

Graduate Institute of Applied Linguistics

Thesis Approval Sheet

This thesis, entitled
TAPS: Checklist for Responsible Archiving of Digital Language Resources
written by
Debbie Chang
and submitted in partial fulfillment of the requirements for the degree of
Master of Arts
with a major in
Applied Linguistics
has been read and approved
by the undersigned members of the faculty
of the Graduate Institute of Applied Linguistics.

Gary F. Simons (Mentor)

Michael Cahill

Stephen Parker

Date Signed

TAPS: CHECKLIST FOR RESPONSIBLE ARCHIVING
OF DIGITAL LANGUAGE RESOURCES

By

Debbie Chang

Presented to the Faculty of
the Graduate Institute of Applied Linguistics
in partial fulfillment of the requirements
for the degree of

Master of Arts
with major in
Applied Linguistics

Graduate Institute of Applied Linguistics
June 2010

© 2010 Debbie Chang
All Rights Reserved

ABSTRACT

TAPS: CHECKLIST FOR RESPONSIBLE ARCHIVING OF DIGITAL LANGUAGE RESOURCES

Debbie Chang
Master of Arts
with a major in
Applied Linguistics

The Graduate Institute of Applied Linguistics, June 2010

Supervising Professor: Dr. Gary F. Simons

Documenting endangered languages has risen in importance as half or more of the world's languages have the potential to become moribund or extinct within this century. A new generation of documentary language materials is being created in mostly digital formats, while older "legacy" materials are also being digitized. Digital archiving is necessary to preserve these materials for the long-term, but holds particular challenges, requiring sound technical implementation, planning, and infrastructure. This thesis develops the TAPS Checklist, which was intended to help depositors of language materials assess digital language archives based on (1) areas of special concern to linguists and language communities (Target and Access) and (2) recommended best practices for the long-term preservation of digital information (Preservation and Sustainability). TAPS was tested at nine digital archives. Results suggest that digital language archives are providing necessary services, but many lack resources for adequate preservation planning and for ensuring sustainability.

To my parents, Dr. Myron and Vivian Chang, who have given me so much.

And to my Lord Jesus Christ, who made all things,

makes all things possible,

and is making everything new.

ACKNOWLEDGEMENTS

The work represented in this thesis was funded by a grant from the National Science Foundation (BCS-0723864, *OLAC: Accessing the World's Language Resources*) awarded to the Graduate Institute of Applied Linguistics (GIAL) in Dallas, Texas.

Many individuals gave generously of their time to make this thesis possible. I thank my thesis adviser, Gary Simons, for helping me to better focus and organize my ideas, and for spurring me on to clearer thinking and writing. I thank my other committee members, Steve Parker and Mike Cahill, for their excellent contributions and encouragement. I thank those who kindly met with me and participated in evaluations of their archives: Jeremy Nordmoe, Vurnell Cobbey, and Joan Spanne at the SIL Language and Culture Archives, Heidi Johnson at AILLA, Lydia Motyka at the FDA, Mary S. Linn at the Division of Native American Languages (NAL) within the Sam Noble Oklahoma Museum of Natural History, Nick Thieberger at PARADISEC and Kaipuleohone, and David Nathan at ELAR. I thank Joan Spanne and Will Reiman for their feedback and suggestions during the development of the TAPS Checklist; Wayne Dye for his input on TAPS and assistance as a depositor at PARADISEC; Bob Conrad for independently testing TAPS; and Lynn Landweer for sharing her expertise in sociolinguistics. Finally, I thank GIAL faculty and staff, Grace Community Church, friends, and family for their support.

17 May 2010

TABLE OF CONTENTS

Abstract	<i>iv</i>
Dedication	<i>v</i>
Acknowledgements	<i>vi</i>
List of Tables	<i>xi</i>
List of Figures	<i>xii</i>
List of Acronyms	<i>xiii</i>
Chapter 1: Introduction	1
1.1 The Scope of Language Endangerment	2
1.2 The Significance of Language Diversity	5
1.3 Responses	7
Chapter 2: The Current Language Archiving Landscape	12
2.1 Archival Institutions	12
2.1.1 Focus of the Archive's Collection	
2.1.2 Scope of the Archive's Collection	
2.1.3 Submitter Restrictions	
2.1.4 Institutional Affiliation	
2.1.5 Open vs. Dark Archives	
2.2 Issues Confronting Linguists	25
2.2.1 The Ideal	
2.2.2 Understanding Copyright, Transfer of Ownership, and Informed Consent	
2.2.3 Right of Privacy and Language Rights	
2.2.4 Metadata	
2.2.5 Case Studies	

2.3 Language Communities and Archives	42
2.3.1 Stemming the Tide	
2.3.2 Access from the Perspective of the Language Community	
2.3.3 Ongoing Relationship	
2.3.4 Discoverability of the Resources	
2.3.5 Reaching Communities with Language Resources	
2.3.6 Access Restrictions	
2.4 Challenges of Digital Preservation for Archives	52
2.4.1 Challenges of Digital Preservation: the Human Side	
2.4.2 Challenges of Digital Preservation: the Technical Side	
2.4.3 Organizational Fitness of Digital Language Archives	
Chapter 3: Review of Tools for Assessing Archival Practice	57
3.1 A Clear Call for Standards: the Task Force on Archiving of Digital Information	57
3.2 Towards Standardization: the <i>Reference Model for an Open Archival Information System (OAIS)</i>	58
3.3 Towards Standards for Archiving Cultural Heritage Resources: Trusted Digital Repositories (TDR)	61
3.4 Measuring Compliance: <i>Trustworthy Repositories Audit and Certification (TRAC)</i>	62
3.5 Managing Risk with Internal Audits: <i>Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)</i>	66
3.6 Further on toward Certifying Trustworthiness: the <i>Catalogue of Criteria for Trusted Digital Repositories</i>	69
3.7 Certification of Data: <i>Data Seal of Approval (DSA)</i>	71
3.8 The Need for a Tool to Assess Digital Language Archives	73

Chapter 4: Development and Use of the TAPS Checklist	77
4.1 Methodology	77
4.1.1 The Development of the TAPS Checklist	
4.1.2 Uses of the TAPS Checklist	
4.2 Target	82
4.2.1 Item 1: Mission Statement	
4.2.2 Item 2: Submission Criteria	
4.2.3 Item 3: Designated Communities	
4.2.4 Item 4: Ongoing Relationship	
4.3 Access	87
4.3.1 Item 5: Discoverability	
4.3.2 Item 6: Fixed Identifiers	
4.3.3 Item 7: Reach	
4.3.4 Item 8: Access and Use Restrictions	
4.3.4.1 <i>Copyright and Transfer of Ownership</i>	
4.3.4.2 <i>Access Restrictions</i>	
4.3.4.3 <i>Conditions of Use</i>	
4.4 Preservation	96
4.4.1 Item 9: Evidence of Long-Term Planning	
4.4.2 Item 10: Preservation Strategies	
4.4.3 Item 11: Integrity	
4.4.4 Item 12: Authenticity	
4.5 Sustainability	104
4.5.1 Item 13: Adequate Infrastructure	
4.5.2 Item 14: Financial Sustainability	
4.5.3 Item 15: Disaster Preparedness	
4.5.4 Item 16: Succession Plan	
4.6 Limitations of the TAPS Checklist	109

Chapter 5: Results and Conclusions	111
5.1 Archives Evaluated with TAPS	111
5.2 TAPS Checklist Scores	116
5.3 Relative Strengths and Weaknesses of Digital Language Archives	122
5.4 Relative Strengths and Weaknesses of Global versus Regional Archives	124
5.5 Findings and Conclusions	127
5.6 Recommendations for Further Research	133
5.7 Concluding Remarks	135
Appendix A TAPS (Target, Access, Preservation, and Sustainability): Checklist for Responsible Archiving of Digital Language Resources	136
Appendix B TAPS Checklist Evaluations	140
Appendix B-1 AILLA	
Appendix B-2 ELAR	
Appendix B-3 FDA	
Appendix B-4 Kaipuleohone at UH	
Appendix B-5 NAL at SNOMNH	
Appendix B-6 PARADISEC	
Appendix B-7 SIL	
Appendix B-8 UCSD Melanesian Archive (Mandeville Special Collections Library)	
Appendix B-9 UVA Albert and Shirley Small Special Collections Library	
References	187

LIST OF TABLES

- Table 2.1 Typology of Digital Language Archives Showing Institutional Affiliation
- Table 2.2 Typology of Digital Language Archives Showing Focus
- Table 4.1 Distribution of TAPS Checklist Items among Ten Basic Characteristics of Digital Preservation Repositories
- Table 5.1 Typology of Participating Language Archives Showing Focus
- Table 5.2 Overall and Average Scores for Target, Access, Preservation, and Sustainability
- Table 5.3 Results by Items in Target
- Table 5.4 Results by Items in Access
- Table 5.5 Results by Items in Preservation
- Table 5.6 Results by Question for Sustainability
- Table 5.7 Relative Strengths and Weaknesses of Digital Language Archives
- Table 5.8 Summary of Data Points Used in Statistical Analysis
- Table 5.9 Archives Sorted by Global versus Regional Scope

LIST OF FIGURES

- Figure 1.1 Comparative Percentages of Languages and Their Corresponding First Language Speakers
- Figure 1.2 Examples of Legacy Materials
- Figure 2.1 Materials at AILLA—to Be Digitized
- Figure 2.2 Materials at AILLA— “Handle with Care”
- Figure 4.1 An Access Station at NAL
- Figure 4.2 Materials Being Checked for Authenticity at NAL
- Figure 5.1 Time-activated Security Badge

LIST OF ACRONYMS

AILLA	Archive of the Indigenous Languages of Latin America
AIATSIS	Australian Institute of Aboriginal and Torres Strait Islander Studies
ALMA	African Language Materials Archive
ANA	Administration for Native Americans
ANLC	Alaska Native Language Center
APS	American Philosophical Society
ASEDA	Aboriginal Studies Electronic Data Archive
BLC	Berkeley Language Center
CCSDS	Consultative Committee for Space Data Systems
CRDO	Centre de Ressources pour la Description de l'Oral
CRL	Center for Research Libraries
DANS	Data Archiving and Networked Services
DCC	Digital Curation Centre
DoBeS	Dokumentation Bedrohter Sprachen
DPE	Digital Preservation Europe
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
DSA	Data Seal of Approval
DSAA EB	Data Seal of Approval Assessment Editorial Board
ELAR	Endangered Languages Archive
FEL	Foundation for Endangered Languages
HRELP	Hans Rausing Endangered Languages Project
ISO	International Organization for Standardization
LSA	Linguistic Society of America
NAA	National Anthropological Archives
NAL	(Division of) Native American Languages
NARA	National Archives and Records Administration
NESTOR	Network of Expertise in long-term STOrage and long-term availability of digital Resources
OAIS	Open Archival Information System
OCLC	Online Computer Library Center, Inc.
OLAC	Open Language Archives Community
OTA	Oxford Text Archive
PARADISEC	Pacific And Regional Archive for Digital Sources in Endangered Cultures
RLG	Research Libraries Group
SCOIL	Survey of California and Other Indian Languages
SNOMNH	Sam Noble Oklahoma Museum of Natural History
SOAS	School of Oriental and African Studies
TAPS	Target, Access, Preservation, and Sustainability (Checklist for Responsible Archiving of Digital Language Resources)
TRAC	Trustworthy Repositories Audit and Certification
UNESCO	United Nations Educational, Scientific and Cultural Organization

Chapter 1: Introduction

Languages of wider communication are increasingly replacing local languages on a global scale (Harrison 2007:5). If current indicators hold, it is estimated that half or more of all the world's languages spoken today will be dead or dying by the end of this century at a rate of one language per every two weeks (see section 1.1). The rate and extent of language endangerment makes language documentation one of the most urgent tasks for linguists today. Language documentation and description¹ include the traditional grammar, phonology, lexicon, and morpheme-level interlinear text corpus of descriptive linguistics, and the high-quality audio and video records of the language in a variety of communicative events, emphasized in the emerging discipline of documentary linguistics. The materials created by such efforts may then become the primary source material for language revitalization, the evidence to support descriptive claims, and the raw data for further analysis in a variety of disciplines.

This study is concerned with the long-term preservation of digital language documentation of endangered and dying languages. It is hoped that these irreplaceable records find trustworthy archival homes since institutional archives have the best chance of preserving and maintaining access to digital materials beyond the lifespan of their creators. This chapter introduces the scope of the study. Chapter 2 develops a typology of digital language archives and reviews the digital archiving landscape from the

¹ The term “language documentation” can be understood to mean both documentation and description of a language. See section 2.2 for further discussion of what language documentation encompasses.

perspectives of linguists, language communities, and archives. Chapter 3 surveys existing standards that establish best practice and the tools that assess the trustworthiness of digital archives. Chapter 4 describes the development of the TAPS Checklist, designed to aid linguists and other depositors in choosing an archival home for language materials, and explains the identified best practices in detail. TAPS is an acronym for the four parts of the checklist: Target, Access, Preservation, and Sustainability; see appendix A for the complete TAPS Checklist. Chapter 5 reports the findings and draws conclusions from using TAPS in visits and interviews with digital language archives.

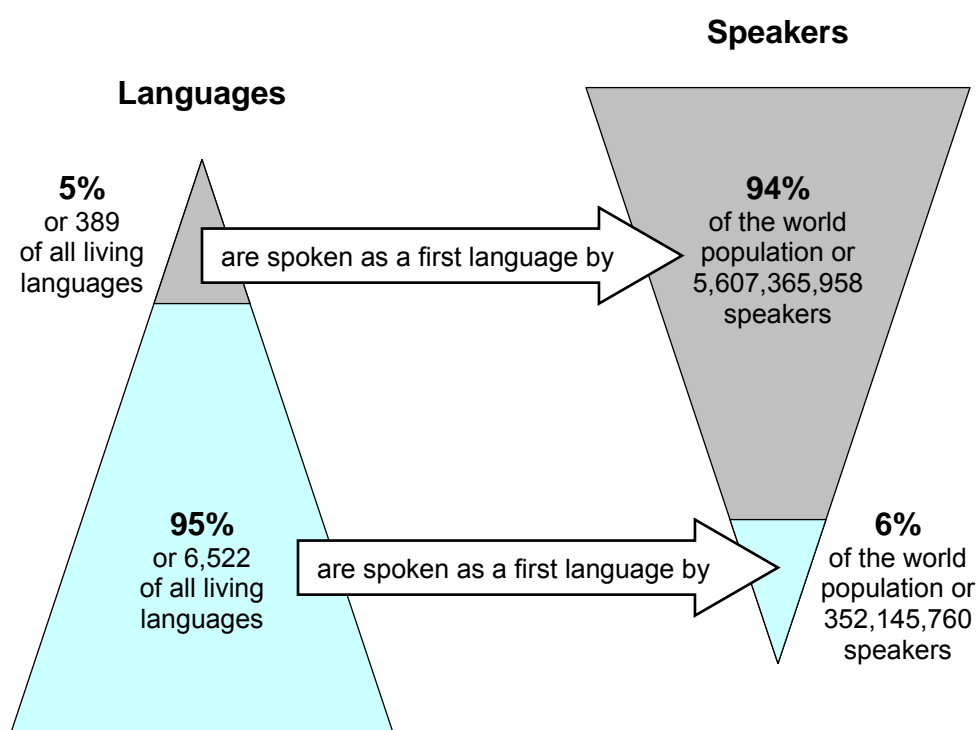
1.1 The Scope of Language Endangerment

Of the 6,000 to 7,000 languages spoken in the world today,² only 85 languages have 10 million or more first language speakers, accounting for almost 80% of the world population, but just 1.2% of all living languages. Furthermore, just 387 languages, or approximately 5% of all living languages, have at least one million first language speakers, accounting for 94% of the world population. The remaining 95% of languages

² Counting the world's languages is a subject of some debate, with some estimates going as low as 3,000 and some as high as 10,000 in recent decades (Crystal 2000:2–11). The definition of a language as opposed to dialect is not always a straightforward one, is dependent on the criteria by which dialects are differentiated, and is subject to sociopolitical and purely scientific considerations. It should be noted, however, that there are language isolates that are not related to any other known languages or language families, and the development of creoles and pidgins may form new natural languages over time. The work of surveying all of the world's languages is also a task yet to be completed. The sixteenth edition of the *Ethnologue* (Lewis 2009:7) numbers the world's living languages at 6,909, the first year that the number has decreased in a succeeding edition (it was 6,912 in the fifteenth edition (Gordon 2005)). Language extinction may be catching up with the ongoing survey of the world's languages. The most recent tally reflected the addition of 163 languages previously unidentified (80 were split and 83 were new varieties not previously associated with another language) and the subtraction of 166 languages (75 were merged with other languages and 91 were recognized as no longer having any remaining speakers). The extinction of 91 languages in the four years between the last two editions of the *Ethnologue* is on pace with estimates of language extinction on the order of one language every two weeks in this century.

are spoken by only 6% of the world's population (Lewis 2009:20). Figure 1.1 illustrates this inverted relationship between languages and their corresponding populations of first language speakers.

