or might start saving the recording session to a second file due to technical limitations. Additionally, a linguist may be running a secondary recording activity such as video or may be taking handwritten notes in their notebook. The audio, video, and text in the notebook together could be considered an aggregate work.

In the context of field linguistics then it is rather easy to consider that a linguist may have a larger effort which correlates to the **Collection level** in an archive, e.g., the output of a grant sponsored field trip lasting several months. Continuing with the hypothetical field trip, many linguists will work on several sub-projects. Projects may center around speech variety (e.g., variation within a language), the speech of a specific place (e.g., multilingualism in a town), a specific speech genre (e.g., talk between parents and baby) or social activity (e.g., harvesting), discourse type such as dialogue, or generic syntax and vocabulary elicitation (e.g., using lists and elicitation tools). It is easy to see that some of these projects might be well-suited for **Series level** descriptions. Within any one of these series, several recording sessions might be undertaken and be aptly suited for description

at the **File unit level**. Every file in a **File unit** would be then considered an **Item**. In citation and referencing a scholar will want to be able to reference each node of the tree from a **collection** down to an **item**.

4.1.2 Bibliographic records and archives

If an archive is to keep track of the items in a collection, it will need to create a **bibliographic record**. This is different from a **bibliographic reference** which is used in the sense of citing and referencing within a document. There is content overlap between a bibliographic record and a bibliographic reference in that the bibliographic record should inform the bibliographic reference. However, a bibliographic record will be the authoritative resource for metadata and usually contain more information than what is needed to create a clear link from a document to an artifact.

Bibliographic records need to be consistent across an archive in order for archives to efficiently manage content and service to archive users. A conceptual model for bibliographic records supports archive management by articulating various management concerns such as rights management, user engagement requirements, and intra-artifact content relationships. *Functional Requirements for Bibliographic Records* (FRBR) (Byrum et al. 2009 [1998], see also Carlyle 2006)[49] is one such widely embraced model in the library sciences. The basic FRBR model is shown in Figure 5. The distinctions in the model guide curators and catalogers in their description of an artifact/item. It informs the creation of various kinds of bibliographic records and the application of metadata to those various kinds of records.[50] FRBR has five important entities: **Endeavour**, **Work**, **Expression**, **Manifestation**, and **Item**. The simple idea is that an Endeavour will produce a work which will have a distinct (but not necessarily unique) intellectual or artistic creation, which is

[49] The FRBR has undergone extensive implementation review in the library sciences and is currently discussed in its evolved state as part of the International Federation of Library Associations and Institutions' (IFLA) *Library Reference Model* (LRM) see Riva, Le Bœuf & Žumer (2017). The LRM contains more associated library management models. In this thesis I limit my discussion to FRBR so I reference FRBR.

[50] For example, copyright claims might be at the expression level or the manifestation level. Someone can take a book from the Public Domain and re-typeset the book creating a PDF for circulation or publication. That person would have copyright to the manifestation they created, but the content would still be in the Public Domain. FRBR allows the scope of the rights claims to be articulated in the metadata records by specifying if the claim is to the expression or to the manifestation.

realized through an expression. Expressions in turn are embodied by a manifestation, which in turn are exemplified by an item.

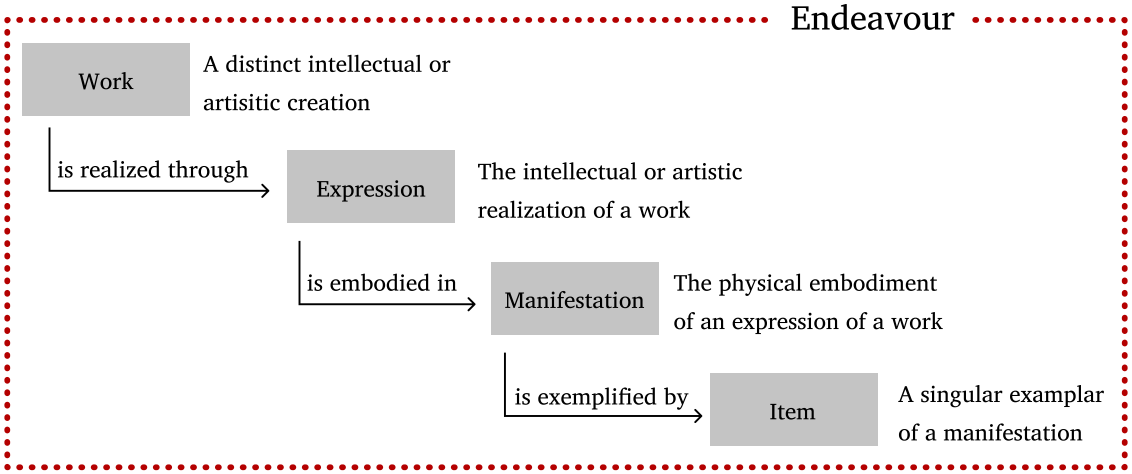

Figure 5. Basic FRBR model

To make a practical application for linguists, an example of an item is a book on a library shelf. Several libraries might have this same book. Each bound book is a unique item. A book such as *Language Typology and Syntactic Description* (Shopen 2007) has multiple editions. Each edition is a different expression. An abridged or translated version of *Language Typology and Syntactic Description* would also be a different expression. A book might have different publication formats such as a PDF version and a print version. These are different manifestations. For a more detailed analysis of how variations of works map to expression and manifestation, see Tillett (2001, 2004, 2009). To the best of my knowledge no attempt has been made to use FRBR with linguistic resources. However, several music catalogues have been recast in FRBR models (Le Boeuf 2005; Vellucci 2007; Riley 2008; Iglesias et al. 2009; Holden 2019). The similarity of this work to language documentation lies in the fact that language documentation materials often have aural and transcribed components, much like western music. Work by Yee (1993; 2007) has also looked at FRBR in the context of video materials, another media type commonly used by language documentation practitioners.

For an application of the FRBR model to audio artifacts, linguists could consider the workflow published in the BOLD method as presented in Reiman (2010). The first recording would be a `.wav` file which is its own work, expression, manifestation, and item. The second recording (i.e., the slow speech) would be a second expression of the same work, but the manifestation of the first recording is interlaced into the manifestation of the second recording. The third recording, which is a translation by another person, is also a new expression. The produced `.wav` file is interlaced into the aggregate of the first two recordings. So in the end one would have an audio recording which was an aggregate work of three different expressions of the same FRBR work, all within the same manifestation. This creates an interesting type of aggregate work, but it is certainly possible for the FRBR model to describe (Shadle 2006; O'Neill, Žumer & Mixter 2015; Coyle 2016:130–135; discuss the complexities of aggregate works).

4.1.3 Metadata standards and archives

Each archive discussed in this thesis participates in the Open Language Archive Community by sending metadata from their holdings to the OLAC aggregator. OLAC metadata is built on two technologies: *Open Archive Initiative Protocol for Metadata Harvesting* (OAI-PMH),[51] a transmission protocol, and *Dublin Core*,[52] a meta-schema for metadata used to increase interoperability between information systems used by libraries and institutional repositories. As a meta-schema, Dublin Core creates general categories. In contrast other schema, such as MARC[53] and institution specific schema, are used by catalogers and curators to provide specific metadata values from controlled sets of options. Dublin Core as a system of description consists of 15 properties known as **elements** and 55 properties known as **terms**. Elements and terms can take a variety of values. Sometimes the values are links or pointers to other records, sometimes the values are free-text, and sometimes values may be elements of a controlled vocabulary.

Dublin Core recognizes nine controlled vocabularies. Of those nine, the framers of Dublin Core include only the elements of one of them within the standard; the rest they

[51] https://www.openarchives.org
[52] https://www.dublincore.org/specifications/dublin-core
[53] https://www.loc.gov/marc

reference as composed by others apart from the standard itself. This speaks to its importance and weight in fully descriptive records. The vocabulary they included is the **DCMIType** vocabulary.[54] It is used to declare a broad type (type-nature) for the resource and prescribes one of the following values: Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text.

Dublin Core usage guidance suggests that an object should be described with the most appropriate value, e.g., if a record represents a collection of artifacts, then it should be described as a **Collection** rather than both **Sound** and **Text**. The guiding principles of Dublin Core also suggest a one-to-one correspondence for item description.[55] This means among other things that three records would be produced, one for the collection, one for the sound artifact, and a third for the text artifact.

The engineering behind Dublin Core makes it favorable for broad search and discovery tasks. These standards were never intended to contain all the elements needed for complete description of language resources. The OLAC metadata standard inherits these biases.

4.1.4 Referencing archival materials

The OLAC metadata standard has been seen by many in the language documentation community as an essential metadata set for language-based resources (Bird and Simons 2001). Participating in the OLAC metadata exchange via the OLAC aggregator by publishing an OAI feed is seen as a best practice. However, if the OLAC metadata standard is viewed as a minimum set of descriptive elements for resources, the risk is that the describer of those resources (usually an institution) will not have enough metadata to do preservation tasks or to support artifact discovery within the archive's user community. One activity of an archive's user community is the referencing of artifacts contained in archives!

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

[55] The One-To-One principle is discussed in the Dublin Core guidelines in section 1.2. https://www.dublincore.org/specifications/dublin-core/usageguide/#whatis

In order to support these activities, archives need to consider the types of items that publishing style sheets acknowledge. These style sheets guide scholars when they reference artifacts and collections. Keeping track of these various item types on their own is a gargantuan task; however, much of the hard work to identify item types has already been done by the CSL development team and has been visually implemented in Zotero (see discussion in Section 3.2).

Table 5 presents the item types currently available via the Zotero user interface, along with the additional item types available via CSL. The additional CSL item types can currently be used in Zotero via the **Extra** field in the user interface.[56] So, even when these values do not appear in Zotero drop down lists, CSL style sheets can use these additional CSL variables to craft a reference in a document.

Table 5. Zotero item types with additional types available via CSL

Zotero		**CSL**
Artwork	Letter	Collection
Audio Recording	Magazine Article	Dataset
Bill	Manuscript	Figure
Blog Post	Map	Musical_score
Book	Newspaper Article	Pamphlet
Case	Patent	Review-book
Conference Paper	Podcast	Treaty
Dictionary Entry	Presentation	
Document	Radio Broadcast	
E-Mail	Report	
Encyclopedia Article	Software	
Film	Statute	
Forum Post	Thesis	
Hearing	TV Broadcast	
Instant Message	Video	
Interview	Web Page	
Journal Article		

In addition to item types fields which do not appear in Zotero's user interface other CSL variables can also be added to the Extra field (see Figure 14 for example).[57] This is done with the pattern **CSL-variable:** ***text value***. However, it is up to the CSL style sheet

[56] See Zotero's user documentation for item types: https://www.zotero.org/support/kb/item_types_and_fields
[57] https://docs.citationstyles.org/en/stable/specification.html#appendix-iv-variables

and the editors who commission them to implement these variables. The first four fields in Table 6 are necessary for archival items in general, while the last two items are necessary for aggregate works in archives. The necessity of using the Extra field is dependent on the kind of material being referenced in the archive; some of these fields may already be present in the Zotero user interface.

Table 6. CSL values useful in crafting archival references

Field	Use
archive	archive storing the item
archive_location	storage location within an archive (e.g., a box and folder number)
archive-place	geographic location of the archive
collection-title	title of the collection holding the item (e.g., the series title for a book)
container-title	title of the container holding the item (e.g., the book title for a book chapter, the journal title for a journal article)
container-author	author of the container holding the item (e.g., the book author for a book chapter)

In the remainder of this chapter, I look at how bibliographic metadata is transferred between institution and end-user during an import to Zotero.[58] Table 7 presents an overview of the archive interfaces and their capabilities to transmit bibliographic metadata to Zotero as of January 2021.

Table 7. Summary of support for Zotero import methods across language archives

	DOI	Embedded metadata	File import
PARADISEC	Collection/Item	None / DOI detection	No
Pangloss (Cocoon)	Collection/Item	Some / DOI detection / unAPI	Yes
Pangloss (Villejuif)	Collection/Item	None / DOI detection	No
SIL L&CA	No	None	No
ELAR (VuFind)	No	Broken COinS	No
ELAR (WordPress)	No	None	No
Kaipuleohone	No	Some DC tags	No

[58] I follow Paterson (2015a:50) in acknowledging "The global levelling of information access through the Internet also enables speakers of endangered languages and academics to engage more fully with each other – rather than, as before, operating in different social circles. Roles such as 'linguist', 'language documenter' or 'endangered language speaker', which might previously have been mutually exclusive, can therefore now be fulfilled by 'academics' and 'native speakers' alike."

For each of the five archives I present a short introduction of the archive, list the technologies used when known, and the set of collections and their items I reviewed in my investigation. I then discuss the various levels of support for the transfer of bibliographic metadata. Discussed methods include import via DOI based APIs, import via embedded metadata in HTML pages, and any options for files which could be downloaded and imported into Zotero. I look at the content of the records imported into Zotero and discuss the totality of transferred metadata, I then look at the metadata captured by Zotero and discuss the record's sufficiency in crafting informative references like those demonstrated in the *APA 6th edition* and *Chicago 17th edition Author-date* style sheets. Where an archive has provided a suggested reference, I contrast that with what can be crafted with Zotero imported information.[59]

4.2 PARADISEC

The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[60] is an institutional archive which delivers services around the digitization of analog language artifacts and access to digital language artifacts. Established in 2003 under the leadership of Linda Barwick and Nick Thieberger, PARADISEC's founding context and evolution is described in their works (Barwick 2003, 2004, 2005; Thieberger 2009; Thieberger & Jacobson 2010; Barwick & Thieberger 2012; Barwick & Harris 2013). Based in Australia, it has been a significant part of the infrastructure supporting language communities in Australia, Papua New Guinea, and the greater Pacific region. The visionary leadership at PARADISEC has been open to accessioning collections from all over the world. This has in part created an operation which is esteemed by many in the scholarly fields of anthropology, (ethno)musicology, linguistics, language documentation, language development, and language revitalization.

Table 8 provides a summary of the import technologies currently available to Zotero users via PARADISEC's website.

[59] It is not uncommon for digital repositories and various online resources to provide a "suggested citation". For the sake of consistent terminology, I suggest that what they really provide is a "suggested reference". The utility of these formatted references is questionable, as formatting decisions ultimately are determined by the publisher of the work containing the formatted reference, not the entity wishing to be referenced.

[60] See footnote 42 in chapter 4.

Table 8. Summary of support for Zotero at PARADISEC

	DOI	Embedded metadata	File import
Collection	Yes	None / DOI detection	No
Item	Yes	None / DOI detection	No
Essence File	No	None - No unique web page	No

4.2.1 Technology infrastructure

For its digital collection management, PARADISEC uses *Nabu*[61] a self-built media content management system (CMS) written in Ruby on Rails.[62] Nabu is, as far as I know, unique among CMS platforms in that it is the only open source content management system that natively provides an OLAC feed.

4.2.2 Collections structure

PARADISEC has an advertised hierarchical arrangement system with the following nodes: *PARADISEC* > *Collection* > *Item* > *Essence Files*. Within a PARADISEC collection there may be many items, and then there may be many essence files per item. These files may include multiple manifestations, e.g., a `.wav` file and a `.mp3` file of the same recording session, where one is a derivative from the other. PARADISEC further lumps multiple manifestations of a single artifact with other creative works under the same item in their structure, e.g., recordings made in different villages on different days.

4.2.3 Collections and artifacts reviewed

In reviewing PARADISEC I looked at three collections: Roger Blench Collection 5 (RB5),[63] Zygmunt Frajzyngier Collection 1 (ZF1),[64] and Greg Anderson Collection 1 (GA1).[65] Although I mention elements from all three collections, for reasons of simplicity in presentation I include examples in this thesis from only RB5.

[61] https://github.com/nabu-catalog/nabu
[62] https://rubyonrails.org
[63] https://catalog.paradisec.org.au/collections/RB5
[64] https://catalog.paradisec.org.au/collections/ZF1
[65] https://catalog.paradisec.org.au/collections/GA1

Each of these collections have different quantities of descriptive metadata. In particular, many of the metadata fields in the PARADISEC public web-view for RB5 and ZF1 are empty—suggesting that they may be under-described collections. The GA1 collection is much more thoroughly described with nearly every field containing some content. As Chapman et al. (2009) discuss, this is not entirely uncommon across archives or institutional repositories because archival content may be imported through various processes, or archives may have several different curation workflows set up where different aspects of the record are enhanced through independent workflows (preservationists, curators, re-users, depositors, producers, etc.). In this thesis I have chosen to review the under-described collection RB5 because it is more informative for our discussion when compared with the collection described in the section on the SIL Language & Culture Archives. In my import experiment, the additional rich metadata provided in GA1 did not affect and were not part of the metadata transferred to Zotero in any way. That is, from Zotero's perspective both RB5 and GA1 collections were equally described related specifically to bibliographic metadata.

Figure 6 shows the top portion of the web-view of the RB5 collection with the list of 104 items in the collection. Some metadata for the collection level is presented on the left. Out of frame are the remaining languages in the collection metadata, an approximate geographical display of where those languages are spoken, the names of those with edit access to the collection, and the "cite as" box which is discussed in Section 4.2.7.

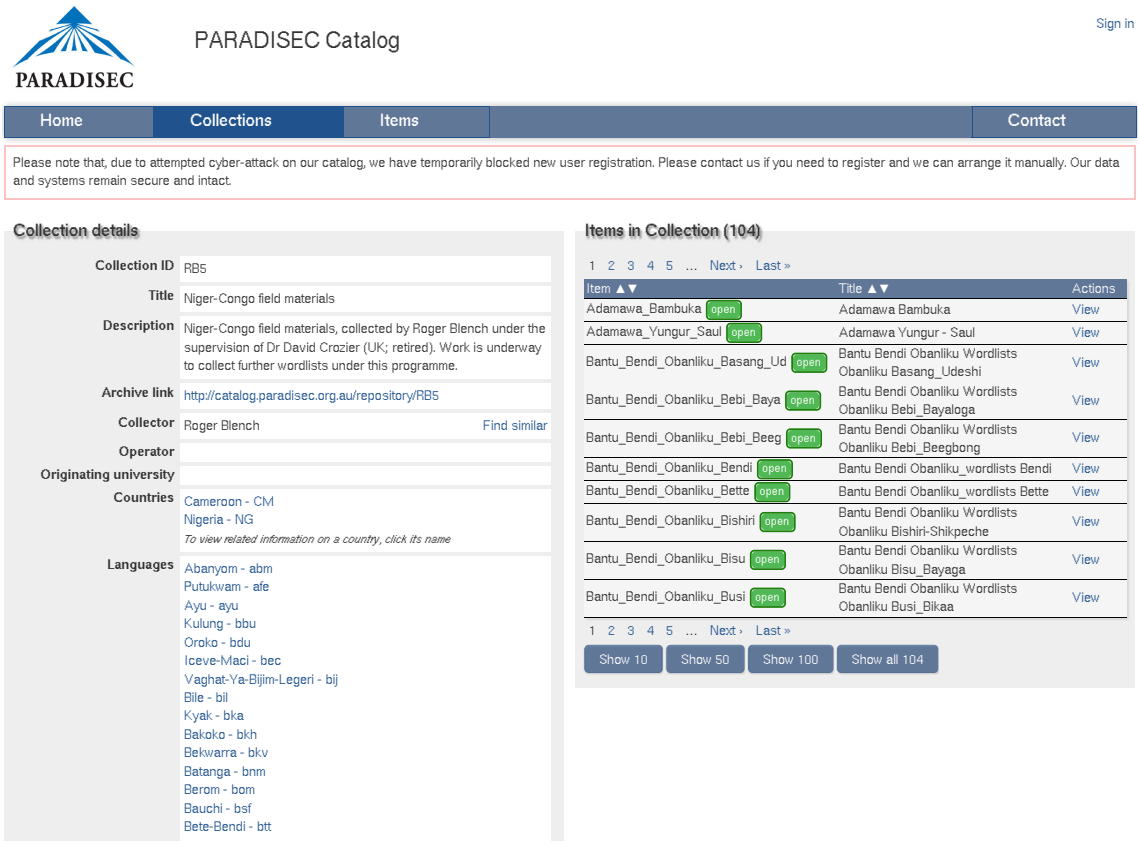

Figure 6. Top portion of the Roger Blench 5 collection display in PARADISEC

To review a single item node within a collection, I chose to look at the item titled *Kamuku wordlists*.[66] It is an item in the RB5 collection and has its own DOI separate from the DOI for the RB5 collection. The item contains 406 different audio artifacts (essence files). Table 9 demonstrates what is possible with unlimited disk space and a constrained hierarchical structure.

Table 9. Component parts: Collection vs. Item

Tier node	**Components**
RB5 Collection	104 Items
Kamuku wordlists Item	406 Essence Files

Figure 7 shows the full page view for the item reviewed (ID: RB5-Kainji_Kamuku_wordlists) while Figure 8 shows the top portion in more detail.

[66] https://catalog.paradisec.org.au/collections/RB5/items/Kainji_Kamuku_wordlists

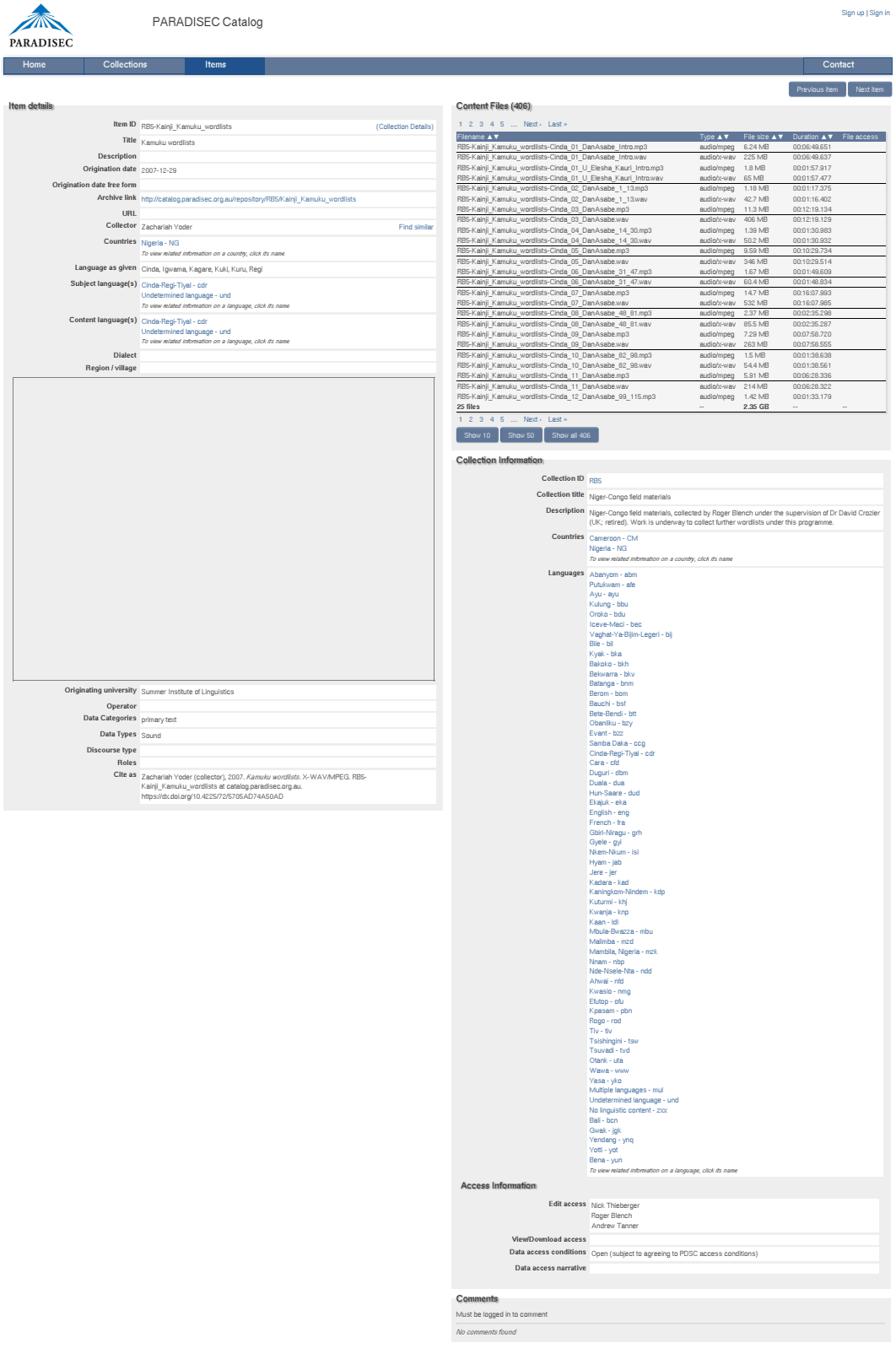

Figure 7. Full view of the Kamuku wordlist an item in the RB5 collection at PARADISEC

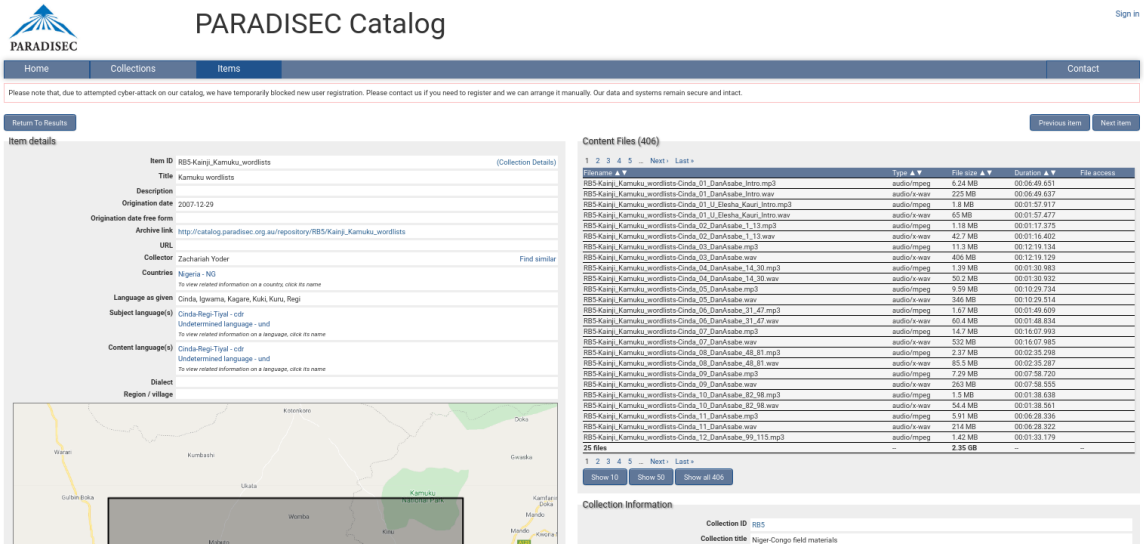

Figure 8. Top portion of the Kamuku wordlist, an item in the RB5 collection at PARADISEC

4.2.4 DOI import

Barwick & Thieberger (2018:137) say "Since its establishment, PARADISEC has always used persistent identifiers for objects in our collections, down to the file-level. In 2016 we added digital object identifiers (DOI) to all collections, items and files." However, essence files do not have their own web pages nor are their alleged DOIs visible via the Nabu web interface. Therefore, as far as end users are concerned, for the crafting of references they might as well not exist. In this section, I show the DOIs investigated (in Table 10), the metadata on file with DataCite, and what Zotero does with the data it receives.

Table 10. PARADISEC DOIs investigated

	DOI
RB5 Collection	10.4225/72/56E977B622032
Kamuku wordlists Item	10.4225/72/5705AD74A50AD
Essence Files	None

4.2.4.1 DOI import from Collection

Recall that Figure 6 shows the RB5 collection metadata that Nabu presents to users. Figure 9 shows the data sent by the DataCite API service to Zotero while Figure 10 shows what Zotero users receive as input when they use the DOI to get metadata from the API service.[67]

[67] All JSON data in this thesis was acquired by using the DataCite API with the command line application CURL. A command similar to the following was used: `Terminal-user:$ curl https://api.datacite.org/works/10.4225/72/56E977B622032` The JSON formatting was converted to standard JSON formatting for editing using: https://jsonformatter.curiousconcept.com. Syntax highlighting and conversion to PDF for inclusion as a figure was done using: https://nsspot.herokuapp.com/code2pdf.

```
{
  "data":{
    "id":"https://doi.org/10.4225/72/56e977b622032",
    "type":"works",
    "attributes":{
      "doi":"10.4225/72/56e977b622032",
      "identifier":"https://doi.org/10.4225/72/56e977b622032",
      "url":"http://catalog.paradisec.org.au/collections/RB5",
      "author":[
        {
          "given":"Roger",
          "family":"Blench"
        }
      ],
      "title":"[Blench Niger-Congo]",
      "container-title":"PARADISEC",
      "description":null,
      "resource-type-subtype":null,
      "data-center-id":"pdsc.repo",
      "member-id":"pdsc",
      "resource-type-id":null,
      "version":null,
      "license":null,
      "schema-version":"3",
      "results":[

      ],
      "related-identifiers":[

      ],
      "citation-count":0,
      "citations-over-time":[

      ],
      "view-count":0,
      "views-over-time":[

      ],
      "download-count":0,
      "downloads-over-time":[

      ],
      "published":"2014",
      "registered":"2016-03-16T15:11:54.000Z",
      "checked":null,
      "updated":"2020-06-11T07:11:07.000Z",
      "media":[

      ],
      "xml":"PD94bWwgdmVyc2lvbj0iMS4wIj8+CjxyZXNvdXJjZSB4bWxucz0iaHR0cDovL2RhdGFjaXRlLm9yZy9zY2hlbWEva2VybmVsL1
    },
    "relationships":{
      "data-center":{
        "data":{
          "id":"pdsc.repo",
          "type":"data-centers"
        }
      },
      "member":{
        "data":{
          "id":"pdsc",
          "type":"members"
        }
      },
      "resource-type":{
        "data":null
      }
    }
  }
}
```


Figure 9. DataCite API supplied JSON response for the RB5 Collection

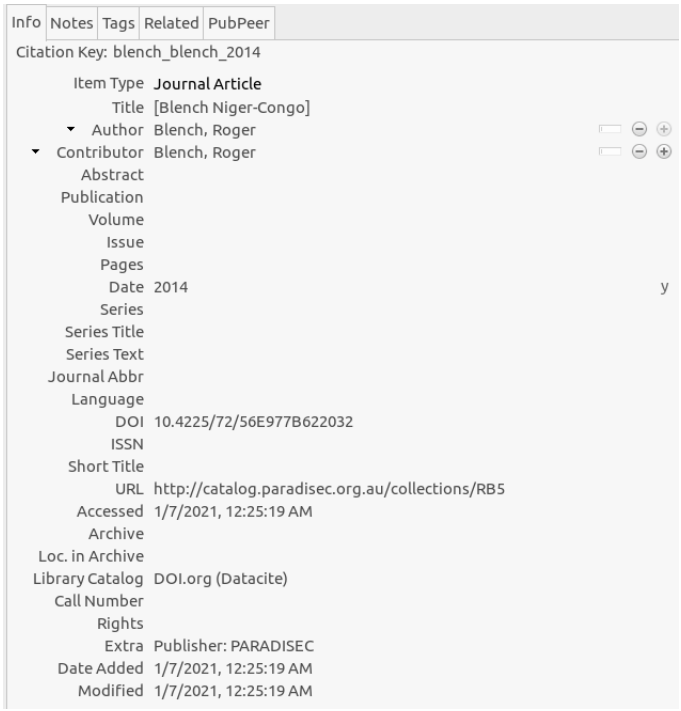

Figure 10. Zotero's interpretation of DataCite API supplied JSON for the RB5 collection

4.2.4.2 DOI import from Item

Recall that Figure 8 shows the item level metadata that Nabu presents to users. Figure 11 shows the data sent by the DataCite API service to Zotero while Figure 12 shows what the Zotero user sees as input when they use the DOI to receive metadata.

```
{
  "data":{
    "id":"https://doi.org/10.4225/72/5705ad74a50ad",
    "type":"works",
    "attributes":{
      "doi":"10.4225/72/5705ad74a50ad",
      "identifier":"https://doi.org/10.4225/72/5705ad74a50ad",
      "url":"http://catalog.paradisec.org.au/collections/RB5/items/Kainji_Kamuku_wordlists",
      "author":[
        {
          "given":"Zachariah",
          "family":"Yoder"
        }
      ],
      "title":"Kamuku wordlists",
      "container-title":"PARADISEC",
      "description":null,
      "resource-type-subtype":null,
      "data-center-id":"pdsc.repo",
      "member-id":"pdsc",
      "resource-type-id":null,
      "version":null,
      "license":null,
      "schema-version":"3",
      "results":[

      ],
      "related-identifiers":[

      ],
      "citation-count":0,
      "citations-over-time":[

      ],
      "view-count":0,
      "views-over-time":[

      ],
      "download-count":0,
      "downloads-over-time":[

      ],
      "published":"2007",
      "registered":"2016-04-07T00:44:41.000Z",
      "checked":null,
      "updated":"2020-06-11T07:17:54.000Z",
      "media":[

      ],
      "xml":"PD94bWwgdmVyc2lvbj0iMS4wIj8+CjxyZXNvdXJjZSB4bWxucz0iaHR0cDovL2RhdGFjaXRlLm9yZy9zY2hlbWEva2VybmVsL1
    },
    "relationships":{
      "data-center":{
        "data":{
          "id":"pdsc.repo",
          "type":"data-centers"
        }
      },
      "member":{
        "data":{
          "id":"pdsc",
          "type":"members"
        }
      },
      "resource-type":{
        "data":null
      }
    }
  }
}
```


Figure 11. DataCite API supplied JSON response for the Kamuku Wordlist

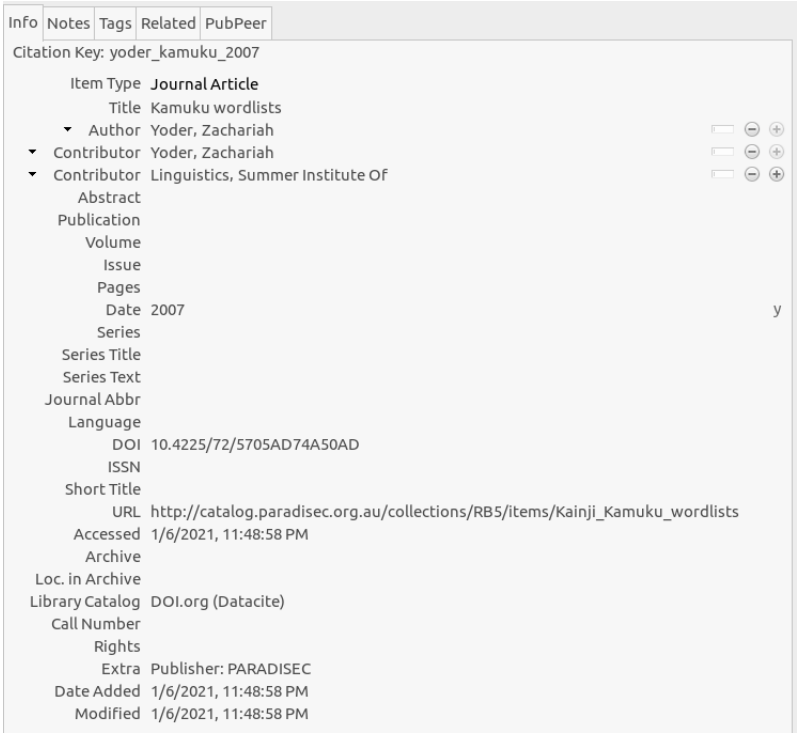

Figure 12. Zotero's interpretation of DataCite API supplied JSON for the Kamuku Wordlist

4.2.4.3 DOI Discussion

There are several observations worth mentioning about the use of DOIs at PARADISEC and the data brought to Zotero through the DataCite API.

The first issue is one of user expectation. Most DOI-users have experienced them in the contexts of referencing an article or a chapter in an edited volume. However, in the PARADISEC context by pointing to collections and items (which are also collections because they are filled with various kinds of artifacts), this situates the DOIs in a context where they point to a portion of a hierarchical structure, which is analogous to the entirety of an edited volume or the series in which the edited volume is published rather than to a chapter within the edited volume.

Second is the particular version of the DataCite metadata schema used by PARADISEC. The JSON data (as presented in Figures 9 and 11) suggests that PARADISEC first sent data to the DataCite database for these records in 2016. At that time the DataCite metadata

schema was at version 3.1 (DataCite 2014). The metadata schema has since evolved and is as of January 2021 at version 4.3 (DataCite 2019). Meanwhile the last time PARADISEC refreshed the data it sent to DataCite for these reviewed collections and artifacts was 2020. Even though it updated the metadata records recently, it still used schema version 3.1. So, something in the pipeline of data from PARADISEC to DataCite is not working in the best interest of DataCite API users. One important issue that is not addressed in PARADISEC's 2020 update to DataCite is the crosswalking of metadata from PARADISEC's application profile[68] to DataCite's metadata schema. That is, the DataCite metadata schema and all the services which come with it can not be useful to PARADISEC's users until PARADISEC undertakes the effort to map DataCite's schema to the existing metadata fields in PARADISEC's metadata schema.

A third issue which directly impacts Zotero users is PARADISEC's choice to not fill in any resource type as can be seen by the value "null" for line 66 of Figure 11. DataCite's resource types are matched to Zotero's item type vocabulary via a JavaScript translator.[69] This means that each DataCite import which is not assigned a value will be imported as a basic document (defaulting to journal article).[70] In the reviewed cases, each PARADISEC object with a DOI should be matched to the DataCite vocabulary **resourceTypeGeneral** with the term **Collection** (DataCite 2019:16). However, this was not the case.

Besides the issues of data interoperability mentioned above, there is a fourth issue of data continuity in the reviewed objects as listed in Table 11. The values of things like dates, titles (names of objects such as collections), and contributor roles are different between the data presented on Nabu and the data presented via DataCite. This can lead to mixed or errant references depending on how a user actually acquires the metadata. This is a less than ideal situation for measuring the impact and use of a collection.

[68] An application profile is the schema that an organization chooses to use. It may contain schema elements from a variety of schema while not implementing any schema completely. For discussion see Heery and Patel (2000). Some have considered OLAC to be an application profile (Hillmann & Phipps 2007).

[69] The exact Zotero translator which interprets the DataCite metadata and converts it to Zotero values can be read from the Zotero Github repository at: https://github.com/zotero/translators/blob/master/Datacite%20JSON.js.

[70] Importing the bibliographic metadata and matching it to the right item type in Zotero impacts how it will appear via CSL in a reference.

Table 11. Data Continuity: Nabu vs. DataCite

	Nabu	**DataCite**
Collection name	Niger-Congo field materials	[Blench Niger-Congo]
Date	2003	2014
Role	Collector	none

As indicated in Table 11, at least three metadata fields are suspect for accuracy.[71] **Collection name**, **date**, and **role** all have different values when one compares the DataCite record and the Nabu record.

In addition to the role variation between Nabu and DataCite,[72] there exists a role variation between Nabu and OLAC. On OLAC[73] Roger Blench is listed as **Compiler**.[74] OLAC defines this term as "The participant is responsible for collecting the sub-parts of the resource together. This refers to someone who creates a *single* resource with multiple parts, such as a book of short stories, or a person who produces a corpus of resources, which may be archived separately." However, on the basis of my personal communication with Roger Blench about this collection, he maintains that he acted more in-line with the OLAC role of **depositor**.[75] Depositor is defined by OLAC as someone who "was responsible for depositing the resource in an archive." **Collector** is a valid MARC relator role.[76] It is defined as someone who "brings together items from various sources that are then arranged, described, and cataloged as a collection. A collector is neither the creator of the material nor a person to whom manuscripts in the collection may have been addressed". OLAC documentation says that the OLAC standard accepts MARC relator roles.[77] However, to the best of my knowledge, these roles have not been added to the OLAC validator.[78] So, if one tries to use MARC roles in their OLAC feed, they are marked as errors. To complicate

[71] Collection ZF1 does not even have a collection title in DataCite.

[72] Zotero categorizes the null it receives from DataCite as **contributor**. Any role for which Zotero does not have an internal match are imported as contributors. Zotero does not export contributors, so they become an application internal way of applying name oriented metadata to a Zotero record.

[73] http://dla.library.upenn.edu/dla/olac/search.html?q=rb5&fq=country_facet%3A%22Nigeria%22

[74] http://www.language-archives.org/REC/role.html#compiler

[75] http://www.language-archives.org/REC/role.html#depositor

[76] https://www.loc.gov/marc/relators/relaterm.html

[77] http://www.language-archives.org/REC/role.html#Role

[78] The validation code for OLAC roles can be found here: http://www.language-archives.org/OLAC/1.1/olac-role.xsd

matters further, DataCite has its own role vocabulary (DataCite 2019:32–35) which is not in alignment with MARC relator roles or OLAC roles.

4.2.5 Embedded metadata in HTML for collection and item levels

No HTML embedded metadata is discovered by Zotero because none is presented in the HTML code (as is shown in Figure 13). Because there is no HTML metadata, Zotero picks up that there is a DOI on the page, and then it fetches the metadata provided via the DataCite API.[80] This results in a Zotero import identical to the DOI import discussed in Section 4.2.4.1 and shown in Figure 10 for the collection and Section 4.2.4.2 and shown in Figure 12 for the item.

```
<!DOCTYPE html>
<html>
<head>
<script type="text/javascript">window.NREUM||(NREUM={});NREUM.info={"beacon":"bam.nr-data.net","errorBeacon":"bam.nr-data.net","licenseKey":"70368890e1","applicationID":"1513229
<script type="text/javascript">(window.NREUM||(NREUM={})).loader_config={licenseKey:"70368890e1",applicationID:"15132299"};window.NREUM||(NREUM={}),__nr_require=function(e,t,r
<script>
  var _rollbarParams = {"server.environment": "production"};
  _rollbarParams["notifier.snippet_version"] = "2"; var _rollbar=["deef9189a000432aa3eb00b91b2842b9", _rollbarParams]; var _ratchet=_rollbar;
  (function(w,d){w.onerror=function(e,u,l){_rollbar.push({_t:'uncaught',e:e,u:u,l:l});};var i=function(){var s=d.createElement("script");var
  f=d.getElementsByTagName("script")[0];s.src="//d37gvrvc0wt4s1.cloudfront.net/js/1/rollbar.min.js";s.async=!0;
  f.parentNode.insertBefore(s,f);};if(w.addEventListener){w.addEventListener("load",i,!1);}else{w.attachEvent("onload",i);}})(window,document);
</script>
<title>Nabu - Niger-Congo field materials</title>
<link href="/favicon.ico" rel="shortcut icon" type="image/vnd.microsoft.icon" />
<link href="/assets/screen-5ae0863da964d3d57a3a518a51b114ad.css" media="screen, projection" rel="stylesheet" type="text/css" />
<link href="/assets/print-c0a14bffb4eda936de85c9ef29d09cb4.css" media="print" rel="stylesheet" type="text/css" />
<!--[if lt IE 8]>
<link href="/assets/ie-d6b11f0862243f01753a7d6759806eb8.css" media="screen, projection" rel="stylesheet" type="text/css" />
<![endif]-->
<script src="/assets/application-d89d708766e552b52282a046b61c269d.js" type="text/javascript"></script>
<script src="https://maps.googleapis.com/maps/api/js?key=AIzaSyAtxUAk_CHTlHR_to5xDCA2kG5YzkVcl4E" type="text/javascript"></script>
<meta content="authenticity_token" name="csrf-param" />
<meta content="Y21zbXQCG6vK5hHfQx4aywDnzdOExSISw/YdfxfOO8c=" name="csrf-token" />
        <!-- Analytical Init: Google -->
        <script type="text/javascript">
          var _gaq = _gaq || [];
          _gaq.push(['_setAccount', 'UA-36256674-1']);
          _gaq.push(['_setDomainName', '']);
          _gaq.push(['_trackPageview']);

          (function() {
            var ga = document.createElement('script'); ga.type = 'text/javascript'; ga.async = true;
            ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-analytics.com/ga.js';
            var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);
          })();
        </script>
<script type="text/javascript">
var Analytical = {
  track: function(page) {
_gaq.push(['_trackPageview', page]);
  },
  event: function(name, data) {
    if (typeof data === 'undefined') { data = {}; }
_gaq.push(['_trackEvent', "Event", name]);
  }
};
</script>
</head>
```

Figure 13. HTML header code from the RB5 collection.

[80] In some cases Zotero will detect both DOIs and HTML metadata. Zotero presents the user a default option. A computer user can choose a non-default option by right clicking the Zotero connector icon in the browser.

4.2.6 File download

PARADISEC does not provide a downloadable metadata file in a common format such as BibTeX, RIS, or MODS at any level—collection, item or essence file.

4.2.7 Complete or sufficient

In this section I present a comparison of what was imported to Zotero with what is needed to craft a reference in both *APA 6th edition* and *Chicago 17th edition Author-date* as discussed in Chapter 1. I then present for comparison the suggested reference provided by PARADISEC. I do this for both the collection level and the artifact level. In this way we can see if the information transferred to Zotero is complete and if the provided suggested reference from the institution is sufficient for an informative reference.

4.2.7.1 Collection

If required to format a reference for the RB5 collection manually in *APA 6th edition*, I would craft it to look like the following:

> Blench, R. (2003). Audio materials. [Set of 104 digital audio sub-collections with different contributors]. Niger-Congo field materials (Collection RB5: https://catalog.paradisec.org.au/collections/RB5). Pacific and Regional Archive for Digital Sources in Endangered Cultures: Sydney, Australia. doi: 10.4225/72/56E977B622032

If required to format a reference for the RB5 collection manually in *Chicago 17th edition* (author-date format), I would craft it to look like the following:

> Roger Blench Deposits. 2003. Niger-Congo field materials, Collection RB5: https://catalog.paradisec.org.au/collections/RB5. Pacific and Regional Archive for Digital Sources in Endangered Cultures, Sydney, Australia. doi: 10.4225/72/56E977B622032

This *Chicago 17th edition* formatted reference is a bit contrived and is not as "clean" as the examples in the Chicago manual of style for two reasons. First, the style manual assumes paper or textual manuscripts are the components of a collection—at least all the examples in the Chicago manual of style indicate such.[81] Second, it is not clear how PARADISEC has applied well-known archival collection management principles such as **Provenance**, **Original Order**,[82] and **Respect des Fonds** as discussed in Hodges & McClurkin (2011:1–2) and DACS (Society of American Archivists 2013:xv–1). The Chicago manual of style's recommendations for collections seem to be biased towards well-described collections within archival industry norms.

PARADISEC's suggested reference for RB5 is given as:

> Roger Blench (collector), 2003. Niger-Congo field materials. Collection RB5 at catalog.paradisec.org.au [Open Access]. https://dx.doi.org/10.4225/72/56E977B622032

There are three comments worth making here. First is about the suggested role in the citation. It's not clear how this role is defined or to which list of roles it references. As a "non-standard" role in publisher style sheets, it is not supported by CSL and therefore also not supported by Zotero.[83] The second thing to note is the presence of *[Open Access]* in the reference. Wikipedia references often include a green unlock icon in the citation indicating that the item linked is open access. However, the open access information is not part of any of the major publication style sheets I have seen, beyond the practice

[81] Part of the rationale given for this is that many publisher style sheets, when citing audio or moving image resources require a separate section from the bibliography called a discography or filmography (Harper 2017:§14.221, §14.262, §15.57).

[82] In my own interactions with PARADISEC as part of research done when looking at archived lexicons (Paterson 2015a), it became clear that the archive was not always preserving the original artifact as it was delivered to them. That is, they would sometimes destroy the integrity of an artifact and only preserve certain components of the artifact even though the entire artifact was necessary for functionality with the generating software. The artifact in question is mentioned in the record at: http://www.language-archives.org/item/oai:paradisec.org.au:DD1-028 . In some cases by viewing item notes via OLAC records, I have seen descriptions where original audio artifacts recorded in lossy formats were converted to lossless file formats. Nabu does not indicate to web-users which digital artifact is the original. Note that items within the RB5 collection may contain both `.wav` (lossless) and `.mp3` (lossy) artifacts. From the Nabu interface it is impossible to tell which of the various file formats are the original recordings, or if they are all original recordings, but from different recording devices.

[83] In Zotero there is a manual way via the Extra field to add custom roles on a per name and per reference basis.

at Wikipedia. However, the open access component is remarkable for a second reason: no item in PARADISEC is truly open access per the widely accepted definition of open access defined in the Budapest Open Access Initiative,[84] which stipulates that open access items must not have any encumbrance to access other than access to the Internet itself. PARADISEC requires users to login and agree to their ethics statement in order to access files.

Finally, I generate references for the PARADISEC RB5 collection via Zotero's DOI import. Zotero's output for *APA 6th edition* is as follows:

> Blench, R. (2014). [Blench Niger-Congo]. https://doi.org/10.4225/72/56E977B622032

[84] https://www.budapestopenaccessinitiative.org

Zotero's output for *Chicago 17th edition* is as follows:

> Blench, Roger. 2014. "[Blench Niger-Congo]." https://doi.org/10.4225/72/56E977B622032.

I find that the information provided by the DOI about the collection makes it challenging to craft a well-formed reference in the style of those recommended by the *APA 6th edition* and the *Chicago 17th edition*. This leads me to believe that an application of Zipf's **principle of least effort** (Zipf 1949)[85] would tempt authors to create poor or non-style-sheet-compliant references for artifacts held at PARADISEC, under the guise that *something is better than nothing* or *nothing is better than something errant*. However, if one were to go full manual mode and create the Zotero record by hand from the Nabu interface using Zotero's **Extra** field as shown in Figure 14,[86] Zotero will export valid *Chicago 17th edition* collection formatted references as shown below.[87]

> Blench, Roger (Depositor). 2003. "Niger-Congo Field Materials." RB5. Pacific and Regional Archive for Digital Sources in Endangered Cultures. Sydney, Australia. https://doi.org/10.4225/72/56E977B622032.

[85] As stated by Case (2002:140), Zipf's principle of least effort is "each individual will adopt a course of action that will involve the expenditure of the *probable least average* of his work".

[86] In Figure 14 the Zotero item type screen for "Book Section" is shown. Any item type option could be used because the item type is from the screen will be overwritten by the one present in the Extra field. However, when choosing a screen, choosing one which has more of the type of fields one might need is advantageous as one needs to place fewer values in the Extra field.

[87] Some terms such as RB5 are not super informative as titles. This should cause every linguist to strive to find informative titles for their works—including collections.

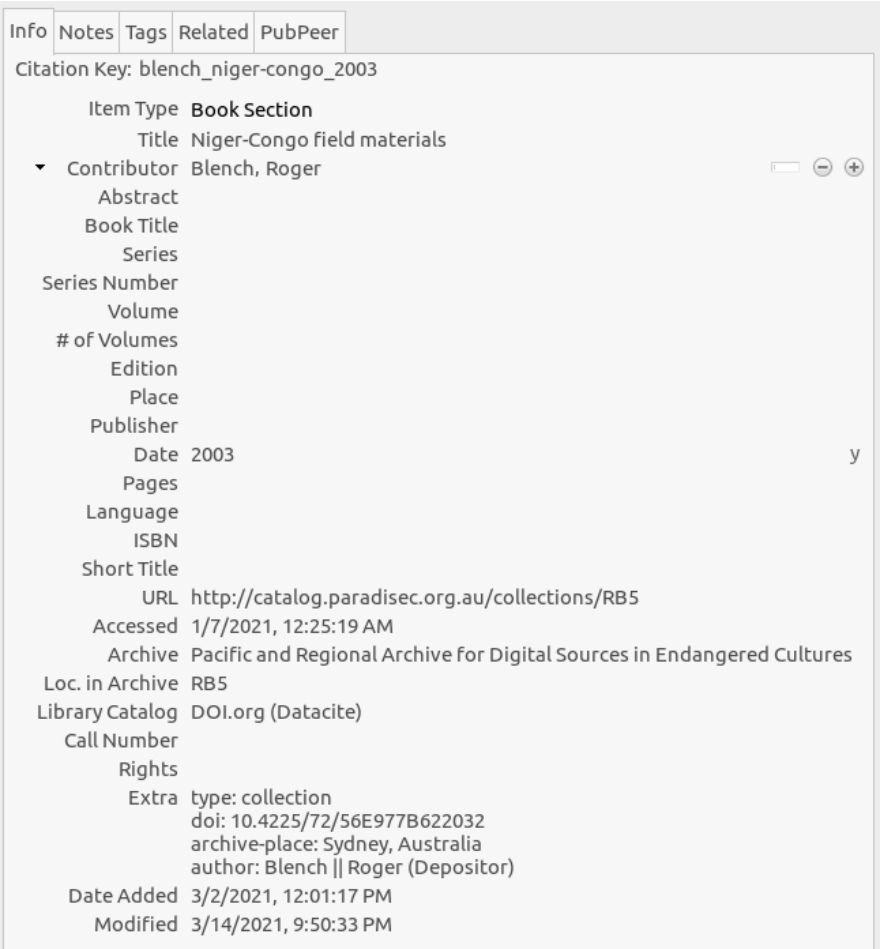

Figure 14. Zotero record for a collection using Zotero's **Extra** field.

4.2.7.2 Item

In this section I look at the sufficiency of metadata for crafting a reference for a component of a collection. Because PARADISEC has chosen to curate the collection in such a manner that the individual essence files do not match with their item node, it is hard to craft an appropriate reference in either *APA 6th edition* or *Chicago 17th edition Author-date*. Many times PARADISEC's item-level nodes are representative of aggregate works—or even more befuddling multiple aggregate works, as is the case with the Kamuku wordlists item. Aggregate works in fact are a type of collection. So, we have a case where PARADISEC's structures are saying we should have a single item, but the archive is presenting multiple items each with multiple artifacts. At this point one could choose

to reference the item as a collection and use strategies as presented in Section 4.2.7.1, or one could choose to focus the reference pattern around implied metadata contained in the names of essence files. However, if one chooses to focus on crafting a reference around the essence file name, then they must reinterpret the item level metadata as a collection and manually adjust any metadata received via the DataCite API—including the DOI. Using locations and events (timestamp ranges) can help archives create aggregate works from ambiguous deposits. The plethora of choices a scholar must make in order to craft a reasonable reference which is both informative and accurate to either *APA 6th edition* or *Chicago 17th edition Author-date* reduces automaticity. Under the *principle of least effort* we should expect this to break the referencing and citation cycle discussed in Chapter 1.

Scholars, when crafting a reference, should take an artifact view of a PARADISEC essence file. However, taking this view does not acknowledge the collection description[88] which is yet to be done and on which the reference should be based.[89]

The following reference is suggested by PARADISEC:

> Zachariah Yoder (collector), 2007. Kamuku wordlists. X-WAV/MPEG. RB5-Kainji_Kamuku_wordlists at catalog.paradisec.org.au. https://dx.doi.org/10.4225/72/5705AD74A50AD

I find the suggested reference by PARADISEC interesting for several reasons. Like the APA style, it attempts to indicate which media type the artifact is. In general this is a good thing. The APA style has given guidance that the kind of information which should go in this slot for physical media is the carrier. These are indicated with expressions like "LP 33 ½", "reel: 7 1/2 ips", or "CD". For digital artifacts, the media carrier slot is used to indicate the technology needed to access the resource. Digital formats can use the

[88] Recall that collection description and collection arrangement are interlinked. An archive that does not reflect the full range of embedded structure has re-interpreted the artifact, and misrepresents its original context to archive visitors.

[89] Two additional thoughts on this issue deserve additional investigation, but I do not address them further. First, what should an archive's suggested reference look like when an artifact is also located in another archive? Traditional manuscripts do not have this issue as there is only one copy. However, digital artifacts could be in two collections at two institutions at the same time or even in two different collections at the same institution at the same time. Second, archivists have long approached collection description as a progressive task. How should references evolve with progressive description practices surrounding language artifacts in language archives?

designator of the set of media types specified by the Internet Assigned Numbers Authority (IANA).[90,91] This is sometimes known as the **MIME type**. The MIME type is Dublin Core and OLAC compatible; therefore, they should be automatically available to any of the archives reviewed in this thesis. However, in the case of the *Kamuku wordlists* this is a collection not a single file. MIME types should only be included in the reference when it is a reference to an individual file. Furthermore, the recommended reference includes the MIME types of "X-WAV/MPEG". This fallaciously indicates two distinct MIME types when only one should be indicated. Second, the reference does not acknowledge that it is part of the RB5 collection except in the construction of the item's ID element. Third, the date is interesting in that the RB5 collection has an earlier date. It seems to me that a span of a few years might be reason to consider a new collection in an archive's holdings, or the date of RB5 should be adjusted to a range style date. If Roger Blench is the Collector of the Collection, what roles are viable for elements of the collection such as this one where Zachariah Yoder is also considered a collector? The curation process should check for conflicts such as overlapping exclusive roles. This could be done programmatically or at least flagged programmatically for manual review and resolved with a curator's clarifying note. I do not find the provided reference satisfying or helpful if I wanted to reference a particular audio file within this "item". The inclusion of information relevant to what technology is required to hear the audio and the inclusion of information about the aggregate work is something that both APA and Chicago require.

4.3 Pangloss

The Pangloss Collection, established in 1994, is described in works such as Thieberger & Jacobson (2010), Michailovsky et al. (2011), Michailovsky et al. (2014), and Vasile

[90] http://www.iana.org/assignments/media-types

[91] In most cases using the IANA registered formats is sufficient. However, in one particularly relevant case a compromise needs to be made. RFC2046 specifies that IANA will host a list of registered Media Types (formerly known as MIME types). This list can be found here: https://www.iana.org/assignments/media-types/media-types.xhtml. It is the list that Dublin Core, and therefore OLAC, reference via the IMT vocabulary: http://purl.org/dc/terms/IMT. The difficulty comes around to this: `.wav` does not have a registered entry in the IANA registry. Even though Mozilla web documentation and browsers support the MIME type `audio/wav` this is not a valid IANA MIME type (see: https://developer.mozilla.org/en-US/docs/Web/HTTP/Basics_of_HTTP/MIME_types/Common_types) Obviously, this could be a problem for language archives which have a tendency to have a significant quantity of `.wav` based audio.

et al (2020).[92] The Pangloss Collection is a special collection (in the library collections management sense) housed "inside of" or on top of the infrastructure of Cocoon (*Collection de Corpus Oraux Numériques*),[93] a national data infrastructure platform supporting open science in France. *The Pangloss Collection* is a presentation and interaction platform for language focused audio and audiovisual artifacts and their derivative or associated works. *The Pangloss Collection* is affiliated with at least a half dozen language labs in the CNRS (*Centre national de la recherche scientifique*) research lab system. Many of these labs also have partnerships with universities and their academic departments. Pangloss contains over 780 hours of recordings in more than 170 languages. Transcription coverage is about fifty percent for the contained audio artifacts.

Contributors to the Pangloss collection have mostly been scholarly researchers affiliated with CNRS-LACITO (*Langues et civilisations à tradition orale*). LACITO is located on the CNRS Villejuif campus at the outskirts of Paris, France. In addition to the interagency collaboration which undergirds Cocoon, the Villejuif campus also has their own campus-based staff supporting the various library and data science needs of the on-campus labs. The Villejuif staff have produced a second interface to Pangloss. The Villejuif display of the artifacts has two modes: a "pro mode" with details designed for a scientific audience and a "default mode" with some details removed in the hopes that a general audience would not be encumbered by details of interest to specialists.[94] I review both the Cocoon interface and the "pro mode" of the Villejuif user interface. Table 12 provides a summary of the import technologies currently available for the two Pangloss interfaces.

Table 12. Summary of support for Zotero at Pangloss.

	DOI	**Embedded metadata**	**File import**
Cocoon Collection	Yes	Some / DOI detection / unAPI	Yes
Cocoon Item	Yes	Some / DOI detection / unAPI	Yes
Villejuif Collection	No	None	No
Villejuif Item	Yes	None	No

[92] See footnote 43 in chapter 4.
[93] https://cocoon.huma-num.fr/exist/crdo
[94] https://pangloss.cnrs.fr/?lang=en&mode=pro

4.3.1 Technology infrastructure

Each Pangloss interface is managed by different designers and IT teams.[95] A variety of technological and archiving agencies are contracted through the national infrastructure system to deliver the interactive experience. The actual software used to deliver the interactive experiences in both cases is not overtly stated. However, through a series of tests I was able to determine that Cocoon uses Jetty[96] as part of the set of technologies used to deliver the service, while the Villejuif interface uses a PHP/MySQL based infrastructure.

4.3.2 Collections structure

The Pangloss holdings have one hierarchical arrangement in the Cocoon interface and another hierarchical arrangement in the Villejuif interface. Web-based user interfaces are the embodiment of philosophical approaches to artifacts. Both Pangloss interfaces embody and facilitate a particular materiality. That is, artifact engagement is suggested by the options of a user interface and at the same time limited by the user interface. User interfaces set the cadence for how members of society interact with artifacts, how they come to let these artifacts inform their activities, and ultimately their beliefs about the artifacts. Because there are two Pangloss user interfaces, there are two philosophical approaches presented.

Within the Cocoon interface, the collection is situated as a special collection. Inside of the special collection, there are individual collections called “Corpora”. Within a corpus, individual items exist. Items can have multiple files. However, unlike the large sets of files presented in a single PARADISEC item (discussed in Section 4.2), Pangloss items have a more limited scope. Items contain only artifacts which are closely related such as a source audio file (usually a `.wav` file), a low fidelity distribution manifestation (`.mp3`) of the source audio file, the source audio file’s transcription, and possibly any instrumental data which was co-produced with the audio recording. Files within items have multiple unique URIs.[97] Corpora are generally based on specific research endeavors and therefore

[95] In fact, a renewed Villejuif-based interface was released while this thesis was being written—and looks really nice.

[96] http://www.eclipse.org/jetty

[97] https://cocoon.huma-num.fr/exist/crdo/identifiers.htm

demonstrate cohesion from that perspective. For instance they might be the combined output of a single scholarly research endeavor, e.g., a single phonetic experiment.[98] The hierarchical arrangement of nodes looks like: *Cocoon* > *Pangloss* > *Collections (Corpora)* > *Item* > *Single File*. From the Cocoon item view, one can access the DOI (and other unique identifiers) of the artifact, a link to the record of the entire collection, and links to associated artifacts. Annotation files, when they exist, have their own DOIs separate from their associated audio files—and are linked to audio files via relational metadata. However, instrumental data files, such as electroglottography files, do not have their own identifiers.

The Villejuif interface presents a "collection of languages" or speech varieties; each body of files for a speech variety is labeled as a "corpus". When a corpus is selected in the user interface one can navigate to a page and interact with the artifacts by listening to them and seeing the transcription when available. The hierarchical arrangement of nodes looks similar to the Cocoon arrangement in terms of nodes, but names are labeled differently: *Pangloss* > *Languages (Corpora)* > *Item* > *Single File*.

The variation in names between presentation interfaces is likely to cause confusion to scholars as they look to craft references, not just for the name variations but also because the collections are not equivalent.

4.3.3 Collections and artifacts reviewed

When evaluating Pangloss I looked at an artifact with the title *SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7* and the collections it is presented in. Within the Cocoon interface it is presented in the *AuCo: corpus audio de langues du Vietnam et des pays voisins* collection,[99] and the *Pangloss* collection. In the Villejuif interface it is presented in the *Vietnamese (Hanoï dialect)* collection.[100]

[98] While I use the term corpus and corpora here when describing Pangloss, I use this term because they use the term. I question the appropriateness of calling these nodes in the hierarchy **corpora**. In addition to the discussion in footnote 8, one complicating factor may be the variation in lexically similar terminology across languages. That is, does **corpus** in French come with the same contextual meaning as **corpora** in English?

[99] https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d

[100] https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro

Pangloss and *AuCo: corpus audio de langues du Vietnam et des pays voisins* are both parallel structures in the Cocoon interface and the artifacts are cross-listed (see Figure 18). The Pangloss collection page within the Cocoon interface is shown in Figure 15. One thing to note is that from this page one can not see the fond/series which would house the collection. Presumably, this is a limitation of the user interface which only lists six series in the collection, not a logical limitation where some set of materials is not listed in the index of the collection.

Figure 16 shows the collection page of *AuCo: corpus audio de langues du Vietnam et des pays voisins* in the Cocoon interface. While the Cocoon page does represent the collection record, from this page one can not see a list of files which are part of the collection. However, by clicking the dark blue button with the text "*Accéder aux 393 enregistrements*", one can find the list of all files in the collection via filters in a faceted view. Otherwise there is no list of all the files within the *AuCo: corpus audio de langues du Vietnam et des pays voisins* collection. Further breakdown of the collection into the contained aggregate works is not possible via the user interface, even though there is an inference in the title of some works that there might be other recordings which are related, e.g., the use of "Speaker M7" in the title implies that their might also be speakers M1 through M6 and artifacts organized around those speakers.[101]

[101] Another example of an aggregate work in the *AuCo: corpus audio de langues du Vietnam et des pays voisins* collection on Cocoon infrastructure includes:
Ferlus, Michel. 2015. *Vocabulary list for Hoa Binh dialect of the Mường language as spoken in the commune of Dan Chu [female speaker], part 1 of 2 (Version 1)*, (2015 July 29). Centre de recherches linguistiques sur l'Asie orientale. https://doi.org/10.34847/COCOON.42B3550F-D8A5-3CBA-AACE-123693C0A194

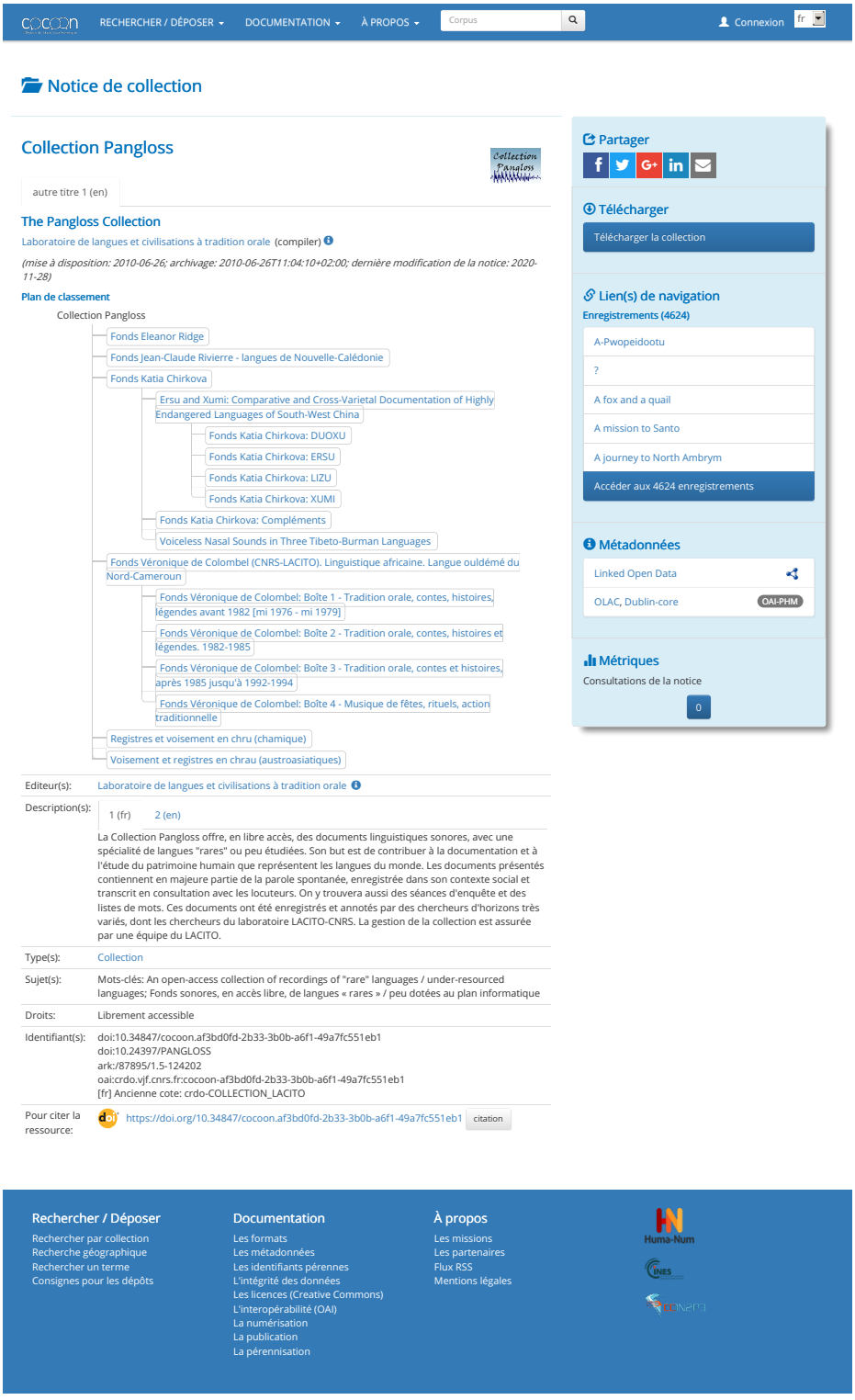

Figure 15. *Pangloss* Collection-level page in the Cocoon interface[102]

[102] https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-af3bd0fd-2b33-3b0b-a6f1-49a7fc551eb1

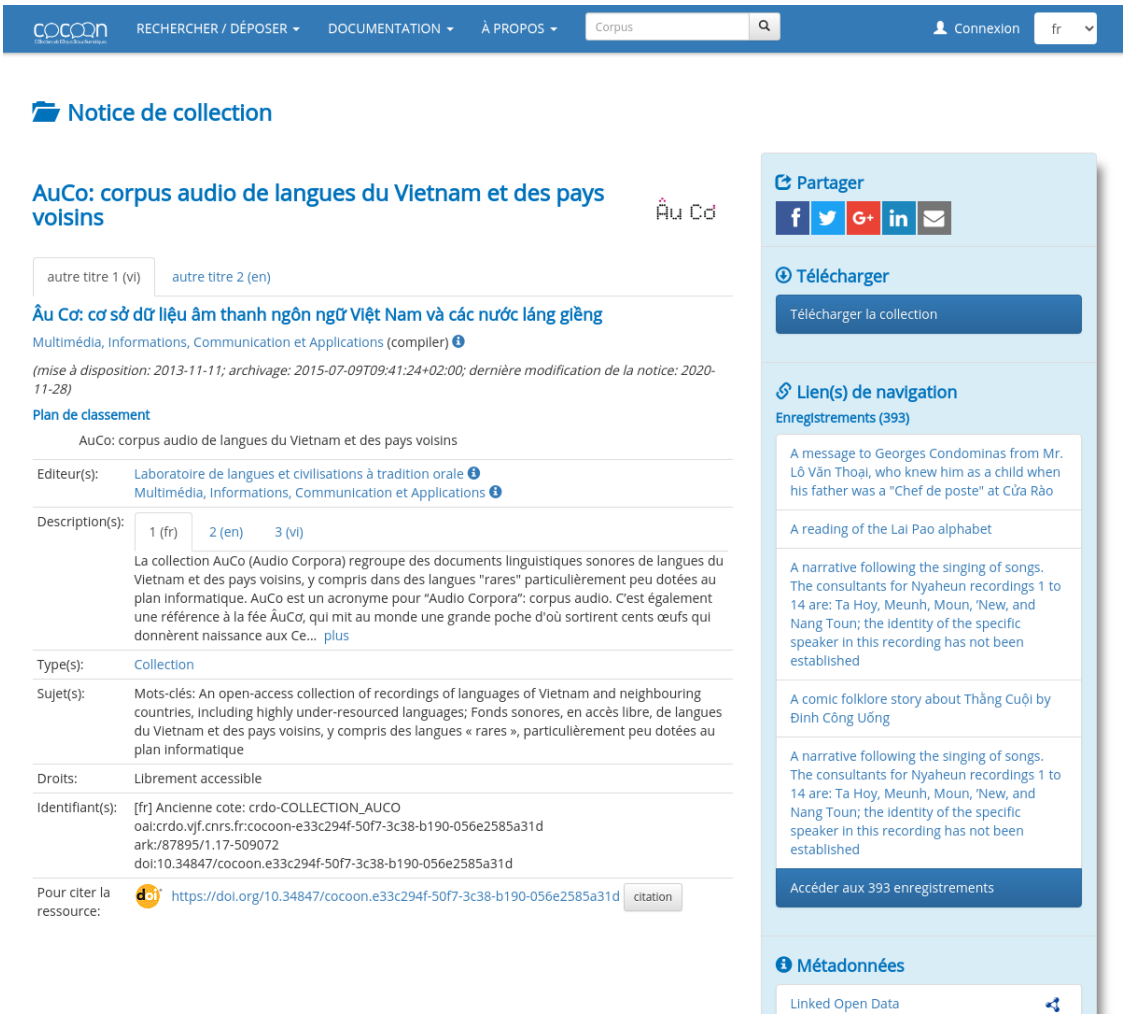

Figure 16. Top of the *AuCo* Collection-level page in the Cocoon interface[103]

Figure 17 shows the collection *Vietnamese (Hanoï dialect)* in the Villejuif interface. In contrast to the Cocoon interface, the Villejuif interface makes no mention of the *AuCo: corpus audio de langues du Vietnam et des pays voisins* collection. The Villejuif interface does list all (presumably all) the contained files in a list containing 21 unique sub-records (some containing the text M1 through M6). That is, the collection description is not really a description of the aggregate of artifacts, but rather a description of the language variety that the files are reported to represent. The description is interesting from a social perspective, but it is not helpful for understanding the artifacts themselves or the context in which they were created. Based on the names of the files, it appears that together they

[103] https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d

form the evidentiary record for an investigation of tone in the Hanoi dialect of Vietnamese. So, albeit in different ways, both interfaces do not acknowledge a crucial component of the aggregate work hierarchical arrangement. Cocoon leaves out the grouping: *Vietnamese (Hanoï dialect)*, while the Villejuif interface leaves out the higher level grouping: *AuCo: corpus audio de langues du Vietnam et des pays voisins*. The hierarchical arrangement in an aggregate work is a crucial contextual component not only for reference formation, but also to inform artifact users of the context of artifact creation and preservation.

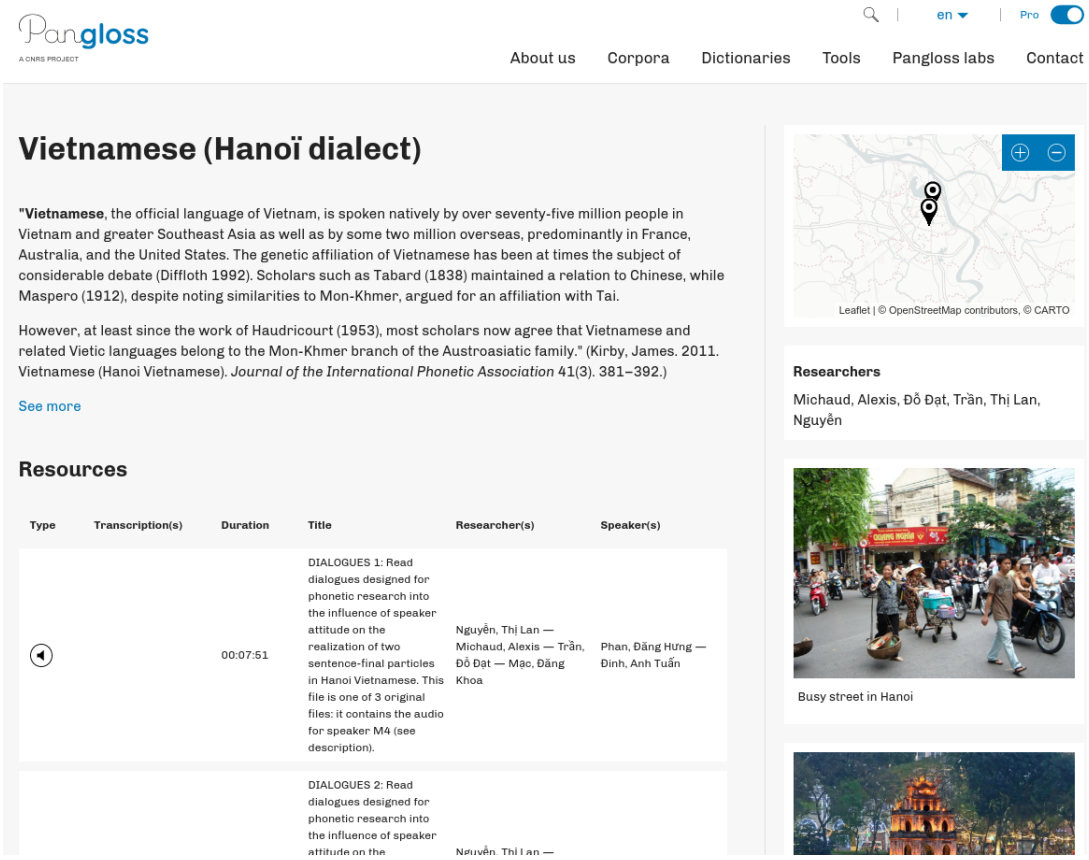

Figure 17. Top of the *Vietnamese (Hanoï dialect)* Collection-level page in the Villejuif interface[104]

When looking at the item view, I chose a component of the collection from speaker M7. Figure 18 shows this item in the Cocoon interface, and Figure 19 shows it in the Villejuif interface.

[104] https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro

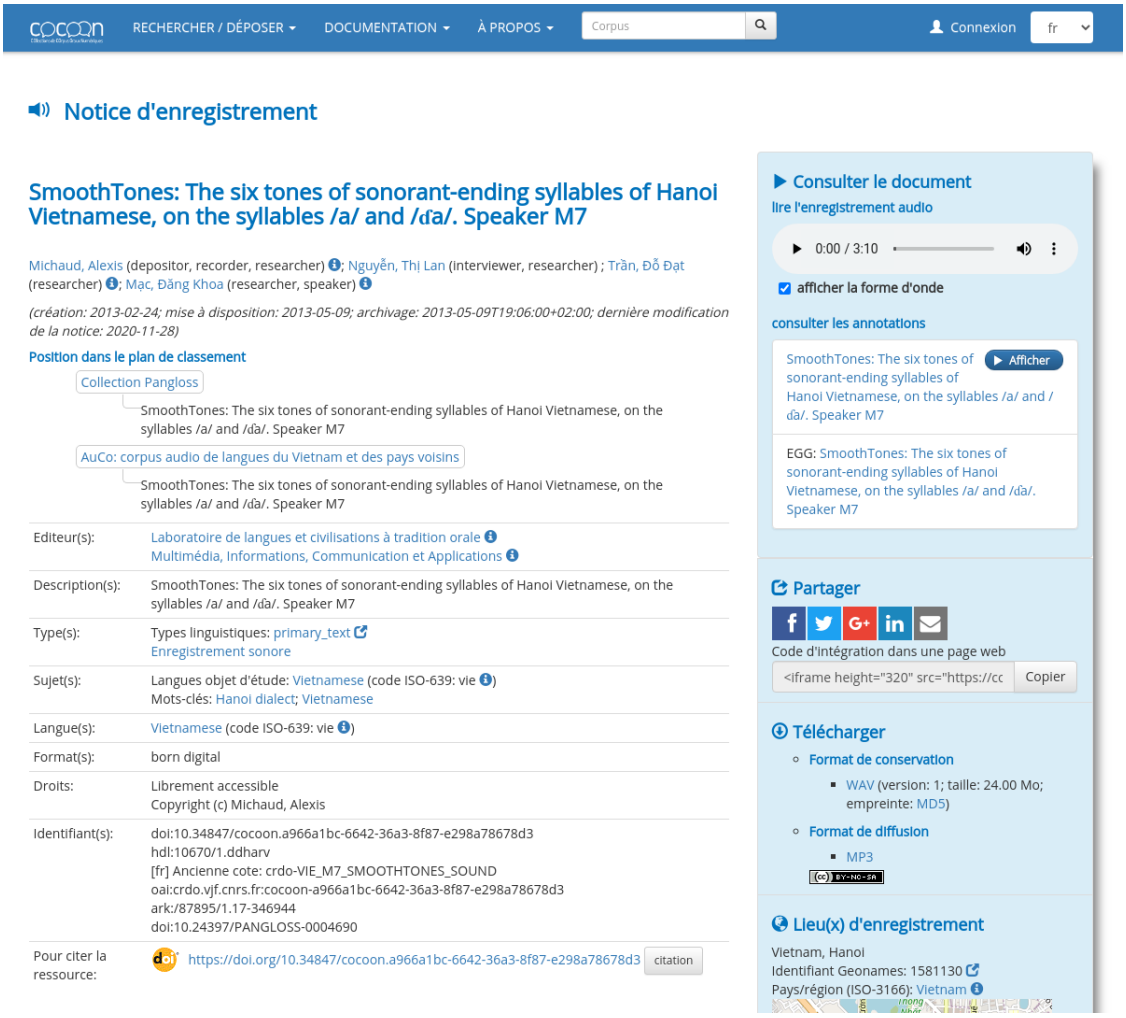

Figure 18. Item-level page in the Cocoon interface[105]

[105] https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3

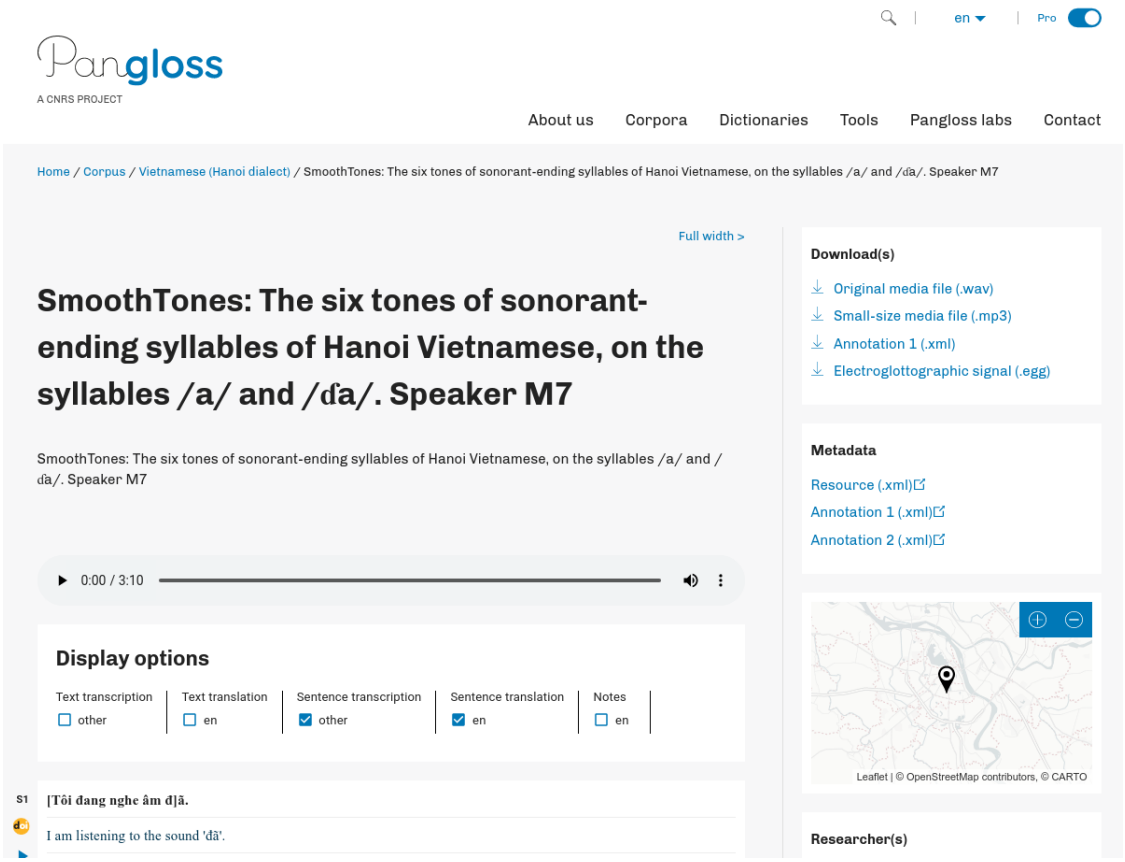

Figure 19. Item-level page in the Villejuif interface[106]

4.3.4 DOI import

The two Pangloss interfaces each assign distinct DOIs to the same artifacts. Additionally, the Cocoon interface assigns DOIs to the collections it recognizes. Villejuif does not assign DOIs to collections.[107] Recall that the visual presentations of "collections" are not equivalent across the two user interfaces. Table 13 displays the DOIs reviewed.

[106] https://pangloss.cnrs.fr/corpus/show?lang=en&mode=pro&oai_primary=cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3

[107] As discussed in Vasile et al (2020), the DOIs that the Villejuif interface assigns to artifacts also use HTML anchor syntax at the end of the resolvable DOI to link directly to segments of the referenced artifact (phrases in a recording). This is not currently implemented in the Cocoon interface. While the direct linking to segments of the phrase with HTML anchors is both technically interesting and creative, the HTML interface is likely ephemeral. I am not sure if it would have been better or not to add the links as a branch to the DOI directly, e.g., 10.24397/pangloss-0004690.001 and 10.24397/pangloss-0004690.002. This could have implications on the metadata which is importable via the DataCite API.

Table 13. Pangloss DOIs

Tier node	DOI
Cocoon collection[108]	10.34847/cocoon.e33c294f-50f7-3c38-b190-056e2585a31d
Villejuif collection	None
Cocoon artifact	10.34847/cocoon.a966a1bc-6642-36a3-8f87-e298a78678d3
Villejuif artifact	10.24397/pangloss-0004690

4.3.4.1 DOI Collection import

Figure 20 contains the JSON data which the DataCite API sends to Zotero when the magic wand tool is used to import the Cocoon collection's bibliographic metadata. The way that Zotero interprets these results is shown in Figure 21.