Figure 1.1: Comparative Percentages of Languages and Their Corresponding First Language Speakers (adapted from Harrison (2007:14); statistics taken from the *Ethnologue*, sixteenth edition (Lewis 2009:20))



Additionally, ten major languages—each spoken by over 100 million people—account for the first languages of almost half of the world's population, while about half of all languages are each spoken by fewer than 10,000 people, and roughly one fifth are spoken by fewer than 1,000 (Lewis 2009:20). Thus, while the average number of speakers per language works out to be 862,572 (*i.e.*, the world population divided by the number of languages in the world), the median population of first language speakers of

all the world's living languages is only 7,560 (*i.e.*, half of the world's languages have more than 7,560 first language speakers, and half have less than 7,560 first language speakers) (Lewis 2009:19).

Population size alone is not an accurate indicator of the endangerment of a given language, however, since speaker figures should always be viewed in the context of the community to which speakers belong (Crystal 2000:20). A language can be termed “unsafe” if its speakers are small in number, but numbers only tell part of the story since relatively small language groups can maintain their language over long periods of time,³ and even languages with initially large numbers of speakers can succumb to sustained pressures contributing to language shift, as in the cases of Breton and Navajo (Krauss 1992:7). However, it is difficult to imagine many communities sustaining everyday use of a language beyond the next generation with fewer than 100 speakers, and at least 6.8% of all living languages are currently in this situation (FEL 2002:155). Some language experts believe that only 10% of all languages are truly “safe,” having official state support and a large number of speakers (Krauss 1992:7). Estimates of language death or “doom” (*i.e.*, languages becoming moribund) to come in this century have ranged greatly, from one quarter⁴ to 90% of all living languages⁵; Crystal (2004:47) and Harrison

³ Population size and language vitality are unevenly related in different parts of the world. In the Pacific islands, a population of 500 speakers correlates with good language vitality, while in the African savannah some linguists consider a language to be endangered if it has less than 20,000 speakers (Crystal 2000:11, 13). As Cahill (2004) notes, it is “when the population is small *and* declining, that language is in danger” [emphasis mine]. It should be noted, however, that catastrophic events can wipe out small language groups more easily than larger ones. See Landweer (2009) for indicators of relative ethnolinguistic vitality, also discussed in section 2.3.1.

⁴ According to Crystal (2000:15), this figure is based on counting languages with 1,000 or fewer speakers as likely to become extinct.

⁵ Krauss (1992:7) extrapolates from available data in Europe and North America and then considers destabilizing forces active in other parts of the world in addition to the assumptions of dominant languages

(2007:7) opt for middle-of-the-road figures of 50% and more than 40% respectively, if current trends continue. It should be noted that groups responsible for monitoring the situation are in agreement concerning its severity, but avoid “hard statistics” (Crystal 2000:18).

Like an endangered biological species, an endangered language is one that could become extinct, but may not yet be “doomed” to extinction. The truest indicator of language vitality is the rate of transmission of the language to the next generation of potential speakers. A language heads toward imminent death when children are no longer learning it. Once that happens, the language — regardless of the number of living speakers left — is termed “moribund,” bound for extinction just as a species that is unable to reproduce itself (Krauss 1992:4). The extinction or death of the language follows a couple of generations later with the passing of the last speaker.

1.2 The Significance of Language Diversity

When a language disappears, a people’s unique system of experience and perception of the world is lost (Harrison 2007:24–25). As precious repositories of traditional knowledge, languages efficiently encode information about plants, animals, and natural environments accumulated over generations and even millennia. We may never know the limits of what is possible in human cognition when languages are lost

and cultures apparent in much of the developed world. With regard to the processes causing language endangerment, Landweer (in personal communication 2010) notes that the vast majority of case studies available in 1992 did not include the small languages of the Pacific (19% of the world’s languages). No dominant language or culture reduces these languages to minority status, though they may still be endangered, only for different reasons than those cited by Krauss. A still “radical” estimate of 80% is based on counting languages with 100,000 or fewer speakers as likely to become extinct (Crystal 2000:15).

(Gibbs 2002:84). Some linguistic features are only known through a handful of languages, quite possibly yet to be discovered and documented. And a complete record of the world's languages is needed to test theories of language universals and distributions in language typology, as well as to know how languages evolved and how they reflect the migration of peoples. We do not know what we stand to lose even with the loss of a single language (Harrison 2007:7), and the urgency to document and describe a given language only increases with proximity to extinction (Krauss 1992:8).

Language is also an embodiment of culture and an expression of identity. Hale (1992:36), in championing cultural diversity, notes that language is much more than grammar and is often inseparable from the intellectual productions of its speakers, such as some forms of verbal art (*e.g.*, verse, song, and chant). Without documentation, endangered languages can irrevocably disappear, and with them whole ethnic identities. Groups that become linguistically assimilated often struggle to maintain their ethnic distinctiveness, as with the nomadic Monchak of Mongolia (Harrison 2007:95–97), or face additional pressures to change religious affiliations in order to conform to a national identity, as with minority groups in Malaysia.⁶ A higher incidence of social problems related to the loss of identity is also found among groups and individuals in the process of losing their language.⁷ With language documentation, however, a language may retain “crucial symbolic value,” and “it remains always possible to maintain or establish a limited role for the language institutionalized within the society, [for example] in schools

⁶ Personal communication with Paul Kroeger.

⁷ For example, as noted by Harrison (2007:101), alcoholism and despair have settled on many of the Tofa, a traditionally nomadic people of Western Mongolia, who have largely lost their ties to the land and whose language is moribund.

or ceremonial life” (Krauss 1992:9). Additionally, indigenous languages, if documented, have the potential to be revived where “the will of the people is strong enough” (*e.g.*, Hebrew,⁸ Cornish,⁹ Wôpanâak¹⁰). A mother tongue can become an asset to a people in defining themselves as a distinct ethnicity, which can help them lay claim to ancestral lands and resources (*e.g.*, Hawai’ian,¹¹ Tagbanua¹²).

1.3 Responses

In the last two decades, endangered languages have become a major issue in the field of linguistics and have broken into mainstream media. Responding to a statement endorsed by linguists at the 1992 International Linguistics Conference in Quebec, UNESCO began the Endangered Languages Project. At its website addressing the topic of intangible cultural heritage, an “Interactive Atlas of Languages in Danger of Disappearing” now lists 2,471 or about one third of all living languages as being in

⁸ See Fishman (1991:287-336).

⁹ As briefly noted by Fishman (2000:268), Cornish is an extraordinary case of resuscitation in this century, several generations following the death of the last speaker.

¹⁰ The traditional language of the Wampanoag of present-day Massachusetts, which had been dormant 150 years before the process of reclamation began in 1993 through the efforts of Jessie Little Doe Fermino née Baird (Mifflin 2008).

¹¹ In 1988, amidst growing support for Hawai’ian language and culture, the 10-year old Office of Hawai’ian Affairs sued for compensation for 1.7 million acres of land ceded by the Republic of Hawai’i at its 1898 annexation. Federal and state grants were made available to all persons of Hawai’ian blood regardless of blood quantum (Niedzielski 1992:376-377).

¹² With great tenacity and strong community cohesion, the Tagbanua of the Philippines in 1998 obtained a Certificate of Ancestral Domain Claim (CADC), a landmark in the struggle of indigenous peoples nationwide to reclaim their ancestral territory. The CADC is a provision under DAO 1993-02 (DENR (Department of Environment and Natural Resources) Administrative Order 02), which recognized the inherited rights of the indigenous cultural communities (Dalabajan 2001:177). After the Indigenous People’s Rights Act (IPRA) became law in 1997, the Tagbanua successfully obtained a Certificate of Ancestral Domain Title in 2001. They are, according to the Tagbanua Foundation’s chairman, Rodolfo Aguilar, “a living example of how IPRA can be used successfully by indigenous peoples” (Ferrari and de Vera 2004).

danger.¹³ All in the same year, 1995, three major endangered organizations were established: an International Clearing House of Endangered Languages at the University of Tokyo, an Endangered Language Fund in the U.S., and a Foundation for Endangered Languages in the U.K. (Crystal 2000:vii–viii). Since then, language endangerment has been the focus of several more key funding programs, new degree programs and summer institutes, has been presented as the theme of numerous session topics of major linguistics conferences,¹⁴ established as a concern of several international commemorative dates,¹⁵ and featured in popular media. The National Geographic Society and the Living Tongues Institute for Endangered Languages sponsored an expedition to raise awareness of language endangerment and revitalization in 2007 (Dobrin *et al.* 2007:60–61). They identified five major “hotspots” of language endangerment on four different continents: the Pacific Northwest Plateau of North America, Central South America, Central Siberia, Eastern Siberia, and Northern Australia.¹⁶ *The Linguists*, described by PBS as “a ‘hilarious and poignant’ documentary chronicling two scientists’ race to document languages on the verge of extinction,” premiered at the Sundance Film Festival in 2008.¹⁷ Today, at least one digital language archive, the Endangered Language Archive at SOAS, is “plagued” with requests from journalists seeking information about the last known speaker of a language, often published as human interest stories (Nathan 2009:6).

¹³ [<http://www.unesco.org/culture/ich/index.php?pg=00206>]

¹⁴For example: [<http://www.anu.edu.au/linguistics/nash/el.html>]

¹⁵ Since 1999, February 21 has been commemorated as UNESCO’s International Mother Language Day (Crystal 2004:2). 2001 was named the European Year of Languages; in the same year, September 26 was established as the World Day of Languages (Crystal 2004:1–2).

¹⁶ [<http://www.nationalgeographic.com/mission/enduringvoices/>]

¹⁷ [<http://www.pbs.org/thelinguists/>]

As interest and documentation of endangered languages has increased, groups and institutions must responsibly curate the growing body of new language resources as well as a backlog of legacy materials—older, “at risk” materials typically not gathered for the purposes of a language documentation corpus but valuable nonetheless as the only documentation available in some languages. These materials are often in need of digitization for the purposes of access and preservation as well as to counter



Figure 1.2: Examples of Legacy Materials. A few examples of the diverse formats of legacy materials awaiting digitization at the Archive of Indigenous Languages of Latin America (AILLA).

deterioration. Records of the world’s languages are increasingly being made and preserved through digital media, useful both to the language community in various language development applications and to the scientific community for many types of analyses. These digital records hold the promise of less costly repurposing and more rapid dissemination, retrieval, and duplication. The challenge of digital media, however, is their abundance of formats and their ephemeral nature (Simons 2006). Who has not experienced the frustration of encountering a broken link on the Internet or of finding a

digital file that is not as it was originally “saved” or that is altogether unreadable due to corruption of the digital medium, changes in proprietary computer programs and platforms, or changes in readily available hardware devices? Digital retrieval mechanisms, software, hardware, and storage media are inherently short-lived, obsolesce more quickly, and more irretrievably than previous forms of media such as paper or audio tape (Lavoie 2004:47).

Ideally, digital language resources are handed over to an archive which preserves the data and ensures that they remain accessible and discoverable for years to come. In the recent past, however, few institutions and national science foundations had formal strategies in place for managing vast quantities and different types of digital data; nor did many have archives committed to maintaining digital objects indefinitely. In a 2002 *Scientific American* article, Steven Bird warned “[endangered] languages may be recorded only to be lost again as the digital recording succumbs to obsolescence” (Gibbs 2002:84).

Since 2002, a number of digital archives specializing in language resources have been established and further developed, but their effectiveness and longevity have yet to be fully tested. This has been a problem for digital archiving in general, as Clifford Lynch (2003) cautioned, “Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps it will prove to be all too easy to later abdicate.” Fortunately, a number of institutions have undertaken to write guidelines for responsible digital archiving. The necessary technical systems of digital archives and detailed criteria for trustworthiness have been respectively specified by such documents as *The Reference*

Model for an Open Archival Information System (OAIS) (CCSDS 1992) and *Trustworthy Repositories Audit and Certification: Criteria and Checklist* (OCLC and CRL 2007).

The purpose of this thesis is to review the recommended best practices concerning digital archives and apply the findings to language resource archiving in order to develop a tool, called the TAPS Checklist for Responsible Archiving of Digital Language Resources, to inform linguists and other depositors about what is essential in choosing an archival home for preserving language resources. TAPS is an acronym standing for the four parts of the checklist: Target, Access, Preservation, and Sustainability (see chapter 4 for an in-depth discussion of the development of TAPS and its parts). TAPS is not a complete catalog of the responsibilities of an archive, but serves to clarify and simplify the process of choosing an archival home for language resources. It highlights qualities that indicate trustworthiness in archival practice and brings attention to issues of special concern to linguists and language communities. With this tool, these groups can make smart choices as depositors of the world's language resources, facilitate better archiving practices as informed producers, and influence how language archives carry out their responsibilities as informed consumers. In this way, I hope to improve the prospects for the long-term preservation of digital language documentation.

Chapter 2: The Current Language Archiving Landscape

The Information Age, characterized by the digital exchange of information and near-instant access to knowledge, has afforded new and diverse ways to create, describe, and disseminate various types of language documentation. Digital information environments further promise “a rich fabric of scholarly resources, learning materials, and cultural artifacts, seamlessly integrated and readily accessible, organized in ways that facilitate traditional uses and encourage new uses as yet undefined” (Lavoie and Dempsey 2004). The ephemeral nature of digital technologies and the proliferation of digital information have created unique problems in the world of archiving, however, making the documentation done in endangered languages susceptible to loss. As Bird and Simons (2003:558) have noted, “in the very generation when the rate of language death is at its peak, we have chosen to use moribund technologies and to create endangered data.” Technological progress today is coming at the expense of preservation into the future. This chapter discusses the changes that are being faced by digital archives, linguists, and language communities and the challenges they present to preserving digital language documentation.

2.1 Archival Institutions

In the wake of the digital revolution, memory institutions such as libraries, museums, and traditional archives have grappled with issues surrounding digital media that pertain to their central mission to collect, organize, and provide access to

information, and ultimately to pass down this information to succeeding generations as a record of culture. This section discusses emerging models for memory institutions with respect to language archives.

The foremost duty of a digital archive is to ensure that data will be preserved in usable forms well into the future. This definition of archiving may not be shared by all institutions or methods dealing in digital information, such as a web site that has posted data or a weblog with an “archive” of past entries. “Archiving on the web” is a misnomer since the Internet is a means for immediate information dissemination rather than a solution for long-term preservation.

Digitization is another activity that is sometimes confused with archiving. Much focus has been placed in recent years on digitizing “legacy” materials that are in danger of being lost, before the information contained on analog tapes and paper further deteriorates, and to make such materials readily accessible without degrading the physical integrity of the original media. While digitization is a first step in saving and repurposing digital language data, digitization alone is not enough for the long-term preservation of digital information. Digital archives need policies to secure the long-term persistence of digital materials to keep up with the times, taking into consideration technological changes and the changing needs of their user communities. A high level of organizational commitment separates archives from straight digitization projects or “archiving” on a website that has no such structures in place. This commitment to preservation is addressed by mission statement, the first topic of the first category, Target, of TAPS Checklist; see section 4.2.1.

In order to present the landscape of current language archiving, the following sections describe a typology for digital language archives. Several dimensions of digital archives are discussed in sections 2.1.1 through 2.1.5 below.

2.1.1 Focus of the Archive's Collection

The first dimension of an archive is the focus of its collection. Digital language archives by definition contain digital language documentation of certain types as defined by an archive's submission policy. The submission criteria¹⁸ of the archive define the nature of digital (or analog) "objects" it collects. Some archives such as the Endangered Language Archive (ELAR)¹⁹ at the School of Oriental and African Studies (SOAS) within the University of London are purely digital, preferring born-digital submissions from its depositors. Therefore, ELAR does not preserve analog materials indefinitely. The U.S. National Anthropological Archives (NAA)²⁰ at the Smithsonian, on the other hand, holds some digitized photographs and electronic records of its collection, but its main aim is to preserve original physical artifacts and it therefore primarily deals in paper and analog language documentation materials. The NAA still provides copies of audio recordings on cassette tape, sometimes using "newly re-mastered" reel-to-reel audio tape

¹⁸ The submission criteria are often found in information written for potential depositors or donors. A description of the submission policy is also an optional OLAC metadata element <archivalSubmissionPolicy> within the OLAC archive description [<http://www.language-archives.org/OLAC/repositories.html#OLAC%20archive%20description>] and is shown in the "more details" page for each participating archive (<http://www.language-archives.org/archives.php>). A page showing all of the submission policies can be found here: [<http://www.language-archives.org/submission-policies.html>]

¹⁹ [<http://www.hrelp.org/archive/>]

²⁰ [<http://www.nmnh.si.edu/naa/>]

from wax cylinders or aluminum disks.²¹ Some digital archives contain primarily written rather than recorded corpora (OTA²²), or transcripts of spoken communication (TalkBank²³), or may have a focus on video and audio recordings (ELAR, PARADISEC²⁴). The dimension of content addresses what is archived and in what forms. Table 2.1 lists archives that contain language documentation materials and the focus of their collections.

Not all participants in the Open Language Archives Community (OLAC)²⁵ can be properly categorized as digital language archives according to the criterion of focus. For example, the Digital Archive of Research Papers in Computational Linguistics at the University of Pennsylvania is a specialized repository for published journal articles, rather than for language data, and therefore is not considered a digital language archive in the typology presented here. The dimension of focus (along with scope and submitter restrictions discussed in the next two subsections) is addressed in submission criteria, the second item of the TAPS Checklist; see section 4.2.2.

2.1.2 Scope of the Archive's Collection

The second dimension of an archive is the scope of its collection, as defined by the submission policy of the archive. Digital language archives can be categorized in accordance with scope, often along geographic lines. Some are global in scope

²¹ [<http://www.nmnh.si.edu/naa/ordering.htm>]

²² Oxford Text Archive, [<http://ota.ahds.ac.uk/about/>]

²³ [<http://talkbank.org/>]

²⁴ Pacific And Regional Archive for Digital Sources in Endangered Cultures, [<http://www.paradisec.org.au/home.html>]

²⁵ [<http://language-archives.org/archives.php>]

(Kaipuleohone, OTA), while others are regional (ASEDA,²⁶ AILLA²⁷), or concentrate on a specific language (*e.g.*, Caddo Heritage Museum archives, Kiowa Museum archives). Some archives that are global in scope also focus on minority or endangered languages (ELAR, PARADISEC), while others contain language data from majority languages (CRDO²⁸, OTA, TalkBank). The first column of tables 2.1 and 2.2 group archives according to the scope of their collections.

2.1.3 Submitter Restrictions

A third dimension is who is allowed to submit material to the archive. Archives with relatively unrestricted submission policies accept language documentation from a wide range of donors (*e.g.*, TalkBank, PARADISEC), while archives with more restrictive submission policies limit the depositors to a select group who are funded or otherwise affiliated with the organization that sponsors the archive (*e.g.*, DoBeS,²⁹ SIL Language and Culture Archives). Both of these types of archives are dynamic in that they continue to accept new material. Archives with the most restricted submitter policies are static collections; for example Boiste, a repository that is a member of OLAC, contains documentation done by a single, deceased individual. Static collections are not listed in the tables below.

Notably, all regional and heritage language archives listed in tables 2.1 and 2.2 have unrestricted submitter policies. There are also some exceptions for archives with

²⁶ Aboriginal Studies Electronic Data Archive, [<http://www.aiatsis.gov.au/research/projects/aseda.html>]

²⁷ Archive of the Indigenous Languages of Latin America, [<http://www.ailla.utexas.org/site/welcome.html>]

²⁸ Centre de Ressources pour la Description de l'Oral (Center for the Description of Oral Resources)

²⁹ Dokumentation Bedrohter Sprachen (Documentation of Endangered Languages)

few submitter restrictions. The NAA accepts donations of materials, but focuses on contributions from Americans (often famous anthropologists), so an archive with relatively unrestricted policies on who can submit data may nevertheless have limitations or otherwise high standards. And though ELAR primarily serves those who receive funding from its parent organization, it may choose to accept endangered language material from those not associated with their organization. Thus “restricted with exceptions” forms a third category for submitter restrictions. In tables 2.1 and 2.2 below, archives are categorized in the first column according to the scope of their collections, and in the second column archives are sub-categorized according to their submitter restrictions.

2.1.4 Institutional Affiliation

Institutional affiliation forms a fourth dimension of digital language archives. A variety of institutions are involved in archiving digital language data. These institutions are often part of a larger traditional institution, such as a university library, museum, NGO, or research center. Notably, many linguistic archives have connections with the linguistic or anthropology department of a major university. Since maintaining digital materials over the long-term requires an elaborate and costly technical infrastructure, including qualified personnel to run it, it is economically advantageous for a digital language archive to be a part of a larger institution in order to realize economies of scale and be more cost effective by sharing fixed costs with either a parent organization or group of organizations. Some existing digital language archives are natural extensions of

their host institutions and tie into the existing infrastructure. As Lavoie and Dempsey (2004) anticipated, digital preservation mechanisms are being integrated with a wide range of other services that constitute a digital library. Table 2.1 lists the institutional affiliations of digital language archives.

Table 2.1: Typology of Digital Language Archives
Showing Institutional Affiliation

Scope	Submitter Restrictions	Archive	Full Name	Institutional Affiliation	
Global	Few restrictions	CRDO	Centre de Ressources pour la Description de l'Oral	CNRS (Centre national de la recherche scientifique) ³⁰	French government
		NAA	National Anthropological Archives	National Museum of Natural History, Smithsonian Institution	U.S. federal government
		OTA	Oxford Text Archive	Oxford University	University (U.K.)
		Rosetta Project	Rosetta Project Digital Archive	Long Now Foundation	Non-profit corporation (U.S.)
		TalkBank	Same	Carnegie Mellon University, University of Pennsylvania	U.S. private university
	Restricted with exceptions	DoBeS	Dokumentation Bedrohter Sprachen	MPI (Max Planck Institute for Psycholinguistics,), Volkswagen Foundation	Research organization, independent, non-profit foundation (German)
		ELAR	Endangered Languages Archive	SOAS (School of Oriental and African Studies, University of London, HRELP (Hans Rausing Endangered Language Project)	University-based program (U.K.)
		SIL	SIL Language and Culture Archives	SIL International, formerly known as the Summer Institute of Linguistics	Non-governmental organization (U.S.)
	Closed	Kaipuleohone	Same	University of Hawai'i at Mānoa (U.S.)	U.S. state university
	Regional	Few restrictions	AILLA	Archive of Indigenous Languages of Latin America	UT (University of Texas, Austin)

³⁰ National Center for Scientific Research under French Ministry of Higher Education and Research

Table 2.1 continued: Typology of Digital Language Archives
Showing Institutional Affiliation

Scope	Submitter Restrictions	Archive	Full Name	Institutional Affiliation/s	
Regional	Few restrictions	ALMA	African Language Materials Archive	Columbia University/UNESCO (United Nations Educational, Scientific and Cultural Organization)	U.S. private university, international NGO
		ANLC	Alaska Native Language Center	UAF (University of Alaska, Fairbanks)	U.S. state university
		APS	American Philosophical Society Library Archives	American Philosophical Society	Private, non-profit organization
		ASEDA	Aboriginal Studies Electronic Data Archive	AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies), an academic research institution	Australian government
		Institute of Papua New Guinea Studies	Same	—	Papua New Guinea government
		NAL	Division of Native American Languages	SNOMNH (Sam Noble Oklahoma Museum of Natural History), OU (Oklahoma University)	Museum at U.S. state university
		PARADISEC	Pacific And Regional Archive for Digital Sources in Endangered Cultures	University of Sydney, University of Melbourne, and Australian National University	Consortium of universities (Australian)
		SCOIL	Survey of California and Other Indian Languages	U. of California, Berkeley	U.S. state university
		Tjibaou Centre	Tjibaou Centre (New Caledonia) Media Center	—	French government
		Vanuatu Cultural Centre	Same	—	Vanuatu government
Tribal	Few restrictions	Caddo Heritage Museum archives	Same	Tribal government	Non-profit organization (U.S.)
		Kiowa Museum archives	Same	Kiowa Culture Preservation Authority	Non-profit organization (U.S.)

Table 2.2: Typology of Digital Language Archives Showing Focus

Scope	Submitter Restrictions	Archive	Submission Details	Focus	
				Content	Type
Global	Few restrictions	CRDO	Accepts documentation of oral resources.	Oral resources only: recordings of speech and annotations	Digital
		NAA	Accepts donations of historical and contemporary anthropological materials that document the world's cultures and the history of anthropology.	Field notes, journals, manuscripts, correspondence, photographs, maps, physical anthropological data, sound recordings, film, video, and other media	Physical and digital (primarily paper-based)
		OTA	Collects, catalogues, preserves and distributes high-quality digital resources for research and teaching.	Textual literary and linguistic resources in more than 25 different languages; some databases and spoken resources in audio and video files	Digital
		Rosetta Project	Initial collection efforts focused on assembling basic "descriptive components" for each language to create an unique archival physical product	Basic descriptive components including general metalinguistic information, phonology, grammar, numbers, lexical data in the form of Swadesh word list, a parallel text (Genesis 1–3), glossed vernacular texts, maps, and orthographic information. Over 1,500 languages were represented on the on the Rosetta Disk.	Digital
		TalkBank	Collections center around a number of subfields studying communication, fostering fundamental research in the study of human and animal communication	Audio and video	Digital
	Restricted with Exceptions	DoBeS	Depositors are usually DoBeS-funded teams.	Documentation of languages in their cultural setting	Digital
		ELAR	Most depositors are funded by HRELP.	Materials that relate to endangered languages	Digital
		SIL	Most depositors are SIL members. Contains new and legacy materials.	Books, journal articles, dissertations, and academic papers about languages and cultures; references for materials written in minority languages	Physical and digital

Table 2.2 continued: Typology of Digital Language Archives Showing Focus

Scope	Submitter Restrictions	Archive	Submission Details	Focus	
				Content	Type
Global	Restricted	Kaipuleohone	Depositors are currently limited to UH staff and students.	Mostly digitized text; also audio and video recordings, photographs, notes, dictionaries, transcriptions	Digital
Regional	Few restrictions	AILLA	Accepts all language materials in and about the indigenous languages of Latin America	Audio, video, image, and text materials;	Digital
		ALMA	Accepts language materials published in African languages	E-books and other original materials	Digital
		ANLC	Serves as a repository for substantial, significant collections relating to identified documentary or pedagogical projects in or on all of the Alaska Native languages and related languages	Published and unpublished materials	Digital and analog
		APS Archives	Accepts materials pertaining to American anthropology, with an emphasis on Native American languages and cultures	Books, manuscripts, personal papers, works of art on paper, digital objects, and other materials	Analog and digital
		ASEDA	Accepts materials relating to Aboriginal and Torres Strait Islander histories and cultures	Photographs, sound recordings, film and video, ephemera, unpublished works (such as manuscripts, personal papers, diaries, field books and theses)	Physical and digital
		Institute of Papua New Guinea Studies	Accepts materials related to Papua New Guinean language and culture	Mainly archives on PNG history, includes commercially recorded music cassettes	Physical and digital
		NAL at SNOMNH	Collection concentrates on Native languages of Oklahoma; also includes Native languages of North America and endangered languages world-wide	Audio and video recordings, manuscripts, books, and teaching curriculum, lesson plans and materials	Digital and analog

Table 2.2 continued: Typology of Digital Language Archives Showing Focus

Scope	Submitter Restrictions	Archive	Submission Details	Focus	
				Content	Type
Regional	Few restrictions	PARADISEC	Accepts materials from linguists, ethnomusicologists and ethnographers. It cannot always accept material that requires conversion or more detailed archival accession as it is under-resourced.	Mainly digitized audio tapes; also textual materials, dictionaries, grammars, articles and other digital objects; as of December 2009, 614 languages from 60 countries were represented	Digital
		SCOIL	Accepts indigenous language materials from the western United States	Primarily field notebooks, indigenous language texts, and word lists; also secondary material derived from these field notebooks, such as file slips, dictionaries and edited texts	Physical and digital (mainly paper-based)
		Tjibaou Centre (New Caledonia) Media Center	Accepts materials to New Caledonian language and culture	Written documents, pictures, audio, video focusing on Oceanic cultures, especially Kanak	Physical and digital
		Vanuatu Cultural Centre	Accepts materials related to Vanuatuan language and culture	Audio, audiovisual, and photographic records of the customs, culture and traditions of Vanuatu	Physical and digital
Tribal	Few restrictions	Caddo Heritage Museum Archives	Accepts materials related to Caddo language and culture	Documents, research materials, photographs, and sound and video recordings	Physical and digital
		Kiowa Museum Archives	Accepts materials related to Kiowa language and culture	Physical artifacts, recordings, and other materials including video	Physical and digital

2.1.5 Open vs. Dark Archives

A fifth and final dimension is whether the archive is “open” or “dark.” In an open archive, descriptive metadata³¹ about the resources are made known to the public (*e.g.*, via the Internet). The resources themselves may have associated restrictions and costs, however, so “open” does not equate to “free.” In case of resources that are not yet ready to be accessed by anyone other than the original depositor, metadata for those resources may be kept hidden for a set time in an otherwise open archive.