[108] To be clear the DOI in the table is the DOI for the collection *AuCo: corpus audio de langues du Vietnam et des pays voisins*. The *Pangloss* collection DOI within the Cocoon interface is 10.34847/cocoon.af3bd0fd-2b33-3b0b-a6f1-49a7fc551eb1. The *AuCo* corpus is the one that was primarily tested and presented here. The differences between the two collections as far as reference generation is concerned pertain to the tokens passed between software, not the types of tokens passed between software, e.g., both have dates though *AuCo* is 2013 and *Pangloss* is 2010.

```
{
  "data":{
    "id":"https://doi.org/10.34847/cocoon.e33c294f-50f7-3c38-b190-056e2585a31d",
    "type":"works",
    "attributes":{
      "doi":"10.34847/cocoon.e33c294f-50f7-3c38-b190-056e2585a31d",
      "identifier":"https://doi.org/10.34847/cocoon.e33c294f-50f7-3c38-b190-056e2585a31d",
      "url":"https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d",
      "author":[
        {
          "literal":"Multimédia, Informations, Communication et Applications"
        }
      ],
      "title":"AuCo: corpus audio de langues du Vietnam et des pays voisins",
      "container-title":"Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale",
      "description":"La collection AuCo (Audio Corpora) regroupe des documents linguistiques sonores de langues du Vietnam et des pay
      "resource-type-subtype":"recordings",
      "data-center-id":"inist.humanum",
      "member-id":"jbru",
      "resource-type-id":"collection",
      "version":"1",
      "license":null,
      "schema-version":null,
      "results":[

      ],
      "related-identifiers":[

      ],
      "citation-count":0,
      "citations-over-time":[

      ],
      "view-count":0,
      "views-over-time":[

      ],
      "download-count":0,
      "downloads-over-time":[

      ],
      "published":"2013",
      "registered":"2020-11-28T21:36:04.000Z",
      "checked":null,
      "updated":"2020-11-28T21:36:04.000Z",
      "media":[

      ],
      "xml":"PD94bWwgdmVyc2lvbj0iMS4wIiBlbmNvZGluZz0iVVRGLTgiPz4KPHJlc291cmNlIHhtbG5zOnhzaT0iaHR0cDovL3d3dy53My5
    },
    "relationships":{
      "data-center":{
        "data":{
          "id":"inist.humanum",
          "type":"data-centers"
        }
      },
      "member":{
        "data":{
          "id":"jbru",
          "type":"members"
        }
      },
      "resource-type":{
        "data":{
          "id":"collection",
          "type":"resource-types"
        }
      }
    }
  }
}
```


Figure 20. JSON response for the Cocoon collection *AuCo* from the DataCite API

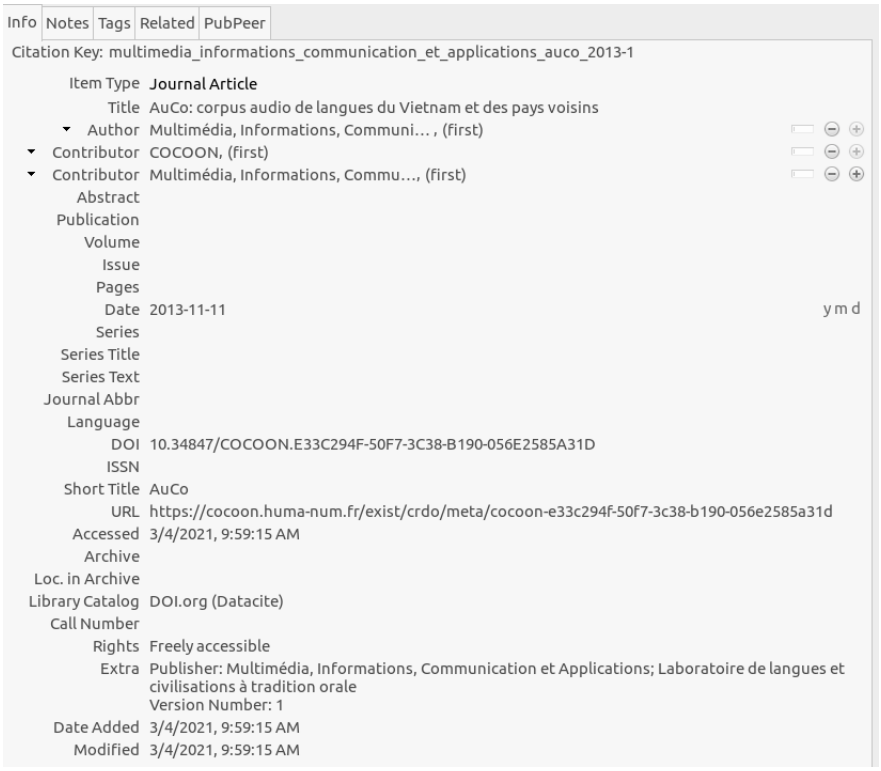

Figure 21. Zotero's interpretation of DataCite API supplied JSON for the Cocoon collection

There are three issues with the data imported via DataCite. First, recall from Table 5 that additional CSL variables can be added to Zotero's **Extra** field. In this case **type: collection** is not invoked to tell Zotero and CSL to use collection templates in style sheets. It should be. Second, the name in the Author field is not recognized as an institutional name. It is an open question as to whether the institutional name is really the best choice here or if this collection should have a depositor's name or some other role (filled by a person) instead of the institution. Third, the data should likely be the range of dates from the production of the artifacts in the collection rather than a single date. As indicated in the Extra field, DataCite provides a version number for the collection. Versioning a collection is a good idea,[109] as versions are in general a great way to handle provenance issues (including notes which apply specifically to a given version). However, no version number is indicated in the user web-based interface and so was not expected upon import. A scholar building this reference manually would not know to include a version number.

[109] I specifically advocate for semantic versioning: https://semver.org

4.3.4.2 DOI Artifact import

Figure 22 contains the JSON data which the DataCite API sends to Zotero when the magic wand tool is used to import the specified audio artifact's bibliographic metadata. However, no record is produced.


```
{
  "data":{
    "id":"https://doi.org/10.34847/cocoon.a966a1bc-6642-36a3-8f87-e298a78678d3",
    "type":"works",
    "attributes":{
      "doi":"10.34847/cocoon.a966a1bc-6642-36a3-8f87-e298a78678d3",
      "identifier":"https://doi.org/10.34847/cocoon.a966a1bc-6642-36a3-8f87-e298a78678d3",
      "url":"https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3",
      "author":[
        {
          "literal":"Michaud, Alexis"
        },
        {
          "literal":"Nguyễn, Thị Lan"
        },
        {
          "literal":"Trần, Đỗ Đạt"
        },
        {
          "literal":"Mạc, Đăng Khoa"
        }
      ],
      "title":"SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7",
      "container-title":"Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale",
      "description":"SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7",
      "resource-type-subtype":"primary_text",
      "data-center-id":"inist.humanum",
      "member-id":"jbru",
      "resource-type-id":"sound",
      "version":"1",
      "license":"http://creativecommons.org/licenses/by-nc-sa/2.5/",
      "schema-version":null,
      "results":[

      ],
      "related-identifiers":[

      ],
      "citation-count":0,
      "citations-over-time":[

      ],
      "view-count":0,
      "views-over-time":[

      ],
      "download-count":0,
      "downloads-over-time":[

      ],
      "published":"2013",
      "registered":"2020-11-28T19:30:29.000Z",
      "checked":null,
      "updated":"2020-11-28T19:30:29.000Z",
      "media":[

      ],
      "xml":"PD94bWwgdmVyc2lvbj0iMS4wIiBlbmNvZGluZz0iVVRGLTgiPz4KPHJlc291cmNlIHhtbG5zOnhzaT0iaHR0cDovL3d3dy53My5vcmcvMjAwMS9
    },
    "relationships":{
      "data-center":{
        "data":{
          "id":"inist.humanum",
          "type":"data-centers"
        }
      },
      "member":{
        "data":{
          "id":"jbru",
          "type":"members"
        }
      },
      "resource-type":{
        "data":{
          "id":"sound",
          "type":"resource-types"
        }
      }
    }
  }
}
```


Figure 22. JSON response for the Cocoon artifact from the DataCite API

Zotero does not produce a record when the Cocoon artifact's DOI is queried. The reason for this is that the DataCite JavaScript translator file depends on the JSON data containing a schema value (see line 32 in Figure 22 where the value is "null"). This is a known bug in some data associated with some DOIs in the DataCite data repository. Apparently, archives and Zotero teams are waiting on DataCite engineers to fix this problem.[110]

The Villejuif based interface uses a separate set of DOIs. Figure 23 contains the JSON data which the DataCite API sends to Zotero when the magic wand tool is used to import the Villejuif audio artifact's bibliographic metadata. The way that Zotero interprets these results is shown in Figure 24.

[110] https://github.com/zotero/translators/issues/2018

```
{
  "data":{
    "id":"https://doi.org/10.24397/pangloss-0004690",
    "type":"works",
    "attributes":{
      "doi":"10.24397/pangloss-0004690",
      "identifier":"https://doi.org/10.24397/pangloss-0004690",
      "url":"https://pangloss.cnrs.fr/corpus/show?lang=en&mode=pro&oai_primary=cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3",
      "author":[
        {
          "given":"Thị Lan",
          "family":"Nguyễn"
        },
        {
          "given":"Alexis",
          "family":"Michaud"
        },
        {
          "given":"Đỗ Đạt",
          "family":"Trần"
        },
        {
          "given":"Đăng Khoa",
          "family":"Mạc"
        }
      ],
      "title":"SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7",
      "container-title":"Pangloss",
      "description":"SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7",
      "resource-type-subtype":"(:unkn)",
      "data-center-id":"inist.lacito",
      "member-id":"jbru",
      "resource-type-id":"sound",
      "version":null,
      "license":null,
      "schema-version":"4",
      "results":[

      ],
      "related-identifiers":[

      ],
      "citation-count":0,
      "citations-over-time":[

      ],
      "view-count":0,
      "views-over-time":[

      ],
      "download-count":0,
      "downloads-over-time":[


      ],
      "published":"2013",
      "registered":"2019-01-30T13:24:40.000Z",
      "checked":null,
      "updated":"2021-02-11T11:48:52.000Z",
      "media":[

      ],
      "xml":"PD94bWwgdmVyc2lvbj0iMS4wIiBlbmNvZGluZz0iVVRGLTgiPz4KPHJlc291cmNlIHhtbG5zPSJodHRwOi8vZGF0YWNpdGUub3JnL3NjaGVtYS9rZXJuZV
    },
    "relationships":{
      "data-center":{
        "data":{
          "id":"inist.lacito",
          "type":"data-centers"
        }
      },
      "member":{
        "data":{
          "id":"jbru",
          "type":"members"
        }
      },
      "resource-type":{
        "data":{
          "id":"sound",
          "type":"resource-types"
        }
      }
    }
  }
}
```


Figure 23. JSON response for the Villejuif artifact from the DataCite API

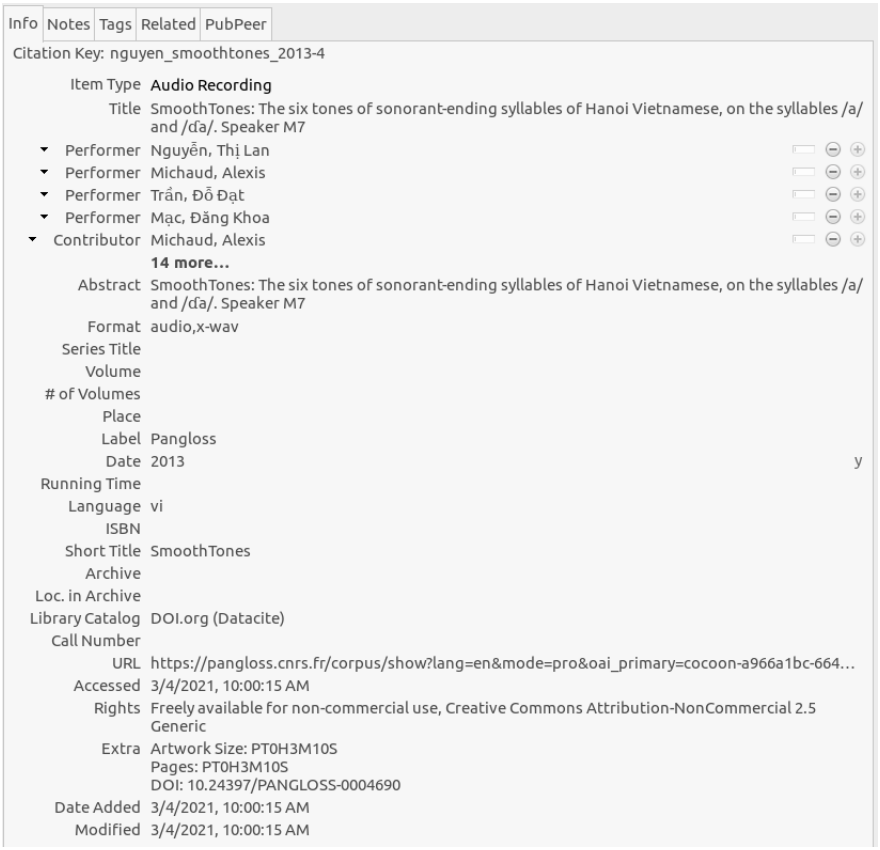

Figure 24. Zotero's interpretation of DataCite API JSON for the Villejuif item

The interpretation of the JSON results as shown in Figure 24 highlights several issues in the DataCite JavaScript translator file. The first is the interpretation of the time value as **artwork size** and **pages** as can be seen in the Extra field within Figure 24. The issue is that the translator was not coded to handle this use case and needs to be extended to cover extent for audio artifacts.[111] The second is the interpretation of the language value.[112] The valid BCP47[113] code for Vietnamese is **vi** as is shown in Figure 24. The issue is that the value should be **en** for English to match the typographical style in which the reference will be published, i.e., a scholar would create a reference in English within

[111] See line 201 in the translator file:
https://github.com/zotero/translators/blob/master/Datacite%20JSON.js#L201

[112] I did not check the role of the MODS language value. It may be that the MODS language value is accurate within the MODS record for what it is supposed to describe. That is, the Zotero MODS translator may be over aggressive, importing language codes into Zotero fields when the purposes of those language codes are not fully aligned.

[113] Best Current Practice 47 is an IETF standard for how to indicate languages in digital documents (Philips & Davis 2009). https://tools.ietf.org/html/bcp47

a paper written in English. Zotero's language field (and its value) is only currently useful for turning Zotero's capitalization on or off. Any value other than blank and **en** turn off capitalization. Ideally, this would be further developed to pass the language value to CSL so that one could have control over which language variant of a CSL file is used for a specific reference. It would impact typical locale variables such as capitalization patterns, date formats, ordinals indicator, and which style of quote marks are used. CSL locale identifiers follow BCP47, the IETF standard for indicating languages in computing contexts. However, CSL only supports a limited number of locales. These can be referenced in the CSL wiki documentation.[114] The Zotero language field is not useful to linguists who are trying to track which language a resource is about. The language a resource is about belongs in the keywords area as a specialized subject term. Third, the collection value "Pangloss" is put in the **Label** field. The label for audio recordings is mapped to CSL field **publisher**. In this case *Pangloss* is not the publisher, rather it is the collection name. It is not clear that there is a publisher other than the research lab.

Still missing from the Zotero record are values for CSL variables used to indicate the aggregate work of which the audio is a part, the archive which houses the artifact, and the location of the archive. These can be placed in the Zotero **Extra** field using the appropriate CSL variables taken from the CSL specification and are indicated in Table 6 in Section 4.1.4.

4.3.5 Embedded metadata

There are two different methods for triggering Zotero to detect metadata in web pages. The first is via the use of HTML meta tags. The second is via the unAPI. The Cocoon interface has both some embedded HTML metadata and utilizes the unAPI. The Villejuif interface uses neither embedded HTML metadata nor the unAPI.

4.3.5.1 HTML

The Cocoon collection has some minimal embedded HTML metadata. This is presented across two formats. Figure 25 lines 14–17 show Dublin Core meta tags while lines

[114] https://github.com/citation-style-language/locales/wiki

18–22 show HTML5 meta tags. Other Dublin Core metadata tags are possible but the archive has chosen not to use them. These same tags and distinctions can also be seen in the code from the Cocoon item page as shown in Figure 27 (same line numbers apply).

```
1  <head>
2  <link rel="icon" href="/favicon.ico" type="image/x-icon" />
3  <link rel="shortcut icon" href="/favicon.ico" type="image/x-icon" />
4  <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
5  <link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
6  <link rel="stylesheet" type="text/css" href="/ext/jquery-ui-1.12.1.min.css" />
7  <link rel="stylesheet" type="text/css" href="/ext/bootstrap-3.3.7/css/bootstrap.min.css" />
8  <link rel="stylesheet" type="text/css" href="/ext/bootstrap-3.3.7/css/bootstrap-theme.min.css" />
9  <link rel="stylesheet" type="text/css" href="/exist/crdo/css/cocoon.css" />
10 <link rel="unapi-server" type="application/xml" title="unAPI" href="/crdo_servlet/un-api" />
11 <link rel="stylesheet" type="text/css" href="/ext/swiper-3.4.2/css/swiper.min.css" />
12 <meta charset="utf-8" />
13 <meta name="viewport" content="width=device-width, initial-scale=1" />
14 <meta property="DC:author" content="Michel Jacobson" />
15 <meta property="DC:copyright" content="CNRS/COCOON" />
16 <meta property="DC:publisher" content="CNRS/COCOON" />
17 <meta property="DC:subject" content="CRDO, COCOON, openarchive, OAI, OLAC, Linguistique, Archives orales, Archives sonores" />
18 <meta name="author" content="Michel Jacobson" />
19 <meta name="copyright" content="CNRS/COCOON" />
20 <meta name="publisher" content="CNRS/COCOON" />
21 <meta name="subject" content="CRDO, COCOON, openarchive, OAI, OLAC, Linguistique, Archives orales, Archives sonores" /><title>CRDO/COCOON: metadata: </title>
22 <meta name="description" content="La collection AuCo (Audio Corpora) regroupe des documents linguistiques sonores de langues du Vietnam et des pays voisins, y compris dans des langues &#34;rares&#34; particulièrement peu dotées au plan informatique. AuCo est
23 <script type="application/ld+json" class="nocomment">{
24   "@context":"http://schema.org",
25   "@type":"DataSet",
26   "name":"AuCo: corpus audio de langues du Vietnam et des pays voisins",
27   "alternateName":[
28     "Âu Cơ: cơ sở dữ liệu âm thanh ngôn ngữ Việt Nam và các nước láng giềng",
29     "AuCo: Audio Corpora of languages of Vietnam and neighbouring countries"
30   ],
31   "description":[
32     "La collection AuCo (Audio Corpora) regroupe des documents linguistiques sonores de langues du Vietnam et des pays voisins, y compris dans des langues \"rares\" particulièrement peu dotées au plan informatique. AuCo est un acronyme pour "Audio Corpora": corpu
33     "The AuCo collection hosts audio recordings of language of Vietnam and neighboring countries, including data of endangered/little-endowed languages. AuCo stands for AudioCorpora; it is also a reference to ÂuCơ, a fairy who bore an egg sac that hatched a hundred c
34     "Cơ sở dữ liệu ÂuCơ thu thập và lưu trữ các bản thu âm bằng các ngôn ngữ trên lãnh thổ Việt Nam và cả các nước láng giềng, bao gồm những ngôn ngữ đang có nguy cơ tuyệt chủng và những ngôn ngữ nghèo dữ liệu. Tên dự án là AuCo: AudioCorpora, Cơ sở dữ li
35   ],
36   "url":"https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d",
37   "sameAs":[
38     "http://purl.org/pol/crdo.vjf.cnrs.fr/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d",
39     "https://doi.org/10.34847/cocoon.e33c294f-50f7-3c38-b190-056e2585a31d"
40   ],
41   "datePublished":"2013-11-11",
42   "keywords":[
43     "An open-access collection of recordings of languages of Vietnam and neighbouring countries, including highly under-resourced languages",
44     "Fonds sonores, en accès libre, de langues du Vietnam et des pays voisins, y compris des langues « rares », particulièrement peu dotées au plan informatique"
45   ],
46   "creator":{
47     "@type":"Person",
48     "name":"Multimédia, Informations, Communication et Applications"
49   },
50   "contributor":{
51     "@type":"Person",
52     "name":"Multimédia, Informations, Communication et Applications"
53   },
54   "publisher":[
55     {
56       "@type":"Organization",
57       "name":"Laboratoire de langues et civilisations à tradition orale"
58     },
59     {
60       "@type":"Organization",
61       "name":"Multimédia, Informations, Communication et Applications"
62     }
63   ]
64 }</script>
65 </head>
```

Figure 25. Cocoon collection HTML metadata

Zotero reads this metadata and imports it to create a Zotero record as shown in Figure 26.

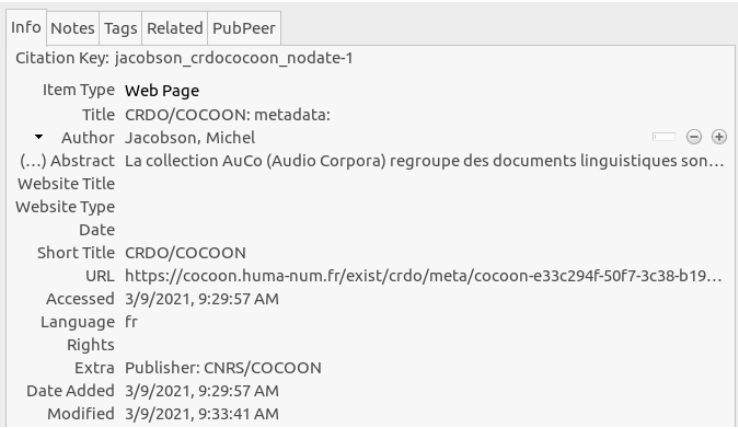

Figure 26. Cocoon item metadata import via embedded HTML metadata

The same analysis was performed for the audio artifact. The embedded metadata is shown in Figure 27. Zotero read the metadata and created the Zotero record shown in Figure 28.

```
<head>
  <link rel="icon" href="/favicon.ico" type="image/x-icon" />
  <link rel="shortcut icon" href="/favicon.ico" type="image/x-icon" />
  <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
  <link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
  <link rel="stylesheet" type="text/css" href="/ext/jquery-ui-1.12.1.min.css" />
  <link rel="stylesheet" type="text/css" href="/ext/bootstrap-3.3.7/css/bootstrap.min.css" />
  <link rel="stylesheet" type="text/css" href="/ext/bootstrap-3.3.7/css/bootstrap-theme.min.css" />
  <link rel="stylesheet" type="text/css" href="/exist/crdo/css/cocoon.css" />
  <link rel="unapi-server" type="application/xml" title="unAPI" href="/crdo_servlet/un-api" />
  <link rel="stylesheet" type="text/css" href="/ext/swiper-3.4.2/css/swiper.min.css" />
  <meta charset="utf-8" />
  <meta name="viewport" content="width=device-width, initial-scale=1" />
  <meta property="DC:author" content="Michel Jacobson" />
  <meta property="DC:copyright" content="CNRS/COCOON" />
  <meta property="DC:publisher" content="CNRS/COCOON" />
  <meta property="DC:subject" content="CRDO, COCOON, openarchive, OAI, OLAC, Linguistique, Archives orales, Archives sonores" />
  <meta name="author" content="Michel Jacobson" />
  <meta name="copyright" content="CNRS/COCOON" />
  <meta name="publisher" content="CNRS/COCOON" />
  <meta name="subject" content="CRDO, COCOON, openarchive, OAI, OLAC, Linguistique, Archives orales, Archives sonores" />
  <title>CRDO/COCOON: metadata: SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /Éa/. Speaker M7</title>
</head>
```

Figure 27. Cocoon item HTML metadata

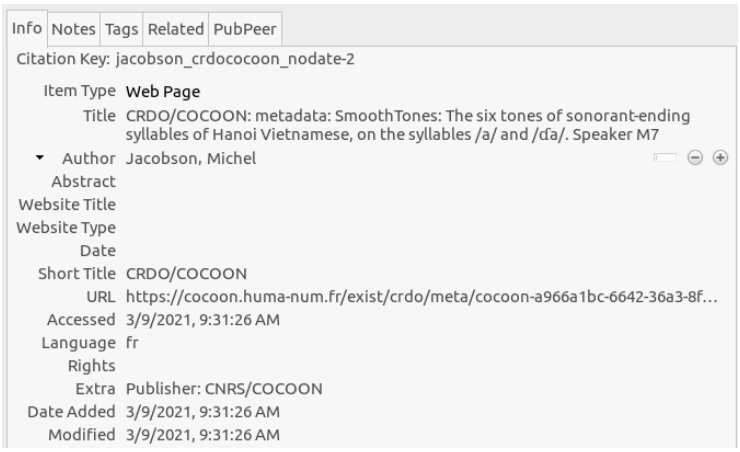

Figure 28. Cocoon item metadata import via embedded HTML metadata

There are several things to note about the bibliographic data imported by Zotero. The first is the perspective of the metadata. This can come into clearer perspective if we ask the question: *does the metadata apply to the web page or to the artifact that the page represents?* If the metadata applies to the web page itself, then it is mostly correct. However, the date and the website title are missing, and the title of the collection was not clearly presented. Using Dublin Core metadata, while not wrong, is unnecessary and not likely parsed by search engines looking at the web page as a web resource (as opposed to an archive or

scholarly resource). However, if the perspective is taken that this metadata is supposed to describe the resource the page is about, then much of the content is errant. Embedded metadata in HTML can be improved by making the page itself transparent to search engines and only presenting metadata on the artifacts and collections. Site based metadata can be presented to search engines on web page(s) at the root of the site.

The Villejuif interface does not have any scholarly embedded metadata as is shown in Figure 29 (collection) and Figure 31 (artifact). This means that Zotero recognizes both collection pages and artifact pages as only the web pages of the user interface. Zotero does pick up the DOIs on the artifact page and can then import the metadata from DataCite. Figure 30 shows the end result of the metadata processed from the collection page shown in Figure 29.

```
<head>
 <link rel="icon" type="image/png" href="https://pangloss.cnrs.fr/favicon.png">
 <meta charset="utf-8">
 <meta name="viewport" content="width=device-width, initial-scale=1.0, minimum-scale=1.0">
  <title>Pangloss Collection | Hanoi Dialect corpus</title>
 <link rel="stylesheet" type="text/css" href="https://pangloss.cnrs.fr/dist/main-pro.css">
 <script src="https://pangloss.cnrs.fr/dist/main.js"></script>
 <script>const CURRENTLANGUAGE = "en"</script>
 <script>
  // CONSTANTS
  const HOMEURL = 'https://pangloss.cnrs.fr/';
  const APIURL = 'https://pangloss.cnrs.fr//api/';
  const ASSETSURL = 'https://pangloss.cnrs.fr//assets/';
  const MODULESURL = 'https://pangloss.cnrs.fr//modules/';
  const URLPARAMETERS = '{"corpus":"Hanoi dialect","lang":"en","mode":"pro"}';
 </script>
</head>
```

Figure 29. Villejuif collection HTML metadata

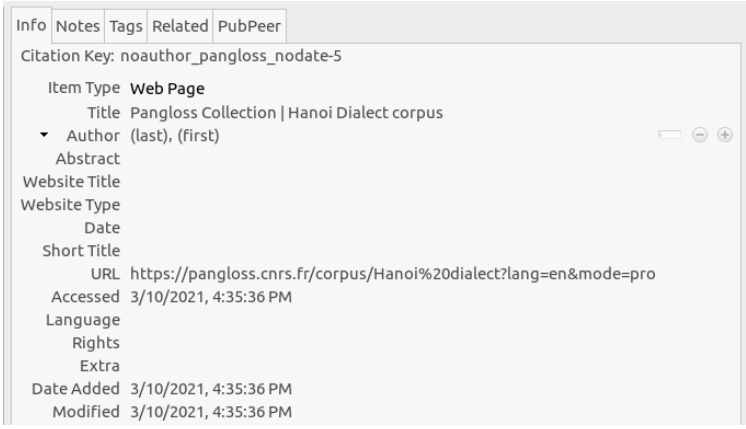

Figure 30. Villejuif collection record in Zotero from HTML metadata

```
<head>
 <link rel="icon" type="image/png" href="https://pangloss.cnrs.fr/favicon.png">
 <meta charset="utf-8">
 <meta name="viewport" content="width=device-width, initial-scale=1.0, minimum-scale=1.0">
  <title>Pangloss Collection | Vietnamese (Hanoi Dialect) corpus - SmoothTones: The six tones of sonorant-ending syllables of Han
 <link rel="stylesheet" type="text/css" href="https://pangloss.cnrs.fr/dist/main-pro.css">
 <script src="https://pangloss.cnrs.fr/dist/main.js"></script>
 <script>const CURRENTLANGUAGE = "en"</script>
 <script>
  // CONSTANTS
  const HOMEURL = 'https://pangloss.cnrs.fr/';
  const APIURL = 'https://pangloss.cnrs.fr//api/';
  const ASSETSURL = 'https://pangloss.cnrs.fr//assets/';
  const MODULESURL = 'https://pangloss.cnrs.fr//modules/';
  const URLPARAMETERS = '{"lang":"en","mode":"pro","oai_primary":"cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3","oai_sec
 </script>
</head>
```

Figure 31. Villejuif artifact page HTML metadata

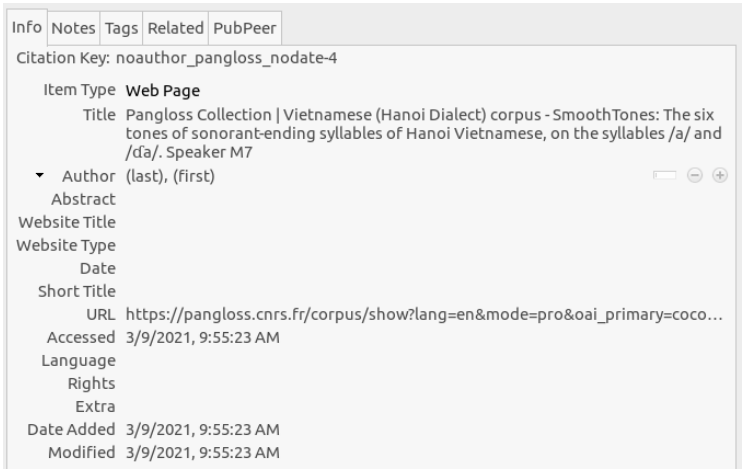

Figure 32. Villejuif artifact page record in Zotero from HTML metadata

In both the collection and the item level records created in Zotero, the HTML metadata is sparse. It does not matter if one is trying to craft a reference for the web pages themselves or for the artifact (or collection) which the pages represent. In both cases more metadata would be needed for Zotero to make a useful record. This is a case which certainly impacts relevance ranking in search engine results.

4.3.5.2 unAPI

Introduced at the end of Section 2.2, the unAPI is an HTML embedded link to a metadata description for the record described by a web page. Figures 25 and 27 show the head of each HTML document. Line 10, replicated in example (1), declares that there is an unAPI server at the location `https://cocoon.huma-num.fr/crdo_servlet/un-api`. Later on in the body section of the HTML code (not shown in this thesis) there is an abbreviation element with the ID of the artifact of which the associated record is to be queried. It appears as example (2).

(1)
```
<link rel="unapi-server" type="application/xml" title="unAPI" href="
/crdo_servlet/un-api" />
```

(2)
```
<abbr class="unapi-id" title="cocoon-e33c294f-50f7-3c38-b190-056e2
585a31d"> </abbr>
```

The unAPI crawler (built into Zotero) queries the root of the unAPI server at `https://cocoon.huma-num.fr/crdo_servlet/un-api`. It receives a response which tells the crawler which metadata formats the unAPI server can provide. The XML response from the unAPI server is shown in Figure 33, indicating that in the Cocoon case, MODS XML is the only bibliographic metadata format provided by the server. The unAPI crawler then crawls the location in example (3) reading the MODS XML file. The MODS file crawled is presented in Figure 34. The results of how Zotero interprets the MODS file are presented in Figure 35. Comparison with Figure 21 shows the different ways that Zotero is told to interpret the referenced object (collection). Zotero interprets the DOI metadata for the same collection as a **Journal Article** item type, whereas Zotero interprets the MODS import and creates a record with the **Document** item type.

(3) `https://cocoon.huma-num.fr/crdo_servlet/un-api?id=cocoon-e33c294f-50f7-3c38-b190-056e2585a31d&format=mods`

```
<?xml version='1.0' encoding='UTF-8'?>
<formats>
<format name='mods' type='application/xml' docs='http://www.loc.gov/standards/mods/v3/mods-3-5.xsd'/>
</formats>
```

Figure 33. Metadata types available via the unAPI server

```
<?xml version='1.0' encoding='UTF-8'?>
<mods xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xsi:schemaLocation='http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-5.xsd' xmlns:xlink="http://www.w3.org
   <titleInfo>
      <title xml:lang="fr">AuCo: corpus audio de langues du Vietnam et des pays voisins</title>
   </titleInfo>
   <titleInfo type="translated">
      <title xml:lang="vi">Âu Cơ: cơ sở dữ liệu âm thanh ngôn ngữ Việt Nam và các nước láng giềng</title>
   </titleInfo>
   <titleInfo type="translated">
      <title xml:lang="en">AuCo: Audio Corpora of languages of Vietnam and neighbouring countries</title>
   </titleInfo>
   <name>
      <namePart>Multimédia, Informations, Communication et Applications</namePart>
      <role>
         <roleTerm type="text">compiler</roleTerm>
      </role>
   </name>
   <typeOfResource collection="yes">mixed material</typeOfResource>
   <originInfo>
      <publisher>Multimédia, Informations, Communication et Applications</publisher>
      <publisher>Laboratoire de langues et civilisations à tradition orale</publisher>
      <dateIssued encoding="w3cdtf">2015-07-09T09:41:24+02:00</dateIssued>
   </originInfo>
   <physicalDescription>
      <form authority="marcform">electronic</form>
      <extent>502 items (speech recordings and transcriptions)</extent>
   </physicalDescription>
   <note xml:lang="fr">La collection AuCo (Audio Corpora) regroupe des documents linguistiques sonores de langues du Vietnam et des pays voisins, y compris dans des langues "rares" particulièrem
   <note xml:lang="en">The AuCo collection hosts audio recordings of language of Vietnam and neighboring countries, including data of endangered/little-endowed languages. AuCo stands for AudioC
   <note xml:lang="vi">Cơ sở dữ liệu ÂuCơ thu thập và lưu trữ các bản thu âm bằng các ngôn ngữ trên lãnh thổ Việt Nam và cả các nước láng giềng, bao gồm những ngôn ngữ đang có nguy cơ tuyệ
   <subject>
      <topic xml:lang="fr">Fonds sonores, en accès libre, de langues du Vietnam et des pays voisins, y compris des langues « rares », particulièrement peu dotées au plan informatique</topic>
   </subject>
   <subject>
      <topic xml:lang="en">An open-access collection of recordings of languages of Vietnam and neighbouring countries, including highly under-resourced languages</topic>
   </subject>
   <identifier type="oai">oai:crdo.vjf.cnrs.fr:cocoon-e33c294f-50f7-3c38-b190-056e2585a31d</identifier>
   <identifier type="ark">http://cocoon.huma-num.fr/exist/crdo/ark:/87895/1.17-509072</identifier>
   <location>
      <url usage="primary display">http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d</url>
      <url>Ancienne cote: crdo-COLLECTION_AUCO</url>
   </location>
   <accessCondition type="restriction on access">Freely accessible</accessCondition>
   <recordInfo>
      <recordContentSource>Cocoon (COllections de COrpus Oraux Numeriques)</recordContentSource>
      <recordOrigin>MODS record was transformed from a OLAC record (Based on Qualified dublin-core metadata) using XQuery specific code from Cocoon.</recordOrigin>
      <languageOfCataloging>
         <languageTerm authority="iso639-2">eng</languageTerm>
      </languageOfCataloging>
   </recordInfo>
</mods>
```

Figure 34. Cocoon collection unAPI response

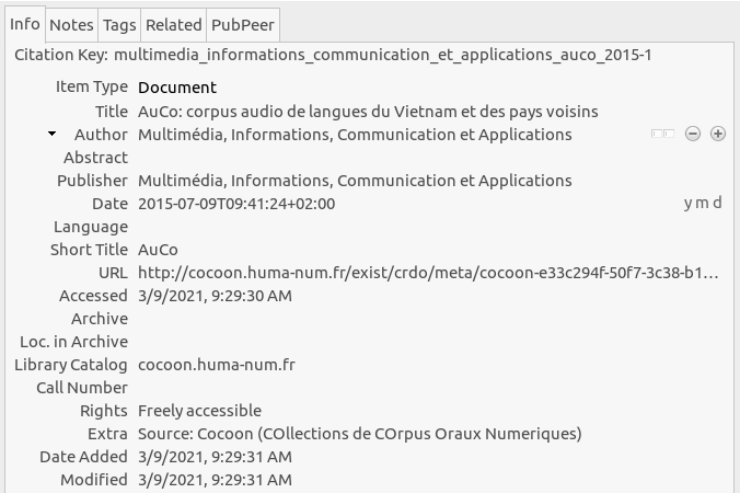

Figure 35. Zotero's interpretation of Cocoon collection unAPI metadata

In similar fashion example (4) shows the URL for the MODS record for the audio artifact queried in Section 4.3.4.2 and shown in Figure 24. The MODS XML is shown in Figure 36 and how Zotero interprets that response is shown in Figure 37.

(4) `https://cocoon.huma-num.fr/crdo_servlet/un-api?id=cocoon-e33c294f-50f7-3c38-b190-056e2585a31d&format=mods`

```
<?xml version='1.0' encoding='UTF-8'?>
<mods xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xsi:schemaLocation='http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-5.xsd' xmlns:xlink="http://www.w3.o
   <titleInfo>
      <title xml:lang="en">SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /đa/. Speaker M7</title>
   </titleInfo>
   <name>
      <namePart>Michaud, Alexis</namePart>
      <role>
         <roleTerm type="text">depositor</roleTerm>
      </role>
      <role>
         <roleTerm type="text">researcher</roleTerm>
      </role>
      <role>
         <roleTerm type="text">recorder</roleTerm>
      </role>
   </name>
   <name>
      <namePart>Nguyễn, Thị Lan</namePart>
      <role>
         <roleTerm type="text">researcher</roleTerm>
      </role>
      <role>
         <roleTerm type="text">interviewer</roleTerm>
      </role>
   </name>
   <name>
      <namePart>Trần, Đỗ Đạt</namePart>
      <role>
         <roleTerm type="text">researcher</roleTerm>
      </role>
   </name>
   <name>
      <namePart>Mạc, Đăng Khoa</namePart>
      <role>
         <roleTerm type="text">researcher</roleTerm>
      </role>
      <role>
         <roleTerm type="text">speaker</roleTerm>
      </role>
   </name>
   <typeOfResource>sound recording-nonmusical</typeOfResource>
   <genre authority="marcgt">speech</genre>
   <genre authority="olac" authorityURI="http://www.language-archives.org/OLAC/1.1/olac-linguistic-type.xsd">primary_text</genre>
   <originInfo>
      <publisher>Multimédia, Informations, Communication et Applications</publisher>
      <publisher>Laboratoire de langues et civilisations à tradition orale</publisher>
      <dateCreated encoding="w3cdtf" keyDate="true">2013-02-24</dateCreated>
      <dateIssued encoding="w3cdtf">2013-05-09T19:06:00+02:00</dateIssued>
      <place>
         <placeTerm type="text">Vietnam, Hanoi</placeTerm>
      </place>
      <place>
         <placeTerm type="code" authority="iso3166">VN</placeTerm>
      </place>
   </originInfo>
   <language>
      <languageTerm type="code" authority="iso639-3">vie</languageTerm>
      <languageTerm type="text">Vietnamese</languageTerm>
   </language>
   <physicalDescription>
      <form authority="marcform">electronic</form>
      <internetMediaType>audio/x-wav</internetMediaType>
      <extent>PT0H3M10S</extent>
   </physicalDescription>
   <note xml:lang="en">SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /đa/. Speaker M7</note>
   <subject authority="iso639-3">
      <topic>vie</topic>
   </subject>
   <subject authority="iso639-3">
      <topic>vie</topic>
   </subject>
   <subject>
      <topic>Hanoi dialect</topic>
   </subject>
   <subject>
      <topic>Vietnamese</topic>
   </subject>
   <relatedItem xlink:href="http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-ec7e2fd7-f50d-461a-be2f-d7f50de61a20"/>
   <relatedItem xlink:href="http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-052abec8-87f7-352c-b7ae-d2fd96444b8b"/>
   <relatedItem xlink:href="http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-af3bd0fd-2b33-3b0b-a6f1-49a7fc551eb1"/>
   <relatedItem xlink:href="http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d"/>
   <identifier type="oai">oai:crdo.vjf.cnrs.fr:cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3</identifier>
   <identifier type="ark">http://cocoon.huma-num.fr/exist/crdo/ark:/87895/1.17-346944</identifier>
   <location>
      <url usage="primary display">http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3</url>
      <url>http://cocoon.huma-num.fr/data/michaud/crdo-VIE_SmoothTones_M7.mp3</url>
      <url>Ancienne cote: crdo-VIE_M7_SMOOTHTONES_SOUND</url>
      <url>doi:10.24397/PANGLOSS-0004690</url>
      <url>http://cocoon.huma-num.fr/data/archi/masters/346944.wav</url>
   </location>
   <accessCondition>Copyright (c) Michaud, Alexis</accessCondition>
   <accessCondition type="restriction on access">Freely accessible</accessCondition>
   <accessCondition type="use and reproduction">
      <url>http://creativecommons.org/licenses/by-nc-sa/2.5/</url>
   </accessCondition>
   <recordInfo>
      <recordContentSource>Cocoon (COllections de COrpus Oraux Numeriques)</recordContentSource>
      <recordOrigin>MODS record was transformed from a OLAC record (Based on Qualified dublin-core metadata) using XQuery specific code from Cocoon.</recordOrigin>
      <languageOfCataloging>
         <languageTerm authority="iso639-2">eng</languageTerm>
      </languageOfCataloging>
   </recordInfo>
</mods>
```

Figure 36. Cocoon item unAPI response

Figure 37. Zotero's interpretation of Cocoon item unAPI metadata

This particular implementation of the unAPI import depends on Zotero's MODS translator for the conversion of the archive presented data. Other implementations of the unAPI could use BibTeX, RIS, or any other file translator that Zotero has. It is possible that the MODS metadata import could be improved by improving the translator. In general the same issues apply as were previously described in Table 6, e.g., the language values are not relevant for this artifact, the place is not the place of the publisher, and the running time of the artifact is not transferred to the record. I did not try to validate the Cocoon MODS files or run tests on other MODS files to examine the robustness of the Zotero translator. Such a validation check should be conducted before assuming that the Zotero software is faulty.

4.3.6 File download

No “download method” for bibliographic file download exists in either the Cocoon interface or the Villejuif interface. However, two of the types of bibliographic files (RIS and BibTeX) which users who seek files would expect to be able to download are available in text boxes via the Cocoon interface. A user can copy these outputs to their computer clipboard and paste from clipboard into Zotero, bypassing the download step. So, in my evaluation, I assess the Cocoon interface as having download files. Figure 38 shows the interface for copying the BibTeX code for a collection. Zotero will import this as a journal article as shown in Figure 39 because journal articles are the default document type in Zotero. The default document type is invoked because the BibTeX code accurately uses the BibTeX item type `@misc` when referencing a collection. BibTeX does not recognize an independent item type for collections. It is a limitation of the BibTeX data format. BibTeX also does not have a unique item type for audio artifacts so these also use `@misc`. This can be seen in Figure 40. Therefore, Zotero reads the BibTeX code and assigns the generic **Journal Article** type to both types.