In contrast to an open archive, a “dark” archive is not intended to be visible to the public and supports little or no access to archived materials—that is, the archive provides no real-time, online access to the content by anyone except the repository staff (Caplan 2004:2). A dark archive functions as a repository for information that can be used as a failsafe during disaster recovery and simplifies many aspects of preserving digital information (Caplan 2004:3). To obtain a baseline of attainable best practice in preservation, I visited a dark archive, the Florida Digital Archive (FDA), which serves the libraries of the public universities of Florida (see appendix B-3 for the evaluation of the FDA using TAPS).

Digital language archives generally can be characterized as “open” archives since it is expected that they will provide access services as well as secure the long-term viability of archived materials. Thus, the archives that we are concerned with here are open archives that contain language documentation materials (particularly on endangered

³¹ Metadata is structured data about data, and has several types. Descriptive metadata, also known as context information, allows digital objects to be identified according to content through a machine search; see section 2.2.4 below.

languages and minority languages that may become endangered), accept new submissions from depositors, may be global, regional, or tribal in their scope of collection, and may be affiliated with a wide range of larger institutions. The overall “fit” of a given archive to the set of materials to be archived is addressed in the Target component of the TAPS Checklist; see section 4.2.

The challenges of preserving digital information are many. The next sections, 2.2 through 2.4, attempt to highlight major challenges faced by the three key players in digital language archiving: language communities, scholars³², and the archives themselves.

2.2 Issues Confronting Linguists

The field of linguistics has seen documentary linguistics come into prominence in the last decade largely in response to the perceived need to document endangered languages. The task of compiling a language documentation corpus, described more thoroughly elsewhere (Himmelmann 1998, Woodbury 2003, Gippert *et al.* 2006), is the topic of a number of degree programs (notably SOAS and the University of Hawai’i at Mānoa) as well as new courses and seminars offered by university linguistics departments and schools (such as the Graduate Institute of Applied Linguistics, the author’s own graduate school). With the rise of documentary linguistics, a new focus has been placed on closer representations of primary data that make analytic claims verifiable or reproducible. This has led to a renewed emphasis on corpora and archival practice

³² Members of language communities are also represented in this category. To simplify this discussion, the concerns of the language community and scholars are considered separately.

(Garrett *et al.* 2008). In light of these developments, linguists need to adopt new practices to do high-quality language documentation and to ensure that their recordings of speech practices and other primary linguistic data will be preserved.

The data-driven nature of language documentation projects holds the promise that they will be of interest to a wide-range of groups and individuals (Himmelman 1998:171), perhaps no more so than a language community whose speech practices have been documented and whose language may be endangered or become extinct. As noted in chapter 1, an endangered language is any language for which there is a possibility that parents will no longer be passing it on to their children, causing the language to become moribund, and then inevitably to die out. A language in common use among children today can become endangered if there are pressures that could cause language shift within the next century. In California, where many languages have died out or are in “crisis,” language communities attempting to revitalize heritage languages are most interested in basic speech events that answer such questions as:

How do people greet each other? What are the “rules” of conversation?

What kinds of small-talk do they do? What are the colloquialisms that they use? What role do facial expressions and gestures play in conversation?

How does conversational style differ depending on sociolinguistic factors?

(Hinton 2005:25)

Hinton notes that for many languages which have ceased to be spoken altogether now, these questions will never be answered (Hinton 2005:25). Traditional field linguists mainly concerned with language description have collected a wealth of wordlists,

dictionaries, grammars, phonologies, and interlinearized texts, but what is largely missing from the record was natural conversation, which was impossible to capture without sound or video recording, or clear guidelines on the value of “raw” linguistic data. Legacy linguistic materials gathered for a particular analytic format or research goal are still invaluable and should be preserved. Language documentation, in contrast, is concerned with compiling, commenting on, and archiving the speech practices of a community (Himmelmann 1998:165) and the increasing ease of audio and visual recording enable linguists to capture communicative events with supporting information. Given the concern to document endangered languages, the products of language documentation are also worthy of being archived in their own right.

The hope is that the primary data that linguists collect will be used by the people who need it most urgently, particularly speaker and heritage communities, and may be used in ways that were not originally anticipated. Because the documentation of potentially endangered languages will be useful for generations to come, and since the life spans of individual linguists as well as the technologies used are limited, institutional solutions that will outlive individuals and provide expertise in dealing with digital data over time are needed. However, linguists experience a number of common barriers to archiving a corpus of language data in archival institutions. But if linguists allow their material to lie dormant and un-archived, they are deciding how communities will use it, in essence coming between a community and its language. A responsible language archive can help to bring the two together (Garrett *et al.* 2008). The TAPS Checklist was designed to help linguists and other depositors of language data to identify responsible

language archives. The challenges that linguists face in finding an archival home for their data and implications for the TAPS Checklist are discussed in Sections 2.2.1 through 2.2.5.

2.2.1 The Ideal

The ideal language documentation project begins with a linguist (or team of linguists) qualified to do the documentation³³, who from the beginning communicates with an archive that can provide guidance as to the best methods and equipment for carrying out the project and that will commit to the long-term preservation of the resultant data. The benefits of consulting with an archive before fieldwork is done are many. The work of the linguist is likely to be better organized, increasing the chance that the material will be submitted to the archive and accessioned sooner. Talking with an archive also increases the chance of receiving grant funding for a project. Many granting agencies will require archiving plans, and even where it is not a requirement, having plans to archive reflects best practice (Garrett *et al.* 2008).

The first category, Target, of the TAPS Checklist was designed to address the preliminary conditions of deposit. Item 1, mission statement, establishes the archive's commitment to long-term preservation; see section 4.2.1. Questions regarding the

³³ Himmelmann notes that “the task of language documentation is not an easy one. Ideally, the person in charge of the compilation speaks the language fluently and knows the cultural and linguistic practices in the speech community very well. This, in general, implies that the compiler had lived in the community for a considerable amount of time. Furthermore, the compiler should be familiar with a broad variety of approaches to language and capable of analyzing linguistic practices from a variety of points of view. These demands will only rarely be met by a single individual. Hence, the compilation of high-quality language documentation generally requires interdisciplinary cooperation as well as close cooperation with members of the speech community.” I would add a level of technical expertise with recording equipment and methods of language documentation are also necessary.

general dimensions of the archive addressed in sections 2.1.1 through 2.1.6 should then be investigated. The linguist should determine the specific conditions of deposit and submission criteria of the archive: the requirements for the informed consent of members of the language community or governing body of the language community (see section 2.2.3 and 2.2.4), what types of data and formats the archive supports, and so forth. Item 2 of the TAPS Checklist, submission criteria, verifies the focus, scope, and submitter restrictions of the archive; see sections 2.1.1–2.1.3 and 4.2.2. The needs of the language community and other potential users of the archive should also be assessed, and the linguist should be aware of what services the archive provides to those potential users. Item 3 of the TAPS Checklist, designated communities, guides depositors in considering their desired audience; see section 4.2.3.

It should also be noted that language communities that are actively involved in the design of a documentation project from the very beginning can shape the essential aspects of the project (Himmelmann 1998:188). The involvement of the community in this way may be desirable since communicative events are organized in culture-specific ways (Himmelmann 1998:177). Dobrin and others (2007:64–65) note the “singularity” of languages and cultures and how likewise, the documentation of each must be approached uniquely. Similarly, the language community may also choose to interface with the archive in culture-specific, and sometimes self-determining, ways. Item 4 of the TAPS Checklist, ongoing relationship, addresses the archive’s responsibility towards the language community; see section 4.2.4.

Issues pertaining to rights of privacy, language rights, and copyright (discussed below in sections 2.2.2 and 2.2.3) should also be discussed with the language community at the beginning of a project. The linguist should determine what restrictions on access and use the archive allows, which can assure a language community that the data will not be misused. Item 8 of the TAPS Checklist thus addresses access and use restrictions; see section 4.3.4.

More than one archive may be consulted as needed, and more than one archive may be ultimately entrusted with different parts of a language documentation project, perhaps on the basis of supported data types, or geographic specialization which has bearing on how an archive may serve a community.

In many cases, however, the linguist has not had the benefit of talking to an archive in advance, or especially in cases of legacy materials, archiving the primary data was not an original goal and informed consent for the purpose of publishing a language documentation corpus was not obtained from individuals in the language community. The interests, rights, and protection of individual contributors and the language community as a whole must take precedence over scientific interests. Sections 2.2.2 and 2.2.3 below define and outline some of the major issues of copyright, language rights, and the right of privacy, and the inherently sensitive nature of some types of language data; section 2.2.4 addresses metadata; and section 2.2.5 highlights a few case studies.

Finally, counter to the sense of urgency regarding the need to document and “save” endangered languages as an absolute good, Grinevald (2003:60-61) raises the concern that sometimes doing no fieldwork on an endangered language is best. In her

discussion of the unique challenges of documenting endangered languages with respect to the interests of the speakers, she notes that a good rule of thumb for fieldwork should be to at least not leave the field worse off than it was before, and to allow time for situations to evolve and working relationships to take place. Before embarking on a documentation project, Grinevald (2003:61) recommends feasibility studies, initial reconnaissance trips, and networking and consulting with others who have done fieldwork with a given community or neighboring communities. In light of the shortage of trained linguists available to document all of the endangered languages worldwide, linguists are encouraged to focus their attention on communities that are seeking their help and expertise.

2.2.2 Understanding Copyright, Transfer of Ownership, and Informed Consent

Copyright is a type of legal protection for intellectual property, which refers to “creations of the mind: inventions, literary and artistic works, and symbols, names, images, and designs used in commerce.”³⁴ Copyright in particular covers “literary and artistic works such as novels, poems and plays, films, musical works, artistic works such as drawings, paintings, photographs and sculptures, and architectural designs. Rights related to copyright include those of performing artists in their performances, producers of phonograms in their recordings, and those of broadcasters in their radio and television programs.” The purpose of copyright is to protect the right of authors and creators to benefit from the sale or distribution of their works and thereby promote creative

³⁴ [<http://www.wipo.int/about-ip/en/>]

expression. It does so by inherently vesting in the creator of a work the right to control its reproduction, distribution, adaptation, translation, performance, and public display.

Copyright does not protect facts or ideas, only the fixed and tangible expression of facts or ideas (SEM 2001:14).

Copyright law gives a speaker or performer in a recording the inherent right to control the distribution of that recording. For this reason, it is necessary for linguists to get informed consent from those members of the language community whom they have recorded before they are free to distribute those recordings. It should also be noted that multiple copyrights may apply in the same recording. Thus informed consent agreements need to be worked out prior to filling out an archive deposit agreement. Informed consent from speakers or authors of material should be properly obtained on the field and should document: (1) what uses of the materials are authorized by the appropriate persons and (2) that the linguist is authorized to archive the materials (SEM 2001:20). In informed consent, the performer is not transferring copyright, but simply granting a “license” to authorize reproduction, distribution, adaptation. Likewise, the archive obtains permissions to perform preservation functions on and provide access to materials.

The terms of deposit as expressed in an archive’s deposit form will define whether ownership is transferred or maintained. Archives that contain physical objects as well as digital files may require or encourage the donor to transfer ownership of materials (*e.g.*, reel-to-reels or cassette tapes) to them (*e.g.*, NAA and APS—both caretakers of physical objects as well as digital media). Transfer of ownership is not a particularly relevant concept for digital materials, since they are easily duplicated and not

usually bound to a particular physical manifestation. More purely digital archives (*e.g.*, AILLA, ELAR) do not seek out or routinely obtain transfer of ownership. Even when an archive owns tapes, however, it does not own the intellectual property that is on them.

2.2.3 Right of Privacy and Language Rights

The individuals who participate in language documentation have the right to privacy: they have the right to ask that their identity not be disclosed and to insist that a particular text not be shared if it compromises their privacy. Speakers and performers have to consent to the distribution of the materials provided by them. In addition, the compiler of the documentation must make certain that no data are included that may be harmful to an individual or upset the speech community (*e.g.*, bad mouthing, gossip, *etc.*), even if this possibility is not foreseen by the contributors themselves. As *Protocols for Native American Archival Materials* stipulate, “For Native American communities the public release of or access to specialized information or knowledge—gathered with and without informed consent—can cause irreparable harm” and “[p]rivacy rights extend to groups in some situations” (First Archivists Circle 2007:10).

The linguist may need to negotiate with individual contributors on the form of transcriptions that are made public (Himmelmann 1998:173). They may be embarrassed by the “uhms” and false starts and prefer that the transcripts of data be edited to look more like the writing found in newspapers or textbooks. However, editing to make spontaneous communicative events appear more like written texts eliminates the possibility of analysis in a variety of frameworks such as discourse and conversation

analysis, or interactional sociolinguistics. A possible compromise would be to publish the edited version as printed text but to store the recording and transcript of the original communicative event in the archive to be accessed exclusively for further scientific inquiry.

The linguist may also need to navigate issues impinging on the rights and privacy of the language community as a whole—most often represented by its political and cultural leadership (Himmelman 1998:172). There are basically two motives for a language community to restrict the extent and public availability of language data: (1) its linguistic practices involve secret aspects and taboos, and (2) the community wishes to prevent the exploitation, ridicule, or improper portrayal of its culture. Himmelman (1998:173) notes that regarding the first motive, the public documentation of such practices could reveal the secrets or lead to the violation of a taboo and thereby negatively affect the language community. Exploitation is normally understood as having to do with economic rights such as profit sharing. Regarding the motivations of preventing ridicule and improper portrayal, these interests are similar to the interest of the moral right recognized in continental European copyright law that protects against defamation of an individual's reputation by derivative versions of his work that he does not agree with. That is,

Independently of the author's economic rights, and even after the transfer of the said rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory

action in relation to the said work, which would be prejudicial to his honor or reputation (Lieberman 2001, quoting Article 6 of the Berne Convention).

Archives that have an overall open access policy to materials may not accept the deposit of sensitive materials like this, though most digital language archives have options for depositing materials with access restrictions in place. Even so, archives like AILLA caution against the deposit of sensitive materials.³⁵ All authors and performers should have a say on the final, publicly accessible version of their contributions, and often the language community as a whole will also want some control over the further processing and distribution of the material.

2.2.4 Metadata

Metadata is “data about data” that is structured in specific ways and has several types, including descriptive (describes resources for purposes of discovery and identification), structural (how digital objects may form compound objects), and administrative (information, such as rights management metadata and preservation metadata, that help an archive manage resources). Descriptive metadata, also known as context information, allows digital objects to be identified and searched according to content. Currently, text remains the principal means for searching and browsing collections, even when they contain documents in other media (Witten and Bainbridge 2002:79). For language materials, it is incumbent upon linguists to provide detailed descriptive metadata concerning the context of communicative events. The choice of

³⁵ [http://www.ailla.utexas.org/site/five_con.html]

metadata format should be made in consultation with the archive where the linguist intends to deposit materials.³⁶ Particularly with regard to language communities:

For every language and every speaker of that language, it will be of great benefit to both scholarship and to the descendants of the speaker for the collector to record a good deal of personal information about the speaker. This could include a life history, at least a short one and, it is hoped, a long one (Hinton 2005:25).

It is this kind of metadata that makes language resources especially useful to language communities. Making materials discoverable with such metadata is one way to counter a prevailing sentiment that academic people have been content to study indigenous communities without returning useful materials to the people they have studied (Yamamoto 1998:115).