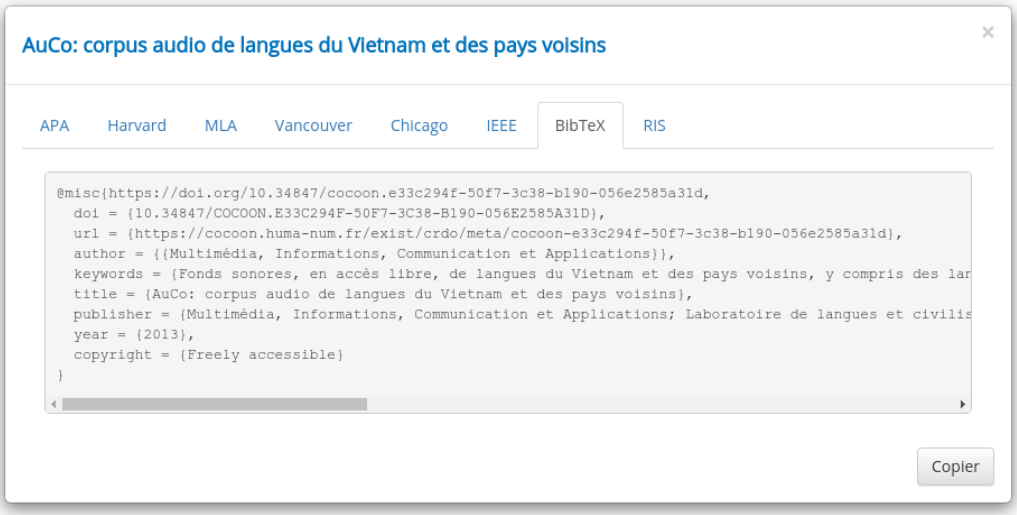

Figure 38. Metadata download options via the Cocoon interface as presented on a collection

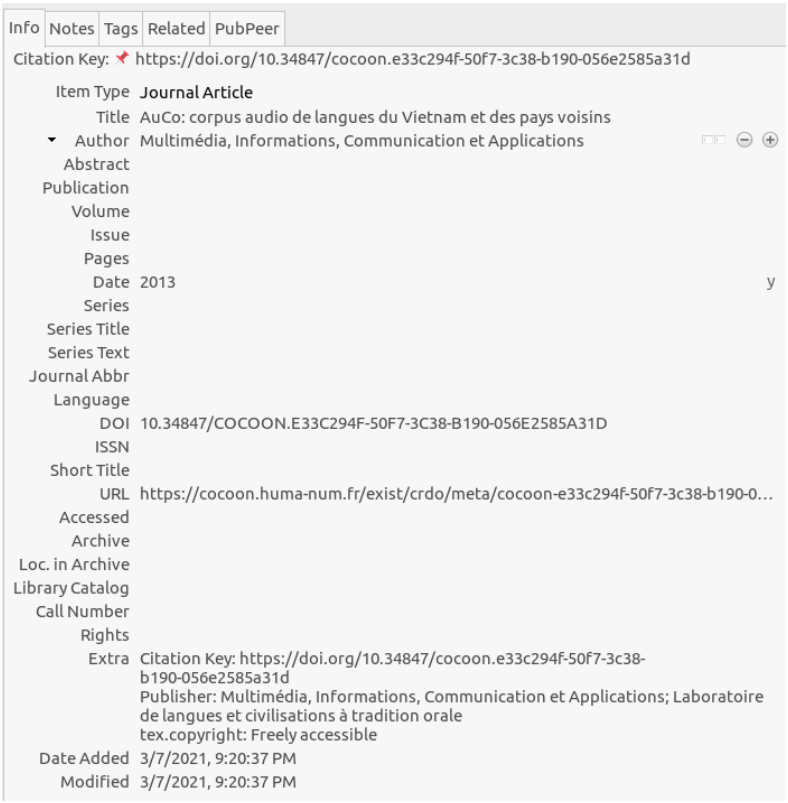

Figure 39. BibTeX import from the Cocoon interface as presented on a collection

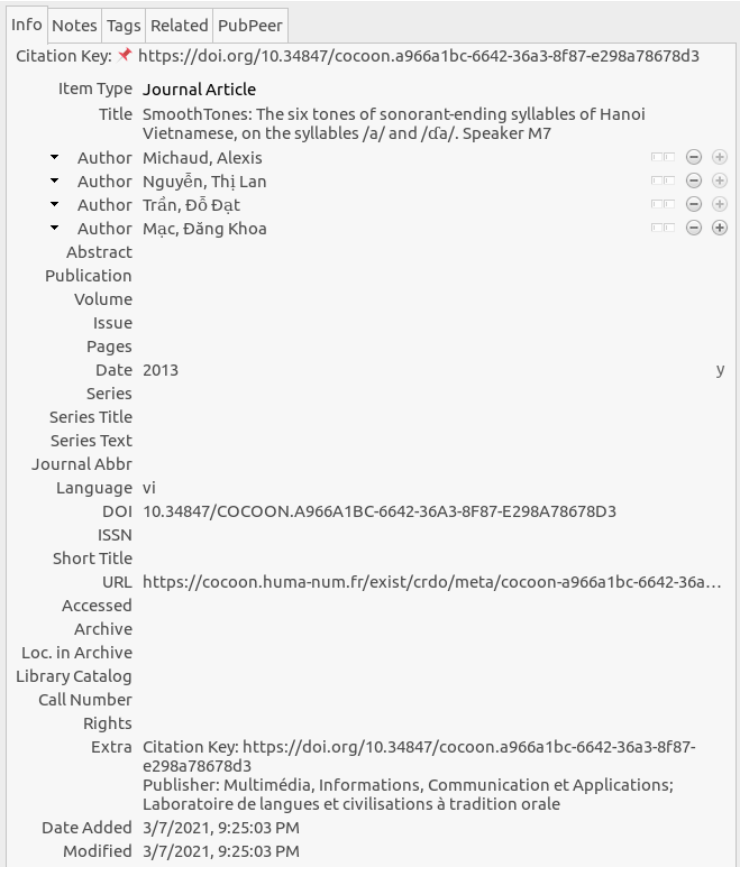

Figure 40. BibTeX import from the Cocoon interface as presented on an item

BibTeX and RIS are different data formats in the scope of their schema. RIS is generally considered to have a broader scope because it has more item types. This means that what is brought into Zotero from the archive is different than when importing a BibTeX file. The results of the RIS import can be seen in Figure 41 and Figure 42 where the Zotero record is shown for both collection and item levels. Note in Figure 41 the item type is again reduced to the default type of journal article but in Figure 42 the item type is recognized as an audio artifact.

Figure 41. RIS import from the Cocoon interface as presented on a collection

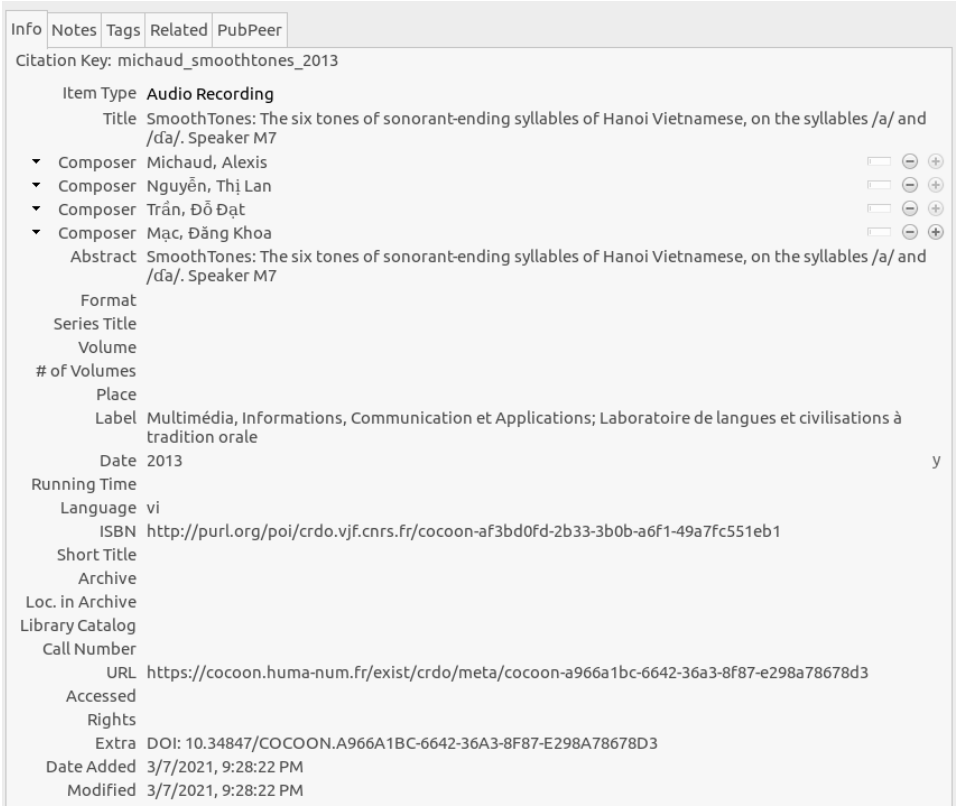

Figure 42. RIS import from the Cocoon interface as presented on a item

4.3.7 Complete or sufficient

In this section I present a comparison of what was imported to Zotero with what is needed to craft a reference in both *APA 6th edition* and *Chicago 17th edition Author-date* as discussed in Chapter 1. If there was a suggested reference for either the artifact or the collection I also discuss it. In this way we can see if the information transferred to Zotero is complete and if the provided suggested reference from the institution is sufficient for an informative reference. Of all five archives reviewed, Pangloss was the most prolific in its options to provide bibliographic metadata to archive users. For this reason, there are several subsections where each subsection addresses a single combination of style sheet (APA vs. Chicago), item type (collection vs. artifact), and interface (Cocoon vs. Villejuif). However, none of the methods included the total run time for the audio artifact transferred to Zotero—an important component when using *Chicago 17 Author-date* formatted references (Harper 2017:920 §15.57).

4.3.7.1 APA Collection—Cocoon

This section presents the variation created when using the various inputs from the Cocoon interface when trying to craft an *APA 6th edition* collection reference. For ease of reference the APA 6th edition collection reference template from VandenBos (2010:212) is provided again:

> Author, A. A. (Year, Month Day). Title of material. [Description of material]. Name of collection (Call number, Box number, File name or number, etc.). Name and location of repository.

Given the information at hand about the collection via the web interface, I would follow the APA template and manually craft a reference something like the following:[115]

> Multimédia, Informations, Communication et Applications. (2013). *AuCo: corpus audio de langues du Vietnam et des pays voisins // AuCo: Audio Corpora of languages of Vietnam and neighbouring countries // Âu Cơ: cơ sở dữ liệu âm thanh ngôn ngữ Việt Nam và các nước láng giềng* [393 artifacts; recordings and transcriptions]. Pangloss (ark:/87895/1.17-509072) Cocoon: *COllections de COrpus Oraux Numeriques*, Paris, France. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D

[115] Two things to note. First, different data sources provided different dates for the collection. Second, the collection has two languages in its title. *APA 6th edition*, is a Roman-script-only reference style, and it says to put translations in square brackets. However, this case is not one where I have translated the title (or someone else has translated the referenced work), rather the collection is issued with a title in multiple languages. *APA 6th edition* is not clear on what to do in cases where an item is issued in multiple languages. Typographically, I could take various approaches: use an em-dash between the two titles or use something like double pipes or double forward slashes. I chose double forward slashes.

Note that none of the following exports based on the bibliographic metadata imported to Zotero produce anything like the above hand crafted reference. The following is the "suggested APA style reference" provided via the Cocoon interface:

> Multimédia, Informations, Communication et Applications. (2013). *AuCo: corpus audio de langues du Vietnam et des pays voisins* (Version 1). Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D

In each of the following boxes the first reference is made exactly as the content was imported to Zotero. The second reference was created after **type: collection** was added to Zotero's Extra field to make the CSL exports use the collection reference pattern. In some cases there was no difference.

The following is the APA formatted output from the DOI input:

> Multimédia, Informations, Communication et Applications. (2013). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D
>
> Multimédia, Informations, Communication et Applications. (2013). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D

The following is the APA formatted output from the HTML input. Note that it varies significantly from the others as it likely is a reference to the web page rather than the object the web page presents:

Jacobson, M. (n.d.). *CRDO/COCOON: Metadata:* Retrieved March 9, 2021, from https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d

Jacobson, M. (n.d.). *CRDO/COCOON: Metadata:* CNRS/COCOON. Retrieved from https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d

The following is the APA formatted output from the unAPI (MODS) input:

Multimédia, Informations, Communication et Applications. (2015, July 9). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. Multimédia, Informations, Communication et Applications. Retrieved from http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d

Multimédia, Informations, Communication et Applications. (2015). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. Multimédia, Informations, Communication et Applications. Retrieved from http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d

The following is the APA formatted output from the RIS input. Notice that the name was imported as a personal name rather a corporate name and is therefore shortened as such:

> Multimédia, I., Communication et Applications. (2013). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D
>
> Multimédia, I., Communication et Applications. (2013). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D

The following is the APA formatted output from the BibTeX input:

> Multimédia, Informations, Communication et Applications. (2013). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D
>
> Multimédia, Informations, Communication et Applications. (2013). *AuCo: Corpus audio de langues du Vietnam et des pays voisins*. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D

These automated references fail to provide information about three components of the collection reference: the archival institution, the multi-lingual title of the collection, and the description of the collection. Arguably a fourth dynamic is missing, any information about the fonds level and below and their identifiers. Fonds are called series in Figure 4.

4.3.7.2 APA Collection—Villejuif

This section presents the variation created when using the various inputs from the Villejuif interface when trying to craft an APA 6th edition collection reference. Recall that the two collections (*AuCo: Corpus audio de langues du Vietnam et des pays voisins* and *Vietnamese (Hanoï dialect)*) do not represent the same thing (they have different content) and that

the Villejuif based interface did not assign DOIs to its collections, nor did it provide a wide range of metadata download options. Given the information at hand about the collection *Vietnamese (Hanoï dialect)*, I would follow the APA template and craft a reference something like the following:

> Michaud, A., Đỗ Đạt, T., Thị Lan, N. (2013). Vietnamese (Hanoï dialect) [21 recordings; recordings and transcriptions]. Pangloss. Centre André-Georges Haudricourt laboratoire LACITO, Villejuif, France. https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro

The following references are crafted by using the metadata imported to Zotero by looking at the website metadata. The first reference is made exactly as the content was imported to Zotero (not changing the item type). The second reference was created after **type: collection** (changing the item type to collection) was added to Zotero's Extra field to make the CSL exports use the collection reference pattern. The second reference does not output the date retrieved, even though the data is present. This is due to the patterns dictated by the style sheet as implemented via CSL.

> Pangloss Collection | Hanoi Dialect corpus. (n.d.). Retrieved March 9, 2021, from https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro
>
> Pangloss Collection | Hanoi Dialect corpus. (n.d.). Retrieved from https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro

Neither of the two automated APA outputs are anything like the information needed to make a well-formed reference. In this case it is not the fault of Zotero, but rather poverty of the information available to it.

4.3.7.3 Chicago Collection—Cocoon

This section presents the variation created when using the various inputs from the Cocoon interface when trying to craft a Chicago 17th edition collection reference. For ease of reference the Chicago 17th edition example reference from Harper (2017:§15.54) is provided again:

> Egmont Manuscripts. Phillipps Collection. University of Georgia Library.

Given the information at hand about the collection via the web interface I would follow the Chicago template and craft a reference something like the following:

> AuCo: Corpus audio de langues du Vietnam et des pays voisins // AuCo: Audio Corpora of languages of Vietnam and neighbouring countries // Âu Cơ: cơ sở dữ liệu âm thanh ngôn ngữ Việt Nam và các nước láng giềng. Cocoon: COllections de COrpus Oraux Numeriques. Laboratoire de Langues et civilisations à tradition orale & Laboratoire Ligérien de Linguistique. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D

None of the following exports based on the bibliographic metadata imported to Zotero produce anything like the above handcrafted reference. The Cocoon interface does provide access to a Chicago style pre-formatted reference, but it is not in the author-date variety of Chicago and the specific edition of Chicago was not specified, so it is not listed here. In each of the following boxes the first reference is made exactly as the content was imported to Zotero. The second reference was created after **type: collection** was added to Zotero's Extra field to make the CSL exports use the collection reference pattern. In some cases there was no difference. One difference to note is that those items which have been told to use the **type: collection** still use quotes whereas the handcrafted reference follows the patterns presented in the Chicago 17th edition and does not include quotes. This is likely a bug in the Chicago 17th edition CSL file. The following is the Chicago formatted output from the DOI input:

Multimédia, Informations, Communication et Applications. 2013. "AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins," November. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D.

Multimédia, Informations, Communication et Applications. 2013. "AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins." Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D.

The following is the Chicago formatted output from the HTML input. Note that it varies significantly from the others as it likely is a reference to the web page rather than the object the web page presents:

> Jacobson, Michel. n.d. “CRDO/COCOON: metadata:” CNRS/COCOON. Accessed March 9, 2021. https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d.
>
> Jacobson, Michel. n.d. “CRDO/COCOON: metadata:” CNRS/COCOON. Accessed March 9, 2021. https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d.

The following is the Chicago formatted output from the unAPI (MODS) input:

Multimédia, Informations, Communication et Applications. 2015. “AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins.” Multimédia, Informations, Communication et Applications. http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d.

Multimédia, Informations, Communication et Applications. 2015. “AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins.” Multimédia, Informations, Communication et Applications. http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-e33c294f-50f7-3c38-b190-056e2585a31d.

The following is the Chicago formatted output from the RIS input:

Multimédia, Informations, Communication et Applications. 2013. “AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins.” https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D.

Multimédia, Informations, Communication et Applications. 2013. “AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins.” Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D.

The following is the Chicago formatted output from the BibTeX input:

> Multimédia, Informations, Communication et Applications. 2013. “AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins.” Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D.
>
> Multimédia, Informations, Communication et Applications. 2013. “AuCo: Corpus Audio de Langues Du Vietnam et Des Pays Voisins.” https://doi.org/10.34847/COCOON.E33C294F-50F7-3C38-B190-056E2585A31D.

4.3.7.4 Chicago Collection—Villejuif

This section presents the output of the Villejuif interface when trying to craft a Chicago 17th edition collection reference. The Chicago 17th edition example reference is presented at the top of Section 4.3.7.3. Given the information at hand about the collection via the web interface I would follow the Chicago template and craft a reference something like the following:

> “Hanoi Dialect Corpus.”. Pangloss Collection. Laboratoire de langues et civilisations à tradition orale, Villejuif, France. https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro.

Note that none of the following exports based on the bibliographic metadata imported to Zotero produce anything like the above handcrafted reference. In the following box the first reference is made exactly as the content was imported to Zotero. The second reference was created after **type: collection** was added to Zotero’s Extra field to make the CSL exports use the collection reference pattern. In this case there was no difference. The following is the Chicago formatted output from the HTML input (which is the only input for collections in the Villejuif interface):

"Pangloss Collection | Hanoi Dialect Corpus." n.d. Accessed March 9, 2021. https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro.

"Pangloss Collection | Hanoi Dialect Corpus." n.d. Accessed March 9, 2021. https://pangloss.cnrs.fr/corpus/Hanoi%20dialect?lang=en&mode=pro.

4.3.7.5 APA artifact—Cocoon

This section presents the variation created when using the various inputs from the Cocoon interface when trying to craft an APA 6th edition audio artifact reference. Note that the audio artifact is found in an archival collection. There is not an example pattern for this in APA so we need to extract the pattern from the example for an item in a collection and an audio artifact. For ease of reference the APA 6th edition item in a collection and an audio artifact templates from VandenBos (2010:213, 209) and Skutley (2012:25) are provided again. The APA 6th edition examples of referenced collections components (VandenBos 2010:213):

Frank, L. K. (1935, February 4). [Letter to Robert M. Ogden]. Rockefeller Archive Center (GEB series 1.3, Box 371, Folder 3877), Tarrytown, NY.

Berliner, A. (1959). Notes for a lecture on reminiscences of Wundt and Leipzig. Anna Berliner Memoirs (Box M50). Archives of the History of American Psychology, University of Akron, Akron, OH.

The following is the APA 6th edition audio recording reference template (VandenBos 2010:209, Skutley 2012:25):

Writer, A. A. (copyright year). Title of song [Recorded by B. B. Artist if different from writer]. On *Title of album* [Medium of recording: CD, mp3, record, cassette, etc.]. Retrieved from http://xxxxx (Date of recording if different from song's copyright date)

Given the information at hand about the audio artifact via the web interface I would follow the APA guidance and craft a reference something like the following:[116]

> Michaud, A., Nguyễn, T. L., Trần, Đ. Đ., & Mạc, Đ. K. (2013). *SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7* [Vietnam, Hanoi]. AuCo: corpus audio de langues du Vietnam et des pays voisins // AuCo: Audio Corpora of languages of Vietnam and neighbouring countries // Âu Cơ: cơ sở dữ liệu âm thanh ngôn ngữ Việt Nam và các nước láng giềng [audio/.wav] (Pangloss, ark:/87895/1.17-509072). Cocoon: *COllections de COrpus Oraux Numeriques*, Laboratoire de Langues et civilisations à tradition orale, Laboratoire Ligérien de Linguistique & Multimédia, Informations, Communication et Applications. Paris, France. Retrieved from http://cocoon.humanum.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3

My handcrafted reference shows the audio artifact, the aggregate work (collection), the archive, the location in the archive, and the archive's location. The reference contains enough information to find the resource either by contacting the institution or through digital means. Note that this format is very similar to other aggregate work references like a chapter in an edited volume.

Recall that DOI import did not work for this item. So I skip it and move to present the output on the basis of the HTML input. Note that it varies significantly from the others as it likely is a reference to the web page rather than the object the web page presents:

> Jacobson, M. (n.d.). CRDO/COCOON: Metadata: SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7. Retrieved March 9, 2021, from https://cocoon.humanum.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3

[116] I find that the location of the medium of recording can be confusing. It is often the case that one can only access a particular unit of audio by first accessing its aggregate work which will be in a specific medium. However, with digital aggregate works, it is possible that the medium should be an attribute of the title (the artifact/auido track) instead of the aggregate because aggregate works might have different mediums.

The following is the output on data from the unAPI input. Note that with the unAPI import that names were not brought into the Zotero record with their first-name/last-name distinctions, therefore they are treated as corporate names by Zotero:

> Michaud, Alexis, Nguyễn, Thị Lan, Trần, Đỗ Đạt, & Mạc, Đăng Khoa. (2013). *SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7*. Vietnam, Hanoi: Multimédia, Informations, Communication et Applications. Retrieved from http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3

The following is the output on data from the RIS input. Note how unlike the import via the unAPI the names are imported correctly, however there is a difference in how the URL is presented:

> Michaud A., Nguyễn T. L., Trần Đ. Đ., & Mạc Đ. K. (2013). *SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7*. Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.A966A1BC-6642-36A3-8F87-E298A78678D3

The following is the output on data from the BibTeX input. Note how the names are also not well imported to Zotero, but at least one does not have to manually type the freakishly long DOI:

> Michaud, Alexis, Nguyễn, Thị Lan, Trần, Đỗ Đạt, & Mạc, Đăng Khoa. (2013). *SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7*. https://doi.org/10.34847/COCOON.A966A1BC-6642-36A3-8F87-E298A78678D3

4.3.7.6 APA artifact—Villejuif

This section presents the variation created when using the various inputs from the Villejuif interface when trying to craft an APA 6th edition audio artifact reference. Recall that the Villejuif interface does not have any embedded metadata, and it indicates to Zotero that it needs to get bibliographic metadata via the DOI and DataCite API. The Zotero crafted, APA formatted, bibliographic reference is presented below:

> Nguyễn T. L., Michaud A., Trần Đ. Đ., & Mạc Đ. K. (2013). *SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7* (p. PT0H3M10S) [Audio,x-wav]. Pangloss. https://doi.org/10.24397/PANGLOSS-0004690

The APA formatted reference based on the HTML Zotero reads is presented below. Notice the impact of passing variables (content between ampersands) via the URL.

> Pangloss Collection | Vietnamese (Hanoi Dialect) corpus—SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7. (n.d.). Retrieved March 15, 2021, from https://pangloss.cnrs.fr/corpus/show?lang=en&mode=pro&oai_primary=cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3&oai_secondary=cocoon-052abec8-87f7-352c-b7ae-d2fd96444b8b&optionTextTranscriptions=&optionTextTranslations=&optionSentenceTranscriptions=other&optionSentenceTranslations=en&optionWordTranscriptions=&optionWordTranslations=&optionMorphemeTranscriptions=&optionMorphemeTranslations=&optionNotes=

4.3.7.7 Chicago artifact—Cocoon

This section presents the variation created when using the various inputs from the Cocoon interface when trying to craft a Chicago 17th edition audio artifact reference.

Chicago 17th edition does not contain any unpublished research related audio artifacts, so I have replicated their examples here for archival audio artifacts:

> Coolidge, Calvin. [1920?]. "Equal Rights" (speech). In "American Leaders Speak: Recordings from World War I and the 1920 Election, 1918-1920." Library of Congress. Copy of an undated 78 rpm disc, RealAudio and WAV formats, 3:45. http://memory.loc.gov/ammem/nfhtml/.
>
> Holiday, Billie, vocalist. 1958. "I'm a Fool to Want You." By Joel Herron, Frank Sinatra, and Jack Wolf. Recorded February 20,1958, with Ray Ellis. Track 1 on Lady in Satin. Columbia CL 1157, 33x/3 rpm.

Given the information at hand about the audio artifact via the web interface I would follow the Chicago 17th edition guidance and craft a reference something like the following:

> Michaud, Alexis, Thị Lan Nguyễn, Đỗ Đạt Trần and Đăng Khoa Mạc. (2013). "SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7." In "AuCo: corpus audio de langues du Vietnam et des pays voisins // AuCo: Audio Corpora of languages of Vietnam and neighbouring countries // Âu Cơ: cơ sở dữ liệu âm thanh ngôn ngữ Việt Nam và các nước láng giềng." Pangloss, ark:/87895/1.17-509072. Cocoon: *COllections de COrpus Oraux Numeriques*, Laboratoire de Langues et civilisations à tradition orale, Laboratoire Ligérien de Linguistique & Multimédia, Informations, Communication et Applications. Paris, France. WAV copy, 00:03:10. Retrieved from http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3

My handcrafted reference shows the audio artifact, the aggregate work (collection), the archive, the location in the archive, and the archive's location. The reference contains enough information to find the resource either by contacting the institution or through digital means. Note that this format is very similar to other aggregate work references

like a chapter in an edited volume. I have not included the DOI because I have included the ARK ID which is shorter, and I have included the archive's URL.

Recall that DOI import did not work for this item. So I skip it and move to present the output on the basis of the HTML input. Note that it varies significantly from the others as it likely is a reference to the web page rather than the object the web page presents:

> Jacobson, Michel. n.d. "CRDO/COCOON: metadata: SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7." CNRS/COCOON. Accessed March 9, 2021. https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3.

The following is the output on data from the unAPI input. Note that with the unAPI import that names were not brought into the Zotero record with their first-name/last-name distinctions, but because they have commas in their names and they are listed in the Zotero record with the role performer they look like they fit with the desired style output:

> Michaud, Alexis, Nguyễn, Thị Lan, Trần, Đỗ Đạt, and Mạc, Đăng Khoa. 2013. SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7. Vietnam, Hanoi: Multimédia, Informations, Communication et Applications. http://cocoon.huma-num.fr/exist/crdo/meta/cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3.

The following is the output on data from the RIS input. The names are imported as composers on an audio recording item type. Chicago does not list those roles:

> SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7. 2013. Multimédia, Informations, Communication et Applications; Laboratoire de langues et civilisations à tradition orale. https://doi.org/10.34847/COCOON.A966A1BC-6642-36A3-8F87-E298A78678D3.

The following is the output on data from the BibTeX input. Note how the names are also not well imported to Zotero, but at least one does not have to manually type the freakishly long DOI:

> Michaud, Alexis, Nguyễn, Thị Lan, Trần, Đỗ Đạt, and Mạc, Đăng Khoa. 2013. "SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7." https://doi.org/10.34847/COCOON.A966A1BC-6642-36A3-8F87-E298A78678D3.

4.3.7.8 Chicago artifact—Villejuif

This section presents the variation created when using the two inputs from the Villejuif interface when trying to craft a Chicago 17th edition audio artifact reference. Recall that the Villejuif interface does not have any embedded metadata and indicates to Zotero that it needs to get bibliographic metadata via the DOI and DataCite API. The Zotero crafted, Chicago 17th edition formatted, bibliographic reference using the DOI import method is presented below:

> Nguyễn Thị Lan, Michaud Alexis, Trần Đỗ Đạt, and Mạc Đăng Khoa. 2013. SmoothTones: The six tones of sonorant-ending syllables of Hanoi Vietnamese, on the syllables /a/ and /ɗa/. Speaker M7. Audio,x-wav. Pangloss. https://doi.org/10.24397/PANGLOSS-0004690.

The Chicago 17th edition formatted reference based on the HTML Zotero reads is presented below. Notice the impact of passing variables (content between ampersands) via the URL.

> “Pangloss Collection | Vietnamese (Hanoi Dialect) Corpus - SmoothTones: The Six Tones of Sonorant-Ending Syllables of Hanoi Vietnamese, on the Syllables /a/ and /ɗa/. Speaker M7.” n.d. Accessed March 15, 2021. https://pangloss.cnrs.fr/corpus/show?lang=en&mode=pro&oai_primary=cocoon-a966a1bc-6642-36a3-8f87-e298a78678d3&oai_secondary=cocoon-052abec8-87f7-352c-b7ae-d2fd96444b8b&optionTextTranscriptions=&optionTextTranslations=&optionSentenceTranscriptions=other&optionSentenceTranslations=en&optionWordTranscriptions=&optionWordTranslations=&optionMorphemeTranscriptions=&optionMorphemeTranslations=&optionNotes=.

4.3.7.9 Pangloss summary

Even though Pangloss offers more methods than any of the other five archives survived by which to transfer bibliographic metadata, each method fails in some respect.

Sometimes it is clearly the limitation of the transmission format (as in the BibTeX cases), but in other cases there is a general lack of clarity on the part of publishing style sheets on what data component must be included for success, e.g., neither Chicago nor APA provide examples for how to reference audio artifacts which are part of field recordings. Style sheet ambiguities make technical implementations difficult.

4.4 SIL Language & Culture Archives

The SIL Language & Culture Archives[117] is SIL International's corporate repository for language related artifacts. SIL International is an NGO with its head offices in Dallas, Texas. The SIL Language & Culture Archives also has a physical holdings location in Dallas, Texas. The hierarchical nature of holdings at the L&CA is highly influenced by SIL's corporate architecture, which prior to the 2000's primarily functioned under a "franchise" business model, e.g., SIL-PNG, SIL Sudan, SIL Peru, etc., with a common supporting organization known as SIL International. SIL was founded in 1934 and so has a long corporate history (Aldridge & Simons 2018), but the existence of national and regional level business units within the SIL family of organizations has varied over time. The practice for decades was for each business unit to be responsible for its own language artifacts, as deemed necessary—including artifact inventory, retention, deaccession, and preservation. Today, this has a significant impact on how collections are managed and how the hierarchical system as a whole is structured. Prior to 1999 the SIL collections were managed as a bibliography rather than an archive. The bibliography was chiefly concerned with the preservation of references rather than artifacts. When viewed as a whole, the organization's historical practices and the natural development into an institutional repository have created a great deal of variation in the availability of items listed in the various collections. Nordmoe (2018)[118] suggests that the collections contain a 3:2 ratio of public to private content and 20:80 ratio of unpublished to published materials across a scoping of 68,000 + items. This means many artifacts in the SIL Institutional Repository are formally published somewhere, and many others are not available at all via open access venues.

Table 14 provides a summary of the import technologies currently available at SIL's Language & Culture Archives.

[117] See footnote 44 in chapter 4.
[118] Jeremy Nordmoe is the Director of the SIL Language & Culture Archives.

Table 14. Summary of support for Zotero at L&CA

	DOI	Embedded metadata	File import
Item	No	None / DOI detection	No

Metadata for items in the L&CA are not queryable via DOI. The archive interface does not provide any HTML embedded bibliographic metadata (Dublin Core, EPrints, Highwire Press, or COinS), nor any downloadable files of bibliographic metadata. This makes the task for collecting bibliographic metadata completely manual.

4.4.1 Technology infrastructure

Since 2011 SIL has used *DSpace*[119] for managing digital holdings. A data feed containing metadata approved for public view is provided to sil.org and other SIL websites which are served via *Drupal*.[120] Some bitstreams[121] from the holdings are also available at these sites. These Drupal based websites become the primary means of end-user engagement with the SIL Language & Culture Archives catalogue/content. Examples include:

- Mexico: https://mexico.sil.org/resources
- Cameroon: https://www.silcam.org/resources
- Peru: https://peru.sil.org/resources
- PNG: https://pnglanguages.sil.org/resources

In 2011–2012 designers on the sil.org project suggested the use of Drupal modules to make bibliographic metadata available via HTML embedding and downloadable files.[122] However, this work was not carried out as no clear return on investment (to SIL as a whole or specifically the L&CA) for the funds required to implement or maintain features could be articulated. In that same project, designers following Bird and Simons (2003a)

[119] Currently running version 6.3.
[120] https://www.drupal.org
[121] Bitstreams is a technical concept used in DSpace for the bits of a particular file. It is explained more in Section 4.6.2.
[122] The *biblio* module was originally suggested, as that was the model of choice for delivering this kind of functionality via Drupal. Biblio has now been superseded by *bibcite*: https://www.drupal.org/project/biblio; https://www.drupal.org/project/bibcite

suggested implementing a stable URI system for artifact referencing (such as the Handle[123] or DOI systems). However, counter use cases were provided in the design evaluation process. The biggest issue was *what should happen when an archive (or organization) wishes to repress, deaccession, or remove an artifact (inclusive of its associated records and relations)?* With an externally registered URL system, anonymity is relinquished. DOIs and Handles are resolvable identifiers. SIL does have an item ID system, where each item has a numeric ID, but it is not related to a stable URI structure.

4.4.2 Collections structure

Within the public-facing web presentations that SIL offers no collection information is shown to users. The L&CA uses DSpace as its back end (Nordmoe 2018), but since the DSpace structures are not visible, they are not discussed in this section.

4.4.3 Collections and artifacts reviewed

The L&CA does not display collection level data via `sil.org`, the public interface for the L&CA's holdings. So in this case all a user has to go on is what is presented about the artifact. The archive does link between related works when they are related via the Dublin Core **hasPart**/**isPartOf** relationships.[124, 125] A good example of how the L&CA links resources in aggregate works can be seen in the listing for Hartell (1993a) which not only has its subsections listed but also has a link to the French translation (Hartell 1993b). So, while the L&CA does not list collections, it does list aggregate works in particular cases. The case of Hartell's publications exhibits the fact that the L&CA does this both when aggregate works are in a part-whole relationship and when aggregate works are part of the same FRBR work, but are separate FRBR expressions.

I chose to review L&CA item 52216 (Figure 43). The artifact's title is *Audio Wordlist: Kamuku Survey, Cinda dialect, Danasabe village*.[126] It is not indicated to be part of any larger work or have any related entries. In fact, as discussed later in Section 6.3.2.2 this

[123] https://www.handle.net
[124] http://purl.org/dc/terms/hasPart
[125] http://purl.org/dc/terms/isPartOf
[126] Titles for works in language research, especially field work elicited audio, are notoriously difficult to craft informatively.

item comprises part of the same work reviewed in Section 4.2.3 as part of PARADISEC's *Kamuku wordlists*.

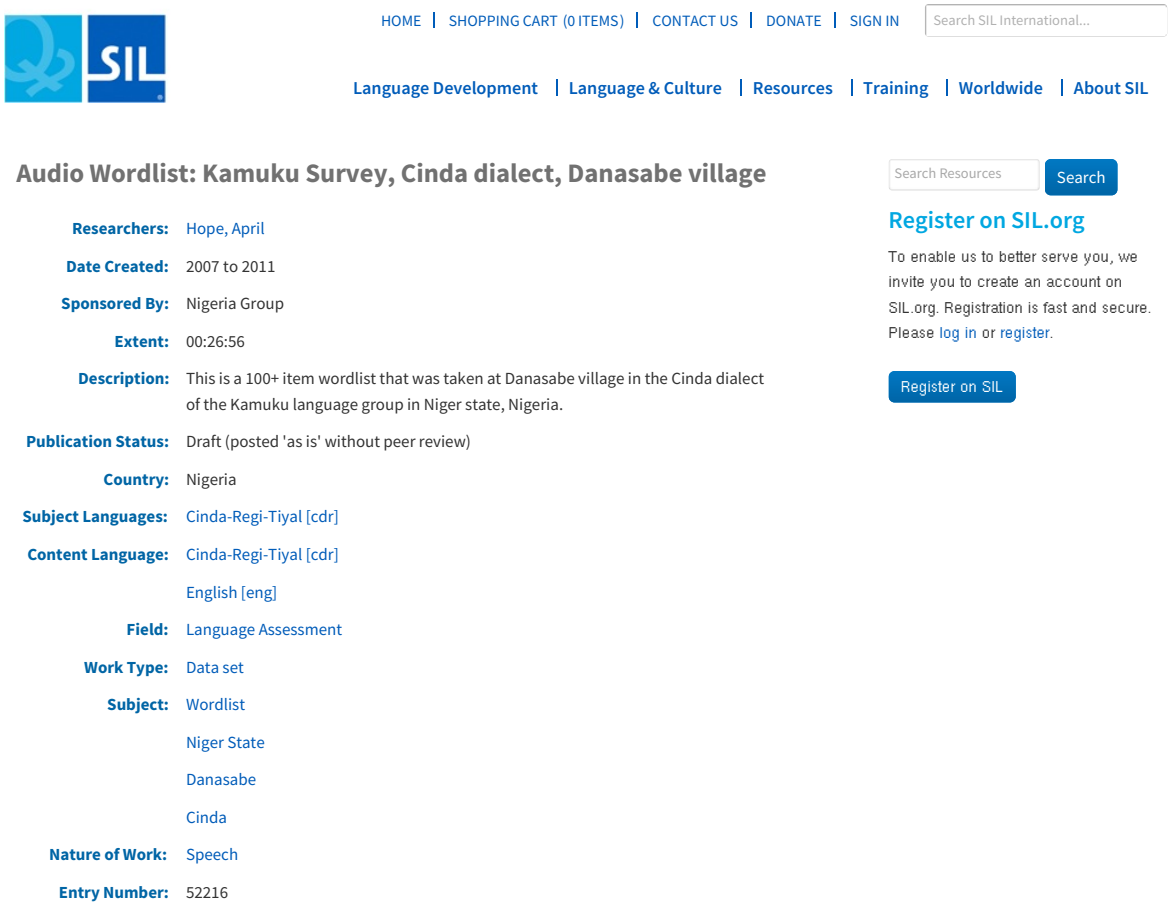

Figure 43. SIL L&CA item 52216 as seen on sil.org

4.4.4 DOI import

As a publisher SIL International does not subscribe to DOI services nor do any of its national level entities. However, many SIL staff publish in venues where DOIs are provided. These DOIs are listed in L&CA records when known. Zotero will use these DOIs if detected;[127] however, the metadata from these DOIs are not sourced from the L&CA. The reviewed artifacts had no DOIs.

[127] This can be tested with L&CA item 52727: https://www.sil.org/resources/archives/52727

4.4.5 Embedded metadata in HTML

As is seen in Figure 44 lines 8–12, SIL is meticulous about supplying HTML embedded metadata for Open Graph (a metadata schema created by Facebook and used by Twitter),[128] on each page of `sil.org`.[129] The Open Graph metadata can be identified with the `og:` prefix which starts the metadata property value. The L&CA provides no metadata for Google Scholar or other scholarly search indexers via embedded HTML metadata. Zotero detects the Open Graph metadata and categorizes the archive page as a web page on import. The imported record is displayed in Figure 45. If a DOI exists in the content of the web page, Zotero can see it and import the metadata from the DOI publisher. However, a user will need to right click the Zotero connector icon in the browser and choose the import via DOI option as is illustrated in Figure 3 in Chapter 2.

```
1  <head>
2    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
3  <link rel="shortcut icon" href="https://www.sil.org/sites/default/files/sil-favicon.png?qpgghx" type="image/png" />
4  <meta name="viewport" content="width=device-width, initial-scale=1, maximum-scale=1, minimum-scale=1, user-scalable=no" />
5  <meta name="generator" content="Drupal 7 (https://www.drupal.org)" />
6  <link rel="canonical" href="https://www.sil.org/resources/archives/52216" />
7  <link rel="shortlink" href="https://www.sil.org/node/62887" />
8  <meta property="og:site_name" content="SIL International" />
9  <meta property="og:type" content="article" />
10 <meta property="og:url" content="https://www.sil.org/resources/archives/52216" />
11 <meta property="og:title" content="Audio Wordlist: Kamuku Survey, Cinda dialect, Danasabe village" />
12 <meta property="og:updated_time" content="2020-12-02T05:44:01-06:00" />
13 <meta property="article:published_time" content="2013-02-01T20:11:05-06:00" />
14 <meta property="article:modified_time" content="2020-12-02T05:44:01-06:00" />
15   <title>Audio Wordlist: Kamuku Survey, Cinda dialect, Danasabe village | SIL International</title>
16 </head>
```

Figure 44. HTML code from SIL L&CA item 52216

[128] https://ogp.me
[129] Some non-relevant JavaScript was removed to clarify the presentation.

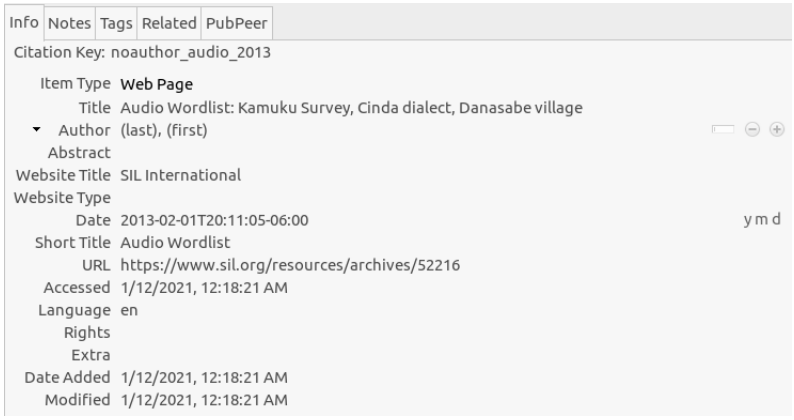

Figure 45. SIL L&CA item 52216 as imported from the publicly accessible `sil.org`

4.4.6 File download

The Language and Culture Archive does not provide a downloadable metadata file in a common format such as BibTeX, RIS, or MODS at any level.

4.4.7 Complete or sufficient

In general crafting a reference to be style compliant for either *APA 6th edition* or *Chicago 17th edition Author-date* based on the information the L&CA made available is difficult for several reasons, including that the medium is unknown, the date is not precise, the collection is unstated, and the context within another aggregate work is unstated. Zotero only sees the web page and not the work it represents.

4.4.7.1 Collection

Since the L&CA does not make collection information available, it is impossible for the average person who engages with the archive to actually create a collection citation.

4.4.7.2 Item

The HTML embedded metadata pulled into Zotero as shown in Figure 45 is insufficient to craft an *APA 6th edition* or *Chicago 17th edition Author-date* style reference. To craft an APA or Chicago reference one would need to look at the web page and manually build the reference in Zotero. If one tries to use the data imported to Zotero the reference will

follow the **Web Page** item type formats for the chosen CSL style. The *APA 6th edition* is shown below.

> Audio Wordlist: Kamuku Survey, Cinda dialect, Danasabe village. (2013, February 1). Retrieved January 12, 2021, from SIL International website: https://www.sil.org/resources/archives/52216

In addition to manually building the reference in Zotero, one would also need to assume that the reviewed artifact is indeed a single artifact and not an aggregate work (in a **isPartOf** relationship with another item). To do this one would just use the APA's audio recording template and Chicago's audio style template (both are shown in Chapter 1).

A handcrafted, APA 6th edition-ish style reference might look like the following. Note that there is no place in the APA template for an audio recording shown in chapter 1 to describe the archive, so I have diverged from the APA and added the archive's name and location in between the medium and the URL.

> Hope, A. (2007–2011). Audio Wordlist: Kamuku Survey, Cinda dialect, Danasabe village. [Medium unknown]. SIL Language & Culture Archives, Dallas, Texas. Retrieved from https://www.sil.org/resources/archives/52216

A handcrafted Chicago 17th edition audio recording references for a single item Author-date format might look like the following:

> Hope, April. 2007–2011. "Audio Wordlist: Kamuku Survey, Cinda dialect, Danasabe village". Language & Culture Archives. Unknown formats, 00:26:56. SIL Language & Culture Archives ID: 52216. https://www.sil.org/resources/archives/52216.

4.5 Endangered Languages Archive

The Endangered Languages Archive[130] is a digital archive embedded at the School of Oriental and African Studies (SOAS), London. It was started in 2004 and launched in 2005. It has been described in works by Munro & Nathan (2005), Wittenburg (2007), and Nathan (2011, 2013b, 2013a). Major contributions to ELAR come from scholars who are funded via the Endangered Language Program (also at SOAS).

Table 15 provides a summary of the import technologies currently available at ELAR.

Table 15. Summary of support for Zotero at ELAR

	DOI	**Embedded metadata**	**File import**
VuFind Collection	No	COinS	No
VuFind Item	No	COinS	No
WordPress Collection	No	None	No
WordPress Item	No	None	No

4.5.1 Technology infrastructure

During the initial stages of this study, the user interaction with the collections catalogue was managed via *VuFind*[131] an open source library catalog management tool. However, a new user interface was put in place in February 2021 as ELAR moved its collection management infrastructure to Preservica[132] with a *WordPress*[133] front end.

4.5.2 Collections structure

A significant motivation for structural segmentation in the hierarchical organization system at ELAR is grant or funding allocation, e.g., see the metadata displayed in Figure 46. The artifacts collected under the auspices of a funded project are compiled into a collection. A collection then contains bundles which may contain one or more artifacts; these artifacts may in some cases have multiple manifestations. Commonly **bundles**, the

[130] See footnote 45 in chapter 4.
[131] https://vufind.org/vufind
[132] https://preservica.com
[133] https://wordpress.org

only sub-grouping concept in-between a file and a collection, are groupings of artifacts which were created together or are derivatives of some common artifact. Both collections and bundles can have names, but names are not unique, therefore the collection ID or record ID (for bundles) is important (collection IDs and record IDs are distinct from deposit IDs). The hierarchical system under VuFind looks like: *SOAS Library* > *ELAR* > *Collections* > *Bundles* > *Files*. However, with the move to Preservica, it is not clear how dependent ELAR is on SOAS or University London library services and long term affiliation.[134]

4.5.3 Collections and artifacts reviewed

In reviewing ELAR I browsed a few collections but it was quicker and more convenient to just read the HTML code. Examples presented here show the same view in both the VuFind and the WordPress interfaces. They present the *Cicipu documentation*[135] collection (shown in Figure 46) and the *228 word list*[136] bundle (shown in Figure 47) which is part of the *Cicipu documentation* collection.

[134] It is not outside of archival industry norms for archives (or special collections) to move between institutions or to rotate their position within management structures of the same organization. A good example of this can be found in the history of `arxiv.org`. The repository started at Los Alamos National Laboratory (Butler 2001), moved to Cornell's Computer Science Department, moved again to be managed under Cornell's University Library and then moved back to Cornell's Computer Science Department (CornellCIS 2018).

[135] https://elar.soas.ac.uk/Collection/MPI97667

[136] https://elar.soas.ac.uk/Record/MPI538104

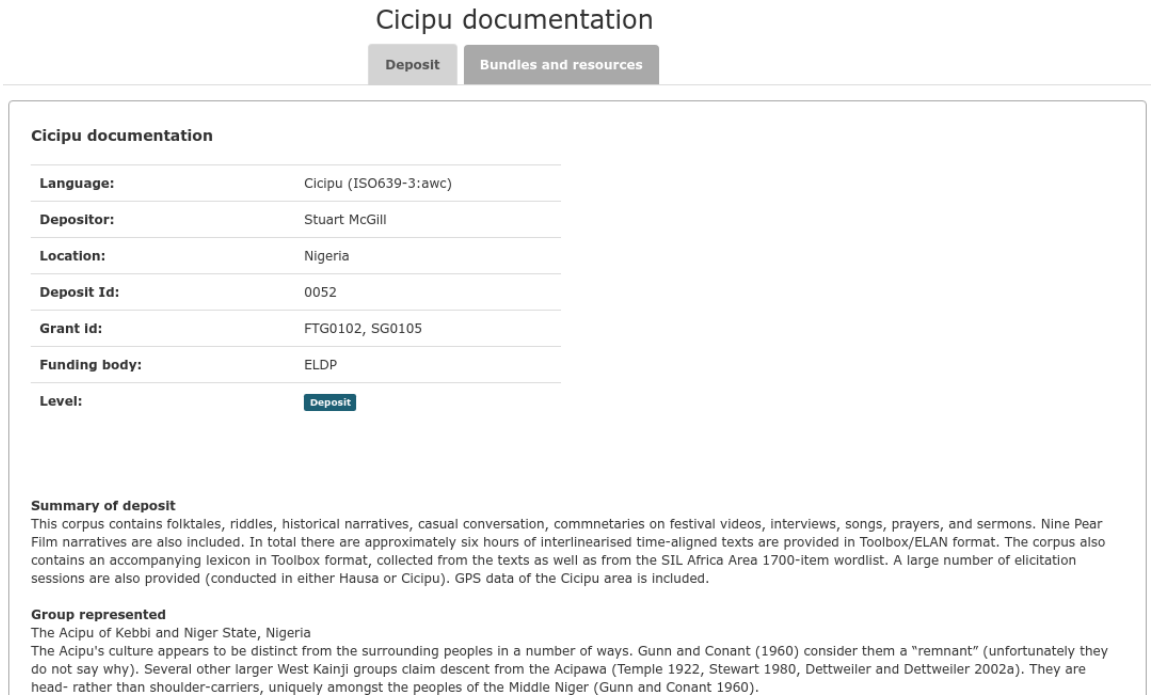

Figure 46. ELAR collection page in the VuFind interface

Notice how in Figure 46 the metadata shown for a collection is not the kind of bibliographic metadata needed for crafting a reference, e.g., no dates are present, the title is not labeled as title, no identifier is provided for this node of the collection, and the name of the creator is not specified. It is, however, helpful for determining other kinds of contextual information about the formation of the collection.

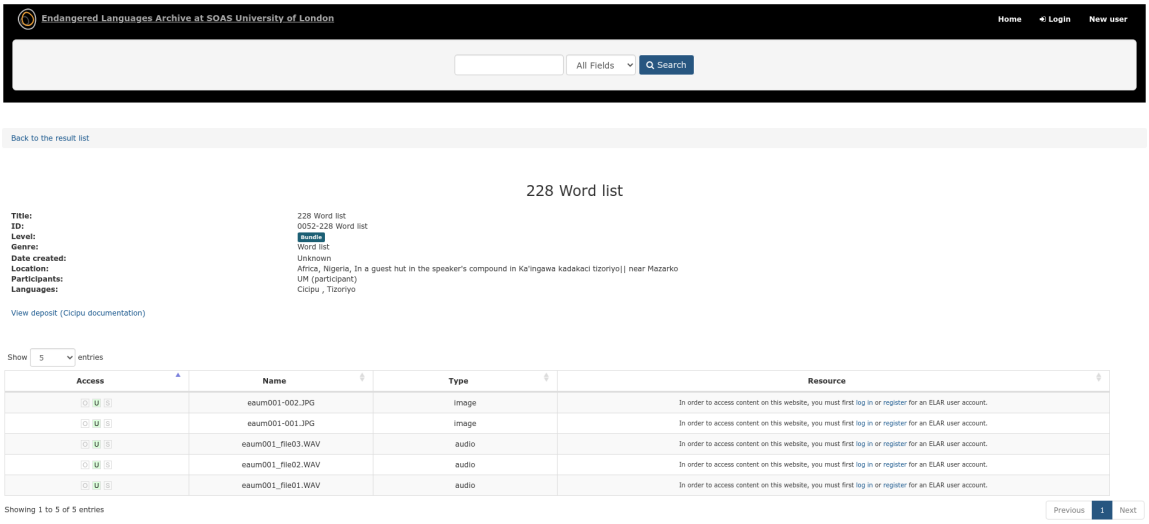

Figure 47. ELAR bundle page in the VuFind interface

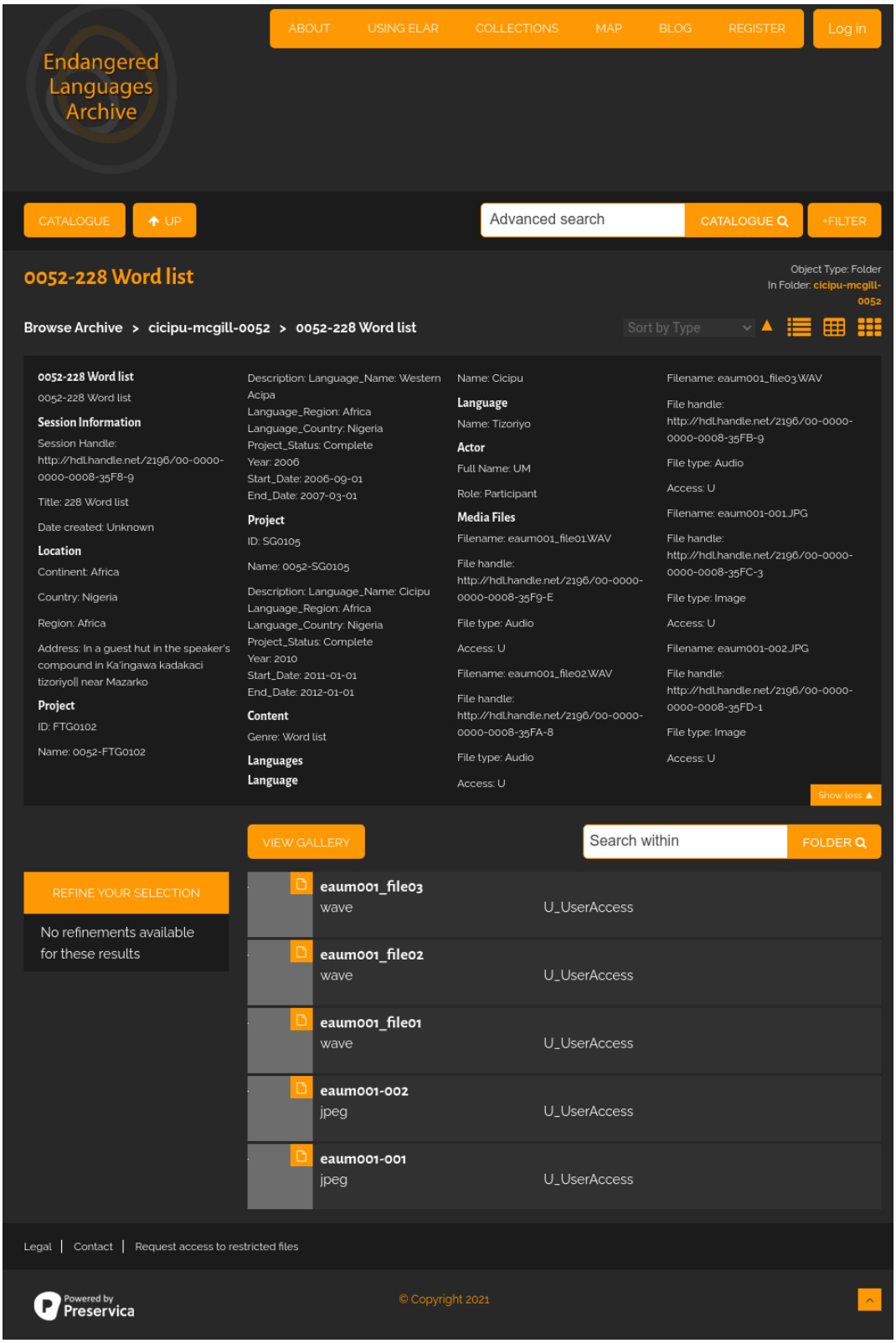

Figure 48. Metadata on ELAR bundle page in the WordPress interface

With the switch in platforms to WordPress, the visual presentation of metadata is much improved as shown in Figure 48. There is one small caveat: in the VuFind interface, collections and bundles had IDs. These IDs were nice and short. They were part of the URLs which were used in ELAR's "suggested reference" pattern where they said: "To refer to any data from the corpus, please cite the corpus in this way:"

> McGill, Stuart. 2012. Cicipu documentation. London: SOAS, Endangered Languages Archive. URL:https://elar.soas.ac.uk/Collection/MPI97667. Accessed on [insert date here].

These URLs still resolve at the moment, but a new URL structure was revealed and each collection, bundle (sometimes called a session), and file now has a persistent URL via the handle system (the handles do not demonstrate the logical structure of the collections). However, because the previous ID which was part of the URL is no-longer visible anywhere on the site, it is hard for people who use references to know if the object they are seeing on their screen is in fact the one which a previous author is referring to in their publication. This underscores the need for references to not just be clear with their digital pointers, but also with their textual descriptions of the collections and artifacts to which they are pointing.

4.5.4 DOI import

ELAR does not use DOIs so there is no option for Zotero users to import metadata via this method.

4.5.5 Embedded metadata in HTML

VuFind's software supports COinS. COinS code from ELAR is shown in Figure 49. This code triggers Zotero's detection of multiple importable objects (referenceable objects) as shown in Figure 50. COinS is designed to describe item types like books, book sections (chapters in an edited volume), journal articles, and dissertations, rather than collections[137] and datasets[138] (in the Dublin Core senses). Therefore, details needed for

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

crafting references for collections and datasets are not included in the COinS specification. This means that when Zotero reads the embedded metadata, reference details are not available to the Zotero user.

```
<span class="Z3988" title="ctx_ver&#x3D;Z39.88-2004&amp;ctx_enc&#x3D;info&#x25;3Aofi&#x25;
2Fenc&#x25;3AUTF-8&amp;rfr_id&#x3D;info&#x25;3Asid&#x25;
2Fvufind.svn.sourceforge.net&#x25;3Agenerator&amp;rft.title&#x3D;228-item&#x2B;word&#x2B;
list&#x2B;and&#x2B;Elicitation&#x2B;of&#x2B;various&#x2B;sentences&#x2B;in&#x2B;
Tidoddimo&amp;rft.date&#x3D;&amp;rft_val_fmt&#x3D;info&#x25;3Aofi&#x25;
2Ffmt&#x25;3Akev&#x25;3Amtx&#x25;3Adc&amp;rft.creator&#x3D;&amp;rft.format&#x3D;Bundle"></span>
```

Figure 49. COinS code at ELAR (line breaks added).

When multiple items are detected on a page Zotero allows the user to choose which they want to import.

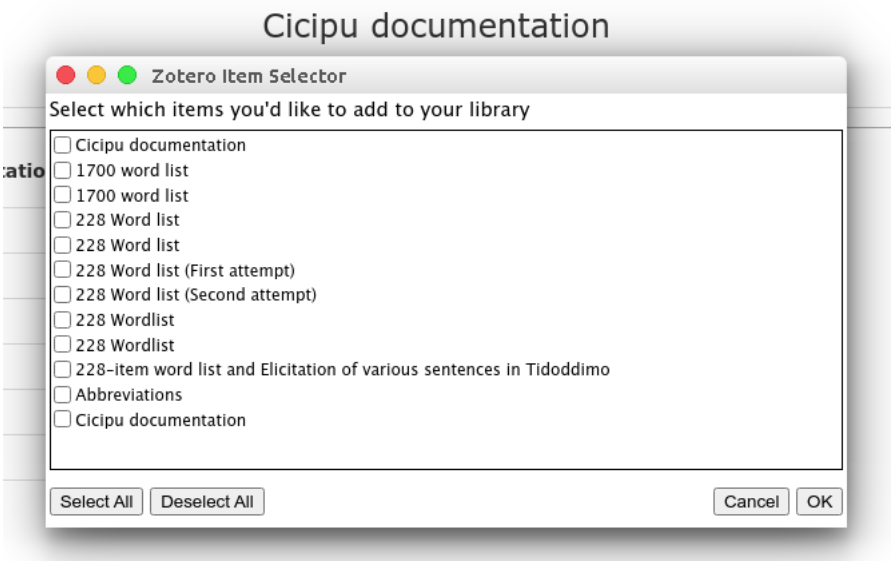

Figure 50. Import choice when multiple items are available

While the pre-2021 VuFind interface provided COinS integration, it was functionally useless as most objects in the archive were not within scope of the type that COinS supports. Therefore, for the types of items reviewed, there was no embedded metadata. The new WordPress interface did away with the COinS support. Currently with the WordPress

interface (shown in Figure 48), there is no embedded HTML metadata for Zotero to find. This means that Zotero will recognize these web pages as web pages, and not detect any artifact records (as shown in Figure 51).

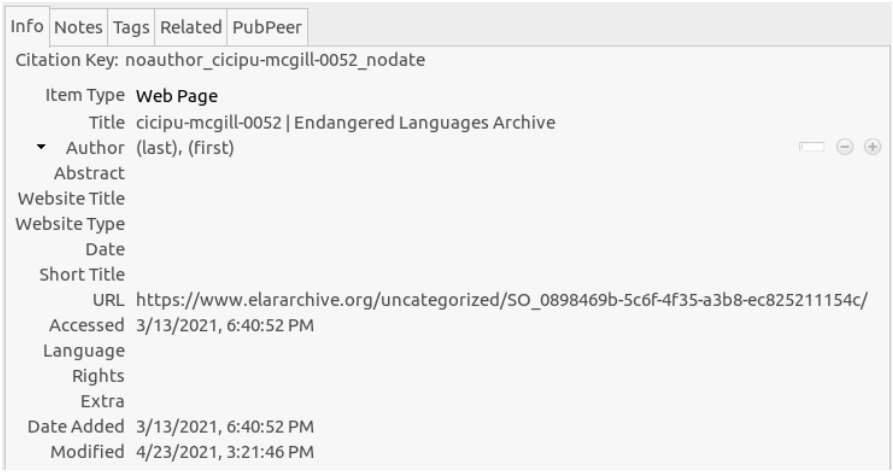

Figure 51. Zotero imports metadata from ELAR as Web Pages

4.5.6 File download

ELAR does not provide a downloadable metadata file in a common format such as BibTeX, RIS, or MODS at any level—collection, bundle, or file.

4.5.7 Complete or sufficient

The following references are crafted from interactions with the WordPress interface. Because Zotero detects ELAR web pages as the Zotero item type web page, only minimal metadata is discovered—even for web page references. Insufficient metadata is made available to both scholarly indexing tools and to Zotero. Given the information at hand about the collection via the web interface, I would follow the APA template and craft a reference something like the following:

> McGill, S. (2006–2007). Cicipu documentation. [Audio and video recordings with some transcriptions]. Endangered Languages Archive, London: SOAS. http://hdl.handle.net/2196/00-0000-0000-0001-7D83-D

The following is the Zotero produced APA reference for the collection based on the embedded metadata (web page title and URL) in the HTML detected by Zotero. Note that absent are discrete identifiable names, roles, dates, location of archive, identifier, and permanent URL. Further, the title is of the web page and not that of the collection.

> Cicipu-mcgill-0052 | Endangered Languages Archive. (n.d.). Retrieved March 13, 2021, from https://www.elararchive.org/uncategorized/SO_0898469b-5c6f-4f35-a3b8-ec825211154c

The following is the Zotero/CSL produced APA reference for the bundle based on the embedded metadata (web page title and URL) in the HTML detected by Zotero. Categorically Zotero detects the same information. Differences include the web page's title is computed differently by WordPress and the URL is different.

> 0052-228 Word list | Endangered Languages Archive. (n.d.). Retrieved March 13, 2021, from https://www.elararchive.org/uncategorized/SO_42f36d5f-44ed-40d0-820c-c28d06b14986

In the Cicipu documentation collection, there are several bundles which have the same title. So, in this case citing a bundle without a handle does not make the reference very clear to a reader. The VuFind interface had bundle IDs but these are not visible in the new WordPress ELAR user interface.

Given the information at hand about the collection via the web interface I would follow the Chicago template and craft a reference something like the following:

> Cicipu documentation. Endangered Languages Archive, London: SOAS. https://www.elararchive.org/uncategorized/SO_42f36d5f-44ed-40d0-820c-c28d06b14986

In contrast, the Zotero/CSL output in Chicago 17th edition from the same embedded metadata (web page title and URL) detected through HTML looks like the following for each collection (recall bundles are collections or aggregate works as well):[139]

> "Cicipu-Mcgill-0052 | Endangered Languages Archive." n.d. Accessed March 13, 2021. https://www.elararchive.org/uncategorized/SO_0898469b-5c6f-4f35-a3b8-ec825211154c.

The following is the Zotero/CSL produced Chicago 17th edition reference for the bundle based on the embedded metadata (web page title and URL) in the HTML detected by Zotero.

> "0052-228 Word List | Endangered Languages Archive." n.d. Accessed March 13, 2021. https://www.elararchive.org/uncategorized/SO_42f36d5f-44ed-40d0-820c-c28d06b14986.

In both cases (the collection and the bundle) what is being referenced is the web page not the collection. Zotero does not have any indication that a collection exists.

4.6 Kaipuleohone

Kaipuleohone[140] was established in 2008 at the University of Hawai'i. It is described by Albarillo and Thieberger (2009) and Berez (2013). It is a collection of ethnographic research materials at the University of Hawai'i. Kaipuleohone hosts a wide variety of research materials collected during the course of scholarly projects based at the university or conducted by university affiliated staff.

Table 16 provides a summary of the import technologies currently available at Kaipuleohone.

[139] In the new WordPress based interface, artifacts contained within bundles do have their own handles which can be used as IDs.

[140] See footnote 46 in chapter 4.

Table 16. Summary of support for Zotero at Kaipuleohone

	DOI	Embedded metadata	File import
Collection	No	None	No
Item	No	Yes - Dublin Core	No

4.6.1 Technology infrastructure

The collection is managed by the linguistics department while the software and infrastructure support is provided via the university library. The digital collection is hosted along with other university special collections using DSpace[141] (the same software that powers the SIL's L&CA discussed in Section 4.4).

4.6.2 Collections structure

One challenge faced by both traditional archives and all digital repositories is the decision on how to arrange the hierarchical systems—including a model which determines which metadata elements should be applied to each node in the hierarchy. Relationships with organizations providing digital infrastructure to special collections may dictate some prerequisites with regards to hierarchical systems including node structures and the ability to apply metadata to nodes. Software choices may lend themselves to certain types of hierarchical structures. The general DSpace architecture applies in the case of Kaipuleohone as it is a sub-community within the set of communities managed by the UHM linguistics department. Their hierarchical structure therefore looks like: *Communities/Sub-Communities* > *Collection* > *Item* > *bitstreams* (see Smith 2002:546 for further details).

4.6.3 Collections and artifacts reviewed

For this project I looked at two collections: *Blust Field Notebooks*[142] and *Marshallese Language from the 1950s*.[143] Web pages representing a collection have a unique ID via the Handle system. The collection page for *Marshallese Language from the 1950s* (shown in

[141] At the time of writing the university was using DSpace version 5.7.
[142] https://hdl.handle.net/10125/33115
[143] https://hdl.handle.net/10125/27416

Figure 52) includes a very helpful description introducing the major contributors to the collection.

Figure 52. Collection description at Kaipuleohone

At the item level, I looked at item BB1-018[144] within the *Marshallese Language from the 1950s* collection (as shown in Figure 53).

[144] https://scholarspace.manoa.hawaii.edu/handle/10125/30788

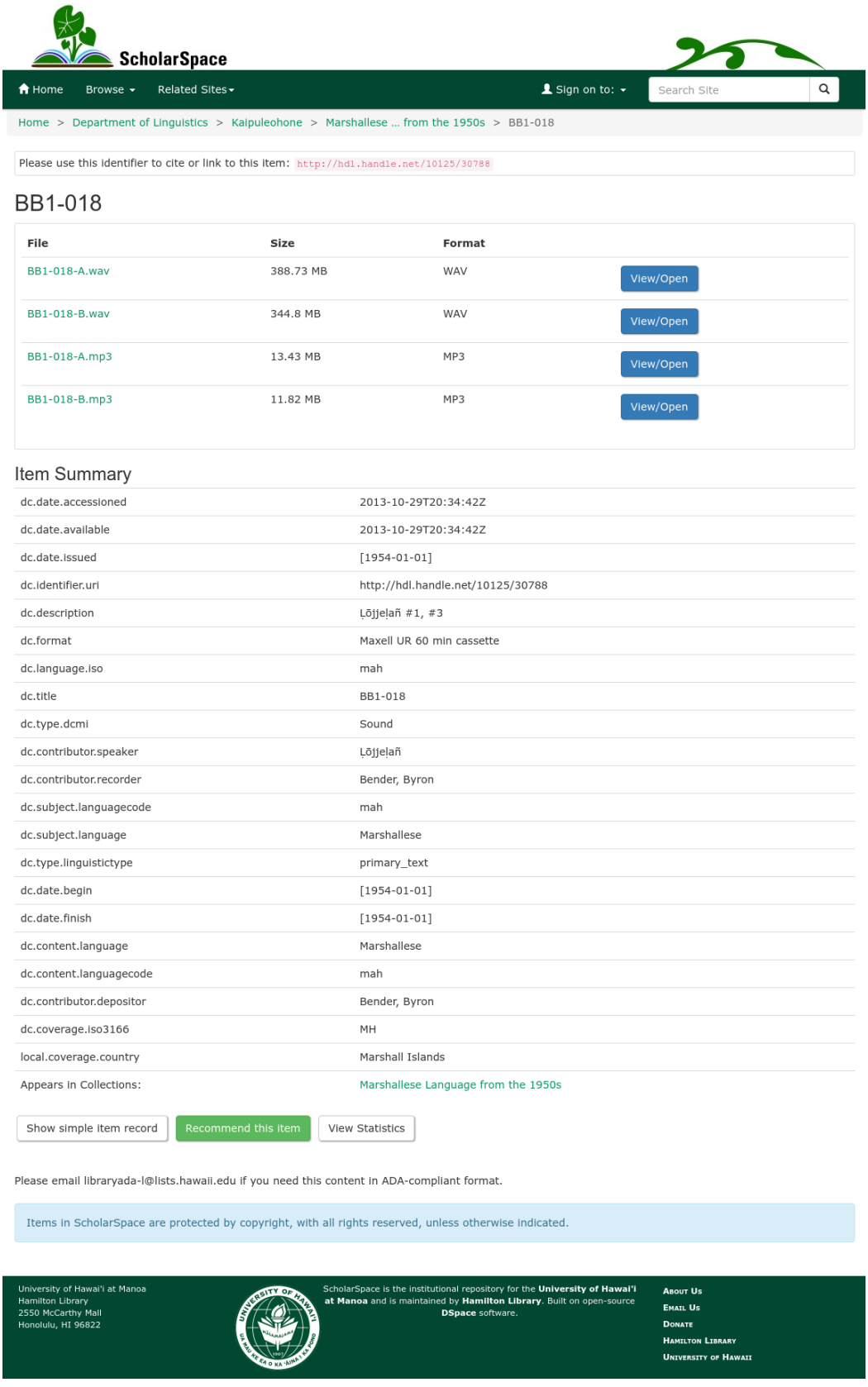

Figure 53. Item view at Kaipuleohone

4.6.4 DOI import

Kaipuleohone does not use DOIs so there is no option for Zotero users to import metadata via this method.

4.6.5 Embedded metadata in HTML

Collection pages do not contain any HTML embedded metadata about a collection. Therefore, they import as simple web pages to Zotero requiring an author to collect their own metadata manually for crafting a reference for a collection.[145]

Item level pages are different. In contrast to all the other archives discussed in this thesis, DSpace at Kaipuleohone provides rich embedded Dublin Core metadata at the item level (DSpace sense of **item**). This can be seen in Figure 54.

```
<!DOCTYPE html>
<html lang="en">
    <head>
        <title>ScholarSpace at University of Hawaii at Manoa: BB1-018</title>
        <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
        <meta name="Generator" content="DSpace 5.7" />

<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DCTERMS.dateAccepted" content="2013-10-29T20:34:42Z" scheme="DCTERMS.W3CDTF" />
<meta name="DCTERMS.available" content="2013-10-29T20:34:42Z" scheme="DCTERMS.W3CDTF" />
<meta name="DCTERMS.issued" content="[1954-01-01]" xml:lang="en_US" scheme="DCTERMS.W3CDTF" />
<meta name="DC.identifier" content="http://hdl.handle.net/10125/30788" scheme="DCTERMS.URI" />
<meta name="DC.description" content="Ļōjjeļañ #1, #3" xml:lang="en_US" />
<meta name="DC.format" content="Maxell UR 60 min cassette" xml:lang="en_US" />
<meta name="DC.language" content="mah" xml:lang="en_US" scheme="DCTERMS.RFC1766" />
<meta name="DC.title" content="BB1-018" xml:lang="en_US" />
<meta name="DC.type" content="Sound" xml:lang="en_US" />
<meta name="DC.contributor" content="Ļōjjeļañ" xml:lang="en_US" />
<meta name="DC.contributor" content="Bender, Byron" xml:lang="en_US" />
<meta name="DC.subject" content="mah" xml:lang="en_US" />
<meta name="DC.subject" content="Marshallese" xml:lang="en_US" />
<meta name="DC.type" content="primary_text" xml:lang="en_US" />
<meta name="DC.date" content="[1954-01-01]" xml:lang="en_US" scheme="DCTERMS.W3CDTF" />
<meta name="DC.date" content="[1954-01-01]" xml:lang="en_US" scheme="DCTERMS.W3CDTF" />
<meta name="DC.contributor" content="Bender, Byron" xml:lang="en_US" />

<meta name="citation_date" content="[1954-01-01]" />
<meta name="citation_abstract_html_url" content="http://scholarspace.manoa.hawaii.edu/handle/10125/30788" />
<meta name="citation_language" content="mah" />
<meta name="citation_pdf_url" content="http://scholarspace.manoa.hawaii.edu/bitstream/10125/30788/1/BB1-018-A.wav" />
<meta name="citation_title" content="BB1-018" />
```

Figure 54. Embedded metadata code from DSpace (non-relevant code removed).

[145] This would be done by indicating the CSL variable in the **Extra** field as indicated in Figure 14. Any CSL style with a pattern defined for **collection** shows output results in authored documents. See discussion about the **Extra** field in Section 4.1.4.

Figure 55. View of Item BB1-018 imported to Zotero via embedded metadata.

However, even with this rich Dublin Core metadata, in the header of the web page, most of the metadata is not transferable to the Zotero user. There are likely several reasons for lack of successful transfer. The most likely reason is the presence of two `DC.type` elements (lines 18 and 23).[146] With two `DC.type` elements, declared Zotero defaults to its default document type—journal article. Daniel Ishimitsu of UHM Hamilton Library suggested (p.c. 2019) that this is because the embedded metadata was chosen to increase the ScholarSpace profile in Google Scholar, and compatibility with Zotero was not a design requirement at the time of implementation. However, even with Google Scholar as the target, several metadata values are just poor form. Dates are not in their declared formats. For example, in Figure 54 lines 12, 24, and 25, dates should not have square

[146] The presence of `priamry_text` in line 23 is a value from OLAC metadata, and should not be included in the HTML header as there are not any other services that read OLAC metadata.

brackets around them per the W3CDTF standard for indicating dates.[147] Square brackets around dates in some library cataloging traditions have been used to indicate that the date was hand-written on the material. However, even if these brackets appear in the source record, they should be removed before being put in the HTML header metadata, as they are not a valid format in W3CDTF, Google Scholar, or useful in bibliographic reference managers such as Zotero.[148] The noise created by using various date formats and techniques like square brackets is a broadly encountered problem by metadata harvesters and is not unique to Kaipuleohone (Toves & Hickey 2014).[149] However, conformity to modern cataloging standards like RDA and clean transformations of metadata from in-house storage practices to declared metadata standards improves interoperability.

Other data points to consider are language tags. There are three types of language tags in Figure 54. The first appears in line 2 `<html lang="en">` as part of the HTML tag.[150] This declares that the foundational language of the document is English. This code is done correctly. However, each of the meta tags in lines 12, 14–26 contain `xml:lang="en_US"`. This is in most cases not needed because it is saying that the language of the metadata is English, which is the default language of the document. However, if the metadata was not in English, then presumably some other value would be able to be put in place of `en_US`. More egregious than being redundant is that the code value is wrong. When used, HTML5 calls for the use of BCP47 codes in this position; `en_US` is a locale code. BCP47 codes have a hyphen/minus sign instead of an underscore.[151] The correct code would be `en-US`. Lines 14 and 19 are the only lines which might benefit from a code in this position, but it does not seem that the software has detected that the values of these fields are not in English. The third language code which appears on line 16 is declared to be in the controlled vocabulary of RFC1766.[152] However, "mah" is not a valid language code

[147] https://www.w3.org/TR/NOTE-datetime.

[148] As is shown in Section 4.6.7.2 Zotero does strip out the square brackets when creating a reference.

[149] Though often described as a point of variation, specific examples from institutional repositories on date variation are hard to pin down as research on metadata quality suffers from a clear definition of "quality", and most metadata quality studies have a broad overview. However, Klinke (2018) found 4,320 date patterns across 472,669 objects in The Museum of Modern Art's records (New York) and Kräutli (2016:76ff) discusses issues with variation and digital collection artifact presentation based on records in the Victoria and Albert Museum collections.

[150] https://www.w3.org/International/tutorials/language-decl

[151] https://unicode-org.github.io/cldr/ldml/tr35.html#Unicode_Locale_Identifier_CLDR_to_BCP_47

[152] https://tools.ietf.org/html/rfc1766

under RFC1766. One possible explanation is that the option to embed metadata comes from the DSpace theme, rather than being specifically coded for this DSpace community and its metadata terms. Within the DSpace community (Kaipuleohone), it is likely that catalogers use ISO 639-3 instead while the DSpace theme (or possibly even the rest of the Hamilton Library) uses RFC1766.[153]

4.6.6 File download

Kaipuleohone does not provide a downloadable metadata file in a common format such as BibTeX, RIS, or MODS at any level—collection, item, or bitstream.

4.6.7 Complete or sufficient

In this section I present a comparison of what was imported to Zotero with what is needed to craft a reference in both *APA 6th edition* and *Chicago 17th edition Author-date* (as previously presented in Chapter 1). Kaipuleohone does not provide a suggested reference; so unlike ELAR and PARADISEC, there is no institutional reference to compare. The artifact reviewed is interesting and different from other audio artifacts discussed in this thesis. In contrast to the others and in line with the additional information given in the collection description (shown in Figure 52), I would consider classifying this audio artifact as an interview. Reference styles like *APA 6th edition* and *Chicago 17th edition Author-date* both have layouts with special considerations for interviews. These interview formats list the roles of those engaged in the recording, the medium of the recording, and the archive. The following references were both produced using the Zotero interview template. If I were to reference these materials, I would craft my reference to look like those presented below. The following is the APA 6th edition interview reference (VandenBos 2010:213–214 #69):

[153] Dublin Core allows for language codes from ISO 639-3, ISO 639-2, and RFC1766. RFC1766 is an earlier language code standard to which Dublin Core makes reference. Dublin Core would have ideally referenced BCP47 rather than a specific RFC document as RFC documents are obsoleted from time to time with more current documents, but this is not likely to happen for another 25 years, at which time ISO 639-3 may have run out of codes for new languages, and a new code standard will be implemented. When crafting technical standards, referencing BCP47 is often preferred to referencing ISO 639-3 or specific RFC documents as it provides a stable reference point and allows for evolution of the RFC documents to occur seamlessly.

Ļōjjeļañ. (1954). BB1-018 (B. Bender, Interviewer) [Audio/wav digitized from cassette tape]. ScholarSpace at The University of Hawai‘i at Mānoa, Honolulu, HI (Kaipuleohone, Marshallese Language from the 1950s). Retrieved from http://scholarspace.manoa.hawaii.edu/handle/10125/30788

The following is the Chicago 17th edition interview reference in the Author-date format (Harper 2017:§14.211):

Ļōjjeļañ. 1954. BB1-018 Interview by Byron Bender. Audio/wav digitized from cassette tape. Kaipuleohone, Marshallese Language from the 1950s. ScholarSpace at The University of Hawai‘i at Mānoa, Honolulu, HI. http://scholarspace.manoa.hawaii.edu/handle/10125/30788.

In the following subsections, the collection and audio artifact references are presented based on the data that Zotero detected.

4.6.7.1 Collection

The following is the APA 6th edition collection reference output based on the data that Zotero detected:

ScholarSpace at University of Hawaii at Manoa: Marshallese Language from the 1950s. (n.d.). Retrieved January 19, 2021, from https://scholarspace.manoa.hawaii.edu/handle/10125/27416

The following is the Chicago 17th edition collection reference from Zotero in Author-date format:

“ScholarSpace at University of Hawaii at Manoa: Marshallese Language from the 1950s.” n.d. Accessed January 19, 2021. https://scholarspace.manoa.hawaii.edu/handle/10125/27416.

Neither of these collection references conform to the style guides. What is shown is basic web page bibliographic metadata. The archive has aptly enabled the relevant pages, creating a fluke where the Zotero output is reasonably close to a reference for a collection. For example, the collection title, the institutional repository, and the university's name all appear in the title of the web page.

4.6.7.2 Item

The following is the APA 6th edition reference output based on the data that Zotero detected:

> Ļōjjeļañ, Bender, B., & Bender, B. (1954). BB1-018.
> Retrieved from http://scholarspace.manoa.hawaii.edu/handle/10125/30788

The following is the Chicago 17th edition reference output from Zotero in Author-date format:

> Ļōjjeļañ, Byron Bender, and Byron Bender. 1954. "BB1-018," January.
> http://scholarspace.manoa.hawaii.edu/handle/10125/30788

These item level references reflect the sparse nature of the metadata actually interpretable by Zotero.

4.7 Review summary

A review of these five archives indicates that the cycle described in Chapter 1 is not as easily completed with archival material as it is with formally published materials such as books and journal articles. The archives reviewed have either chosen to not support bibliographic metadata transfer, or they indirectly support it. To briefly recap, the communicability[154] of metadata between institutions that hold language artifacts and authors that need to reference those artifacts is limited even though there are diverse methods for

[154] Some have labeled this **interoperability**. Interoperability has been cast as the ability to move data out of one application and into another. Another way to look at the concept that term **interoperability** tries to capture is: *How does some data increase in value to an end-user as it travels through a workflow?*

importing bibliographic metadata to Zotero. We have seen that there are some technical issues with Zotero, primarily with data translation. We have also seen that there are data issues with metadata repositories such as DataCite. However, the largest portion of the technical obstacles actually lie with the language archives. The Pangloss collection is the most articulate in the sense that it had the most bibliographic transfer options, and the bibliographic metadata it transferred to Zotero was the most complete. However, even with great user interfaces making the transfer of bibliographic metadata easy, some critical data points are not transferred, e.g., audio duration, and some are transferred in unusable ways, e.g., language identifier.

This thesis has focused on a technical assessment of bibliographic metadata interoperability between authorship tools and archives. While resources could be expended on addressing this important issue, there are other issues which if unaddressed along with the technical issues would likely leave archive users unsatisfied and less likely to reference evidence from language archives. One such issue is arrangement and description. The significant degree of variation in curation practice specifically in regards to how items are classified (assigned an item type) has an impact on what metadata is perceived as necessary for transmission to reference managers such as Zotero. This ultimately impacts authors and the referencing cycle described in Chapter 1.

CHAPTER 5

Example references

In this thesis, I address two types of archive records—the collection record and the audio artifact record, specifically within the context of the archive collection.[155] I would be remiss to mention the sparse nature of bibliographic data transferred between an archive and a reference manager, without also giving some examples of the types of data points that researchers should expect and in fact need to make informative references. I have already provided some of the same examples in chapter 1 and, as relevant, in chapter 4. In this chapter I bring together the templates from the *APA 6th edition* and the *Chicago 17th edition Author-date* styles and contrast them with the exemplars provided by the style sheets of *Language* and *Linguistic Inquiry*. I provide examples for referencing an archive collection, a component in a collection, and an audio artifact. I also juxtapose the style guides with guidance provided by the style sheets of *Language* and *Linguistic Inquiry* and "suggested citation" formats provided by ELAR and PARADISEC.

The following is the APA 6th edition collection reference template (VandenBos 2010:212):

> Author, A. A. (Year, Month Day). Title of material. [Description of material]. Name of collection (Call number, Box number, File name or number, etc.). Name and location of repository.

[155] An author may be tempted to reference a digitally hosted collection as they would a web page. After all, when an archive does not provide collection level metadata on import, Zotero imports the web page's metadata as a web page item. However, there is a conceptual difference between referencing a web page and referencing a collection. Collections may move from institution to institution, and may change Internet locations. A web page is a specific unit of content at a specific point in time.

The following is the APA 6th edition examples of referenced collections components (VandenBos 2010:213):

> Frank, L. K. (1935, February 4). [Letter to Robert M. Ogden]. Rockefeller Archive Center (GEB series 1.3, Box 371, Folder 3877), Tarrytown, NY.
>
> Berliner, A. (1959). Notes for a lecture on reminiscences of Wundt and Leipzig. Anna Berliner Memoirs (Box M50). Archives of the History of American Psychology, University of Akron, Akron, OH.

The following is the APA 6th edition audio recording reference template (VandenBos 2010:209, Skutley 2012:25):

> Writer, A. A. (copyright year). Title of song [Recorded by B. B. Artist if different from writer]. On *Title of album* [Medium of recording: CD, mp3, record, cassette, etc.]. Retrieved from http://xxxxx (Date of recording if different from song's copyright date)

The APA style guide, as opposed to the *Chicago Manual of Style*, provides descriptive templates. Even with the wonderful examples of the *Chicago Manual of Style* one must still infer the categories of information they have deemed necessary.

The following is the Chicago 17th edition collection reference for a whole collection in Author-date format (Harper 2017:§15.54):

> Egmont Manuscripts. Phillipps Collection. University of Georgia Library.

Unlike the APA provided template above, Chicago 17th edition only provides exemplars. I present my analysis of these exemplars within yellow boxes below each lavender box. The following is my analysis of the collection reference pattern based on the Egmont Manuscript exemplar.

> Series or (Sub)series Name. Collection Name. Archive Name.

The following is the Chicago 17th edition collection reference for a single item in a collection in Author-date format (Harper 2017:§15.54):

Dinkel, Joseph, n.d. Description of Louis Agassiz written at the request of Elizabeth Cary Agassiz. Agassiz Papers. Houghton Library, Harvard University.

Family Name, Given Name, Date. Title or Description of the artifact. Collection Name. Archive Name, Archive Institution Name.

The following is the Chicago 17th edition audio recording references for a single item in a collection in Author-date format (Harper 2017:§15.57):

Coolidge, Calvin. [1920?]. "Equal Rights" (speech). In "American Leaders Speak: Recordings from World War I and the 1920 Election, 1918-1920." Library of Congress. Copy of an undated 78 rpm disc, RealAudio and WAV formats, 3:45. http://memory.loc.gov/ammem/nfhtml/.

Holiday, Billie, vocalist. 1958. "I'm a Fool to Want You." By Joel Herron, Frank Sinatra, and Jack Wolf. Recorded February 20,1958, with Ray Ellis. Track 1 on Lady in Satin. Columbia CL 1157, 33x/3 rpm.

Speaker's Family Name, Given Name. Date. "Title" (Type of audio). In "Title of Aggregate work." Archive. Copy of original carrier, current formats, time-based-length. URL.

Family Name, Given Name, role. Date. "Title" By author(s) Given Name Family Name. Recorded Date, with Notable Performer. Position in Aggregate work on Title of Aggregate work. Publisher Publisher-ID, Carrier.

Both APA and Chicago treat archival material as part of an aggregate work. The basic problem solved by the reference is *where can someone find the thing referenced?* Crucial criteria for solving the search problem are identifying the archive and the location in the archive where the artifact is managed. Additionally, they both describe the artifact so that

someone following the reference would be able to identify if the artifact was truly the one referenced (the content at the other end of a URL can change from time to time, just as your friend or spouse may sometimes answer your phone when someone is calling you).

Archives as publishers and distributors of artifacts should be making the bibliographic data available in several ways. First, it should be visually available to human readers of web pages. Second, it should be available in machine readable formats. Additionally, navigation through the tiers of organization of collections to artifacts via web interfaces should help people understand the organization structure of collections. In this way there should be a distinct difference between an exhibit which presents artifacts from a collection for comparison with other artifacts, and a web page which presents artifacts for their publication state. For example, if one wanted to compare the contents of two articles, that comparison could be done in an online tool. Such a comparison should not be the use case which drives the publication of the two articles.

The search and finding component of reference formation has in recent years been undermined by the social forces which take a perspective that the primary role of references is for "attribution"—which has become a necessary building block for career development. The style sheets for *Language* (Dawson 2011) and *Linguistic Inquiry* (Witz 2009) are influential in the academic publishing industry (for linguistics) because they are replicated by other editorial teams (and style sheets). Together they represent an interesting editorial perspective on referencing the evidentiary record. In particular, they reject the philosophical approach to referencing presented by Haspelmath (2014) in *The Generic Style Rules For Linguistics*. When dealing with unpublished materials, it states:

> ... but such unpublished papers should only be cited from recent conferences, if it can be expected that the material will eventually be published.

The implication here is that unpublished works should not be referenced. While there is a philosophical question on what "published" means, I take it here to mean peer-reviewed publication, rather than public access. In 2014, many preprint servers and conferences hosted content in publicly accessible locations. In this context, Haspelmath and Michaelis'

(2014) blog post arguing for the peer review status of annotated corpora (an output of language documentation practice) becomes relevant to archived collections—the implication is that in order to reference archival content, it first needs to be peer-reviewed.[156]

In contrast to my interpretation of Haspelmath and Michaelis (2014), the style sheets for both *Language* and *Linguistic Inquiry* each provide an example for referencing unpublished materials, meaning they find it acceptable to reference unpublished and non-peer-reviewed materials.[157,158]

The following is an example from *Language's* style sheet for referencing a manuscript with an online component:

> SUNDELL, TIMOTHY R. 2009. Metalinguistic disagreement. Ann Arbor: University of Michigan, MS . Online: http://faculty.wcas.northwestern.edu/~trs341/papers.html.

> SURNAME, GIVEN NAME. Year. Title. City-of-author's-institution: Institution-of-author, Manuscript indicator. Online: URL.

The style sheet for the journal *Linguistic Inquiry*[159] heavily relies on Chicago 15th edition (Mahan 2003) and provides an example of referenced material from an archive:

> Zoll, Cheryl. 1998. Positional asymmetries and licensing. Ms., MIT, Cambridge, MA. Available on Rutgers Optimality Archive, ROA 282, http://roa.rutgers.edu.