2.2.5 Case Studies

Though it is clear that archiving for community access is essential in documentation, negotiating access to language documentation sometimes presents problems. Conathan and Garrett (2009:6), who draw from their experiences with regard to informally deposited materials in SCOIL³⁷ and the Berkeley Language Center

³⁶ For linguistic data, the OLAC metadata set ([\[http://www.language-archives.org/OLAC/olacms.html\]](http://www.language-archives.org/OLAC/olacms.html)) is a “shallow and broad” application profile of the Dublin-core metadata standard ([\[http://dublincore.org/\]](http://dublincore.org/)) for describing language resources in general; the ISLE Metadata Initiative (IMDI, [\[http://www.mpi.nl/IMDI/\]](http://www.mpi.nl/IMDI/)) is a proposed “narrow and deep” metadata standard for describing multi-media and multi-modal language resources.

³⁷ SCOIL contains field notes c. 1950 and later. There is an open access assumption: there are no restrictions, but users must register to view digital images and must agree to terms of use.

(BLC),³⁸ usually without agreements, define problem cases as mostly of three types: (1) failures of planning by linguists or archives, (2) linguists seeking to restrict access to archived materials, and (3) heritage communities seeking to restrict access to archived materials.

A failure of planning by the archive could be an “excessively conservative” access policy, for example, when access is restricted until the depositor says otherwise. In one case, a linguist deposited recordings of a severely endangered language, and could not be reached regarding instructions for the use of the materials. When the heritage community sought access for language revitalization purposes, no access was granted for about a year until it was decided that liberal access rules would apply when depositors were inaccessible. Linguists may also commit errors in planning by failing to plan, for example, in the case of linguists who have passed away without specifying what will happen to their materials. If the materials are deposited by the executors of their estate, they may end up being deposited in locations that render the materials inaccessible, possibly in an archive that is far away from the language community and that does not specialize in language material.

Both linguists and language communities may seek to restrict access to language materials. In the case of linguists seeking to restrict language communities from access to archived materials (Conathan and Garrett 2009 cite personal and financial disputes between a linguist and language community), the linguists’ limitations on materials tend

³⁸ The BLC contains audio recordings c. 1950 and later. Conservative access assumptions are made regarding access: no online access or copies made until the depositor is contacted and permits access and copies. To listen to digital audio users must agree to terms of use (no registration).

to prevail and overrule the needs of language community. Conflicts also arise between language communities with related dialects, or the heirs of persons recorded and a language community. Without clear informed consent and agreements on the terms of access, the assertion of heritage community rights is uncertain.

As problems with older material at the language archives at UC Berkeley illustrate, language archives need to have clearly defined policies regarding access conditions; most of these materials were collected without any “contract” between collector and speaker about the future use of the material. Hinton asserts that a crucial task for archives in the future is to ensure that there is a contract with the speaker and collector that makes clear the access conditions (ANA 2005:26). The responsibility of this task naturally falls on the collectors of the material (*i.e.*, the linguist).

Conathan and Garrett (2009:15) recommend that archives have a consistent overall strategy, clear deposit agreements in place, and an understanding of what will happen when those who are allowed to make access decisions are no longer accessible, which will inevitably happen. They also conclude that most actual problems involve “turf disputes,” rather than cultural property issues, except in the sense that all information about a language is the cultural property of the descendants of its speakers. Archives are not well suited to adjudicate such disputes and hence language archives usually let linguists decide. ELAR is developing a “Facebook”-like model for depositors to interact directly with language communities and others who may request access to materials (Nathan in press). Designated representatives of the original depositors and

“default” actions of the archive take on importance once the original depositors become unavailable.

The University of California at Berkeley language archives concludes that, by default, an open access policy for the noncommercial use of materials is desirable; this solves the problem of linguists restricting material. Furthermore, concerns about access are framed in terms of copyright or cultural property, but often neither is applicable, though this may change as groups increasingly define language as their cultural property. Donor agreements are a linguist’s opportunity to place restrictions on access to archival material or to identify culturally sensitive material. Except for materials that are inherently sensitive and should be restricted, archives can have clear policies that make data openly accessible if the original depositors or other designated gatekeepers who place restrictions on the materials are unreachable (*e.g.*, AIATSIS). Another solution is instituting embargo periods³⁹ that restrict access to deposited materials for a fixed amount of time, perhaps to allow the depositor to complete work on publications before the material are released to other scholars or to otherwise protect time-sensitive concerns.

Access restrictions are best framed in consultation with language communities. As the *Protocols for Native American Archival Materials* state (First Archivists Circle 2007:9):

[Libraries and archives must] recognize that the conditions under which knowledge can be ethically and legally acquired, archived, preserved, accessed,

³⁹ This term is borrowed from the world of academic publishing, where embargo periods are used to protect the revenue of publishers. The author’s right to distribute copies for free is restricted until the publisher has had an appropriate amount of time, the embargo period, to recoup costs.

published, or otherwise used change through time. Some materials may have been collected or later restricted by a donor in contravention of community rights and laws or of contemporary federal laws or professional ethics. In all of these cases the rights of a Native American community must take precedence.

Dauenhauer and Dauenhauer (1998:91) note that in language communities of Southeast Alaska, there is a “real and legitimate fear” that traditional ethnic materials will be misused, leading many elders and communities in the direction of secrecy. The concern is not only with regard to ownership, however, but stewardship and transmission. Unless tapes are worked with, they are in danger of deteriorating and fading (Dauenhauer and Dauenhauer 1998:92), and for some language communities, there will soon be no one left to understand and document the content. Their sense of urgency is apparent as they write, “We appreciate the fear of desecration, but we believe that the risks of sharing information are less dangerous at the present time than the risk that it may otherwise be lost forever.” Joel Sherzer, a key figure in the foundation of AILLA, echoes this sentiment (2002:11):

As time goes by, scholars as well as indigenous communities will regret more and more the loss of such materials and value those cases in which it was properly archives, preserved, and made available to others. While I would be the first to recognize that archiving such materials involves ethical and political issues which must be carefully attended to, not to archive and preserve such material is also an ethical and political act in which in many cases can never be undone.



Figure 2.1: Materials at AILLA—to Be Digitized. Materials in the process of being archived at AILLA. Post-it note reads, “From Joel’s office—scan when time permits.”



Figure 2.2: Materials at AILLA—“Handle with Care.” Materials in the process of being archived at AILLA. Boxes are marked “handle with care.”

2.3 Language Communities and Archives

Language death has occurred throughout human history, but today, languages of wider communication are replacing local and minority languages at unprecedented rates and on a global scale (Crystal 2004:50), especially where minority languages are in competition with dominant languages and cultures.⁴⁰ It doesn't take long for a language to disappear once the will to continue with it leaves its community (Crystal 2004:51), though it often takes two generations after the one which failed to pass its language on for members of a community to regret the loss of their language (Crystal 2004:61).

According to Crystal, the first generation in the midst of language shift, struggling to stabilize or establish a new language and social position, is typically not concerned, but their children often acutely regret the loss of their heritage. Harrison records that some last speakers of languages are “resigned to fate,” or think of language shift as “progress,” but last speakers may also regret the loss of their language (2007:9):

Svetlana D., one of the last speakers of Tofa, told me in 2001: ‘The other day my daughter asked me, ‘Mom, why didn’t you teach us Tofa?’ ... I don’t know why. Such a beautiful, difficult language! Now it is all forgotten.’

In some Southeastern Alaskan communities, in which the native languages are moribund unless the observed trends slow down or reverse, many in the community violently object to predictions of language death, and express denial or anger when the topic is broached (Dauenhauer and Dauenhauer 1998:71–72). Dauenhauer and Dauenhauer note that in

⁴⁰ Landweer in personal communication.

these communities, the loss of language or culture is an “extremely emotional issue” to face, involving “the same stages of grief that one experiences in the process of death and dying.”

Once a language has lost its community of native speakers, the task of resurrecting it is “hugely difficult” but not impossible (Crystal 2004:51), as attested by a number of Native American and Australian heritage communities that are engaged in reclaiming their languages from archives (Crystal 2004:57). When an undocumented language dies, it is as if the language had never been (Crystal 2004:49), but when adequate language documentation is available, it can be the basis for revitalization efforts (Hinton *et al.* 2002, Hinton and Hale 2001).

2.3.1 Stemming the Tide

The process of language loss is often unmarked by a landmark decision or a single catastrophic event, though these can be factors. Trends that contribute to the worldwide demise of languages interact in complex ways in disparate parts of the world. Such factors include effects of globalization, modernization, urbanization, and nationalism; political, social, military, religious and economic pressures; privileges accorded to languages of wider communication, ignorance or rejection of multilingualism, disparities in language prestige (Rahman 2003:10); resettlement; increased contact with outsiders leading to a change of marital pattern from endogamy to exogamy;⁴¹ decimation of populations by disease, warfare, genocide (Krauss 1992:6), natural disaster, high infant

⁴¹ Landweer in personal communication.

mortality and low birth rate, group demoralization (Cahill 2004); social, economic, or habitat destruction, deforestation, desertification, displacement, demographic submersion, language suppression in forced assimilation or assimilatory education, and media bombardment—especially television (Krauss 1992:6).

While the odds may seem stacked against them, language communities can work to reverse the tide and then work toward that end. A variety of factors can interact to reinvigorate languages, even those that are spoken by small communities that have seen significant declines in their population and in the use of their language. Eight indicators of relative ethnolinguistic vitality developed by Landweer (2009:1–4) include (1) level of access to the outside world and outsiders; (2) stable domain use of the language; (3) frequency and type of code switching; (4) marriage and immigration patterns that support the language; (5) social networks that support the language; (6) the internal and external recognition of the group as a unique community; (7) language prestige; and (8) access to a stable and acceptable⁴² economic base. Yamamoto notes nine overlapping factors that help to maintain and promote small languages (Yamamoto 1998:114, Crystal 2000:143-144), including the existence of a dominant culture in favor of linguistic diversity, the training of native speakers as teachers, the involvement of the speech community as a whole in revitalization efforts, and the creation and strengthening of environments in which the language is used. Solidarity and increased language use can also be gained by modern means such as the Internet even when a community has been scattered (Crystal 2004:89–91). In several case studies, Cahill basic access to healthcare

⁴² For swidden agriculturalists, this means perceived adequacy of land resources (Landweer in personal communication).

and medicine (*e.g.*, vaccination programs); ability to cope with pressures from outside cultures; the acceptance and support of multilingualism in schools, government, and other public domains; the establishment of language learning venues; the development of a standard orthography, literacy, and publication in the language, with subsequent boosts in the prestige of the language; and restored hope and peace that comes through spiritual renewal (Cahill 2004).

Perhaps the first step in a community being able to stem the tide is to comprehend what is at stake: as Oliver Wendell Holmes wrote, “every language is a temple in which the soul of those who speak it is enshrined” (Crystal 2004:59); or as Ezra Pound stated, “the sum of human wisdom is not contained in any one language, and no single language is capable of expressing all forms and degrees of human comprehension” (Crystal 2004:59). The leader of one language community that has taken action to achieve its own goals for sustaining its language and culture, Pandikar, wrote “I am now in pain. I see my ‘soul’ right in front of my eyes, dying” (Pandikar 2003:3) and with regard to one of the aboriginal languages in Malaysia, “if [it] is lost in Malaysia, it will be lost forever. God will not create it again” (Pandikar 2003:4). However, once a community realizes that its language is in danger, it can introduce measures which can revitalize it.

For many languages, the movement towards the revitalization and perpetuation of endangered or dying languages is a critical race against time (Slaughter 2005:27). Wurm (1991:7) notes that among the possible effects on a language with speakers in contact with another language with speakers who are “culturally more aggressive and more powerful” is that the language may be profoundly altered or replaced over time. The

language may lose distinctness (in its vocabulary, structure, and domains of use), lose a number of its characteristics rooted in the traditional culture of its speakers (becoming in some ways “an imitation of the language of the culturally more aggressive people”), or disappear and be replaced (entirely or in a modified or “pidginized-creolized” form of it). Slaughter (2005:26) of the Indigenous Language Institute in Santa Fe, New Mexico identifies three main priorities: (1) identifying the remaining speakers and documenting their language and knowledge; (2) transferring and quickening the pace of transfer of their language and knowledge to the community; and (3) deepening the knowledge of the language in the community. Measures to accomplish objectives (2) and (3) include bilingual school programs, language classes, summer workshops, creative uses of multimedia, webcasts, and language fairs, and depend on the involvement of the community and the availability of adequate resources. Objective (1), linguistic documentation, supports development of a “critical mass” of pedagogical materials for learners and teachers. The “vast array of Native American phrasebooks and dictionaries, workbooks, reference grammars, curriculum materials, reading materials and workbooks” are being produced mostly by local native teachers and education staff and in more extreme situations are developed from archival holdings (Hinton 2005:25). The revival of Wôpanâak—a language in the northeast region of the United States that hadn’t been written or spoken for nearly 150 years—was aided by the rich documentation discovered by Jessie Little Doe Fermino who found a huge body of native written documents including letters to the Massachusetts Legislature in the 1700s, pleading with lawmakers

to keep white settlers from taking their land, and the first Bible printed in the New World (John Eliot's translation of the King James Bible printed at Harvard in 1663).⁴³

Aside from raw linguistic materials, Hinton suggests that language teaching and learning materials produced for Native American languages should be actively archived since language programs come and go, and materials developed under favorable conditions are liable to be lost when conditions are unfavorable. Since heritage languages are now often taught as a second language, more tools, new skills and materials are required (Slaughter 2005:26).

2.3.2 Access from the Perspective of the Language Community

As discussed in the preceding section, language documentation provides the raw material for revitalization projects for languages on the brink of extinction. Archived materials are invaluable to people trying to keep or regain their languages and cultural traditions. The wordlists and dictionaries, grammars and texts collected in the past are often the only materials left with which the communities can work to learn and attempt to re-establish their languages (Hinton 2005:25). Experience has shown that the main users of digital language documentation are the descendants of the speakers and other members of the indigenous language and heritage communities. For instance, the Berkeley Language Center archives⁴⁴ report that more than 90% of their users are from language and heritage communities, seeking materials for language and cultural maintenance and revitalization (Garrett *et al.* 2008:20,43, Hinton 2005:24–25).

⁴³ [<http://archive.southcoasttoday.com/daily/11-00/11-19-00/a03sr014.htm>]

⁴⁴ [<http://blc.berkeley.edu/index.php/blc/pages/collections/>]

Much of the work of linguists, ethnographers, musicologists and anthropologists on indigenous languages lay dormant in overcrowded archives, museum basements and off-campus warehouses (Kipp 2005:24). These linguistic resources can find new and valuable life in the hands of the Native people from whose heritage languages they derived (Hinton 2005:24). From the perspective of language revitalization, the first priority is to ensure that there are ample resources for the learners and teachers of the languages (Slaughter 2005:27). However, just collecting a language documentation corpus is not enough; the language must spread out and into the community and families and be used in order to be revived (ANA 2005:49). Therefore discovering and accessing language documentation are the key issues in this discussion. Linguists and other depositors of language materials are encouraged to consider their intended audience and to ensure that the archive caters to that audience. This issue of designated communities is item 3 of the TAPS Checklist; see section 4.2.3.

2.3.3 Ongoing Relationship

Currently there is a void in many indigenous communities regarding information and material on their languages. The bulk of language documentation lies dormant in academic archives (ANA 2005:23). The creation of an accessible and reciprocal connection between tribal communities and the archives is called for in order to bridge the gap. The broadly defined subject of ongoing relationships, the fourth item on the TAPS Checklist, includes revenue sharing as one of the possible ways a language community can benefit from archived materials; see section 4.2.4.

Once an archive has successfully processed historical and legacy materials so that they are in usable formats, contacts with a language community engaged in cultural and language revitalization should be maintained to keep them informed of work done by scholars in recent years (ANA 2005:23). One archive may not be able to accommodate all the holdings that a community may be interested in, and the community may not be interested at present, which brings us to the next issue of needing to be able to discover resources as they are needed, regardless of where they are located.