[156] Currently there is no standard method or accepted practice for peer-reviewing an archive collection of language materials. Within the tradition of archiving, collections would simply have their descriptions enhanced, but for linguists the question remains: *How does one argue that enhancing the description of a collection has intellectual merit worthy of attribution sufficient enough to use as evidence in hiring and tenure processes?*

[157] Note that neither example is of an audio artifact, nor of an archival collection.

[158] Neither of these examples provide enough for a third party to create a CSL style sheet (both reference objects which would have the DCMIType **text**) for that publication without additional consultation with the editors. Style sheets, which are freely accessible via Zotero, are often constrained by the lack of sufficient concrete examples provided by editors in their publishing style guides. CSL style sheets are further discussed in Section 3.2.

[159] *Linguistic Inquiry* is interesting in that many authors use Optimality Theory. Some of the founding documents presenting Optimality Theory were not formally published for many years but were referenced from the Rutgers Optimality Archive, a sort of early preprint server used by discussants of Optimality Theory. This firmly established the need to reference textual manuscripts within the formal publishing literature.

Surname, Given Name. Year. Title. Manuscript indicator., Institution-of-author, City-of-author's-institution. Available on Archive, Archive ID, URL to archive.

Some of the reviewed archives provide pre-formatted references for users to copy and incorporate into their works.[160] The utility of these formatted references is questionable, as formatting decisions ultimately are determined by the publisher of the work containing the formatted reference, not the entity wishing to be referenced. The examples below are replicated from items or collections reviewed in chapter 4. The VuFind interface to ELAR provided the following statement and reference: "To refer to any data from the corpus, please cite the corpus in this way:"

McGill, Stuart. 2012. Cicipu documentation. London: SOAS, Endangered Languages Archive. URL:https://elar.soas.ac.uk/Collection/MPI97667. Accessed on [insert date here].

ELAR's suggestion creates an overgeneralization if one wants to reference a particular artifact within the collection. Citing and referencing strategies together need to be able to address issues of granularity. Pangloss takes an interesting approach by assigning HTML anchor tags at the phrasal level of oral texts. How well authors will notice these extensions to the URL structure remains to be seen. PARADISEC's "suggested citation" consists of the following:

Roger Blench (collector), 2003. Niger-Congo field materials. Collection RB5 at catalog.paradisec.org.au [Open Access]. https://dx.doi.org/10.4225/72/56E977B622032

It is interesting that the role is provided. However, it is not entirely clear that this is necessary if DACS practices are followed for naming collections. If collections are given appropriate item type status in reference strategies, then additional role information is

[160] Tangentially, it is interesting that many of entities use "recommended citation" or "cite as" when they really provide a reference.

redundant. Generally, item type status is distinguished by position and content of elements in the reference.

CHAPTER 6

Theory and applications

In Chapter 4, we see that there are some technological barriers to the frictionless flow of bibliographic metadata from language archives to authors who would reference the materials. However, the lion's share of barriers actually arise out of curation and archival practices. Changing these patterns of behavior in archives would take an immense amount of work. So, *Is it a problem worth solving?* Or in other words, *Is distribution of bibliographic metadata from archives for archived content worth solving?*

In order to address the question, I apply an economic model and cast archives and their clients within that model. The economic model builds upon a critical observation by Bourdieu (1977:645)—that language can be analyzed as both an object of understanding and an instrument of action. As objects of understanding, language and language artifacts are subject to commodification and influences of economic exchanges just like other goods. Commodification of language is well discussed in the literature (see among others: Ganahl 2001; Dobrin et al. 2009; Hemphill & Blakely 2015; Holborow 2018; Guo et al. 2020; Škevin Rajko & Šimičić 2020). Less discussed is the influence language archives have in the ecology of language as a commodity. An economic model applied to commodities traded by language archives must take into account that which language documentation practitioners have long been aware—language archives have multiple audiences (Holton 2012; Woodbury 2014). Generally these audiences are described in contrastive categories such as academics and language "community" members as illustrated in Figure 56.

Figure 56. Archive audiences as described in the literature

When researchers and language "community" members are cast as the various sides to which a business or central platform must cater it appears as if there is a genuine two-sided market with two separate groups both of which may either be contributors of language artifacts or consumer/re-users of language artifacts. However, turning from how language archive audiences are described in the language documentation literature, it is also possible to categorize these same audiences as consumers and producers. This situates the audiences within the terms more commonly used for digital content delivery platforms, but also lends to an analysis of the transactions as a single-sided market due to the unidirectional transfer of language artifacts (from producer to consumer).[161] I illustrate this in Figure 57.

Figure 57. Archive audiences by content relationship

Due to the multiplicity of transactions in which audience members may fill multiple roles, an analysis depending on a single-sided market is not comprehensive enough to

[161] Even if we were to take this unidirectional view of exchange, it we still need to acknowledge that consumers seek out and consumer language artifacts for different reasons. That is, as discussed more later, the jobs-to-be-done that they each seek to do are different. This minimally suggests that there are distinct audiences serviced by archives.

capture the complexity of the market activity. In this work, I follow Paterson and Nordmoe (2013) in situating archives as the platform in a two-sided market because it provides an economic model that helps us see the interactions of the multiple audiences.[162] Some well-known two-sided markets include credit card companies (Evans 2011b) and Facebook (Evans 2011a:379).

Credit card companies attempt to reduce friction between businesses and their clients. To do this, they enlist two classes of customers: businesses and people. They tell businesses that their card is used by X number of people for the purpose of offering that market to businesses. Businesses value the reduction of friction during the financial transaction, so they sign up to accept the credit card as payment. Businesses in turn agree to give a portion of their transactions to the credit card company in exchange for less friction at the point of sale. Credit card companies then tell people that X number of businesses are accepting their card. The greater the quantity of card holders, the more businesses are interested in accepting the card; the greater the quantity of businesses accepting the card, the more people are willing to use the card instead of alternatives. The more transactions occur using the card, then the more revenue the credit card company makes. In this way, the credit card company is the financial transaction platform.

Facebook has a similar model where the social networking site is the platform. Facebook offers a product to some people to use for free—a communication tool. It then charges other parties to use the platform in certain ways—to sell advertisements. Facebook acts as a matchmaker to connect businesses with people. It positions its platform so that it can offer two products: the eyeball-time of the categories of people in whom businesses are interested *to those businesses* and the advertisements that businesses create *to the people*.[163]Just as a language community is made up of individuals, so is any business

[162] This is asserted in section four of the poster (26178-1.pdf) and illustrated from an L&CA perspective on page 27 of the accompanying handout (26178-2.pdf).
https://scholarspace.manoa.hawaii.edu/bitstream/10125/26178/2/26178-2.pdf#page=27
The choice to situate language archives within the model of two-sided markets has some similarities with prior work situating journal publishers as two-sided markets (McCabe & Snyder 2007, Jeon & Rochet 2010). The primary similarity is that both language archives and journal publishers are platform owners acting as content receivers and distributors, often with different audiences. Situating language archives within the model of two-sided markets also has similarities with internet content delivery markets which are often two-sided markets (Zhang et al. 2014).

[163] The second offer is possible because people are looking for person-to-person connections. Facebook offers the potential for these connections for free, but then uses the desire for these connections to place paid offerings from businesses.

that advertises on Facebook. Any individual may interact with the platform in either type of transaction on any given day. On Monday morning a business owner might submit an advertisement through Facebook's marketing tools. And then on Tuesday, that same business owner, may be scrolling through her neighborhood Facebook group and come across an advertisement for the local bakery, click the link, get the coupon, and continue the series of exchanges available via the platform. It is the multiplicity of transaction types, their directions, and the roles which any particular individual may fill that makes the two-sided market analysis most helpful.

The two-sided market (Parker & Van Alstyne 2000; Rysman 2009) is an economic model with three classes of actors. First, there is **the business** which positions itself as a **platform**; the business then interacts with two different classes of customers (**Class1** and **Class2**).[164] The business sets up its transactions such that it is a "matchmaker" or "middle man" between the two classes of customers. As discussed by Rysman (2009:126–129), technical definitions of a two-sided market vary. Rysman (2009:127) states:

> ... the literature on two-sided markets could be seen as a subset of the literature on network effects. However, papers on two-sided markets tend to focus on the actions of the market intermediary, particularly pricing choices, whereas papers on network effects typically focus on adoption by users and optimal network size.

I take this to mean that one could also apply insights from literature on **value networks** to language archives. However, because I am specifically looking at a platform type business I have chosen the term **two-sided market**. Broadly cast, the business in a two-sided market will seek to charge a total amount for a service provided. Generally that service involves connecting two parties which would otherwise not normally be able to conduct a transaction on their own. The business will split the total sum of what they charge across the parties from the transaction—the charged split may be 0–100 where one class of customer is charged nothing while the other class of customer covers all the financial costs.

[164] It is often the case in two-sided markets that actor class distinctions are made on the basis of role in a given transaction, e.g., consumers may also be producers therefore switching classes between transactions. YouTube is a great example of actor class crossing; content producers often also watch videos as well—the person is the same, but the transactions with the market vendor (YouTube) are different.

A two-sided market is also generally distinguished from a single-sided market because it both "buys" and "sells" the product—it has transactions going both ways between classes of actors. Language archives fit the two-sided market definition because they both "buy" and "sell" language artifacts. That is, the terms of acquisition are often different between the receiving/"buying" of language artifacts and the giving/"selling" of language artifacts. Additionally, those who deposit language artifacts are often not able to directly engage with consumers of language artifacts.

Within the two-sided market model, I adopt a view of the transaction as put forward by Clayton Christensen. Christensen presents the idea that people hire products to do a job (Christensen 2011; Christensen et al. 2016). That is people have a job-to-be-done, and need a way to solve that problem. The thing they buy is what they perceive is the solution; their true placement of value is in the solution—as manifested in the object(s) they acquire to solve the job-to-be-done task. By way of example, a homeowner with a clogged toilet does not buy a plunger for the aesthetic nature of the plunger. In fact the better analysis is that they are not even buying a plunger, rather they think they are buying the capability to create a working toilet. This is often evidenced by the frustration encountered when the plunger fails to fix the clogged toilet and the person makes a second trip to their local hardware store to purchase a plumber's snake. The job-to-be-done has not changed; what the homeowner is still buying is the capability to unclogged the pipe.

Academic linguists have a job-to-be-done. It is to create products that can and will be referenced. For language "community" language-artifact users, the most frequently perceived job-to-be-done is to transform commodified *objects of language* into a socially viable communication system which functions as an *instrument of action* between language users.

People who engage with a two-sided market do so because they believe that the business (running the platform) will provide them with a solution to their job-to-be-done. Osterwalder & Pigneur (2010) and Osterwalder et al. (2015) label the business's proposed solution as the **value proposition**. Osterwalder et al. (2015) and Strategyzer (2017) map the value proposition to a customer's jobs-to-be-done and motivations. The mapping between the value proposition and the jobs-to-be-done creates an identifiable link where a

class of actor finds value in the solution a business offers. The tools described in Osterwalder et al. (2015) and Strategyzer (2017) help businesses discover this link and the deep motivational reasons why people find value in an offering. When a potential customer starts to see the value in an offering, they develop a **reason-to-believe**. Maintaining that reason-to-believe is an important goal for a corporation if they are going to turn single transaction customers into multi-transaction customers. While Osterwalder and Christensen's terminologies are designed to be widely applicable, the same general idea about value was previously articulated for the museum industry in Bitgood (2006:464) where he states the following about value.

> The general value principle (Bitgood 2005; 2006) argues that the value of an experience is calculated (usually without awareness) as a ratio between the benefits and the costs. We attend to things that are perceived as beneficial (such as satisfying curiosity, enjoyment) only if the costs are perceived as low in relation to the benefits. This means the value of an experience may change even if the perceived benefits stay the same. That is, if the costs (time, effort, and so on) are perceived to be high, the value of the experience is lower than it would be if the costs were perceived to be low. You are likely to choose to earn $100 if it takes only one hour and involves an activity you enjoy doing; but you are less likely to engage in the same activity if it takes 30 hours. The value of $100 is discounted if it takes too much time to earn it. (Bitgood 2006:464)

I consider the user interaction with a website—the main interaction platform for many digital language archives—a designed experience. In this way, exchange via an archive either for depositing or receiving is crafting the perception of value by Class1 and Class2 actors. It is defining the value proposition.[165]

In Figure 58, I illustrate the contrast in the flow of revenue and value between a typical business (one sided market) and a two-sided market. The traditional business analysis approach is to conceptualize value and revenue in a single sequential stream. Each actor in the chain is alleged to increase the value in some way to the end customer, and in exchange each actor in the chain receives "value" packaged in their desired format. As described earlier, these assumptions are not present in a two-sided market. For the two-sided market, I maintain the label **value** following Osterwalder et al. (2015), and I apply

[165] There may be non-website based interactions with digital archives, but these too are designed interactions.

it to what the platform returns to Class1 and Class2 actors. Osterwalder et al. (2015) further discusses the customer-perceived positive value in terms of **gains** while Bitgood's (2006:464) costs are discussed in terms of **pains**.[166]

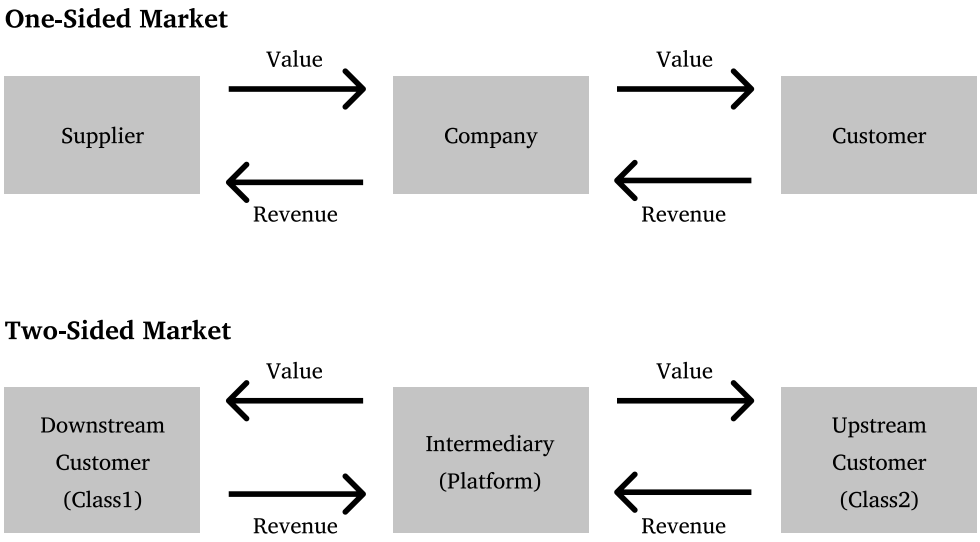

Figure 58. Value vs. revenue flow in markets

Both Osterwalder et al. and Bitgood are careful to note that value may be a functional change in situation or an emotional and social change in the way that the job-to-be-done is accomplished. I intentionally use the label value in Figure 58 for what is returned to archive customers. This is not just to emphasize these social and emotional components of the customers motivations to engage in the transaction but also to provide a loosely articulated category for the kind of thing returned to archive customers.[167] In this sense, then, value is a crucial concept here; value is objectified; it is what we buy (acquire). We may "buy" (or acquire) it with our money, our time, or our work. The thing we are finding our value in may or may not be tangible, e.g., we might exchange something to get the value of language knowledge.

[166] I am unaware of any published studies about archives or repositories which have used the methodologies presented in Osterwalder et al. (2015) to evaluate customer interactions. Wilms et al. (2020) use *Social Exchange Theory* (Homans 1958), while Davis & Connolly (2007) do not present their results in the context of a theory.

[167] Some suggest that the thing returned to the customer is a commodity; however, as seen with the toilet example, the homeowner was really acquiring a capability.

While business and economic models generally look at financial transactions, I employ the insights of the two-sided market transactional model to describe the flow of information. That is, I let information and cultural heritage artifacts take the place of money (currency) as value is passed between the three classes of actors. Some value language artifacts for their innate properties—as collectors of fetish objects. However, the more frequent case is that people value them as part of their intangible cultural heritage. They are a medium of access to the capability of knowing about a heritage culture or a worldview through the language artifacts of a particular language. The object is a medium of exchange for a capability. As such, the medium can be traded for other mediums or capabilities—both tangible and intangible. In this way, I see language artifacts as a type of currency because they are exchanged among actors in the model, and they are traded for capabilities. I follow Paterson and Nordmoe (2013) in broadly casting language researchers as one class of customer, language users as the second, and the language archive as the business or platform owner. However, I extend their limited description of the rationale for the two-sided market by acknowledging that a non-exclusionary analysis indicates that the distinction between Class1 and Class2 actors is that Class1 are language-resource creators, while Class2 are language-resource consumers. That is, this additional perceptive lends strength to the use of the two-sided market model. Both researchers and language users are consumers and/or creators. Archives hold the platform for interactions across time between various interested stakeholders in language artifacts.

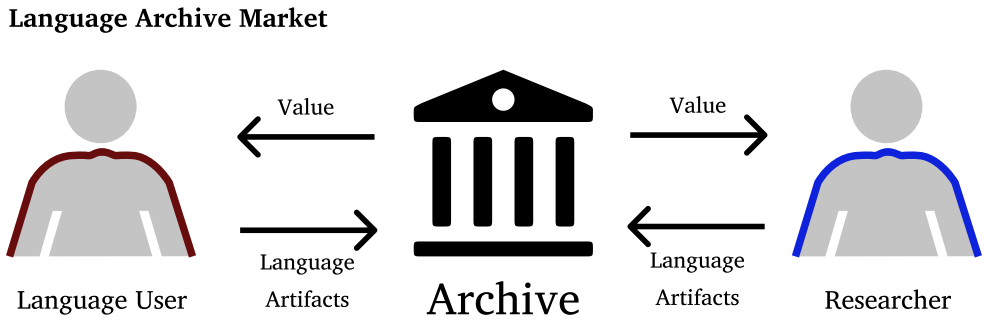

Figure 59. The language archive as a two-sided market

By situating the language archive in the context of a two-sided market, I position my analysis in terms of value generation and artifact acquisition (revenue).[168] This stands in contrast to more traditional models of content archives which focus on content stewardship under the Open Archival Information System (OAIS) model.[169] Under OAIS, archives are responsible to deliver the content they received in the same condition they received it in. One of the business observations which comes with a two-sided market analysis is that when there is a decrease in the platform's perceived value by customers in one type of exchange, it impacts the ability of the platform to provide value to customers in a different type of exchange. This can come in several ways such as brand value or number of users with which to make value exchanges. In the two-sided market model, language artifacts are like a currency in that they are attributed value and they can move fluidly around within the network of artifact and language users. However, language artifacts are unlike currency in the Keynesian sense because it is not necessarily true that the value of artifacts goes down when they are in greater circulation. In fact, the opposite is often assumed to be true: the more widely distributed a language artifact is circulated, the greater the general awareness is of the artifact and the greater its social impact! By situating archives as the platform owner, it challenges the perspective that archives are passive depositories in the data life cycle. Rather they are active agents with a responsibility to shape the marketplace to create value and ensure their own sustainability within the network of users.

By focusing on value generation, it allows one to ask questions of the archive like: *In what ways does the platform need to evolve in order to continue generating ongoing value?* If we ask the original question about whether the problem is worth solving, the question can be re-framed in the context of this new question. That is, *What value is lost when the free flow of bibliographic metadata is hindered by friction, or when the description and arrangement of materials in the archives do not support efficient and informative citation and reference generation?* Does the threat of lost value in turn threaten the very existence of language archives? If it does threaten the existence of language archives, then it might

[168] I discuss value further in 6.2. From the economic models, *value* is what one obtains in exchange for what they give up.

[169] https://www.iso.org/standard/57284.html

be a concern worth addressing. I propose that it does threaten their existence. When references are hard to make, academics leave the network for other solutions as shown in Paterson (2015a) where academic linguists have opted to use backup and social sharing tools instead of archives. Using other kinds of tools instead of archives, lowers the value of the whole network.

In the remainder of this chapter I address the value exchange between archives and language users (Section 6.1), the value exchange between archives and linguists in academia (Section 6.2), and how archives can maintain the ability to deliver value (Section 6.3).

6.1 Value exchange between archives and language users

Language archives inherently have limited interactive value for language users, especially in the context of language revitalization work. The reason for this is that language archives inherently treat language artifacts as objects. Revitalized language is not a set of actions which can be done with objects but rather an instrument of action—more specifically the restoration of an instrument of action.

Much of the current discussion in language documentation venues about language-user access to archival content and language revitalization centers around the process of creating and using objects—language artifacts. To the extent that artifact stewardship is what is common practice in language revitalization, archives can and should have a role in the exchange of language artifacts between creators and consumers (generally Class1 and Class2 actors in the two-sided market model). However, there is strong evidence that consumers are mostly interested in interacting with content once it is outside of the archive. Therefore, in some workflows it does not make sense to ever put the content into an archive in the first place. For example, the *Mukurtu CMS*[170] (presented in Christen et al. 2017) is heralded by Kanahele and Holton (2021) as a great solution for communities of the Pacific as a way to facilitate user engagement with cultural heritage artifacts. Unfortunately, these artifacts from the online Mukurtu communities are not also stored

[170] https://mukurtu.org

in archives.[171] In Thailand, language documentation efforts are turning to commercial services like SoundCloud (Ferreira et al. 2021) not just to facilitate community access but in lieu of archiving.[172] In these cases, secondary points of distribution become the referenceable location, often with contributor roles removed—instead of a language archive, where contributor roles can be found. It is not clear that much value is transferred from archive to language user at the current time—at least as language archives are currently set up with their current web interfaces.[173] This does not mean that value is not transferred from archives to some class of consumer, only that there is a middle man between language archives and language users who must "enhance" or add value to an existing artifact to make it a consumable. For instance, a technologist may extract content from archives and recast it in something like a *Mukurtu* instance.

In contrast to the model where a technologist is situated between a general public-facing portal presenting artifacts and an archive's user interface, The Digital Library Services team at the University of Cape Town chose to use *Omeka-S*[174] to create a display of the artifacts in the *ǂKhomani San | Hugh Brody Collection*, one of the special collections at the library (Jones & Muftic 2020a, 2020b). This display has elements which were tailored for topics of interest to the language communities represented in the collection. By working from within the context of the archive rather than extracting the artifacts from the

[171] Mukurtu or Omeka are both web-platform content management systems which enable the creation of online exhibits. Exhibits are an important public engagement tool in the repertoire of museums. Quigley (2019:§1.3) discusses both museum exhibits and catalogues as tools of public engagement and as scholarly works. Quigley notes that within the museum sector exhibitions are considered ephemeral, whereas catalogues are more enduring but both are crucial for carrying out the mission of the museum. The role of digital language exhibits and the digital language museum as a distinct entity from the digital language archive are not broadly discussed in language documentation circles. Historically, museums such as the Peabody Museum have had a significant impact on the production of language and culture descriptions. They still play a significant role as centers of research and in shaping the collective consciousness about specific languages and the people who speak those languages. Specific language museums are a more recent innovation, e.g., National Museum of Language (established 1997), Museu da Língua Portuguesa in São Paulo (established 2006), Canadian Language Museum (established 2011), Mundolingua in Paris, Museum of Languages in Leiden (established 2015). Museums in the language preservation context are not undiscussed, for example, Kuzmin (2013) lumps them together with archives and libraries under the broad title of "memory institutions". However, Thieberger (2013) presents the museum function for exhibit creation as independent from digitization and archival functions. For additional discussion on language museums see Lehmann (1992, 2001), Crystal (2004), and Sönmez et al. (2020).

[172] In the presentation there is a stated hope or plan that these materials will be archived at some institution. Unfortunately, the linking between secondary distribution channel and primary preservation channel does not currently exist.

[173] Web interfaces at language archives are not designed to engage with audiences who are interested in using language as an instrument of action.

[174] https://omeka.org/s

holding institution and recasting them in a self-hosted website, the archive also now has a much richer collection description and can continue to deliver the enriched contextualized content as part of its service to content consumers.

6.2 Value exchange between archives and linguists in academia

Value is transferred from archives to linguists in three primary ways. In what follows I list the three ways, with the remainder of this section focused on the third way value is transferred. First, archives provide a place to deposit media generated during research efforts. The act of depositing fulfills many modern grant requirements—that is, a researcher can continue to receive grants because they have deposited their artifacts in an archive, per the terms of the last received grant. Second, linguists can act in the role of content consumer and access content from other researchers and conduct further research with those artifacts. Third, the archive provides an end point or distribution point for other producers to reference existing language work.

The significance of this third aspect of realized value—reference counts—should not be overlooked. In 2020, I was indirectly involved with a group of linguists who were struggling with how to fit three names (Principal Investigator, Researcher/Recorder/Postdoc, and Speaker) into a referencing format for a set of annotated corpora they wished to release as interactive resources. The importance of the name accessibility and name prominence across publication styles was shaping the way that the scholars were approaching the organization of their collections and artifacts.

Both Berez-Kroeker et al. (2018) and Haspelmath (2014) argue for changes in referencing practices in order to favorably impact (for scholars) the metrics by which scholars are evaluated.

> Unfortunately most linguists do not know how to go about advocating that "data work" be given the same kind of attribution as "analysis work" in hiring, tenure and promotion cases. (Berez-Kroeker et al. 2018:11)

> The strongest justification for simple rules is that the references should be automatically parsable (e.g. by Google Scholar), and correct and complete author names should be extractable. In the modern age, this is crucial for scientometric and hence career-building purposes. (Haspelmath 2014:footnote 16)

The perceived value transfer is one where mid-career scholars do not have frequently referenced works because they made early-career choices to engage in language documentation practices, and then the returned value (in terms of defensible indicators of scholarly influence) is not enough to sustain them while navigating career paths into their late-career stages. The effects of career academics' desire to impact indicators demonstrating their scholarly influence should not be under appreciated. Let's cut to the chase, "linguistic social work" (Newman 2003:11) is not without value transfer or expectations. After working with a linguist, a language "community" might have a few resources in their language, a new orthography, or recordings of their traditional stories in an archive, but now the piper needs to be paid. The economic currency of academia is reference counts.[175]

In this regard, well-crafted references are highly valued by academic authors, but they are only a token used towards career development. When archives strive to reduce transactional friction related to metadata interoperability, they increase the value of their platform to all parties in the system. If a researcher has a choice on which archive to deposit materials, it is going to be the one with the least amount of friction and the greatest return for their effort. The archive that can boast "more people reference our collections than any other archive" will have academics favor submitting to their holdings.

As references to artifacts and "data" are valuable to linguists in academia, questions are raised: *what is an artifact,* and *what constitutes data?* In the ongoing conversation

[175] In some ways, reference counts (also refereed to in the literature as "citation counts") are not like currency in that they are not directly exchangeable once awarded. However, Piwowar et al. (2007) refer to references as currency saying: "A currency of value to many investigators is the number of times their publications are cited. Although limited as a proxy for the scientific contribution of a paper, citation counts are often used in research funding and promotion decisions..." Reference counts are estimated by Diamond (1986) to have a United States Dollar (USD) value between $50–$1300 in 1984 dollars per reference depending on the organization of employment and academic discipline. Mueller-Langer & Watt (2018) evaluate the $3000 USD price for hybrid open access publication charged by some publishers and equate it with reference counts. Their results suggest that those who pay a $3000 open access fee value a single reference to their work at least $3,278 USD; a price point at which they suggest pre-print and post-print deposits in institutional repositories make a better investment without a statically significant difference in reference counts generated per publication (though they estimate that there may be as much as an 8 percent increase in reference counts for open access publications).

around citing and referencing linguistic artifacts, Berez-Kroeker et al. (2018) cast their position statement on reproducibility in linguistic science "with regard to facilitating a culture of proper long-term care and citation of linguistic ***data*** sets" (emphasis added). Berez-Kroeker et al. (2018) and *The Tromsø Recommendations for Citation of Research Data in Linguistics* (2019) merge the concepts of "data" and evidence. By merging these concepts a new narrative is projected: if something is not an opinion published in a journal, then it must be "data", or worse, anything at the other end of a reference is "data". I would like to remind us that evidence comes in different types which are based on different contexts of production, modes, and variation in carrier types (paper, CD, DVD, reel-to-reel, etc.). It does not serve us well to think of everything in the archive as "data"—even if data is the term commonly used for digital information. The unitary view of data is countered by computer scientists, another class of scientists that Berez-Kroeker et al. (2018) point to as having ongoing discussion about practices of citation and referencing. They have struggled with the definition of data—specifically contrasting it with software. They state in Katz et al. (2016):

> Software is data, but it is not just data. While "data" in computing and information science can refer to anything that can be processed by a computer, software is a special kind of data that can be a creative, executable tool that operates on data.

Just as the computer scientists had to tease apart the difference between data and software, our discipline of linguistics needs to tease apart the difference between data and evidence. Our discipline is in jeopardy of conflating the concepts of data and evidence. I propose sticking with a definition of data that is compatible with Dublin Core DCMIType term: **Dataset**.[176] The DCMIType definition is *Data encoded in a defined structure.* This definition appears to be rather broad. In practice, though, it narrows based on other guidance provided. A dataset is perceived to not be classifiable as one of the other DCMITypes (such as Sound, MovingImage, Text, and Collection, etc.) and is exemplified in the DCMIType description: *lists, tables, and databases. A dataset may be useful for direct machine processing*.

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

As a co-author of *The Tromsø Recommendations for Citation of Research Data in Linguistics*, I certainly agree with the tenet that the evidentiary record should be cited and referenced. However, I am concerned that conflating the concepts of data and evidence does not support the "scientific" goal of improving citation and referencing.[177]

The role of evidence is acknowledged in Berez-Kroeker et al. (2018:7). While the main topic of their position paper is about the referencing of data, they make a heretofore under-explored connection between description and arrangement of the evidentiary record when they say:

> This of course presumes that the data themselves are also properly preserved, discoverable, and accessible.

I find it necessary to restate this as follows, replacing "data" with "evidence":

> This of course presumes that the ***evidence*** is also ***properly preserved***, discoverable, and accessible.

So what does *properly preserved* mean? If there can be properly preserved evidence, is there then such a thing as un-properly or improperly preserved evidence? And then what does this mean for citation and referencing? The state of archival records suggests that the notion of "properly preserved evidence" has no community consensus among linguists—perhaps there should be. One might assume that proper preservation also includes proper curation and artifact description, which in the case of references would include categorizing an artifact with an item type beyond the overused term **data**. Preferably, the used item type would also align with how the artifact would fit into various style sheets.

[177] It is possible that the weighting of reference counts by type would cause linguists to desire more detailed description of the language artifacts they deposit. That is, if their deposits are all lumped together/typed alike they may loose credit in a weighted system, e.g., a low fidelity/accuracy but highly used dataset for historical comparative linguistics might garner more reference counts than a hand-annotated (e.g., for part of speech) and translated collection of audio and video recordings, yet both might appear to be referenced as "data" and therefore falsely evaluated as equivalent. If nothing else, linguists should be interested in the type description of their language-artifact deposits for the sake of how they might be referenced and thereby communicate the nature of their work. Dictionaries are clearly not journal articles, but is an elicited word list "data" or an "interview"?

It is this idea of proper curation, or an idealized state of curation, which leads to the discussion of varied curation practices in Section 6.3.

6.3 Archives: maintaining the ability to deliver value

Many linguists and language documentation practitioners use the term "archive" to denote a place which collects or holds collections of language artifacts, in essence a depository. I propose that depositories come in two types: **Archives** and **Repositories**. Both offer value to people who seek to solve challenges related to the long term stewardship of artifacts they have amassed or created; however, the two come with important variations on the types of results they provide. *Archives* have ongoing curation, cataloging, preservation, and enrichment schemes or plans for their artifacts. In contrast, *repositories* ingest and outgest (distribute) with minimal ongoing changes to the artifacts with which they were entrusted (or for that matter the arrangement or description of those artifacts)—they faithfully persist[178] (i.e. preserve) the artifacts.[179] For the person committing an artifact to a depository one possible question to ask then is: *By choosing this depository, do I archive the language artifacts or do I merely persist them?* Both archives and repositories address the issue of artifact persistence, but the expected life of the artifact is different. Both have business plans, but they seek to provide value to different audience classes. The different classes of audiences hold different perspectives on what good stewardship means.

Broadly, archives take an active role in the evolution of their collections, while repositories seek to let others find and generate value from their collections. By casting archives as active agents in the marketplace for language commodities, we can see that archives derive their value not as a depository, but as the center of a knowledge economy and an artifact-centric economy. However, remaining in the center is difficult due to sociological evolution and cultural evolution within the groups of Class1 and Class2. An archive needs to adapt to remain relevant and sustainable. The question becomes how should

[178] **Persist** is a transitive verb in computer science. The following example definition is taken from Wikipedia: Non-orthogonal persistence requires data to be written and read to and from storage using specific instructions in a program, resulting in the use of persist as a transitive verb: *On completion, the program persists the data.* https://en.wikipedia.org/wiki/Persistence_(computer_science)

[179] Often repositories and their architecture assume system-to-system communication as the primary means of engaging with artifacts. System-to-system interaction often happens over an API, and is generally a business-to-business type of interaction rather than a person-to-business type of interaction.

the platform be managed to continuously provide value to customers in either Class1 or Class2. For this, a business model can help. A business model clearly articulates "the rationale of how an organisation creates, delivers and captures value" (Osterwalder & Pigneur 2010:14). Archives need to be more than collections of features around media on servers. For the organizations to generate value, their goods need to be in circulation. The arrangement of artifacts and the description of artifacts according to archival principles, such as those articulated in DACS, give archives the organizational capacity and operational capability to leverage their holdings to create value exchanges with new and existing audiences alike. This is an important concept for businesses who need to adjust to evolutionary market spaces. I return to points I first made in Section 4.1 by suggesting that *arrangement* and *description* according to DACS enable archives to generate enduring value, a key element of *properly preserved*. When we look at the issue of citation and referencing, good references are designed to give the reader context about the artifact being referenced.[180] Guidance from the discipline of archiving about good arrangement and description is also designed to give users context about the artifact with which they are interacting.

6.3.1 Arrangement

Across the collections and artifacts reviewed, there was variation in arrangement. Pangloss appears to follow the recommendations of the French national organization for audio preservation for cultural heritage records (Marcadé et al. 2014) which align well with the tiered structure presented in DACS. All of the other archives reviewed followed a different pattern of arrangement. That is, arrangement of materials within language archives in many cases does not follow best practices or industry norms within the broader discipline of archival content management. For example, Table 17 is replicated from Kung et al. (2021:§'Step 7: Arrangement & Discoverability'). It appears in an online course presenting an introduction to language archiving. The table shows the two-tier hierarchy which is employed by many language archives and the various terms they use

[180] They also list the archive. In many ways they are an advertisement for the archive. This should add to the reasons that an archive might want content it stewards to be cited and referenced.

to describe those tiers.[181] Three of the archives in their list are also discussed in this thesis (PARADISEC, ELAR, and Kaipuleohone). The list shows that the issues are wider spread than just the archives that I have reviewed. Each of the listed archives has a two-tiered structure, even though the names of the structures vary from archive to archive.

Table 17. File arrangement at some language archives

Repository	Collection	Folder
PARADISEC	Collection	Item
ELAR	Deposit	Bundle
AILLA	Collection	Resource
TLA	Corpus	Session
CLA	Collection	Item
Kaipuleohone	Collection	Item
Dataverse Repositories	Dataverse	Dataset

The final line in Table 17 gives us an indication as to why arrangement principles like those illustrated in Figure 4 (Section 4.1.1) and described in Hodges & McClurkin (2011:1–2) are eschewed. Digital artifact management tools like *Dataverse*[182] and DSpace (used by Kaipuleohone and L&CA) do not easily facilitate the required hierarchical structures. DSpace has already been briefly mentioned in Section 4.6.1 in regard to Kaipuleohone; the L&CA also uses DSpace as a back-end storage and management solution. In the L&CA's case, the SIL hierarchy system is essentially an historical organizational structure mapped to a current organizational structure via the DSpace notion of **communities**. This seems to be the typical application of communities in DSpace (Davis & Connolly 2007, Tramboo et al. 2012:§2.1). DSpace communities can be recursive, containing (sub)communities. Communities can contain multiple collections. Collections contain **items**. Items can have one or more **bitstreams** (files). Most metadata in DSpace is stored at the item level. Depending on the bibliographic model for record keeping, aggregate works may have a single DSpace item level record inclusive of all the parts as bitstreams, or it may contain no bitstreams and only contain non-literal pointers (links)[183] with a **hasPart** relationship to another item level record which contains the artifact (bitstream). The

[181] All abbreviations in Table 17 are listed in the abbreviations list in the front matter.
[182] https://dataverse.org
[183] For a discussion on literals and non-literals in bibliographic contexts see Coyle (2008).

general DSpace hierarchy looks like: *Communities* > *Sub-Communities* > *Collection* > *Item* > *bitstreams*. Arguably, the software suites have been designed to facilitate the repository model of artifact stewardship, not an archival model of stewardship. DSpace was first designed to only facilitate the storage of text based documents, and in its architecture model the "logical" work (analogous to the FRBR notion of work) was to be the equivalence of the DSpace **item**, and all variants of the item were to be stored with the item (Smith 2002:546). For text based documents this might be something like a Microsoft Word file and a PDF variant, or it might mean a preprint and its publication version are stored as separate bitstreams under the same DSpace item. However, lines between a new work and a variant become less obvious in language documentation materials, e.g., is a transcription file a variant of a recording? Language archives are also very aware of rights issues, e.g., one manifestation or expression of a work might have different rights or licenses applied to it than other expressions and manifestations. The application of rights metadata to artifact records can be seen as descriptive work, indeed description and arrangement work together to provide the organizational capacity to capture and deliver value. When language archives implement models from library science which account for these issues rather than remaining subject to the technological imperialism of software, it allows for depositories to leverage their holdings for audience engagement. It also allows for clearer referencing of artifacts at expression and manifestation levels.

The internal storage architecture of repository software is only one way that the depository's holdings are impacted. A second way is through the ingest process. DSpace has a process-based role-aware ingest workflow. Sometimes metadata application profiles can be complex making metadata entry laborious. Ingest workflows for one type of content from a particular class of depositor doesn't always make sense for another type of content from a different class of depositor. This can lead to the need for multiple workflows or for complex workflows. There have been numerous pieces of software written to facilitate the ingest of new content to DSpace (Nordmoe 2011, Weiland et al. 2019, Yumi & Kelly 2020).[184] The constrained and laborious process of ingesting artifacts to repository

[184] In many ways the tool chain described by Nordmoe (2011), Weiland et al. (2019), Yumi & Kelly (2020) places DSpace's deployment and interaction in the system-to-system interaction type rather than a person-to-system. One interpretation is that DSpace leans towards facilitating repositories more than archives.

software has two impacts: first, the human psychological response is to lump artifacts together under the same record to avoid extra work; second, it means that the relational metadata between records is generally missing (left undone) or must be crafted by hand by platform administrators after both artifacts have been ingested to the repository. Linguists and language documenters don't just make one or two artifacts, rather they create entire collections from fieldwork stints lasting weeks or months. The metadata relating artifacts to each other can be immense and provide deep descriptive insights to the collection. These descriptive insights comprise a key way that archive users derive value. Hsu et al. (2015) point out that rich metadata is one way that people evaluate satisfaction with archives and is foundational to perception of an archive's value.

In the OAIS model, repositories are removed from requirements for interacting with people because they interact with systems (systems for ingest and distribution). Their customers are no longer the content users but rather the system administrators of the systems which interact with the OAIS compliant repository. Therefore, the class of users (systems) in a repository framework does not derive value in the same way as a class of users (people) would because they are looking at the content through different sets of requirements. One result is that inter-artifact relationships are not as valued in repository oriented systems. To illustrate the contrasting way that repositories approach artifact handling, consider the use of *SayMore*,[185] SIL's software for organizing fieldwork recordings and metadata about those recordings (Moeller 2014). As a language documentation practitioner and and early tester of *SayMore*, I engaged in discussions with the corporate architect at the L&CA on how *SayMore* outputs (essentially zip files containing sets of artifacts) should be ingested to the depository for management by the archive. It was suggested by the architect that the entire package (containing all the artifacts and metadata) should be ingested as a single DSpace item; metadata then only needs to be managed for the collective of the `.zip` file and not the component parts. In a repository type model this makes a lot of sense. The principle is usually described in the following ways: "Garbage in garbage out", "Faithfulness to the input", and "seek to manage the fewest number of units while

[185] https://software.sil.org/saymore

delegating interaction and information management processes to other systems and architecture". This is the OAIS model, and in a sense it is exactly and only what DSpace implements (OAIS is first discussed in the introduction of Chapter 6): ingest is done by one system, long term storage is done by another system (DSpace), and presentation and access is done by a third system as illustrated in Figure 60.

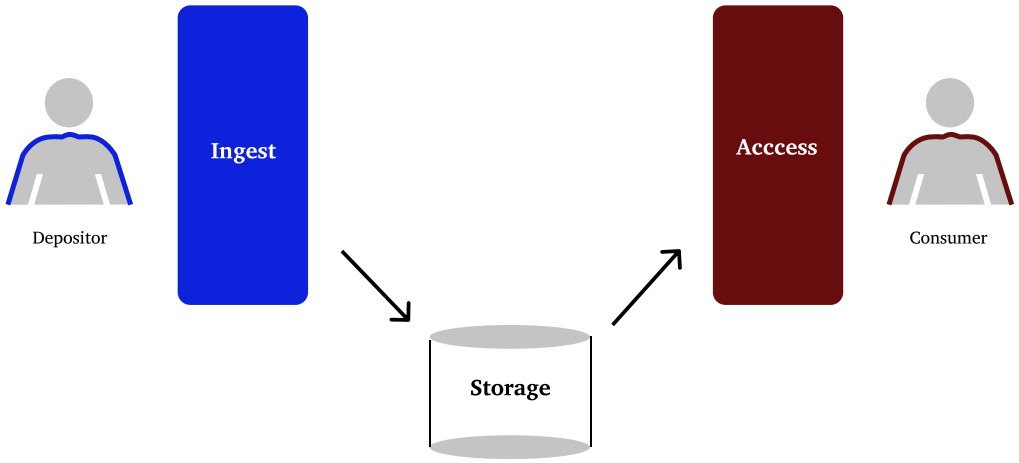

Figure 60. Simplified OAIS model

However, DSpace in its out-of-the-box configuration, only works smoothly for artifacts which are a single work, expression, manifestation, and item. If the item is a collection which will be displayed by another system, then there is no loss as long as the deposited package also contains metadata and instructions for the down-stream systems. However, if the storage system is also the browsing and managing system, then there are issues with communicating inter- and intra-resource relations. For example, DCMIType collection exploration is really difficult in the default DSpace interface.

To summarize, software concepts appear to be a driving force in determining the number of tiers within a collection, and subsequently how many artifacts are uniquely described within a collection at language archives. The workflows imposed by the software dictate how the institution arranges their collections impacting their ability to deliver and capture value. To support value generation via referencing and citation, institutions need to implement cataloging processes which allow for the description of unique artifacts, along with the description of each tier of the organizational structure of the collection.

We can see some of the impacts of arrangement practices within the artifacts and collections reviewed in this thesis. For example, Figure 53 is the Kaipuleohone record for an "item". It presents two `.wav` files and two `.mp3` files. These are presumably different manifestations of the same expression. Furthermore, the record reports in its metadata that the format of the artifact the record is about as "Maxell UR 60 min cassette". We can also see this information about **format** replicated in Figure 54 on line 15. However, this record is not a record for the actual Maxwell cassette, rather it is a record for the digitized content. So, the record producer (Kaipuleohone in this case) has conflated a third manifestation (the one on the cassette tape) into the record. Additionally, we know that the record description contains "Ḷōjjeḷañ #1, #3". Presumably, #1 and #3 are separate works by virtue of being separate narrations. So we have a record which covers three manifestations and two different works. This becomes very confusing for citation and referencing.

Best practice (see Koelling 2006 and Wijesundara & Sugimoto 2018) calls for one record to be created for the physical object—the cassette—and another record for each work on the recording. In this way, a one-to-one principle applicable at the manifestation level (manifestation-to-record) is preserved.[186] It may be that #1 and #3 are each a full side (as in Side A and Side B) of the cassette, but recordings could be carried out across sides of the cassette (i.e., #1 may have a duration of 38 minutes and #3 only 10 minutes), especially if the cassette was not the original carrier and the cassette was itself a transfer from a reel-to-reel recording. The digitization process and any post-editing process would determine exactly how many files make up a single work, and how many artifacts should be created from the digitization of this particular cassette. The digitization process is not described in the record, nor is a link to the description of the process provided. The conflation of works and manifestations into a single record makes it difficult for the institution to deliver value to its audiences. This directly impacts discovery and referencing, and ultimately the perception of the archive as a source of value by its audiences (for a more in depth discussion on archive value perception see Hsu et al. 2015).

[186] See footnote 55 in chapter 4.

The conflation of manifestations into a single record at Kaipuleohone reveals a complex issue which still needs wider consensus and work among those participating in the language artifact archiving and publishing communities. That issue is, *what are the components of an archival collection?* As far as I can tell there is no discussion in the academic literature on how hierarchical systems used to organize materials in a collection, aggregate works, and bibliographic record models such as FRBR could or should interact when linguistic and language documentation artifacts are described. [187] When we compare the organizational systems of several archives, we can see that each has its own practice with regard to the types of content included in hierarchical nodes smaller than the **archival collection**. Within the discipline of archiving, the term **collection** can be applied within different frameworks to mean different things. Variation in the use of the term collection is illustrated in Figure 61.

[187] There has been discussion of FRBR application more broadly across cultural heritage collections (among others see Nicolas 2005, Daquino & Tomasi 2015, Decourselle et al. 2015, and Farrokhnia 2019). For an overview of various models, not just FRBR, see Farrokhnia (2019:§2.2.4). There are also ample discussions on how FRBR applied to time based media collections such as film and audio collections. FRBR has also been discussed in the context of music collections which have complex relations between annotations and performance. See Section 4.1.2 for specific citations.

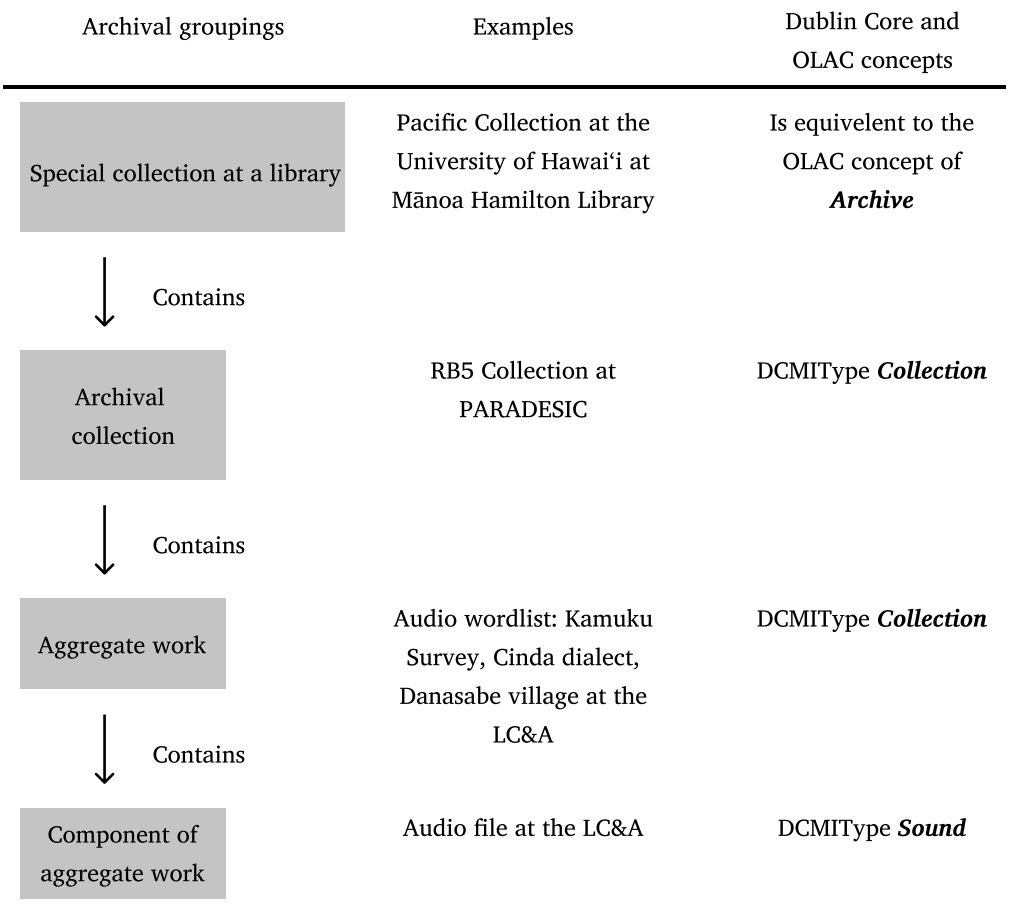

Figure 61. Collection concepts

The motivating factors in choosing these hierarchical structures are often tied to the financial notion of sustainability. That is, each archive is in an ongoing journey to demonstrate their social value in order to secure their operational funding. In many cases, funding is not provided by either Class1 or Class2 actors who are deriving value from an archive's information content. For example, if we compare the larger public narratives of organizations like SIL (the parent organization that the L&CA serves) and ELAR we can quite easily see that the organizational structures that ELAR brings forward in its choice of hierarchical structures are the funded projects. That is, there is a high degree of institutional value in the creation of a quantifiable and identifiable "object" (the collection) which can be presented as a return-on-investment for the donated funds to the Endangered Languages Program. In contrast, the narrative SIL puts forward to its funders is not project based but rather draws on a framed narrative where artifacts are presented rela-

tive to their associated language. By presenting the artifact and hiding the organization structure (collections) which helped to breathe the artifact into existence, the language-community/artifact frame maintains a continuity with frames presented to SIL's primary funders, while giving the organization the flexibility to manage projects through a variety of means.

Across archives, the shallow depth in hierarchical systems can create complications for easily transmitting bibliographic data consistently. However, the presentation is not atheoretical, rather language archives prioritize presentational views which communicate to their funders over other views which might impact organizational structures within collections.[188]

6.3.2 Description

Across the archives reviewed in Chapter 4, there is substantial variation in curation practice for describing language artifacts. As mentioned in Section 6.3.1 when the hierarchy is conflated, it impacts the ability to effectively describe artifacts so as to accurately communicate their context. There is a lot of discussion on the kinds of metadata that linguists should keep when creating collections of language artifacts (Nathan & Austin 2004, Himmelmann 2006, Bergqvist 2007, Lüpke 2010, Good 2011, Austin 2013). Linguistics is after all a metadata-hungry discipline.

Some metadata suggested for linguists to generate relates the language analysis nature of the content contained in artifacts. Some metadata is needed for artifact preservation and management, still other metadata elements are helpful in the understanding the type-nature of the artifact. The works of Nathan and Theiberger sometimes use the terms **thick** and **thin** to draw contrast between artifact description and linguistic content exposition. As they use the term, **thin** relates to metadata which describes the artifact. The term **thin** can be interpreted as drawing parallels with minimalism and in these contexts, under-description of the artifacts. So an outstanding question is how well described are the artifacts in collections at language archives? There are several ways that the quality of a

[188] By views here, I mean web page presentations. That is, the web pages of the archive are influenced by the design need to craft the narrative for the funders rather than organization of content for artifact-reuse.

description can be measured, but quality metadata is notoriously hard to define.[189] One could take the perspective that the quality of a collection description is the completeness of the record description, on the basis of the number of metadata elements provided to the curator at the time the collection was deposited for initial description. This would include any description for the tiers (series, sub-series, file unit), and any specific elements required or mandated by the institutionally adopted metadata application profile. OLAC provides a basic quality judgment on submissions of participating archives when it assigns a rating to the archive. It does this by ranking the average number of fields provided by the archive against the number of fields in the OLAC application profile (see Hughes 2004 for further discussion).[190] More recent work has been done to review the descriptive quality of language archive holdings, (Burke & Zavalina 2019, 2020a, 2020b). This work looks at the textual content of descriptions. As far as I can tell no-one has reported on the accuracy of language archive metadata when compared with the semantic intent of metadata elements,[191] e.g., are collection records described with the DCMIType collection, and do records with the DCMIType collection actually refer to collections?

Accuracy in metadata builds trust between artifact re-users and holding institutions. Depositories which function as archives are expected to be responsible for their metadata and curation practice, including more consistent application and adherence to semantics; depositories which function as repositories often require depositors to provide the metadata about a resource. This allows for more opportunities for variation with regard to semantics.

Open Language Archive Community vocabularies are built on top of Dublin Core (Bird and Simons 2003b). This means that the core vocabularies of Dublin Core apply to records, not just the OLAC vocabularies. All of the archives in this thesis are participating members of OLAC; however, OLAC does not prescribe any model for the application of Dublin Core metadata to institutional records. Dublin Core can be applied in a variety of ways including hierarchically (with cascading levels of description and association between elements) or flat (where all elements are applied to a single record).

[189] For a more in-depth discussion on metadata quality assessment see discussion in Shreeves et al. (2005).
[190] http://www.language-archives.org/metrics/compare
[191] By semantics I mean the purpose for which the metadata schema element was designed.

Dublin Core approaches the type-nature of an object by suggesting the use of the DCMIType vocabulary. The suggestion to use the DCMIType vocabulary seems rather strong to me because the framers of Dublin Core saw fit to also provide this vocabulary. They did not provide any other vocabularies, but rather referenced each of the other vocabularies. This is not to say that DCMIType is the only way to describe the type-nature of an object but it is within the same set of metadata that many in linguistics and language documentation find foundational via the authority of the OLAC metadata standard which builds upon Dublin Core.

The DCMIType vocabulary only contains twelve terms (listed in section 4.1.3). Yet still language archives seem to use the terms in different ways. None of the archives reviewed have to-date submitted any of the reviewed collections to OLAC.

Across the spectrum of archival activities related to the preservation of cultural heritage materials, there are many metadata schema and application profiles with more specific relationships than what is provided for in Dublin Core.[192] While these other schema can in some contexts provide added value to collections described with Dublin Core, none of them helps define the "it". By "it" I primarily mean the type-nature of the thing being described as exemplified in the DCMIType vocabulary. To connect the "it" to both description and referencing practice (reference styles in publications generally form patterns based on the type-nature of items referenced), it is helpful to ask the following questions:

- What is the "it" that archives have? That is, what is the type-nature of items held at an archive?
- What is the "it" that archives don't have? That is, what items held at an archive have the wrong DCMIType applied or no DCMIType applied?

[192] Some of the more well-known metadata schemas used within the cultural heritage preservation community include: Conceptual Reference Model (CIDOC-CRM) used by museums; Europeana Data Model (EDM) for use by the data aggregator Europeana (7,000 institutions use this); Categories for the Description of Works of Art (CDWA); Lightweight Information Describing Objects (LIDO); and PCore used by audio archives. Some of these focus on describing content, others on format/carrier.

- What is the "it" that people think they are getting? That is, what type-nature do consumers expect when they investigate a record with a specific DCMIType label? And what is the consumer hoping to get—e.g., a collection from Mali, a transcribed narrative viewable in ELAN,[193] or set of files arranged for use in big data phonetic analysis?
- What is the "it" that Zotero recognizes? That is, how does Zotero interpret an item based on the declared type-nature?

By asking these four questions, we can look at the value proposition that archives are making to Class1 and Class2 actors, the value that Class1 and Class2 actors are expecting to receive, and how the artifact's bibliographic metadata will translate to formatted references via Zotero.

These questions can guide archive administrators as they make decisions about how and why to apply DCMIType terms consistently. In the following sections I discuss DCMIType as a fundamental element of the description of records.

6.3.2.1 Where have all the collections gone?

In this thesis I have focused on collections and audio artifacts. In the DCMIType vocabulary both **collection** and **sound** are valid terms. We should be able to find the collections reviewed in this thesis within the OLAC set of records. Figure 62 indicates that there are 771 records in OLAC for collections. This is a vast under representation if we consider that aggregate works (including collections like RB5) should have a DCMIType of **collection**.

[193] https://archive.mpi.nl/tla/elan

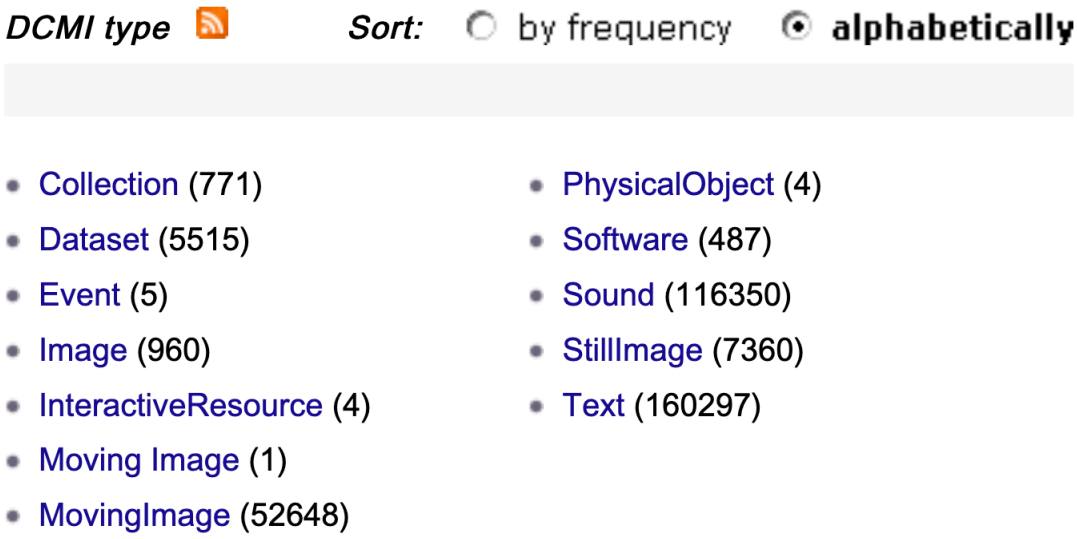

Figure 62. OLAC DCMITypes on 20 March 2021[194]

As shown in Figure 63, the collection records which are listed in OLAC are provided by only seven of the participating archives. Of the archives reviewed in this thesis, Pangloss and PARADISEC were the only ones to provide collection records to OLAC. All of the PARADISEC-provided records were part of a single collection, and it seems to be an administrative error that the DCMIType **collection** was applied to each item within the collection.[195] The Cocoon collection records do represent aggregate works as an archive would conceive of a collection—an administrative unit for managing artifacts. Figure 64 shows one such collection record. However, even though Cocoon does provide these collection records, no relationship data is provided to the records of the constituent parts. Further, the specific collections I review in Chapter 4 are for some reason not included in OLAC.

[194] http://dla.library.upenn.edu/dla/olac/browse.html?browse=dcmi_type_facet&
OLAC does not persist its data, so updates by archives to the aggregator will alter the numbers presented. OLAC metadata showing these figures is persisted in Zenodo as Paterson (2021).

[195] The alternative view is that there is only one collection within PARADISEC which is correctly noted for DCMIType at their item level. It seems that many of the items in the concerned collection are aggregate works and therefore would accurately be described with the DCMIType term **collection**. The collection in question is: https://catalog.paradisec.org.au/collections/TNS1 .

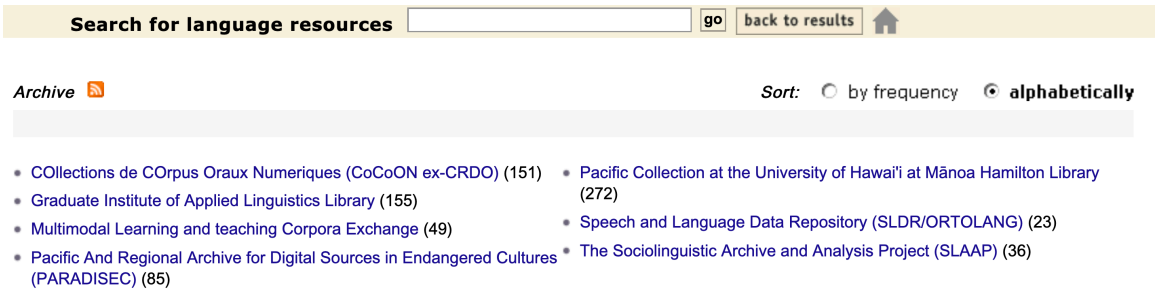

Figure 63. OLAC Collections by archive on 20 March 2021

This shows us that the “it” that archives have when measured as collections is not well shared with OLAC and may not be well described at all. The “it” that archives are saying they have is something other than aggregate works. In a similar vein then the “it” that archives do not have is the inverse.

OLAC Record

oai:crdo.vjf.cnrs.fr:cocoon-b3fb7859-9bda-3345-b359-61f9bfda6bf4

Metadata

Title:	Vietnamese Attitudes (VnA)
Access Rights:	Freely accessible
Alternative Title:	Vietnamese Attitudes (VnA): Audio-video-electroglottographic corpus for the study of attitudes in Vietnamese Vietnamese Attitudes (VnA): Corpus audio, vidéo et électroglottographique pour l'étude des attitudes en langue vietnamienne Vietnamese Attitudes (VnA): Bộ dữ liệu tiếng nói-hình ảnh phục vụ nghiên cứu thái độ cảm xúc trong tiếng Việt
Contributor (depositor):	Mac, Dang Khoa
Date Available (W3CDTF):	2015-02-26
Date Issued (W3CDTF):	2015-05-15T18:50:14+02:00
Description:	Le corpus Vietnamese Attitudes (VnA) est un corpus audio-visuel pour l'étude des attitudes simulées en langue vietnamienne en vue de leur caractérisation et de la synthèse de la parole expressive en vietnamien. Il comporte un ensemble d'enregistrements vidéo de 125 phrases composées de une à huit syllabes. Ces phrases ont été produites en chambre sourde par deux locuteurs (un homme et une femme, originaires de Hanoi) avec 16 attitudes ou expressions : déclaration, question simple, exclamation de surprise neutre, exclamation de surprise positive, exclamation de surprise négative, évidence, doute/incrédulité, autorité, irritation, ironie sarcastique, mépris, politesse, admiration, maternelle séduction et familière. Les phrases contiennent des syllabes portant les différents tons (2, 3, 4, 5, 6, 5b, 6b) en position de début, milieu et fin permettant l'étude de l'interaction des tons lexicaux avec la prosodie. Une répétition parmi les trois effectuées par le locuteur masculin comporte également des signaux électro-glotto-graphiques.
Identifier:	Ancienne cote: crdo-COLLECTION_VN_ATTITUDE
Identifier (URI):	https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-b3fb7859-9bda-3345-b359-61f9bfda6bf4 https://doi.org/10.34847/cocoon.b3fb7859-9bda-3345-b359-61f9bfda6bf4 https://cocoon.huma-num.fr/exist/crdo/ark:/87895/1.17-498728
Language:	Vietnamese; Vietnamien
Language (ISO639):	vie
Publisher:	Multimédia, Informations, Communication et Applications
Subject:	Corpus audio-visuel pour l'étude des attitudes simulées en langue vietnamienne Vietnamese language Vietnamien
Subject (ISO639):	vie
Type (DCMI):	Collection

OLAC Info

Archive:	COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO)
Description:	http://www.language-archives.org/archive/crdo.vjf.cnrs.fr
GetRecord:	OAI-PMH request for OLAC format
GetRecord:	Pre-generated XML file

OAI Info

OaiIdentifier:	oai:crdo.vjf.cnrs.fr:cocoon-b3fb7859-9bda-3345-b359-61f9bfda6bf4
DateStamp:	2020-11-28
GetRecord:	OAI-PMH request for simple DC format

Search Info

Citation:	Mac, Dang Khoa (depositor). 2015. Multimédia, Informations, Communication et Applications.
Terms:	area_Asia country_VN dcmi_Collection iso639_vie

Inferred Metadata

Country:	Viet Nam
Area:	Asia

http://www.language-archives.org/item.php/oai:crdo.vjf.cnrs.fr:cocoon-b3fb7859-9bda-3345-b359-61f9bfda6bf4
Up-to-date as of: Wed Mar 10 18:12:48 EST 2021

Figure 64. Cocoon collection record in OLAC on 20 March 2021

6.3.2.2 The item type identification crisis

Categorizing artifacts into identified item types seems to be rather difficult for not just linguists but publishers as well.[196] We have already seen in footnote 6 in Chapter 1 how linguists have had challenges distinguishing a corpus from a collection. We have also seen in Section 6.2 how data scientists have struggled with the distinction between data and software. A review of the item types registered with DataCite shows that a significant portion of the DOIs which point to "data" are actually pointing to textual objects like journal articles and conference proceedings (Robinson-Garcia et al. 2017).[197] In this way we can say there is sometimes a confusion between "data" and a conference paper or any other form of grey literature which might find itself in an institutional repository that assigns DataCite DOIs to content.

Reference works which are very often multi-part, multi-contributor aggregate works have suffered from terrible confusion as well. Zotero provides specific item types for articles or sections of reference works because some style sheets require specific information regarding these resources. Reference works are common in linguistics and include works like: *The Blackwell Companion to Phonology*, the *International Encyclopedia of Linguistics*, the *World Atlas of Language Structures*, the *Ethnologue*, *Linguistic Minorities in Europe Online*, and the *Glottolog*. Their creation and use has traditionally been understood to be in the domain of text and published as collections of articles which are a part of a whole. Recently when looking to reference Škevin Rajko & Šimičić (2020) a part of *Linguistic Minorities in Europe Online*, the publisher, *De Gruyter*, has told the CrossRef API that the work article is a **dataset**. This impacts referencing because Zotero gets its metadata for propagating the reference via the CrossRef API.

Lest we think this is some clerical error in the mis-categorization of reference works, let us consider some remarks from one of the editors of the *Glottolog*. These comments were provided in the context of a discussion across two threads on Github where referencing

[196] Linguists are often content providers and metadata generators for OAIS model repositories, meaning that these systems are susceptible to the item type identification crisis.

[197] It is also acknowledged that for some, like linguists who focus on text based corpora, e.g., the Brown Corpus of Standard American English (Francis & Kučera 1961), textual objects are the focal objects in their investigations.

strategies and item types were being discussed.[198,199] In the first quote we can see that the user *haspelmath*, an editor at the Glottolog, in the Glottolog issue tracker on Github, presents the understanding that the Glottolog is a published book and not a web page (Zotero detects all Glottolog pages except the base URL `https://glottolog.org` as a web page), while in the second quote we can see that the same user the next day has the understanding that the Glottolog is a dataset.

> Glottolog is a published book, not a "webpage". You could order a physical copy through some print-on-demand service, but more to the point, Glottolog is of course a special kind of book: A highly structured database that does not make much sense on paper. Maybe such books should eventually get a special Bib category, but for the time being, "book" is the closest match. "Webpage" would send the wrong message. (*haspelmath* Glottolog Github issue 536)

> Yes, the way I understand it, the Glottolog *IS* the dataset. The webpage is just a particular presentation of the data. Maybe we should somehow make this clearer. (*haspelmath* Glottolog Github issue 535)

An FRBR approach to describing the Glottolog via DCMIType would allow for the work to have multiple expressions to match its various versions, and for each expression to have three manifestations: a book with the DCMIType **Text**, a web page with the DCMIType **InteractiveResource**,[200] and a dataset with the DCMIType **Dataset**. So in some cases we have publishers who have challenges distinguishing between parts and wholes, and in other cases we have publishers who have challenges distinguishing between manifestations, and yet other cases we have publishers who have troubles distinguishing an item type.

The item type identity crisis appears even in the artifacts reviewed for this thesis. For example, we have seen how in the Kaipuleohone artifact record (Figure 53) is listed as **Sound**, It is clear that there are at least two separate works #1 and #3, indicating

[198] https://github.com/glottolog/glottolog/issues/535
[199] https://github.com/glottolog/glottolog/issues/536
[200] http://purl.org/dc/dcmitype/InteractiveResource

the record represents a collection type artifact. Minimally, the one-to-one principle for Dublin Core is not followed. However, this is not the only record which suffers from mis-alignment of DCMITypes.

Careful investigation of the artifacts reviewed under the L&CA (item ID: 52216) and PARADISEC (*Kamuku wordlist*) show that they come from the same recording session. That is, they are the same work, expression, and manifestation. Unlike PARADISEC, the L&CA has broken up the artifacts which comprise the *Kamuku wordlist* into several unlinked records.[201] By breaking up the recordings into their aggregate works, the L&CA has provided access to more contributors and their roles than the PARADISEC record offers. This can be viewed as an overall increase in description. Neither Nabu (PARADISEC)[202] nor the L&CA provide DCMIType information via their public interfaces. However, in their OLAC feeds, they do provide DCMIType information. In both cases the OLAC records are listed as **Sound**, not **Collection** which would be more appropriate given the Dublin Core definition: *A collection is described as a group; its parts may also be separately described.* In the L&CA case, they do not use the DCMIType vocabulary within their DSpace instance. In lieu of the DCMIType vocabulary, the L&CA uses a custom vocabulary to modify the Dublin Core type element. They call the vocabulary "the scholarly work vocabulary". It appears to be heavily influenced by or an extension of the *EPrints Type Vocabulary Encoding Scheme*. [203,204] Within this custom vocabulary, the L&CA record indicates that the set of audio files is a "dataset".

Specifically in regards to how items are classified (assigned an item type), the difference between **dataset** and **collection** can have an impact on what metadata is perceived as necessary for transmission to reference managers such as Zotero.

[201] Unlinked at least from the view of the artifact shown in the review. Other artifacts, such as L&CA item:82881 which is a transcription of the elicited wordlist, link to the audio, but no reverse relationship is shown in the audio record.

[202] The Nabu interface does not show this metadata. One needs to visit the item record at the OLAC website to see this metadata: http://www.language-archives.org/item/oai:paradisec.org.au:RB5-Kainji_Kamuku_wordlists

[203] http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Type_Vocabulary_Encoding_Scheme

[204] It is possible that the DCMIType is derived in an XML transform from some combination of fields, as no value of "sound" exists in the record (see Figure 66). Another explanation might be that the current item type is not the same as what appears on OLAC as of January 2021, L&CA had not updated their OLAC feed since 2013. Also note that the OLAC text types differ between the archives (PARADISEC: primary text; L&CA: lexicon). See: http://www.language-archives.org/item/oai:sil.org:52216.

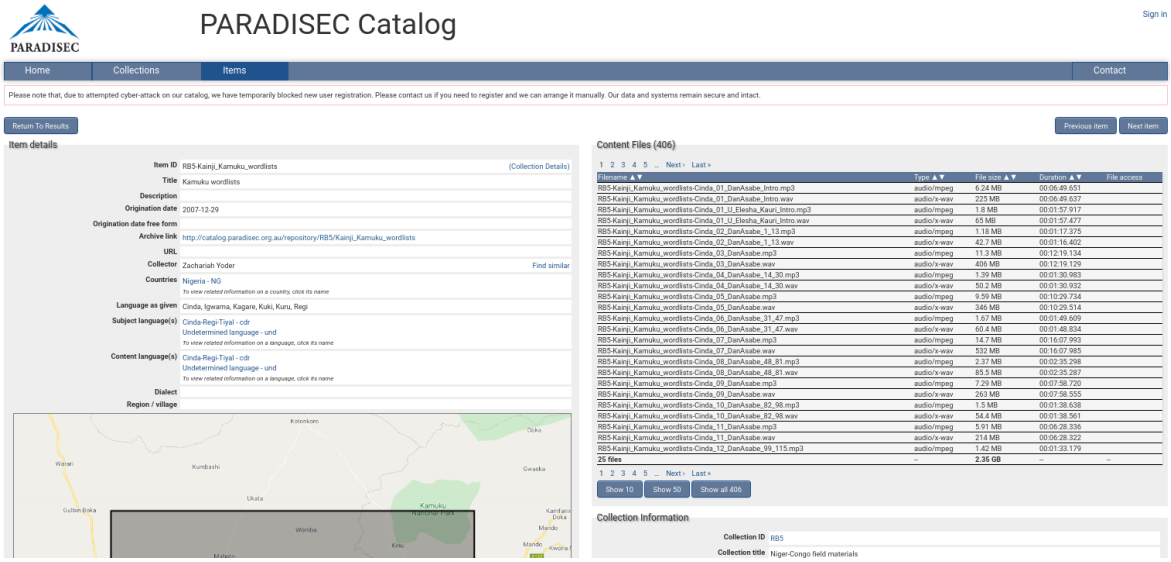

Figure 65. Kamuku wordlist at PARADISEC

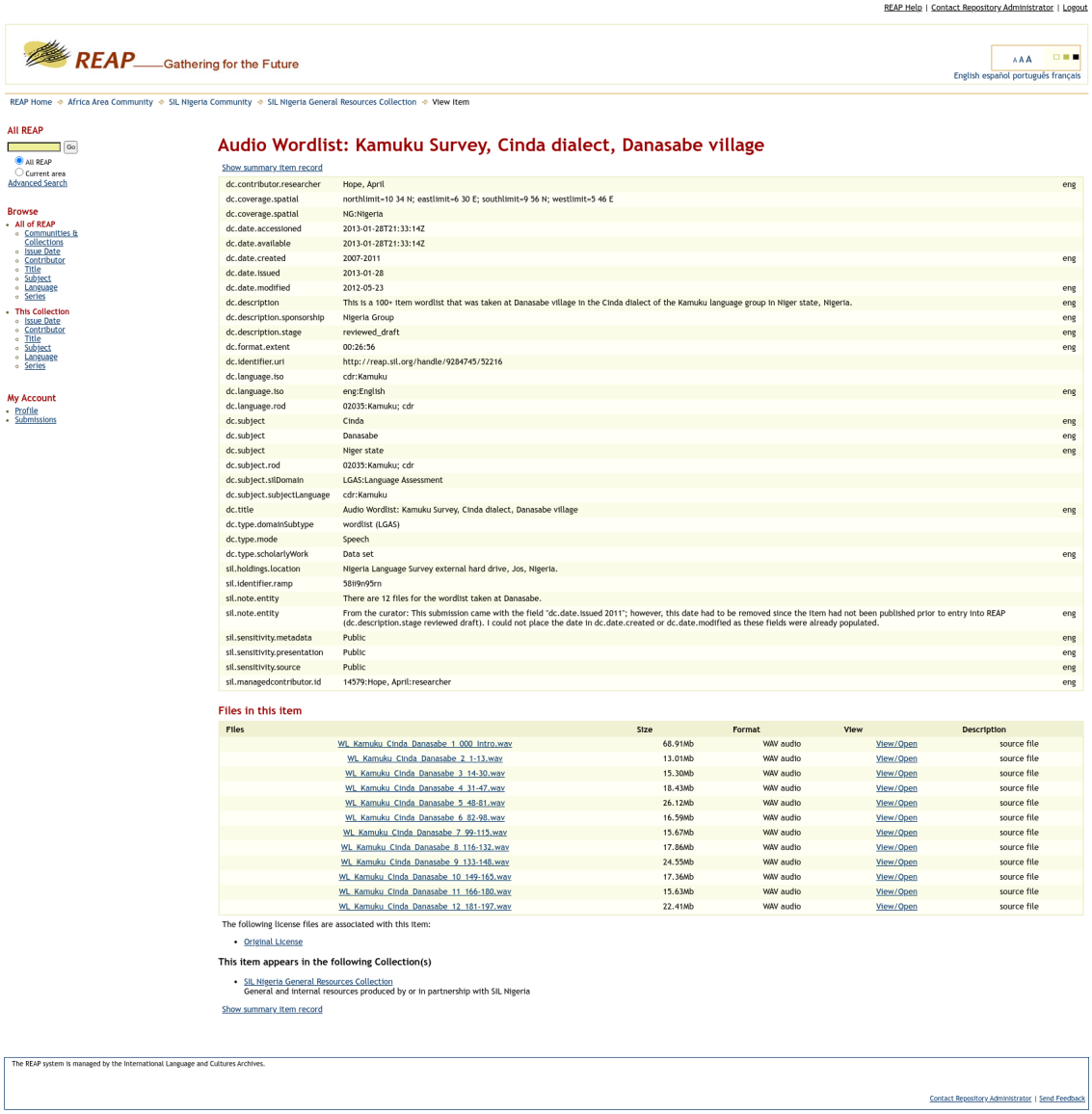

Figure 66. SIL L&CA item 52216 as seen in SIL's DSpace interface

6.3.2.3 Identifying the item for referencing

One final consideration in the structure of what is communicated between archive and reference manager is the **item type**. Getting people and systems to know "what" they have is an important part of generating value in communities of re-users. Understanding artifacts and generating community value are significantly enhanced by identifying item type. Zotero items types (as shown in Table 5) are displayed in the application's user

interface and pulled from the item types in CSL specification. The item types are only added to the CSL specification on the basis that some style sheet makes a significant distinction to warrant a new item type. In this way the categories match what reference style sheets (all 9,000+ that CSL supports) call for, not what Zotero users might consider as independent types of items. Items types are informed by reference patterns in style sheets, but also consider the impacts on how a user should reference the same content in another style sheet. Some well articulated style sheets like Chicago or APA specify citation and reference formats for each of the item types in Table 18.

It seems that there is a great deal of confusion, a total lack of awareness, or a total lack of empathy for existing item types as required by style sheets. Item types are those things which pattern together for a common reference. For instance many of us are familiar with a book or a journal as an item type and how they are referenced differently. Some of us may be familiar with a video as an item type, but are we careful to distinguish the difference between a film and a video as some style sheets call for? From time to time there are new item types which must be considered, such as the poster presentation which is not a panel presentation nor a conference paper. But within the disciplines of language documentation and linguistics do archives under-describe their holdings? For instance, some audio recordings may be better cited as an **interview**, a distinct item type acknowledged and prescribed associated roles by some style sheets, rather citing the artifact as a generic **audio recording** with OLAC specific roles such as **speaker**. If we treat the interview as "data" and merely reference it as data, we risk losing the opportunity to clearly tell the reader something important about the nature of the artifact we reference. OLAC even assents to the importance of describing text types with the *Discourse Type Vocabulary*. However, the relevant term **dialogue** is perhaps a bit broader than the CSL item type **interview**. Keep in mind, OLAC, built upon Dublin Core, was designed for artifact discovery, not considering the requirements of bibliographic metadata.

Table 18. Zotero item types with additional types available via CSL

Zotero		**CSL**
Artwork	Letter	Collection
Audio Recording	Magazine Article	Dataset
Bill	Manuscript	Figure
Blog Post	Map	Musical_score
Book	Newspaper Article	Pamphlet
Case	Patent	Review-book
Conference Paper	Podcast	Treaty
Dictionary Entry	Presentation	
Document	Radio Broadcast	
E-Mail	Report	
Encyclopedia Article	Software	
Film	Statute	
Forum Post	Thesis	
Hearing	TV Broadcast	
Instant Message	Video	
Interview	Web Page	
Journal Article		

The quality of a collection's description is important to most archivists because it is often representative of how helpful an archive can be to the archive's clients. Common practice at many non-language-archives is to progressively increase the description details of a collection; this may take years or decades, as there are many collections and few archivists. Additionally, finding someone with the interest and the domain knowledge required to evaluate artifacts in ways which will enrich descriptions is often challenging. The discussion of the evolution in arrangement and description of a collection as "curation" stands in contrast to tasks I have previously discussed under the label of "curation". The main difference being that my previous discussion has been focused on the establishment of new collections, whereas now I am discussing the revision of records or the enhancement of existing records. Broadly conceived, curation tasks are related to the activities of preservation, digitization, and format transformation which may be necessary for language artifact use (Weber 2021).[205] Most visitors (or clients) to an archive must accept the metadata provided at its face value—having no other reference point

[205] Activity done to an existing collection is distinct from assembling a collection. In both cases the format of artifact may need to undergo conservation activities to enable further use.

upon which to evaluate the accuracy (quality) of the metadata. With language archives, the situation is especially dire, because current social practice in documentary linguistics does not greatly reward curation activities—improving the metadata and records of collections. Peer-reviewed journal articles about collections are more easily translated into career building blocks, e.g., examples include Salffner (2015), Caballero (2017), Gawne (2018), Oez (2018), Franjieh (2019), Holley-Kline (2019). So rather than working with an archive to enrich the descriptions of the holdings, it is seen as more rewarding (career-wise) to publish an article about the holdings. The trend to move collection descriptions out of the archive's purview and into the journal-sphere has led to meta-discussion on what should be included in these types of articles, e.g., Sullivant (2020), Fitzgerald (2021). However, the separation of archival content from description (via external publications) reinforces the OAIS or repository-type business model that many language archives seem to have adopted. A more integrated approach to collection description should be pursued; however this requires a different management approach by language archives. Such an integrated approach would need to recalibrate peer-review in such a way as to make sure the effort expended during peer-review was not spent on an article about the collection, but rather on the arrangement and description of the content of the collection. Such a recalibration would not only reward newly created collections, but could also serve to facilitate the description of older materials in "legacy" collections. Nordmoe (2013) discusses the losses encountered because the social reward system for career linguists values the acquisition of new artifacts over the reuse of already existing artifacts and the preservation of production contexts (of existing artifacts). Weber (2021) describes the lack of scholarly acknowledgment for work conducted with legacy materials; a sentiment articulated outside of the field of linguistics (Suarez & Tsutsui 2004; Thessen et al. 2019). A progressive and open peer-review model would acknowledge the contributions of curation. The much needed curation of linguistic collections impacts the ability to craft informative references. Among the collections consulted for this thesis, the ZF1 collection at PARADISEC is in need of curation. Among other things the DataCite API returns a notice "PLEASE PROVIDE TITLE" as its title. An academic system that acknowledges curation could encourage cross-generational collaborations with emeritus professors over

their fieldwork notes and recordings. The added attention to curation would also facilitate review and correction of bibliographic metadata errors such as those which occur in the ZF1 collection.

CHAPTER 7

Forward movement

We have seen that there are technical challenges inhibiting the easy reuse of bibliographic metadata between archives and authors who reference artifacts. There are two things our community of language artifact creators, stewards, and consumers can do to improve the transmission of bibliographic metadata between archives and tools such as reference managers. The first is to strengthen the relationship between archives and software tool developers. The second is to change the status quo of language archives in how they share bibliographic metadata about their holdings.

A contribution of developer time, such as sponsoring specific development goals to be conducted by the Zotero team, or developing code outside the core project (but licensed in a comparable way) and contributed to the Zotero project for distribution with Zotero would widely benefit language-artifact-using-communities.[206] Additional engagement by Zotero end-users with the CSL and Zotero developers, describing interactive challenges in the citation of archival materials in language artifact situations, would help those developer teams understand the contexts in which their products are used.[207] I have found these teams to be very responsive and knowledgeable of current issues in scholarly communication. They are priority driven, meaning that there are constraints on their activities. They evaluate the cost of work to be done (financial and man hours) and evaluate it in the context of the impact any change will have upon end-user expectations. When engaging with any open source project, important factors affecting successful implementation include:

[206] Such development could be undertaken in several ways, and need not be solely conducted as part of linguistic research. Rather development could be part of library science, digital humanities, or computer science education or research activities. Generally the development is not arduous, and could be the collaboration of several people as part of a course level project. Currently recognition and reward are limited outside of personal satisfaction and motivation.

[207] For example, adding use cases where using MARC relator roles in the citation can be added to the discussion here: https://discourse.citationstyles.org/t/feedback-csl-1-0-2-media-roles/1669.

clearly communicating the use case, the needed changes, and the breadth of the proposed impact among the product end-users, and possibly financing for developer time. Zotero is an open source project. It is also a modular project. Components such as translators (custom page scrapers written in JavaScript) and import plugins (for instance importing *Lameta*[208] records or OLAC-OAI records[209] to make bibliographic records) need not be written by the current Zotero developers. Contributions of code can come from a variety of global locations, individuals, teams, and sponsoring agencies.

The second way to contribute to better transmission of metadata between language artifact archives and the users of those artifacts is to take responsibility for the transmission process as a discipline, building into the repositories the technologies needed to communicate with tools like Zotero. In many cases the metadata is already part of the archive's records. It is only a matter of making it available, not just to Zotero but also to search engines like Google Scholar. This is an important step in leading our communities into a publishing practice where the evidentiary record is cited and referenced. The exact technological approach each archive needs to take will be different, depending on their descriptive and arrangement practices and the current state of their web platform. The front door (interface) of digital archives is constantly changing. For example, during the research phase of this thesis Pangloss and ELAR both launched new web interfaces. PARADISEC is looking forward to a new interface later in 2021 (Thieberger p.c., 2021). In that new interface they plan on making some Dublin Core Metadata available in the HTML code, some of which can be seen in Figure 67 lines 14–20.

[208] https://sites.google.com/site/metadatatooldiscussion/home
[209] https://forums.zotero.org/discussion/52296/open-language-archives-community-olac-translator-needed

```
<!doctype html>
<html ng-app="pdsc" ng-controller="MetadataCtrl">
  <head>
    <meta charset="utf-8">
    <base href="/">
    <meta name="description" content="">
    <meta name="viewport" content="width=device-width">
    <meta name="referrer" content="no-referrer-when-downgrade">

    <!-- namespace declarations -->
    <link rel="schema.dc" href="http://purl.org/dc/elements/1.1/" />

    <!-- item metadata -->
    <meta name="dc.title" content="{{item.title}}" />
    <meta name="citation_public_url"  content="{{item.identifier[1]}}" />
    <meta ng-repeat="contributor in item.contributor" name="dc.creator" content="({{contributor.role}}) {{contributor.name}}" />
    <meta name="dc.date" content="{{item.date}}" />
    <meta name="dc.description" content="{{item.description}}" />
    <meta name="dc.rights" content="{{item.rights}}" />
    <meta name="dc.publisher" content="PARADISEC" />
  <link href="/viewer/styles.2911294991e978a1456b.css" rel="stylesheet">
</head>
```

Figure 67. Preliminary code for new HTML embedded metadata at PARADISEC

Catching the winds of change and navigating the archive towards the center of the value exchange remains the ongoing challenge of language archives.

References

Albarillo, Emily E. & Nicholas Thieberger. 2009. Kaipuleohone, the University of Hawai'i's Ethnographic Archive. *Language Documentation & Conservation* 3(1). 154–181. http://hdl.handle.net/10125/4422

Aldridge, Boon & Gary F. Simons. 2018. Kenneth Pike and the making of Wycliffe Bible Translators and SIL International. (February 2, 2018). Christianity Today — Christian History. https://www.christianitytoday.com /history/2018/february/kenneth-pike-sil-wycliffe.html

Allinson, Julie. 2008. Describing Scholarly Works with Dublin Core: A Functional Approach. *Library Trends* 57(2). 221–243. doi:10.1353/lib.0.0034

Altman, Micah. 2013. Connecting Research, Publications, and Evidence: The Lifecycle and Institutional Ecology of Data. Paper presented at: Public Access to Federally-Supported Research and Development Data and Publications: Data. Washington DC, (16th–17th May 2013). http://vimeo.com/ 71358834 (2014-05-21).

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Bradley McDonnell, Chris Cieri, Stefano Coretta, Koenraad De Smedt, Lindsay Ferrara, Robert Forkel, Susan Smythe Kung, Alon Lischinsky, Sebastian Nordhoff, Hugh J. Paterson III, Hiram Ring, Nick Thieberger & Margaret E. Winters. 2019. *Tromsø recommendations for citation of research data in linguistics.* Online: Research Data Alliance Linguistic Data Interest Group. doi:10.15497/rda00040

Arlitsch, Kenning & Patrick S. O'Brien. 2012. Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech* 30(1). 60–81. doi:10.1108/07378831211213210

Arlitsch, Kenning & Patrick S. O'Brien. 2013. Google Scholar and Institutional Repositories. *Improving the visibility and use of digital repositories through SEO,* 79–94. (LITA guides), Chicago, Illinois: ALA TechSource, an imprint of the American Library Association.