2.3.4 Discoverability of the Resources

Many types of materials relating to a heritage language may have value in a revitalization effort. Most of the archived work on heritage languages, however, was completed generations ago and there is a distinct void in information readily available to language communities about how their language was studied, reported or used by others. Sometimes the name of the linguist who was involved can be used to identify a collection and determine where it is archived (ANA 2005:24).

Adequate and properly indexed metadata assists discovery and piecing together the context. Resources created today, then, should incorporate standardized metadata. It is helpful to compile scattered collections of a heritage language into one location or index (ANA 2005:24). For locally indexed resources, the language community would first have to know where to search for indexed materials. Posting the indexed resources on the web still presents difficulties. Internet search engines are unreliable for locating language resources since search results lack precision, presenting many unrelated results.

They also lack good recall, not presenting all the desired results. An archive can greatly improve search by using standardized descriptive metadata, which is structured, machine-readable data that describes the resources. The discoverability of resources is further enhanced if the archive participates in an aggregating service like OLAC's that periodically harvests metadata and brings together all known information about a given language into one place. Discoverability is the first topic (item 5) under the Access category of the TAPS Checklist; see section 4.3.1. Continued access after a resource is discovered is aided by the use of fixed identifiers, item 6 on the TAPS Checklist; see section 4.3.2.

2.3.5 Reaching Communities with Language Resources

Once language materials can be reliably located, the language community needs to be able to obtain them. Since many of these communities are small and remote, cost can be an issue. Even "free" digital copies of materials may incur shipping costs. Resources transmitted through the web also require Internet connections, the necessary equipment, and electricity. Partnerships between archives and local cultural centers that can provide copies and local access are one possible solution. The issue of reach is addressed in item 7 of the TAPS Checklist; see section 4.3.3.

2.3.6 Access Restrictions

An archive should be able to accommodate legitimate access restrictions such as language material that includes injurious gossip, or sacred or sensitive content. Some

language communities call for sensitivity in handling ceremonial or gender restricted material, sensitive genealogical material, photographs or recordings of deceased persons. The benefits of open access should be considered the norm rather than the exception in order to avoid undo restrictions. NAL at the Sam Noble Oklahoma Museum of Natural History puts forth reasons for an unrestricted access policy in its *Restrictions on Use Policy*.⁴⁵ First, since the users of archived materials are largely from the language communities concerned, restrictions would more likely limit native peoples than nonnative peoples from accessing language and language teaching materials. Secondly, many of the documented languages have or will soon have no speakers; these people who have lost their language have experienced any number of destabilizing forces and undergone profound changes such that original restrictions may not make sense and should not apply—all that is left of their languages is or will be what was stored in archives. Thirdly, restrictions based on tribal membership are, in most cases, based on blood quantum, but due to intermarriage, direct descendants may not be eligible for the same tribal rolls; restricting collections to specific tribal membership has in some cases unwittingly kept direct heirs from using materials. Finally, restrictions require that an authorized member of the language community or the depositor update who may have access to the collection. When this information does not get updated as is often the case, the material may become completely inaccessible. As mentioned in section 2.2.5, possible solutions to problems with access restrictions is to make clear that restrictions will expire after a fixed time period, or if contacts with the language community or

⁴⁵ [<http://www.snomnh.ou.edu/collections-research/cr-sub/nal/restrictions%20policy.pdf>]

depositors are not kept up. Access and use restrictions are considered in item 8 of the TAPS checklist; see section 4.3.4.

2.4 Challenges of Digital Preservation for Archives

The challenges of digital preservation for language archives are generally subdivided into three areas. Sections 2.4.1 through 2.4.3 address the human issues, technical issues, and organizational fitness concerning digital language archives.

2.4.1 Challenges of Digital Preservation: the Human Side

Jim Gray of Microsoft is credited with saying, “May all your problems be technical.” With respect to archiving, it is possible for an archive to solve difficult technical problems by putting forth more effort; however, an archive’s efforts may not be able to solve problems that arise from the behaviors of others. Digital language archives, like other memory institutions, are changing the way people interact with information and thus they need to decide what information to save, how long to save it, and to anticipate how people will want to interact with information well into the future. When it comes to digital language archives, the rights and interests of researchers, future scholars, speakers, performers, and the larger language or heritage community must also be taken into consideration.

Administrative and political processes can also take enormous amounts of time and often cause some frustration. In the set-up of the Kaipuleohone archive at the University of Hawai’i at Mānoa, for example, numerous meetings were needed to discuss

the details of depositing linguistic data within the existing university library system and in the end, only text-based documents could be fully supported. Continued funding that often depends on decisions made by those outside the circle of immediate stakeholders is also a challenge for the sustained functionality and existence of digital language archives.

It is simple to state a commitment to the long-term persistence of digital materials, but it is complex to articulate precisely what the outcome of preservation should be for future users—the language communities, depositors, and scholars. Since digital content can incorporate features that have no parallel in physical materials, archives must determine how many of these features can or should be preserved for future users of the content. A set of agreed outcomes is the second of thirteen aspects of digital preservation examined by Lavoie and Dempsey (2004). The choice of preservation strategies, as well as issues of access, and funding the services that the archive provides will need to reflect a consensus of all “stakeholders” associated with the archived materials. Communication between the archive and stakeholders is necessary in order to promote consensus on preservation outcomes and make clear the archive’s preservation policies. This ensures that the archive’s commitments match stakeholder expectations. However, it is duly noted that “achieving such a consensus is difficult, and in some circumstances, impossible” (Lavoie and Dempsey 2004). It has been assumed that digital language data should be preserved indefinitely as a valuable part of the human record, but questions regarding the communities for whom the materials are preserved have far-reaching implications (Smith 2004:1). Distinctive qualities of individual languages and the

particular needs of the speaker communities also determine how users prefer to interact with archives (Dobrin *et al.* 2007:7).

2.4.2 Challenges of Digital Preservation: the Technical Side

As an emerging and developing field, the sound implementation of digital archives continues to pose significant technical challenges. Document and media formats continue to proliferate, with improvements in hardware and software promoting obsolescence. While there are recognized “archival” formats that can be supported over time, the prevalence of proprietary working and presentation formats present complications for digital language archives, which serve depositors who will often prefer to retrieve digital objects in their original form. Clearly, simply preserving bit streams is not the same as preserving digital information in usable forms. The ease of creating digital objects, including numerous variants of the same item, may create an “information glut,” which poses challenges for the selection of digital objects for accessioning. Additionally, though video holds great promise in capturing communicative detail, there are not yet agreed upon archival forms for digital video formats, unlike for text and graphics. Digital archives need to be selective about materials they commit to preserve, and define formats that they will support, as defined by their submission criteria, item 2 of the TAPS Checklist; see section 4.2.2.

Digital archives require a level of sophistication in planning and preparing for changes over the long-term, to maintain the viability and machine-readability of digital objects. The problem is two-fold: keeping up with changing technologies (forward

migration of data) and managing physical decay—establishing a regular process of error detection to determine if degradation is occurring, regular copying to new media, and error correcting codes to ensure new generations are faithful copies of the original (Leggett 2005:14).

The third category of TAPS, Preservation, addresses the technical processes of digital archives on which the preservation of archived language resources depend (see section 4.4): item 9 addresses evidence of long-term planning (see section 4.4.1); item 10 addresses preservation strategies (see section 4.4.2); item 11 addresses the continued integrity of the materials (see section 4.4.3); and item 12 addresses the continued authenticity of the materials (see section 4.4.4).

2.4.3 Organizational Fitness of Digital Language Archives

Even with the most elegant technical implementation, effective ways to interface with user communities, and cordial relations with host institutions, the continued existence of any digital language archive depends on the sustainability of the infrastructure and organizational planning that supports it. Organizational fitness is considered in the fourth category of TAPS, Sustainability; see section 4.5.

The quality of the infrastructure of an archive includes both trained staff in adequate numbers, and the “technical infrastructure”—the adequacy of the physical plant, sound computing practices and digital media. Item 13 of the TAPS Checklist addresses the adequacy of this infrastructure; see section 4.5.1.

In order to ensure longevity, digital language archives need to have long-term sources of funding or be otherwise economically viable to pay for adequate staff, facilities, and services. Though the cost of digital storage has dropped dramatically over the years, the file sizes of digital images, video, and multimedia presentations are much larger in comparison to files that are text-based, and thus require much more storage space. Some language collections are geared primarily toward teaching a language and contain enormous amounts of audio and video-based documentation. When dealing with “many gigabytes” of data, server space continues to be a precious commodity (Nathan 2009:2). The maintenance of technical processes and other services that an archive provides tend to rise not only with inflation, but with increases in the size of an archive’s collection (Beagrie *et al.* 2008:53). The economics of digital preservation is an area of recent and ongoing research (BRTF-SDPA 2010, Beagrie *et al.* 2010). Item 14 of the TAPS Checklist addresses the issue of financial sustainability; see section 4.5.2.

Finally, even with adequate infrastructure and funding in the foreseeable future, an archive should have contingency plans in case of disaster or cessation of operations. Item 15 of the TAPS Checklist, disaster preparedness, addresses responsible back-up practices and a disaster recovery plan; see section 4.5.3. Item 16 of TAPS addresses a succession plan, if for any reason the archival institution ceases to exist; see section 4.5.4. The next chapter discusses the establishment of best practices for digital archives, and the tools for assessing archives’ trustworthiness on which the TAPS checklist is based.

Chapter 3: Review of Tools for Assessing Archival Practice

The goal of archiving the results of a language documentation project is to ensure that materials are securely preserved and accessible over the long-term. The basic question the depositor must ask is whether an archive is trustworthy or not to perform these functions. There are several tools available for assessing archival practices that are based on standards for digital archiving. Sections 3.1 through 3.7 discuss the development of these standards that establish best practice for digital archiving, and describe existing assessment tools, including a brief history and summary of the scope of the tools. Since all of these tools are meant to be used by archiving professionals, section 3.8 introduces the need for a simpler tool, the TAPS Checklist, which is aimed at helping linguists and other depositors of language materials to assess the fitness of digital archives for receiving their materials.

3.1 A Clear Call for Standards: the Task Force on Archiving of Digital Information

In December 1994, the Commission on Preservation and Access and the Research Libraries Group (RLG) created the Task Force on Archiving of Digital Information. This Task Force was charged with defining the key issues in digital archiving to ensure “continuing access to electronic digital records indefinitely into the future” (Task Force on Archiving of Digital Information 1996:iii). In their 1996 final report, *Preserving Digital Information*, the Task Force determined that critical to the emerging digital archiving infrastructure were adequate numbers of trustworthy organizations that stored,

migrated, and provided access to digital collections (Task Force on Archiving of Digital Information 1996:40). Furthermore, the Task Force recognized that these trustworthy organizations could not simply identify themselves as such and called for a process of certification for digital archives that would create a “climate of trust” regarding the preservation of digital information (Task Force on Archiving of Digital Information 1996:40). The Task Force did not articulate details of such a certification process, however, since at that time there was no organized “digital preservation community” with common, consensus-driven practices, and no standards on which to base criteria for certification.

3.2 Towards Standardization: the *Reference Model for an Open Archival Information System (OAIS)*

Within the same period of activity as the Task Force on Archiving of Digital Information, significant steps towards common standards in digital archiving were being made with the development of the *Reference Model for an Open Archival Information System (OAIS)*. In 1995, the NASA Consultative Committee for Space Data Systems (CCSDS) began developing the *Reference Model for an Open Archival Information System* as an ‘open’ and public model, in conjunction with archive and library communities (Kenney and Buckley 2005). As a result of these early cooperative efforts, the OAIS Reference Model was already being widely adopted as a starting point in digital preservation efforts before becoming an ISO standard in early 2003 (as ISO 14721:2003). For example, in March 2000, RLG and the Online Computer Library Center (OCLC)

began their collaboration to establish attributes of a digital repository for research organizations, building on and incorporating the emerging international standard of the OAIS Reference Model (RLG and OCLC 2002:i); see section 3.3 below. Some of the early implementers of OAIS included the Networked European Deposit Library (NEDLIB), the National Library of Australia, the CURL⁴⁶ Exemplars in Digital Archives (CEDARS) project in the U.K., the U.S. National Space Science Data Center (NSSDC), and the U.S. National Archives and Records Administration (NARA).

The OAIS Reference Model was designed to create broad consensus on the requirements for an archive to provide long-term preservation of digital information. Permanent or indefinite long-term preservation must consider the impacts of a changing user community, of outliving institutions that house information, and of changing technologies, including support for new media and data formats (CCSDS 2002:1-1). The focus of the model is on digital information, both as the primary form of information held and as the form of supporting information about archived materials. The model can also accommodate information that is inherently non-digital, but does not address the modeling and preservation of such information in detail. The model is applicable to any archive, as it does not specify a particular design or implementation.

⁴⁶ The Consortium of University Research Libraries (CURL) is composed of research libraries in the British Isles whose mission is "to promote, maintain and improve library resources for research in universities [<http://www.ercim.eu/publication/ws-proceedings/DELOS6/cedars.pdf>].

The “Blue Book” version⁴⁷ of the *Reference Model for an Open Archival Information System* is a technical “Recommendation” that establishes a common framework of terms and concepts which can be called an Open Archival Information System (OAIS).⁴⁸ Most importantly for our discussion, the Reference Model defines a “minimal set” of responsibilities for an archive to be called an OAIS, that is, an archive consisting of the people and systems (*e.g.*, policies, procedures, hardware, software) responsible for acquiring, preserving and disseminating the information for a “designated community” (CCSDS 2002:1-1). This distinguishes an OAIS archive from other uses of the term “archive.”

The OAIS Reference Model addresses a full range of archival functions (ingest, archival storage, data management, access, and dissemination), migration of digital information to new media and forms, data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. It also identifies internal and external interfaces to archive functions, as well as some high-level services at these interfaces (CCSDS 2002:1-2). The new and unencumbered set of terms and concepts that the Reference Model introduced allow

⁴⁷ CCSDS documents have nine different possible color designations [<http://public.ccsds.org/about/FAQs.aspx>]. Blue Books designate “recommended standards” and represent the highest level of technical specification [<http://public.ccsds.org/publications/BlueBooks.aspx>]. In August 2009, a “Pink Book” version of the *Reference Model for an OAIS* was released, updating the original issue “based on input from the user community as well as working group-internal review” [<http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx>, available at: <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf>]. Pink Books are CCSDS Draft Recommendation that is an update to a Blue Book that is released for formal review.

⁴⁸ The term ‘open’ in OAIS is used to imply that this Recommendation, as well as future related Recommendations and standards, are developed in open forums. It does not indicate that access to the archive is unrestricted.

archives to be meaningfully compared, provide a basis for further standardization, and promote greater awareness and support of archival requirements (CCSDS 2002:iii).

The OAIS Reference Model continues to have wide acceptance in the digital library community, and has become the authoritative model for best practices in digital archiving. Critics have pointed out that since space agencies were the first to realize the need to make data interoperable and accessible over long periods, and because OAIS was developed by the Consultative Committee on Space Data Systems, the “space bias” occasionally emerges; for example, “the information may in general be submitted using a wide variety of common and not-so-common forms, such as books, documents, maps, data sets, and moon rocks” (CCSDS 2002:3-1). Priscilla Caplan of the Florida Digital Archive (FDA) notes that the OAIS model of content information “probably makes a lot of sense for science and social science datasets,” but that it is difficult to apply to textual or visual cultural heritage information. For example, the analysis of OAIS by the OCLC/RLG Working Group on Preservation Metadata found the distinction between semantic and structural representation information “too hard to make” and omitted it from their implementation of an OAIS (Caplan 2004:7).

3.3 Towards Standards for Archiving Cultural Heritage Resources: *Trusted Digital Repositories (TDR)*

In 2002, the same year that the *Reference Model for an Open Archival Information System* came out, the RLG and the OCLC published *Trusted Digital Repositories: Attributes and Responsibilities (TDR)*, which built upon and incorporated

OAIS as an emerging international standard (RLG 2002:i). *TDR* further “articulated a framework of attributes and responsibilities for trusted, reliable, sustainable digital repositories capable of handling the range of materials held by large and small cultural heritage and research institutions” (OCLC and CRL 2007:1). The document was particularly useful to institutions that were entrusted with the long-term preservation of cultural heritage resources and could be used in combination with the OAIS Reference Model as a digital preservation planning tool, as a basis for an institutional program, and as a method to enable the effective exchange of information and developments between institutions (OCLC and CRL 2007:1). It concentrated on high-level organizational and technical attributes of digital archives and discussed potential models for digital repository certification. It was not prescriptive about the specific nature of rapidly emerging digital repositories and archives, but reiterated the call for certification of digital repositories and recommended the development of a certification program and articulation of auditable criteria (OCLC and CRL 2007:1).

3.4 Measuring Compliance: *Trustworthy Repositories Audit and Certification (TRAC)*

Even before the OAIS Reference Model became an ISO standard, institutions began to declare themselves “OAIS-compliant” to underscore their trustworthiness as digital repositories. What was lacking, however, was an established understanding of “OAIS-compliance” and criteria that could measure compliance (OCLC and CRL 2007:Foreword).

In response to these concerns, another major development in determining the trustworthiness of digital archives was announced with the 2007 release of the *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (OCLC and CRL 2007:2) written by the RLG-National Archives and Records Administration Task Force on Digital Repository Certification and jointly published by the Center for Research Libraries (CRL) and OCLC. Formed in 2003, the joint task force specifically addressed digital repository certification with the goals of developing widely applicable criteria and a certification process to “identify digital repositories capable of reliably storing, migrating, and providing access to digital collections.” In the last two years of the development of the audit and certification criteria and checklist, the RLG-NARA task force received critical contributions from the CRL Auditing and Certification of Digital Archives project, the NESTOR⁴⁹ project (Germany), and the Digital Curation Centre (United Kingdom) (OCLC and CRL 2007:Foreword).

The audit checklist was based on the OAIS Reference Model. The following quote from the Introduction indicates a wide range of intended uses:

Though designed as a set of criteria to facilitate the certification of digital repositories, this document and ... checklist have a number of uses outside of the carefully prescriptive world of certified repositories. Envisioned uses of this

⁴⁹ The German network of expertise in digital long-term preservation. Named after the advisor of the ancient Greeks in Troy, it is “a cooperative project of libraries, archives and museums as well as of leading experts forming a network of expertise in long-term preservation and long-term availability of digital resources” (from the NESTOR homepage: [<http://www.langzeitarchivierung.de/index.php?newlang=eng>]). “The acronym NESTOR was taken from the English version of the official BMBF (Bundesministerium für Bildung und Forschung, German Federal Ministry for education and research) project name ‘Network of Expertise in long-term STorage and long-term availability of digital Resources’” (from the FAQ: [http://www.langzeitarchivierung.de/modules.php?op=modload&name=PageEd&file=index&page_id=23]).

document include repository planning guidance, planning and development of a certified repository, periodic internal assessment of a repository, analysis of services which hold critical digital content on which institutions rely, and objective third-party evaluation of any repository or archiving service (OCLC and CRL 2007:5).

In a press release, James Michalko, Vice President of RLG Programs, stated, “This is a critical time for research institutions tasked with providing long-term access to digital information. *TRAC* will help institutions objectively evaluate responsibilities against capabilities and identify potential risks to digital content held in repositories, archives, and by content providers. It provides the community with a tool to facilitate assessment and understanding, and will enable vital collaboration among repositories.”⁵⁰

The *Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist* represented the work of many experts from an international range of communities in research, governments, data archives, and cultural heritage organizations (OCLC and CRL 2007:3) and represented best current practice and thought about the organizational and technical infrastructure required to be considered trustworthy and capable of certification. Unlike *TDR*, *TRAC* was explicit; it: (1) established a baseline definition of a trustworthy digital repository, specifying the components that must be considered and evaluated as a part of that determination; (2) discussed the envisioned uses of the document, and the principles underlying the application of the criteria; and (3) documented criteria that trustworthy repositories should meet, providing explanations

⁵⁰ [<http://www.oclc.org/research/news/2007-03-12.htm>]

and examples (OCLC and CRL 2007:2). The eighty-eight page document breaks down audit and certification criteria into three parts:

- A. Organizational infrastructure
- B. Digital object management, and
- C. Technologies, technical infrastructure, and security.

These three parts are further divided into 14 subsections. In all, 114 distinct certification criteria are listed among the subsections. The description of the criteria is followed by appendices including a glossary, minimum required documents (detailing policies, procedures, and plans of an archive, preservation planning and strategies, the access and delivery systems of an archive), and explanations of OAIS definitions of understandability to a designated community and “ingest” of objects into an archive. A detailed checklist in table form for audit purposes is also included. While it makes no claim of being exhaustive, it is one of the most comprehensive sets of certification guidelines available to date, incorporating “existing standards and best practices for trustworthy repositories and related digital object management and [it] is applicable for audit and certification activities” (OCLC and CRL 2007:4).

In the 18 months before TRAC was published, three organizations worked to establish a unified, international process for certification. The CRL worked in conjunction with the Digital Curation Centre or DCC (U.K.) in the Auditing and Certification of Digital Archives project, and the certification working group of the Network of Expertise in Long-Term Storage of Digital Resources (NESTOR) project in Germany. The DCC conducted audits of selected digital archives, while NESTOR

developed its own set of standardized criteria. The resulting products, a risk-management tool created by DCC and Digital Preservation Europe (DPE), and NESTOR's *Catalogue of Criteria*, are discussed respectively in sections 3.6 and 3.7 below. The outcome of an international standard for the certification of digital archives is briefly noted at the end of section 3.7.

3.5 Managing Risk with Internal Audits: *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*

In the same month that version 1.0 of *TRAC* was released, DCC and DPE came out with the first iteration of the *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA); the interactive web application⁵¹ has a user's manual in PDF format (DCC and DPE 2009). DRAMBORA was the result of a period of pilot audits of digital archives by DCC from April 2006 to January 2007 (DCC and DPE 2007:25), primarily using the *TRAC* and *Catalogue of Criteria* audit and certification tools as starting points.⁵² These audits exposed that these instruments could not quantify “the extent and effectiveness of organisational compliance” and that “a reliable means for comparing and assessing repositories that are heterogeneous in terms of their scale, scope or mission” remained elusive (DCC and DPE 2007:23).

While holding essential the goal of international consensus on methodology and criteria for auditing digital archives, DRAMBORA “does not attempt to present a

⁵¹ [<http://www.repositoryaudit.eu/>]

⁵² One of the participating archives was the FDA, where I also conducted a site visit (see Appendix B-3). The DCC report may be found here: [<http://www.fcla.edu/digitalArchive/pdfs/DCCfinalreport.pdf>]

comprehensive list of best practice criteria or a benchmark based on specific standards” (DCC and DPE 2007:23). Instead of being another alternative for the assessment of archives, *DRAMBORA* is a toolkit based on principles of risk-based management that “aims to provide a complementary approach that can be used in association with the efforts of both *TRAC* and *NESTOR*” (DCC and DPE 2007:23). *DRAMBORA*, then, facilitates “self-audit” of archives by providing archive administrators with a methodology to comprehensively assess “objectives, activities and assets before identifying, assessing and managing the risks implicit within their organization.”⁵³ While the toolkit offers “a quantifiable insight into the severity of risks faced by repositories right now, and an effective means for reporting these,”⁵⁴ the success of a self-audit depends upon the commitment of a repository—not necessarily an archive—itsself. From the beginning, *DRAMBORA* was intended to be versatile in order to complement other methods of repository audit and certification and sought to address a wide range of repository types, even those that do *not* aim for long-term preservation (DCC and DPE 2007:10).

DRAMBORA Version 1.0 is a 221 page document of written guidelines, with three parts and six appendices. The toolkit itself consists of Part II, Appendix 2, and Appendix 4 (DCC and DPE 2007:17). Part II lays out the audit process and describes the six stages of audit, Appendix 2 incorporates a suite of templates to support the process of conducting the self-audit, and Appendix 4 provides an example of how an audit report might usefully be structured. The self-audit toolkit broadly defines the core functions of

⁵³ [<http://www.repositoryaudit.eu/participate/>]

⁵⁴ [<http://www.dcc.ac.uk/resources/tools-and-applications/drambora>]

a digital repository as eight default “functional classes” which correspond with items in the *TRAC* and NESTOR criteria:

1. acquisition and ingest,
2. preservation and storage,
3. description and metadata management,
4. access and dissemination
5. organization and management
6. staffing
7. finance management
8. technology support and security. (DCC and DPE 2007:47)

These classes are subdivided; the first four are “operational” functional classes specific to archiving, and the last four are “supporting” functional classes which are common to any organization (DCC and DPE 2007:17). The six stages of the audit guide the auditor along a route of analysis similar to that of a potential external auditor (DCC and DPE 2007:27) and the entire self-audit process is estimated to take 24 to 40 hours (DCC and DPE 2007:30).

The DRAMBORA self-auditing procedure would be of value to digital language archives, helping to (1) provide peace of mind with regard to growing, valuable, and at-risk digital collections, (2) increase efficiency by helping to focus and refine operational policies, and (3) perhaps highlight potential opportunities for repository managers to leverage increased development potential by offering a clear way to demonstrate the risks

related to shortfalls in repository funding.⁵⁵ An archive having carried out a DRAMBORA audit would also tend to strengthen the trust of depositors and users, though such an audit may not be a high priority for some archives dealing with a backlog of material to archive, an ongoing need for digitization, and limited funds and personnel. In any case, DRAMBORA is a tool for the managers of archives to improve digital archives from within, rather than being designed for depositors and consumers to make judgments about the organization in question.

3.6 Further on toward Certifying Trustworthiness: the *Catalogue of Criteria for Trusted Digital Repositories*

The NESTOR Working Group on Trusted Repositories Certification in Germany released *Catalogue of Criteria for Trusted Digital Repositories Version 1* (draft for public comment in German and English) in late 2006, and an updated *Version 2* (German only) in 2008 for the purposes of defining “a first catalogue of criteria for trustworthiness and to prepare for the certification of digital repositories in accordance with nationally and internationally coordinated procedures” (NESTOR 2006:6). The document introduces the problems surrounding the long-term preservation of digital objects, and describes key concepts and principles behind the *Catalogue of Criteria* that follows. Like *TRAC*, the *Catalogue of Criteria* was aimed at memory institutions entrusted with providing long-term access to digital materials and was intended for a variety of purposes: “devising, planning and implementing a trusted digital long-term repository” and for internal review

⁵⁵ [<http://www.dcc.ac.uk/resources/tools-and-applications/drambora>]

during any stage of development (NESTOR 2006:4), but rather than being used for external auditing, it was specifically formulated “to serve as an orientation and self-check tool.” Though discussion and standardization within the international context was anticipated, it was compiled mainly for application in Germany (NESTOR 2006:5). The *Catalogue of Criteria* conforms to OAIS terminology, and lists documentation, transparency, adequacy, and measurability as factors pertaining to application of the criteria (NESTOR 2006:4-5). The criteria catalogue itself is divided into three sections, similar to *TRAC*:

- A. Organizational framework,
- B. Object management, and
- C. Infrastructure and security.

These sections include 14 points, most of which are broken out into sub-points, comprising 54 distinct criteria in all. The document closes with the *Catalogue of Criteria* in a compact checklist, formatted for planning and checking purposes, and a glossary (NESTOR 2006:1). The most current version of the document in German is 55 pages in length.

At the times of their respective publications, both *TRAC* and the *Catalogue of Criteria* stopped short of merging into a standard that was adopted by an international standards body. Despite the vision for the development of a certification process to take place “in an international environment and with a unified set of criteria, but with regional implementation, [for example], by country, continent, or geographic region ... small but important differences” emerged among the two sets of criteria. Thus, efforts to form a

unified process of certification was deemed “impractical for geopolitical reasons” (OCLC and CRL 2007:4). It was agreed that CRL, DCC, and NESTOR would act as the audit and certification bodies in their respective regions (North America, U.K., and Germany). Members of the working groups at each organization would also encourage formal collaboration among representatives of other countries to form a “virtual agency” for digital repository audit and certification (OCLC and CRL 2007:4). Currently, a voluntary Working Group at the CCSDS is engaged in developing a new standard based on TRAC for submission to the ISO for approval.⁵⁶

3.7 Certification of Archived Data: *Data Seal of Approval* (DSA)

In 2008, Data Archiving and Networked Services (DANS)⁵⁷ of the Netherlands set forth a “minimum set” of guidelines, which have since been revised,⁵⁸ for certification of archives with their *Data Seal of Approval* or *DSA*.⁵⁹ These guidelines were written in accordance with previous tools for assessing digital data archives noted above (OCLC and CRL’s *TRAC*, NESTOR’s *Catalogue of Criteria*, and DCC and DPE’s DRAMBORA) and form the basis for granting a “Data Seal of Approval” (DANS 2010:4). The seal:

- Gives researchers the assurance that their research results will be stored in a reliable manner and can be reused.

⁵⁶ [<http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying>]

⁵⁷ DANS is an institute of the Royal Netherlands Academy of Arts and Sciences or in Dutch, Koninklijke Nederlandse Academie van Wetenschappen (KNAW), and is also supported by the Netherlands Organization for Scientific Research (NWO). Since its establishment in 2005, DANS has been taking care of storage and continuous accessibility of research data in the social sciences and humanities.

⁵⁸ [<http://www.datasealofapproval.org/?q=node/35>]

⁵⁹ [<http://www.datasealofapproval.org/?q=frontpage>]

- Provides research sponsors with the guarantee that research results will remain available for reuse.
- Enables researchers, in a reliable manner, to assess the repository where research data are held.
- Allows data repositories to archive and distribute research data efficiently (DANS 2010:3).

The stated goal of the *DSA* is to “safeguard high-quality and reliable processing of research data for the future without it entailing new thresholds, regulations or high costs” (DANS 2010:3) and addresses three sets of stakeholders: “research institutions (the data producer), organizations that archive data (the data repository), and users of those data (the data consumer)” (DANS 2010:5). The most current version of the document is 16 pages in length and contains sixteen guidelines that relate to the implementation of the following five criteria (DANS 2010:5):

1. The research data can be found on the Internet.
2. The research data are accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data.
3. The research data are available in a usable format.
4. The research data are reliable.
5. The research data can be referred to.

Three of the 16 guidelines pertain to the data producer, three to the data consumer, and the remaining 10 to the data repository. Archiving institutions are expected to retain the

primary responsibility for safeguarding data and to handle “the overall implementation of the *DSA* in its own specific field.”⁶⁰

The Data Seal of Approval is awarded to data repositories by the Data Seal of Approval Assessment Editorial Board (DSAA EB) for having demonstrated that the data concerned conform to the sixteen guidelines on which the *DSA* assessment procedure is based. The repository is then permitted to display the *DSA* logo, via HTML code, on the front page of its website and in other locations relevant to its communication in the realm of scientific and scholarly research. The DSAA EB would also place a link pointing to the repository on its website datasealofapproval.org. However, the seal is not in itself a mark of a formal external audit or certification; rather the DSAA EB reviews the self-assessment of an archive “on the basis of trust.”⁶¹ The *DSA* gives some basis for claims of the durability of digital data, and more generally, simply promotes the goal of durable archiving.