Austin, Peter K. 2013. Language documentation and meta-documentation. In Mari C. Jones & Sarah Ogilvie, *Keeping Languages Alive: Documentation, Pedagogy and Revitalization,* 3–15. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139245890.003

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

Barwick, Linda. 2004. Turning It All Upside Down . . . Imagining a distributed digital audiovisual archive. *Literary and Linguistic Computing* 19(3). 253–263. doi:10.1093/llc/19.3.253

Barwick, Linda. 2005. Networking digital data on endangered languages of the Asia Pacific region. *International Journal of Indigenous Research* 1(1). 11–16. https://hdl.handle.net/2123/1317

Barwick, Linda & Amanda Harris. 2013. PARADISEC: its history and future. Paper presented at: Research, Records and Responsibility: PARADISEC 10th Anniversary Conference. University of Melbourne, 2 December. https://hdl.handle.net/2123/9833

Barwick, Linda & Nicholas Thieberger. 2012. Keeping records of language diversity in Melanesia: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). In Nicholas Evans and Marian Klamer, eds. *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,* 239–253. (Language Documentation & Conservation Special Publication №5), Honolulu, Hawai‘i: University of Hawai‘i Press. http://hdl.handle.net/10125/4567

Barwick, Linda & Nicholas Thieberger. 2018. Unlocking the archives. *Proceedings of the 2017 XXI FEL conference.* 135–139. Alcanena, Portugal 19–21 October 2017: Foundation for Endangered Languages. https://hdl.handle.net/11343/220007

Berez, Andrea 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:10.1353/ort.2013.0010

Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly & Tyler Heston. 2017. A survey of current reproducibility practices in linguistics journals, 2003–2012. University of Hawai‘i, ms. https://sites.google.com/a/hawaii.edu/data-citation/survey

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. doi:10.1515/ling-2017-0032

Bergqvist, Henrik. 2007. The role of metadata for translation and pragmatics in language documentation. *Language Documentation and Description* 4. 163-73. http://www.elpublishing.org/PID/055

Biber, Douglas. 1993a. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4). 243–257. doi:10.1093/llc/8.4.243

Biber, Douglas. 1993b. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19(2). 219–241. https://www.aclweb.org/anthology/J93-2001

Bird, Steven & Gary F. Simons. 2001. The OLAC metadata set and controlled vocabularies. In Thierry DeClerck, Steven Krauwer & Mike Rosner (eds), *In Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education.* 7-18. Université de Toulouse, France: EACL-ACL; elsnet. https://www.aclweb.org/anthology/W01-1506

Bird, Steven & Gary F. Simons. 2003a. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79(3). 557–582. doi:10.1353/lan.2003.0149

Bird, Steven & Gary F. Simons. 2003b. Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. *Computers and the Humanities* 37(4). 375–388. doi:10.1023/A:1025720518994

Bitgood, Stephen. 2006. An Analysis of Visitor Circulation: Movement Patterns and the General Value Principle. *Curator: The Museum Journal* 49(4). 463–475. doi:10.1111/j.2151-6952.2006.tb00237.x

Boerger, Brenda. 2011. To BOLDly Go Where No One Has Gone Before. *Language Documentation & Conservation* 5. 208–233. http://hdl.handle.net/10125/4499

Bourdieu, Pierre. 1977. The economics of linguistic exchanges. *Social Science Information* 16(6). 645–668. doi:10.1177/053901847701600601

Brander, Gina, Erin Langman, Tasha Maddison & Jennifer Shrubsole. 2019. Evaluating Bibliographic Referencing Tools for a Polytechnic Environment. *Evidence Based Library and Information Practice* 14(2). 4–32. doi:10.18438/eblip29489

Burke, Mary & Oksana L. Zavalina. 2019. Exploration of information organization in language archives. *Proceedings of the Association for Information Science and Technology* 56(1). 364–367. doi:10.1002/pra2.30

Burke, Mary & Oksana L. Zavalina. 2020a. Descriptive richness of free-text metadata: A comparative analysis of three language archives. *Proceedings of the Association for Information Science and Technology* 57(1)-e429. doi:10.1002/pra2.429

Burke, Mary & Oksana L. Zavalina. 2020b. Identifying Challenges for Information Organization in Language Archives: Preliminary Findings. In Anneli Sundqvist, Gerd Berget, Jan Nolin & Kjell Ivar Skjerdingstad, eds. *Sustainable digital communities: Proceedings of the 15th international conference, iConference 2020, Böras, Sweden, March 23–26,* 622–629. (Lecture Notes in Computer Science №12051), Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-030-43687-2_52

Butler, Declan. 2001. Los Alamos loses physics archive as preprint pioneer heads east. *Nature* 412(6842). 3–4. doi:10.1038/35083708

Byrum, John, Suzanne Jouguelet, Dorothy McGarry, Nancy Williamson, Maria Witt, Tom Delsey, Elizabeth Dulabahn, Elaine Svenonius & Barbara Tillett (The IFLA Study Group on the Functional Requirements for Bibliographic Records). 2009 [1998]. *Functional Requirements for Bibliographic Records: Final Report.* ([UBCIM publications, new series] IFLA Series on Bibliographic Control №19.) Munich, Germany: K.G. Saur. http://www.ifla.org/VII/s13/frbr

Caballero, Gabriela. 2017. Choguita Rarámuri (Tarahumara) language description and documentation: a guide to the deposited collection and associated materials. *Language Documentation & Conservation* 11. 224–255. http://hdl.handle.net/10125/24734

Carlyle, Allyson. 2006. Understanding FRBR As a Conceptual Model. *Library Resources & Technical Services* 50(4). 264–273. doi:10.5860/lrts.50n4.264

Case, Donald Owen. 2002. *Looking for information: A survey of research on information seeking, needs and behavior.* Amsterdam: Academic Press.

Chang, Debbie. 2010. *TAPS: Checklist for Responsible Archiving of Digital Language Resources.* Graduate Institute of Applied Linguistics thesis. https://web.archive.org/web/20130617120932/http://www.gial.edu/images/theses/Chang_Debbie-thesis.pdf

Chapman, John W., David Reynolds & Sarah A. Shreeves. 2009. Repository Metadata: Approaches and Challenges. *Cataloging & Classification Quarterly* 47(3–4). 309–325. doi:10.1080/01639370902735020

Christen, Kimberly, Alex Merrill & Michael Wynne. 2017. A Community of Relations: Mukurtu Hubs and Spokes. *D-Lib Magazine* 23(5/6). n.p. http://www.dlib.org/dlib/may17/christen/05christen.html (2021-04-17). doi:10.1045/may2017-christen

Christensen, Clayton M. 2011. *Market Disruptions & Online Learning. [video/MPEG4]. Duration: 4:31. YouTube.* (University Of Phoenix Lecture Series.) Forbes. https://www.youtube.com/watch?v=VmbSpTJXozk (2021-04-17).

Christensen, Clayton M., Taddy Hall, Karen Dillon & David S. Duncan. 2016. Know Your Customers' "Jobs to Be Done". *Harvard Business Review* 94(9). 54–60.

Chudnov, Dan, Peter Binkley, Jeremy Frumkin, Michael J Giarlo, Mike Rylander, Ross Singer & Ed Summers. 2006. Introducing UnAPI. *Ariadne* 48. http://www.ariadne.ac.uk/issue/48/chudnov-et-al/ (2021-01-12).

Chudnov, Daniel & Deberah England. 2008. A New Approach to Library Service Discovery and Resource Delivery. *The Serials Librarian* 54(1–2). 63–69. doi:10.1080/03615260801973448

CornellCIS. 2018. arXiv Looks to the Future with Move to Cornell CIS. Cornell Computing and Information Science. https://cis.cornell.edu/arxiv-looks-future-move-cornell-cis (2021-03-13).

Coyle, Karen. 2016. *FRBR, before and after: a look at our bibliographic models.* Chicago: ALA Editions, an imprint of the American Library Association. https://kcoyle.net/beforeAndAfter/Pt2-978-0-8389-1364-2.pdf (2021-03-06).

Coyle, Karen. 2008. Coyle's InFormation: Literals and non-literals, take 2 (5 September). Coyle's InFormation. http://kcoyle.blogspot.com/2008/09/literals-and-non-literals-take-2.html (2021-01-26).

Crystal, David. 2004. Creating a world of languages. Paper presented at: Linguapax. Barcelona, Spain. https://www.davidcrystal.com/Files/BooksAndArticles/-4857.pdf (2021-04-13).

Daquino, Marilena & Francesca Tomasi. 2015. Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects. In Emmanouel Garoufallou, Richard J. Hartley and Panorea Gaitanou, eds. *Metadata and Semantics Research: Proceedings of the 9th Research Conference, MTSR 2015, Manchester, UK, September 9-11, 2015,* 424–436. (Communications in Computer and Information Science №544), Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-24129-6_37

DataCite Metadata Working Group, Joan Starr, Noémie Ammann, Jan Ashton, Amy Barton, Jannean Elliott, Marie-Christine Jacquemont-Perbal, Merja Karjalainen, Andreas Oskar Kempf, Lynne McAvoy, Elizabeth Newbold, Lars Holm Nielsen, Sebastian Peters, Madeleine De Smaele, Natalija Schleinstein, Wolfgang Zenk-Möltgen, Mohamed Yahia & Frauke Ziedorn. 2014. DataCite Metadata Schema for the Publication and Citation of Research Data v3.1: Documentation. Hannover, Germany: DataCite - International Data Citation Initiative. https://schema.datacite.org/meta/kernel-3.1/index.html (2021-02-27). doi:10.5438/0010

DataCite Metadata Working Group, Madeleine de Smaele, Robin Dasler, Jan Ashton, Sophie Roy, Martin Fenner, Mark Jacobson, Isabel Bernal Martínez, Marleen Burger, Mohamed Yahia, Lisa Zolly, Ted Habermann, Anne Raugh, Violeta Ilik & Andreas La Roi. 2019. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.3. Hannover, Germany: DataCite - International Data Citation Initiative. https://schema.datacite.org/meta/kernel-4.3 (2021-02-27). doi:10.14454/7xq3-zf69

Davis, Philip M. & Matthew J. L. Connolly. 2007. Institutional Repositories: Evaluating the reasons for non-use of Cornell University's installation of DSpace. *D-Lib Magazine* 13(3/4). n.p. http://www.dlib.org/dlib/march07/davis/03davis.html (2021-03-11). doi:10.1045/march2007-davis

Dawson, Hope, (Editor). 2011. *Language Style Sheet.* Washington, DC: Linguistic Society of America. https://www.linguisticsociety.org/sites/default/files/LANGUAGE_journal_style_sheet.pdf (2020-07-23).

Decourselle, Joffrey, Fabien Duchateau & Nicolas Lumineau. 2015. A Survey of FRBRization Techniques. In Sarantos Kapidakis, Cezary Mazurek and Marcin Werla, eds. *Research and Advanced Technology for Digital Libraries. Theory and Practice of Digital Libraries (TPDL) 2015,* 185–196. (Lecture Notes in Computer Science №9316), Poznań, Poland: Springer. doi:10.1007/978-3-319-24592-8_14

Diamond Jr., Arthur M. 1986. What is a Citation Worth? *The Journal of Human Resources* 21(2). 200–215. doi:10.2307/145797

Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. *Language Documentation and Description* 6. 37–52. http://www.elpublishing.org/PID/070

Duong, Khue. 2010. Rolling Out Zotero Across Campus as a Part of a Science Librarian's Outreach Efforts. *Science & Technology Libraries* 29(4). 315–324. doi:10.1080/0194262X.2010.523309

Evans, David S. 2011a. Invisible Engines. In David S. Evans, ed. *Platform Economics: Essays on Multi-Sided Businesses,* 376–381. (SSRN Scholarly Paper), ID 1974020. Rochester, NY: Competition Policy International. https://papers.ssrn.com/abstract=1974020 (2021-03-16).

Evans, David S. 2011b. More than Money. In David S. Evans, ed. *Platform Economics: Essays on Multi-Sided Businesses,* 282–313. (SSRN Scholarly Paper), ID 1974020. Rochester, NY: Competition Policy International. https://papers.ssrn.com/abstract=1974020 (2021-03-16).

Farrokhnia, Maliheh. 2019. *Searching for Cultural Heritage Information: Ontology-based Modeling of User Needs.* (OsloMet Avhandling 2019 №15.) Oslo, Norway: Oslo Metropolitan University. https://hdl.handle.net/10642/7245

Fenner, Martin, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone & Tim Clark. 2019. A data citation roadmap for scholarly data repositories. *Scientific Data* 6(1)-28. doi:10.1038/s41597-019-0031-8

Ferreira, Vera, Buachut Watyam, Siripen Ungsitpoonporn & Mandana Seyfeddinipur. 2021. Planting the seed: A bottom-up approach for archive creation. Paper presented at: PARADISEC@100. Sydney, Australia. https://youtu.be/VpVcAokmPc8

Fitzgerald, Colleen M. 2021. A framework for Language Revitalization and Documentation. *Language* 97(1). e1–e11. doi:10.1353/lan.2021.0006

Francis, W. Nelson & Henry Kučera, (Compilers). 1961. *Brown Corpus of Standard American English.* Providence, RI: Brown University.

Franjieh, Michael. 2019. The languages of northern Ambrym, Vanuatu: A guide to the deposited materials in ELAR. *Language Documentation & Conservation* 13. 83–111. http://hdl.handle.net/10125/24849

Ganahl, Rainer. 2001. Free Markets: Language, Commodification, and Art. *Public Culture* 13(1). 23–38. https://muse.jhu.edu/article/26231

Gawne, Lauren. 2018. A Guide to the Syuba (Kagate) Language Documentation Corpus. *Language Documentation & Conservation* 12. 315–338. http://hdl.handle.net/10125/24768

Gawne, Lauren & Andrea L. Berez-Kroeker. 2018. Reflections on Language Documentation 20 Years after Himmelmann 1998. In Bradley McDonnell, Andrea L. Berez-Kroeker and Gary Holton, eds. *Reflections on reproducible research,* 22–32. (Language Documentation & Conservation Special Publication №15), Honolulu, Hawai'i: University of Hawai'i Press. http://hdl.handle.net/10125/24805

Gawne, Lauren, Andrea Berez-Kroeker, Helene N Andreassen & Eve Okura. 2017. Data Citation in Linguistic Typology. Paper presented at: 12th Conference of the Association for Linguistic Typology (ALT). Canberra, Australia, December 12-14. http://hdl.handle.net/10125/51904

Gawne, Lauren, Berez-Kroeker, Andrea L., Barbara F. Kelly & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11. 157–189. http://hdl.handle.net/10125/24731

Good, Jeff. 2011. Data and language documentation. In Peter K. Austin & Julia Sallabank, eds. *The Cambridge Handbook of Endangered Languages,* 212–234. (Cambridge Handbooks in Language and Linguistics), Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511975981.011

Guo, Shujian, Hyunjung Shin & Qi Shen. 2020. The Commodification of Chinese in Thailand's Linguistic Market: A Case Study of How Language Education Promotes Social Sustainability. *Sustainability* 12(18)-7344. doi:10.3390/su12187344

Harper, Russell David, (Editor). 2017. *The Chicago manual of style.* 17th edn. Chicago, Illinois: The University of Chicago Press.

Hartell, Rhonda L., (Editor). 1993a. *Alphabets of Africa.* Dakar, Senegal: UNESCO and Summer Institute of Linguistics. https://www.sil.org/resources/archives/5133

Hartell, Rhonda L., (Editor). 1993b. *Alphabets de langues africaines.* Dakar, Senegal: UNESCO and Société Internationale de Linguistique. https://www.sil.org/resources/archives/5043

Haspelmath, Martin. 2014. The Generic Style Rules For Linguistics (version December 2, 2014). Zenodo, ms. https://zenodo.org/record/253501 (2020-07-23).

Haspelmath, Martin & Susanne Maria Michaelis. 2014. Annotated corpora of small languages as refereed publications: a vision. Diversity Linguistics Comment. Published (2014-03-11). https://dlc.hypotheses.org/691 (2021-01-07).

Heery, Rachel and Manjula Patel. 2000. Application Profiles: Mixing and Matching Metadata Schemas. *Ariadne* 25. http://www.ariadne.ac.uk/issue/25/app-profiles (2021-03-11).

Hemphill, David & Erin Blakely. 2015. Commodification of Language and Literacy. *Language, Nation, and Identity in the Classroom: Legacies of Modernity and Colonialism in Schooling,* New York, N.Y., USA: Peter Lang US. doi:10.3726/978-1-4539-1343-7

Hillmann, Diane I. & Jon Phipps. 2007. Application Profiles: Exposing and Enforcing Metadata Quality. *DCMI International Conference on Dublin Core and Metadata Applications* DC-2007—Singapore Proceedings. 52–62. Singapore; Dublin, Ohio, USA: Dublin Core Metadata Initiative & National Library Board Singapore. https://dcpapers.dublincore.org/pubs/article/view/866

Himmelmann, Nikolaus P. 1998. Documentary and Descriptive Linguistics. *Linguistics* 36(1). 161–195. doi:10.1515/ling.1998.36.1.161

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel, eds. *Essentials of Language Documentation,* 1–30. (Trends in Linguistics. Studies and Monographs [TiLSM] №178), Berlin; New York: Mouton de Gruyter. doi:10.1515/9783110197730.1

Himmelmann, Nikolaus P. 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation & Conservation* 6. 187–207. http://hdl.handle.net/10125/4503

Hodges, Ann E. & Brenda S. McClurkin. 2011. *Archives and Manuscripts Processing Manual.* Arlington, Texas: University of Texas Arlington—Special Collections Library. https://libraries.uta.edu/sites/default/files/2020-03 /processing-manual.pdf (2021-02-28).

Hoffmann, Roald, Artyom A. Kabanov, Andrey A. Golov & Davide M. Proserpio. 2016. Homo Citans and Carbon Allotropes: For an Ethics of Citation. *Angewandte Chemie International Edition* 55(37). 10962–10976. doi:https://doi.org/10.1002/anie.201600655

Holborow, Marnie. 2018. Language, commodification and labour: the relevance of Marx. *Language Sciences* 70. 58–67. doi:10.1016/j.langsci.2018.02.002

Holden, Christopher. 2019. *The Application of FRBR to Musical Works.* University of North Carolina at Chapel Hill thesis. doi:10.17615/0vzc-kn74

Holley-Kline, Sam. 2019. Isabel T. Kelly's Southern Paiute Ethnographic Field Notes, 1932–1934. *California Archaeology* 11(1). 93–95. doi:10.1080/1947461X.2019.1582131

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts and Paul Trilsbeek, eds. *Potentials of Language Documentation: Methods, Analyses, and Utilization,* 111–117. (Language Documentation & Conservation Special Publication №3), Honolulu, Hawai'i: University of Hawai'i Press. http://hdl.handle.net/10125/4523

Homans, George C. 1958. Social Behavior as Exchange. *American Journal of Sociology* 63(6). 597–606. doi:10.1086/222355

Hsu, Fang-Ming, Tser-Yieth Chen, Chiu-Tsu Fan, Chun-Min Lin & Chu-Mei Chiu. 2015. Factors affecting the satisfaction of an online community for archive management in Taiwan. *Program* 49(1). 46–62. doi:10.1108/PROG-12-2012-0068

HTML5 The Web Hypertext Application Technology Working Group. 2020. MetaExtensions - WHATWG Wiki. June 4, 2020 edn. https://wiki.whatwg.org/wiki/MetaExtensions (2021-01-13).

Hughes, Baden. 2004. Metadata Quality Evaluation: Experience from the Open Language Archives Community. In Zhaoneng Chen, Hsinchun Chen, Qihao Miao, Yuxi Fu, Edward Fox & Ee-peng Lim, eds. *Digital Libraries: International Collaboration and Cross-Fertilization,* 320–329. (Lecture Notes in Computer Science №3334), Berlin, Heidelberg. doi:10.1007/978-3-540-30544-6_34

Idri, Nadia. 2015. Zotero Software: A Means of Bibliographic Research and Data Organisation; Teaching Bibliographic Research. *Arab World English Journal (AWEJ)* 2. 124–133. doi:10.2139/ssrn.2843984

Iglesias, Carlos A., Mercedes Garijo, Daniel Molina & Paloma de Juan. 2009. VMAP: A Dublin Core Application Profile for Musical Resources. In Fabio Sartori, Miguel Ángel Sicilia & Nikos Manouselis, eds. *Metadata and Semantic Research,* 46, 1–12. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-04590-5_1

Jeon, Doh-Shin & Jean-Charles Rochet. 2010. The Pricing of Academic Journals: A Two-Sided Market Perspective. *American Economic Journal: Microeconomics* 2(2). 222–255. http://www.jstor.org/stable/25760393

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. *Language Documentation and Description* 2. 140–153. http://www.elpublishing.org/PID/026

Johnston, Pete. 2010. HTML5, document metadata and Dublin Core. (February 1, 2010). eFoundations. https://efoundations.typepad.com/efoundations/2010/02/html5-metadata-and-dublin-core.html (2021-01-13).

Jones, Kerry & Sanjin Muftic. 2020a. Endangered African Languages Featured in a Digital Collection: The Case of the ǂKhomani San | Hugh Brody Collection. In Rooweither Mabuya, Phathutshedzo Ramukhadi, Mmasibidi Setaka, Valencia Wagner & Menno van Zaanen (eds), *Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL): Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020*. 1-8. Marseille, France: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.rail-1.1

Jones, Kerry & Sanjin Muftic. 2020b. *Endangered African Languages Featured in a Digital Collection: The Case of the ǂKhomani San | Hugh Brody Collection. (Video Presentation)*. South Africa: University of Cape Town. doi:10.25375/UCT.12333893

Kanahele, Alana & Gary Holton. 2021. Mukurtu Hubs and Spokes. Paper presented at: PARADISEC@100. Sydney, Australia. https://www.youtu.be/Wj56fB_gV_U

Karcher, Sebastian. 2016a. Zotero for Data Repositories. Paper presented at: DataCite Webinar, 10 May. Online. https://figshare.com/articles/presentation/Zotero_for_Data_Repositories/3364708/1 (2021-03-11). doi:10.6084/M9.FIGSHARE.3364708.V1

Karcher, Sebastian. 2016b. Zotero for Data Repositories. DataCite. https://vimeo.com/166934111 (2021-03-11).

Katz, Daniel S, Kyle E Niemeyer, Arfon M Smith, William L Anderson, Carl Boettiger, Konrad Hinsen, Rob Hooft, Michael Hucka, Allen Lee, Frank Löffler, Tom Pollard & Fernando Rios. 2016. Software vs. data in the context of citation. *PeerJ Preprints* 4. e2630v1. doi:10.7287/peerj.preprints.2630v1

Klinke, Harald. 2018. TOP20 patterns in the DATE field in the MoMa collection. Note the differences between short and long dash, empty and "n.d."., "c." and "c. "--where there shouldn't be any. We find 1367 patterns: difficult to convert. Experience, suggestions anyone? #DigitalArtHistory #RStudio https://t.co/s9aoAL1UAW. @HxxxKxxx (2018-11-25T21:27Z). https://twitter.com/HxxxKxxx/status/1066805548866289664 (2021-03-17 08:58:13).

Koelling, Jill M. 2006. *Western States Digital Standards Group: Dublin Core Metadata Best Practices Version 2.1.1.* Denver, Colorado: Collaborative Digitization Program at the University of Denver. https://sustainableheritagenetwork.org/system/files/atoms/file/CDPDublinCoreBPs_0.pdf

Kräutli, Florian. 2016. *Visualising Cultural Data.* Royal College of Art dissertation. https://researchonline.rca.ac.uk/1774/

Kung, Susan Smythe, Ryan Sullivant, Elena Pojman & Alicia Niwagaba. 2021. *Archiving for the Future: Simple Steps for Archiving Language Documentation Collections.* New York, NY: Teach Online with Teach:able. https://archivingforthefuture.teachable.com

Kuzmin, Evgeny. 2013. Social Institutions Supporting Linguistic and Cultural Diversity in Cyberspace: Roles, Functions, Responsibilities. In Evgeny Kuzmin & Anastasia Parshakova, *Linguistic and Cultural Diversity in Cyberspace: Proceedings of the 2nd International Conference (Yakutsk, Russian Federation, 12-14 July, 2011),* 32–45. Moscow: Interregional Library Cooperation Centre. http://www.ifapcom.ru/files/News/Images/2013/Yakutsk_web.pdf#page=33

Le Boeuf, Patrick. 2005. Musical Works in the FRBR Model or "Quasi la Stessa Cosa": Variations on a Theme by Umberto Eco. *Cataloging & Classification Quarterly* 39(3–4). 103–124. doi:10.1300/J104v39n03_08

Lehman, Philipp, Philip Kime & Moritz Wemheuer. 2019. *The BibLaTeX Package: Programmable Bibliographies and Citations.* https://ctan.tetaneutral.net/macros/latex/contrib/biblatex/doc/biblatex.pdf

Lehmann, Christian. 1992. Das Sprachmuseum. *Linguistische Berichte* 142. 477–494.

Lehmann, Christian. 2001. Language documentation: a program. In Walter Bisang, ed. *Aspects of Typology and Universals,* 83–97. (Studia Typologica №1), Berlin: Akademie Verlag. doi:10.1524/9783050078892.83

Lüpke, Friederike. 2010. Research methods in language documentation. *Language Documentation and Description* 7. 55–104. http://www.elpublishing.org/PID/082

Mahan, Margaret D. F., (Editor). 2003. *The Chicago Manual of Style.* 15th edn. Chicago, Illinois: Chicago University Press.

Marcadé, Claire, Bernard Guinard, Stéphanie Coulais, Yvon Davy, Eric Desgrugillers, François Gasnault, Véronique Ginouvès, Pierre Marcotte, Mikaël O'Sullivan, Jean-Luc Ramel, Gwenaëlle Sarrat & Laura Thomas. 2014. *Patrimoine culturel immatériel: Traitement documentaire des archives sonores inédites—Guide des bonnes pratiques.* Nantes, France: Fédération des Association de Musiques et Danses traditionnelles. https://halshs.archives-ouvertes.fr/halshs-01065125

McCabe, Mark J. & Christopher M. Snyder. 2007. Academic Journal Prices in a Digital Age: A Two-Sided Market Model. *The B.E. Journal of Economic Analysis & Policy* 7(1)-Article 2. doi:10.2202/1935-1682.1627

Michailovsky, Boyd, Alexis Michaud & Severine Guillaume. 2011. A simple architecture for the fine-grained documentation of endangered languages: The LACITO multimedia archive. *Proceedings of the 2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*. 14–23. Hsinchu City, Taiwan: IEEE. doi:10.1109/ICSDA.2011.6085973

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François & Evangelia Adamou. 2014. Documenting and Researching Endangered Languages: The Pangloss Collection. *Language Documentation & Conservation* 8. 119–135. http://hdl.handle.net/10125/4621

Moeller, Sarah Ruth. 2014. SayMore, a tool for Language Documentation Productivity. *Language Documentation & Conservation* 8. 66–74. http://hdl.handle.net/10125/4610

Mueen Ahmed, Kk & BandarE Al Dhubaib. 2011. Zotero: A bibliographic assistant to researcher. *Journal of Pharmacology and Pharmacotherapeutics* 2(4). 303–305. doi:10.4103/0976-500X.85940

Mueller-Langer, Frank & Richard Watt. 2018. How Many More Cites Is a $3,000 Open Access Fee Buying You? Empirical Evidence from a Natural Experiment. *Economic Inquiry* 56(2). 931–954. doi:https://doi.org/10.1111/ecin.12545

Munro, Robert & David Nathan. 2005. Introducing the ELAR information system architecture. Paper presented at: DELAMAN III. University of Texas at Austin 21–22 November. http://www.robertmunro.com/research/munro05elar.pdf (2021-03-11).

Nathan, David. 2011. Archives as publishers of language documentation: experiences from ELAR. Paper presented at: The 2nd International Conference on Language Documentation and Conservation. The University of Hawai‘i at Mānoa, Honolulu, Hawai‘i. http://hdl.handle.net/10125/5223

Nathan, David. 2013a. Progressive archiving: theoretical and practical implications for documentary linguistics. Paper presented at: The 3rd International Conference on Language Documentation and Conservation. The University of Hawai'i at Mānoa, Honolulu, Hawai'i. http://hdl.handle.net/10125/26115

Nathan, David. 2013b. Access and Accessibility at ELAR, a Social Networking Archive for Endangered Languages Documentation. In Mark Turin, Claire Wheeler & Eleanor Wilkinson, eds. *Oral Literature in the Digital Age,* 21–41. (World Oral Literature Series №2), Cambridge, England: Open Book Publishers. doi:10.111647/OBP.0032.03

Nathan, David & Peter K. Austin. 2004. Reconceiving metadata: language documentation through thick and thin. *Language Documentation and Description* 2. 179–188. http://www.elpublishing.org/PID/029

Newman, Paul. 2003. We has seen the enemy and it is us: The endangered languages issue as a hopeless cause. *Studies in the Linguistic Sciences* 28(2). 11–20. http://hdl.handle.net/2142/11559

Nicolas, Yann. 2005. Folklore Requirements for Bibliographic Records: Oral Traditions and FRBR. *Cataloging & Classification Quarterly* 39(3-4). 179–195. doi:10.1300/J104v39n03_11

Nordmoe, Jeremy. 2011. Introducing RAMP: an application for packaging metadata and resources offline for submission to an institutional repository. *Proceedings of Workshop on Language Documentation & Archiving, SOAS, London, 18 November 2011.* 27–32. London, UK: SOAS. https://www.sil.org/resources/archives/43211

Nordmoe, Jeremy. 2013. Endangered resources: a program for collection and preservation. Paper presented at: The 3rd International Conference on Language Documentation and Conservation. The University of Hawai'i at Mānoa, Honolulu, Hawai'i. http://hdl.handle.net/10125/26116

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

Oez, Mikael. 2018. A Guide to the Documentation of the Beth Qustan Dialect of the Central Neo-Aramaic Language Turoyo. *Language Documentation & Conservation* 12. 430–460. http://hdl.handle.net/10125/24773

Osterwalder, Alexander & Yves Pigneur. 2010. *Business model generation: a handbook for visionaries, game changers, and challengers.* Hoboken, NJ, USA: Wiley.

Osterwalder, Alexander, Yves Pigneur, Gregory Bernarda, Alan Smith & Trish Papadakos. 2015. *Value Proposition Design: How to Create Products and Services Customers Want.* Somerset, New Jersey: Wiley.

O'Neill, Edward, Maja Žumer & Jeffrey Mixter. 2015. FRBR Aggregates: Their Types and Frequency in Library Collections. *Library Resources & Technical Services* 59(3). 120–129. doi:10.5860/lrts.59n3.120

Parker, Geoffrey G. & Marshall W. Van Alstyne. 2000. Information Complements, Substitutes, and Strategic Product Design. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=249585 doi:10.2139/ssrn.249585

Patashnik, Oren. 1988. *BIBTEXing.* (Documentation accompanying BibTeX.) Online: CTAN. https://mirrors.chevalier.io/CTAN/biblio/bibtex/base/btxdoc.pdf (2020-07-19).

Patashnik, Oren. 1998. BIBTEX 101. *TUGboat* 19(2). 204–207. http://raxlues.tug.org/TUGboat/Articles/tb19-2/tb59patash.pdf

Patashnik, Oren. 2003. BibTEX yesterday, today, and tomorrow. *TUGboat* 24(1). 25–30. http://tug2000.tug.org/TUGboat/Articles/tb24-1/patashnik.pdf

Paterson III, Hugh J., (Depositor). 2021. *Open Language Archive Community Nightly for 23 March 2021* (Version 1.0.0) [Data set]. Geneva: Zenodo. doi:10.5281/zenodo.4633463

Paterson III, Hugh J. 2015. Lexical dataset archiving: an assessment of practice. Paper presented at: 4th International Conference on Language Documentation & Conservation. Ala Moana Hotel in Honolulu, HI. February 26th – March 1st. https://hughandbecky.us/Hugh-CV/publication/2015-lexical-database-archiving

Paterson III, Hugh J. 2015. Keyboard layouts: Lessons from the Me'phaa and Sochiapam Chinantec designs. In Mari C. Jones, *Endangered Languages and New Technologies,* 49–66. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/CBO9781107279063.006

Paterson III, Hugh J. & Jeremy Nordmoe. 2013. Challenges of implementing a tool to extract metadata from linguists: the use case of RAMP. Paper presented at: The 3rd International Conference on Language Documentation and Conservation. The University of Hawai'i at Mānoa, Honolulu, Hawai'i. http://hdl.handle.net/10125/26178

Peroni, Silvio & David Shotton. 2012. FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Journal of Web Semantics* 17. 33–43. doi:10.1016/j.websem.2012.08.001

Philips, Addison & Mark Davis, (Editors). 2009. *Tags for Identifying Languages.* (Best Current Practice 47.) Fremont, California: Internet Engineering Task Force (IETF). https://tools.ietf.org/html/bcp47

Piwowar, Heather A., Roger S. Day & Douglas B. Fridsma. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE* 2(3)-e308. doi:10.1371/journal.pone.0000308

Quigley, Aisling. 2019. *Striving to Persist: Museum Digital Exhibition and Digital Catalogue Production.* University of Pittsburgh dissertation. http://d-scholarship.pitt.edu/36659

Ray, Aswini Kumar. 2017. Zotero: Open Source Citation Management Tool for Researchers. *International Journal of Library and Information Studies* 7(3). 238–245. http://ijlis.org/img/2017_Vol_7_Issue_3/238-245.pdf

Reiman, D Will. 2010. Basic oral language documentation. *Language Documentation & Conservation* 4. 254–268. http://hdl.handle.net/10125/4479

Riley, Jenn. 2008. Application of the Functional Requirements for Bibliographic Records (FRBR) to Music. *ISMIR 2008: Proceedings of the Ninth International Conference on Music Information Retrieval.* 439–444. Philadelphia: Drexel University.

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

Robinson-Garcia, Nicolas, Philippe Mongeon, Wei Jeng & Rodrigo Costas. 2017. DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics* 11(3). 841–854. doi:10.1016/j.joi.2017.07.003

Rodríguez-Gairín, Josep-Manuel, Marta Somoza-Fernández & Cristóbal Urbano. 2011. MIAR: hacia un entorno colaborativo de editores, autores y evaluadores de revistas. *El profesional de la información* 20(5). 589–595. doi:10.3145/epi.2011.sep.15

Rueda, Laura. 2016. Zotero for Data Repositories Webinar. (May 17). DataCite Blog. https://blog.datacite.org/zotero-for-data-repositories-webinar (2021-03-11). doi:10.5438/Q2GH-6EGD

Rysman, Marc. 2009. The Economics of Two-Sided Markets. *Journal of Economic Perspectives* 23(3). 125–143. doi:10.1257/jep.23.3.125

Salffner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9. 237–267. http://hdl.handle.net/10125/24639

Salmons, Joseph, (Editor). 2007. *Unified style sheet for linguistics.* Washington, DC, USA: Linguistic Society of America. https://www.linguisticsociety.org/resource/unified-style-sheet

Shadle, Steve. 2006. FRBR and Serials: An Overview and Analysis. *The Serials Librarian* 50(1–2). 83–103. doi:10.1300/J123v50n01_09

Shopen, Timothy, (Editor). 2007. *Language Typology and Syntactic Description.* 2 edn. Vol. 2 of 3. (Grammatical Categories and the Lexicon.) Cambridge: Cambridge University Press. doi:10.1017/CBO9780511619434

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

Simons, Gary F. 2008. Documentary Linguistics and a New Kind of Corpus. Paper presented at: 5th National Natural Language Research Symposium. De La Salle University, Manila, 25 November. https://scholars.sil.org/sites/scholars/files/gary_f_simons/presentation/doc_ling.pdf (2021-03-11).

Skutley, Mary Lynn. 2012. *APA Style Guide to Electronic References.* 6th edn. Washington, DC: American Psychological Association.

Smith, MacKenzie. 2002. DSpace: An Institutional Repository from the MIT Libraries and Hewlett Packard Laboratories. In Maristella Agosti & Costantino Thanos, eds. *Research and Advanced Technology for Digital Libraries. ECDL 2002,* 543–549. (Lecture Notes in Computer Science №2458), Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/3-540-45747-X_40

Society of American Archivists. 2013. *Describing archives: a content standard.* Chicago, Illinois: Society of American Archivists. http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf

Sönmez, Margaret J.-M., Maia Wellington Gahtan & Nadia Cannata Salomone, (Editors). 2020. *Museums of language and the display of intangible cultural heritage.* (Routledge Research in Museum Studies.) London; New York, NY: Routledge. doi:10.4324/9780429491610

Strategyzer. 2017. *Strategyzer's Value Proposition Canvas Explained. [video/MPEG4]. Duration: 3:12. YouTube.* Zürich, Switzerland: Strategyzer. https://www.youtube.com/watch?v=ReM1uqmVfP0 (2021-04-17).

Suarez, Andrew V. & Neil D. Tsutsui. 2004. The Value of Museum Collections for Research and Society. *BioScience* 54(1). 66–74. doi:10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

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

Škevin Rajko, Ivana & Lucija Šimičić. 2020. Language commodification and local sustainability: Molise Croatian as cultural heritage. In Kutlay Yagmur, *Linguistic Minorities in Europe Online.* De Gruyter. doi:10.1515/lme.11473987

Thessen, Anne E., Matt Woodburn, Dimitrios Koureas, Deborah Paul, Michael Conlon, David P. Shorthouse & Sarah Ramdeen. 2019. Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. *Data Science Journal* 18(1)-54. doi:10.5334/dsj-2019-054

Thieberger, Nicholas. 2009. Anxious Respect for Linguistic Data: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Resource Network for Linguistic Diversity (RNLD). In Margaret Florey, *Endangered Languages of Austronesia,* 141–158. Oxford University Press. doi:10.1093/acprof:oso/9780199544547.003.0008

Thieberger, Nicholas. 2013. Multilingualism in Cyberspace – Longevity for Documentation of Small Languages. In Evgeny Kuzmin & Anastasia Parshakova, *Linguistic and Cultural Diversity in Cyberspace: Proceedings of the 2nd International Conference (Yakutsk, Russian Federation, 12-14 July, 2011),* 141–150. Moscow: Interregional Library Cooperation Centre. http://www.ifapcom.ru/files/News/Images/2013/Yakutsk_web.pdf#page=142

Thieberger, Nicholas. 2018. Texts and more texts: corpora in the CoEDL. (February 28, 2018). PARADISEC Blog: Endangered Languages and Cultures. https://www.paradisec.org.au/blog/2018/02/texts-and-more-texts-corpora-in-the-coedl (2021-02-28).

Thieberger, Nicholas & Michel Jacobson. 2010. Sharing data in small and endangered languages: Cataloging and metadata, formats, and encodings. In Lenore A. Grenoble & N. Louanna Furbee, *Language Documentation: Practice and values,* 147–158. Amsterdam: John Benjamins Publishing Company. doi:10.1075/z.158.15thi

Thomson, Alistair. 2016. Digital Aural History: An Australian Case Study. *Oral History Review* 43(2). 292–314. doi:10.1093/ohr/ohw067

Tillett, Barbara B. 2001. Bibliographic Relationships. In Carol A. Bean and Rebecca Green, eds. *Relationships in the Organization of Knowledge,* 19–35. (Information Science and Knowledge Management №2), Dordrecht: Springer Netherlands. doi:10.1007/978-94-015-9696-1_2

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

Tillett, Barbara B. 2009. Definition of Aggregates as Works: Tillett Proposal. Paper presented at: FRBR Review Group. Washington, D.C. https://www.ifla.org/files/assets/cataloguing/frbrrg/aggregates-as-works.pdf (2021-03-05).

Toves, Jenny A. & Thomas B. Hickey. 2014. Parsing and Matching Dates in VIAF. *The Code4Lib Journal* 26-9607. https://journal.code4lib.org/articles/9607

Tramboo, Shahkar, Humma Humma, S M Shafi & Sumeer Gul. 2012. A study on the Open Source Digital Library Software's: Special Reference to DSpace, EPrints and Greenstone. *International Journal of Computer Applications* 59(16). 1–9. doi:10.5120/9629-4272

Trinoskey, Jessica, Frances A. Brahmi & Carole Gall. 2009. Zotero: A Product Review. *Journal of Electronic Resources in Medical Libraries* 6(3). 224–229. doi:10.1080/15424060903167229

VandenBos, Gary R., (Editor). 2010. *Publication manual of the American Psychological Association.* 6th edn. Washington, DC: American Psychological Association.

Vasile, Aurélia, Séverine Guillaume, Mourad Aouini & Alexis Michaud. 2020. Le Digital Object Identifier, une impérieuse nécessité ?L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger. *I2D - Information, données & documents* 2(2). 155–175. https://www.cairn.info/revue-i2d-information-donnees-et-documents-2020-2-page-155.htm doi:10.3917/i2d.202.0155

Vellucci, Sherry L. 2007. FRBR and Music. In Arlene G. Taylor, *Understanding FRBR: What it is and how it will affect our retrieval tools,* 131–151. Westport, Connecticut: Libraries Unlimited.

Verrelli, David I. 2018. Metadata Tags for Academic Publications. Meadowbank, N.S.W., Australia: Division One Academic and Language Services, ms. http://div.div1.com.au/RESOURCES/div/commentary/metadataTags_RevB_L.pdf (2021-01-13).

Wahbeh, Farris. 2009. On "records," papers," and "collection": a DACS case in point – On Archiving Schapiro (August 10). On Archiving Schapiro. https://blogs.cul.columbia.edu/schapiro/2009/08/10/on-records-papers-and-collection-a-dacs-case-in-point (2021-03-05).

Weber, Tobias. 2021. Curation as a distinct academic activity – a perspective from working with legacy materials. Paper presented at: The 7th International Conference on Language Documentation & Conservation (ICLDC), 4–7 March 2021. The University of Hawai'i at Mānoa, Honolulu, HI.

Weiland, Peter, Christiane Baier, Roland Ramthun & Johannes Höhmann. 2019. Reinventing the Wheel? - The Case for the Development of an alternative DSpace Submission Assistant for Psychological Science. Paper presented at: Open Repositories 2019. Hamburg, Germany. https://www.psycharchives.org/handle/20.500.12034/2106

Wijesundara, Chathurangani & Shigeo Sugimoto. 2018. Metadata model for organizing digital archives of tangible and intangible cultural heritage, and linking cultural heritage information in digital space. *Libres* 28(2). 58–80. https://www.libres-ejournal.info/2706

Wilms, Konstantin L., Stefan Stieglitz, Björn Ross & Christian Meske. 2020. A value-based perspective on supporting and hindering factors for research data management. *International Journal of Information Management* 54-102174. doi:10.1016/j.ijinfomgt.2020.102174

Wittenburg, Peter. 2007. Deliverable 3.1 Metadata Integration Report: Distributed Access Management for Language Resources implemented as Specific Support Action. Lund: MPI. https://www.mpi.nl/DAM-LR/deliverables/d3-1-metadata-integration_T24.pdf (2021-03-07).

Witz, Eric, (Editor). 2009. *Linguistic Inquiry Style Sheet.* Cambridge, Massachusetts: MIT Press. https://www.mitpressjournals.org /userimages/ContentEditor/1248106497613/Style%20Sheet%207.6.09.pdf (2020-07-15).

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan and Peter K. Austin, eds. *Special Issue on Language Documentation and Archiving,* 19–36. (Language Documentation and Description №12), London, United Kingdom: SOAS. http://www.elpublishing.org/PID/135

Xia, Fei, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey & Emily M. Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation* 50(2). 321–349. doi:10.1007/s10579-015-9325-4

Yee, Martha M. 1993. The Concept of *Work* for Moving Image Materials. *Cataloging & Classification Quarterly* 18(2). 33–40. doi:10.1300/J104v18n02_04

Yee, Martha M. 2007. FRBR and Moving Image Materials: Content (Work and Expression) versus Carrier (Manifestation). In Arlene G. Taylor, *Understanding FRBR: What it is and how it will affect our retrieval tools,* 117–129. Westport, Connecticut: Libraries Unlimited.

Yumi, Ohira & Visnak Kelly. 2020. Solutions for Capturing Student Work: Using Viero and DSpace to Build non-ETD Student Collections. Paper presented at: Electronic Resources and Libraries 2020. https://rc.library.uta.edu/uta-ir/handle/10106/29044

Zhang, Nan, Tapio Levä & Heikki Hämmäinen. 2014. Value networks and two-sided markets of Internet content delivery. *Telecommunications Policy* 38(5). 460–472. doi:10.1016/j.telpol.2013.03.004

Zipf, George Kingsley. 1949. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology.* Oxford, England: Addison-Wesley Press.