3.8 The Need for a Tool to Assess Digital Language Archives

The concerns raised in chapter 2 with regard to the challenges that linguists, language communities, and digital archives face are only partially addressed by the documentation and tools describing the implementation and operations of trustworthy digital archives discussed above. The tools for assessing archival practice discussed in this chapter have been largely directed toward the archiving professionals themselves, whether for an internal or external review or “audit.” Many of these tools are technical in

⁶⁰ [<http://www.datasealofapproval.org/?q=node/12>]

⁶¹ [<http://www.datasealofapproval.org/?q=node/12>]

nature and not readily accessible to the average depositor. The influential *TRAC: Criteria and Checklist*, including no less than 114 items, is a somewhat unwieldy tool for an informal third-party evaluation of a repository or archiving service. The *Catalogue of Criteria* developed by NESTOR is briefer, but still too long and too technical to be readily accessible. Even if linguists wanted to use these tools to make an assessment of a digital archive, these tools are too complex to be useful. The *Data Seal of Approval* guidelines comprises a “minimum set” based on these and other guidelines, and the corresponding Data Seal of Approval lends confidence to the quality of data in a digital archive, but it lacks a single viewpoint from which to assess criteria, and lacks specificity regarding what the depositor needs to ascertain before entrusting his or her work to the repository (see section 4.2 on Target).

As noted in section 3.6, work is proceeding nationally and internationally towards the goal of digital archive certification. It is hoped that eventually depositors will be able to rely on the trustworthiness of certified archives and “people [will be able to] trust information from a digital repository⁶² as readily as they trust twenty dollar bills from an ATM, without looking inside the shell” (OCLC 2007). With language materials, the concern is not just “counterfeit” information, however. An even greater concern is that the archive will be able to provide access to the originally deposited materials at an unknown date in the future (see sections 4.3 and 4.4 addressing Access and Preservation). Ultimately, the archive needs to establish trustworthiness, and like trustworthy banks, be repositories that will not fail over time (see section 4.5 on Sustainability).

⁶² Synonymous and preferred to “archive” in much of the literature on digital language archiving.

At this time, few digital archives have achieved rigorous certification.⁶³ Only one, Portico,⁶⁴ a dark archive (see section 2.1.5), has attained certification through CRL, which conducted a preservation audit of Portico between April and October 2009, and declared it a trustworthy digital repository as of January 2010 (CRL 2010:2). None of the language archives discussed in chapter 2, however, has participated in the kind of auditing discussed in this chapter. A linguist-depositor currently has no way to know if a given digital language archive is trustworthy.

Even with the advent of certified digital archives, the choice of an archive should take careful consideration. The process of choosing a digital archive can be likened to a “big purchase” in the physical world, such as buying a car or house. Though many cars might be perfectly sound mechanically (and would pass the Texas State emissions and safety testing, for instance), not every car is as appropriate for every driver, and some features are more desirable or less desirable than others. Even more so with a house, many factors dealing with appropriateness—not just being zoned correctly, meeting building codes, and being structurally sound—are taken into account before purchasing (the cost, size, proximity to other locations, and features, to name a few). A further analogy can be drawn to the problem of buying a used car or an older home. The buyer must be very wary of defects that may be invisible to the average buyer, which is why mortgage companies require that the buyer hire an inspector to ensure that the home is really worth the money the bank is about to loan in order to purchase it. The car or home in question then should not only be appropriate, it should be fundamentally sound.

⁶³ [<http://www.fcla.edu/digitalArchive/daInfo.htm>]

⁶⁴ [<http://www.portico.org/digital-preservation/>]

Similarly, many factors go into the selection of an archival home for digital data, even a “certified” one, and like the choice of a car or house, it is one the “purchaser” and future generations of users must live with a long time.

The depositor of digital data for safekeeping in an archive for generations to come thus bears significant responsibility to choose an archive wisely. In some cases this data is the depositor’s life’s work and legacy; in others, it may represent some of the precious and few remaining resources that document a language that is dwindling in numbers of speakers or is already extinct, and thus holds important insights into a people’s cultural heritage and any hope of language revitalization. Furthermore, those who are most intimately acquainted with a particular linguistic community and other aspects of field conditions, and who are in a position to bridge gaps between those communities and the world, should be equipped and empowered to make good choices regarding the treatment of linguistic data from those communities.

Thus a tool is needed to guide a linguist-depositor to probe the key issues of archival trustworthiness. The next chapter explains the development of such a tool, the TAPS (Target, Access, Preservation, and Sustainability) Checklist for Responsible Archiving of Digital Language Resources, which has been designed for linguists, as potential depositors of language data, to assess the archival practices of digital language archives. The TAPS Checklist may be used to establish the trustworthiness of the archive in lieu of certification, and helps the linguist-depositor think through the appropriateness of the archive, which remains a decision each depositor must make individually.

Chapter 4: Development and Use of the TAPS Checklist

The tool introduced in this chapter, the TAPS (Target, Access, Preservation, and Sustainability) Checklist for Responsible Archiving of Digital Language Resources, is designed to assist linguists in evaluating digital archives. It is an application of the guidelines for digital archives presented in chapter 3, and is tailored to the interests of linguists and language communities. Section 4.1 describes the methodology for developing the checklist, sections 4.2 through 4.5 discuss each section of the TAPS Checklist in detail and serves as a “user’s manual,” and section 4.6 addresses the limitations of TAPS.

4.1 Methodology

The procedures used for developing and testing the TAPS Checklist are described in Section 4.1.1. The process owes much to the generosity of the individuals and archives involved. Section 4.1.2 describes the uses and scoring of the TAPS Checklist.

4.1.1 The Development of the TAPS Checklist

The TAPS Checklist was formulated by the author of this thesis through a comparison of components common to trustworthy archives as enumerated in three different tools listed in chapter 3: *TRAC: Criteria and Checklist* (OCLC and CRL 2007), *Catalogue of Criteria* (NESTOR 2006), and the *Data Seal of Approval*, version 1-3 (DANS 2008). I went through the items in each of these tools and grouped them in table

format. From that list, four major categories were identified as being most pertinent for choosing a trustworthy and appropriate digital archive from the perspective of a linguist as potential depositor: Target, Access, Preservation, and Sustainability. These categories form the acronym “TAPS.” Within this framework, the original list was pared down to include the most essential archival functions that could be readily understood and investigated by a non-expert in digital archiving. These were formulated into four questions for each of the major topic areas to create the sixteen items of the TAPS Checklist.

In all, the TAPS Checklist went through fifteen versions during its development. The help of many individuals was invaluable. Throughout the process, drafts were submitted to my thesis advisor, Gary Simons, who was instrumental in guiding and honing the finished product. In the initial stages of development, the Checklist was reviewed by Wayne Dye and William Reiman, individuals who have prepared digital language documentation for archiving. The Checklist was revised based on their input to reflect the particular interests of linguists and language communities, and their understanding of important issues in digital archiving. An interview with Joan Spanne, a specialist in digital archiving, suggested rearrangement of several items and expanding others to include examples.

I conducted the first site visit in September 2009 at the SIL Language and Culture Archive in Dallas, Texas with the archive director, Jeremy Nordmoe, and archivist, Vurnell Cobbey. This visit identified items that benefitted from re-wording, and highlighted items which required personal input from the linguist. Spanne additionally

verified additions and corrections to this initial evaluation using TAPS. In the next site visit, Heidi Johnson at the Archive of Indigenous Languages of Latin America (AILLA) provided feedback and valuable insights regarding issues concerning access resulting in significant revisions to the Access section of the Checklist. These and other site visits I conducted—with Lydia Motyka of the Florida Digital Archive (FDA), which serves the public university libraries in Florida, and Mary S. Linn of the Division of Native American Languages (NAL) archive within the Sam Noble Oklahoma Museum of Natural History at University of Oklahoma—were crucial to determining metrics for TAPS. Dye also consulted as a linguist in using the Checklist to evaluate the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) at which he deposited digital language data. Bob Conrad generously tested the TAPS Checklist at two special collections in university libraries in which he deposited language materials. I conducted the remaining evaluations using TAPS through the Internet. Information on PARADISEC (in addition to that supplied by Dye), Kaipuleohone, and ELAR were gathered online at the archives' websites, and then real-time interviews were conducted. A Skype interview with Nick Thieberger, a key implementer for both the PARADISEC and Kaipuleohone archives, was instructive in the extent of the state of the art possible given limited resources. Finally, a Skype interview with David Nathan, the director of ELAR, underlined the need for archives to serve language communities well. Nathan additionally shared articles, published and in press, that were illustrative of problems and solutions in digital language archiving.

4.1.2 Uses of the TAPS Checklist

The TAPS Checklist is a consensus of archival best practices with a focus on what is practical for a linguist to investigate. In formulating the Checklist, efforts were made to identify and include key activities common to all trustworthy digital archives and issues important to linguistic data and research. These items are intended to be researched first in information published by the archive (*e.g.*, on its website). Gaps are then filled in by correspondence, phone call, or an onsite visit.

Using TAPS, a linguist should be able to discern an archive that is trustworthy from one that is not and eventually find the best archival home for his or her work. This is done by comparing archives to each other across the four categories of Target, Access, Preservation, and Sustainability, so as to determine the relative quality of each archive. TAPS may also be adapted as archives may also be compared item by item, or by a subset of items in the checklist.

Each of the four questions in the categories was rated on a three-point scale with the following metrics:

- Yes = The archive appears to follow best practices. The strongest indicator of this is that the staff is following a written policy or procedure that conforms to best practices. (3 points)
- ? = The archive is in the planning stages of implementing best practices; or the archive partially follows best practices; or the archive is assuming that another entity to which they outsource follows best

practice in implementing the functions indicated, but cannot document it. (2 points)

No = The item is not in the scope of the archive; or the degree to which the archive follows best practice is unclear. (1 point)

No zero value was given in the metric. The absolute differences between the scores and the statistical significance of the scores remain the same whether the point scale begins at 1 or 0.

Since it does not require the specialized knowledge of a formal auditor, the purpose of the TAPS Checklist is not to establish conclusively that the archive is trustworthy (or not), but to establish whether the depositor feels that the archive can be trusted with respect to all the designated communities involved. If the linguist comes away from investigating an archive with significant doubts about its trustworthiness, then we can conclude that significant problems likely are present since they were discoverable by a non-specialist. In the case where the linguist comes away feeling satisfied after investigating an archive, there is always the possibility that there are detailed technical problems that cannot be uncovered by a non-specialist. Only a full formal audit could ultimately guard against that possibility, but TAPS will at minimum be able to help linguists think through issues concerning the digital archiving of their data and steer clear of the clearly untrustworthy archives. In addition to its use by linguists, it is hoped that TAPS will be useful to the archives themselves in identifying significant shortcomings and will contribute to improvement in their trustworthiness. All the archives that I interviewed received the resulting written evaluations (see appendix B) and were given

an opportunity to correct errors and omissions. Many of the individuals that I interviewed indicated that the TAPS Checklist was helpful for them to identify areas that their archives could improve upon.

Each of the sixteen items in the TAPS Checklist was worded to stand alone as much as possible without extensive explanatory notes. However, the following sections “unpack” the contents of the Checklist: the four sections of TAPS and the 16 individual questions are outlined and discussed below in sections 4.2 through 4.5. At the end of explanations for many of the main questions, more detailed questions are included to aid the linguist in probing for the answer to the main question.

4.2 Target

Target refers to the “fit” of the archive with regard to the data to be deposited and the needs of the identified designated communities. This section of the TAPS Checklist, in items 1 through 4 (section 4.2.1 through 4.2.4), addresses the archive’s commitment to maintaining “digital objects” for communities identified by the linguist. The specific questions deal with mission statement, submission criteria, designated communities, and an ongoing relationship to the language community.

4.2.1 Item 1: Mission Statement

Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?

A key characteristic of a trustworthy archive is an explicit mission statement that makes clear its intentions to preserve digital information for the long-term (DANS 2010:11, OCLC and CRL 2007:10). Digitization projects and websites may store and disseminate digital materials in the short-term, but they will not have the infrastructure behind them to guarantee long-term preservation of digital materials and cannot make such claims. This is what differentiates archives from non-archives. A linguist should be wary of any institution that offers to “archive” materials but makes no commitment to long-term preservation in its mission statement.

4.2.2 Item 2: Submission Criteria

Does the material that I want to submit fall within the scope of the archive’s collection policy in terms of content and type (specify: _____)?

Submission criteria pertain to the match between the content and types of the materials you want to deposit and the content and formats of materials that an archive accepts. The linguist should specify the content and type of data he or she wishes to deposit. Since archives have different specializations, determining the scope of the archive’s collection is important to finding a good fit for your materials.

Trustworthy digital archives accept digital objects from the producers based on defined criteria (NESTOR 2006:18, DSA 2009:11). Well-thought out collection policies and guidelines on accepted formats therefore will indicate a high degree of trustworthiness. Support and procedures for digitizing analog and hardcopies of materials should also be well-defined if that is a need for the depositor.

The fit of the archive in terms of the overall collection and kinds of data that it typically handles should be taken into account since a digital object in an “obscure” subject or an “atypical” format (even if it is an “archival” format) may not be as well preserved, supported for access, or maintained in machine readable formats in the long-run as would materials in areas of specialty for the archive. For such reasons, it is ideal for depositors to dialogue with potential archives regarding submission criteria (including file formats) before embarking on a language documentation project.

2. Submission Criteria

In-depth questions: What is the content of material I want to deposit? What formats are they in? Does my content fall within the collection policy of the archive? Do my materials fall within the preferred submission formats of the archive?

4.2.3 Item 3: Designated Communities

Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g., language community, academic community, etc.)?

The OAIS reference model defines a designated community as the group or groups to which the archive aims to make content in the archive accessible (CCSDS 2002:1-10). These are the potential consumers who should be able to understand a particular set of information, and may be composed of more than one user community (CCSDS 2002:1-10). For the purposes of our discussion, I reference the intended user community in the plural as designated communities. Thus, the linguistics community,

the larger scholarly community, the language community, and so forth can all be counted as designated communities.

In order to determine the desired designated community or communities, linguists should first consider which audiences would benefit most from the materials deposited in an archive and who should be able to access the data once it is submitted (the language community, the linguistics community, *etc.*). Linguists should then determine whether those groups are a match to those that the archive is committed to serve. As the archive will be responsible for maintaining services to its designated communities over time, it stands to reason that the changing needs of the desired user communities of the linguist are best served by archives that are already well-positioned to serve those communities.

3. Designated Communities

In-depth questions: What are the desired user communities for the data I want to deposit? Do these fall within the designated communities of the archive? Particularly if the language community is an important user of the deposited materials, does the archive cater to that user community?

4.2.4 Item 4: Ongoing Relationship

Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)

Since it is expected that deposited materials will outlive the depositor who initially acts as a liaison for a language community, it is often desirable for the archive to

interact directly with the language community. If this seems important to the language community as a condition for depositing the materials, then the linguist must ensure that the archive is ready to accept this responsibility. Additionally, certain archives may be able to set up revenue sharing structures for the language community if there are any revenues to be generated from the use of deposited materials. While this may not seem a likely scenario with purely linguistic data, it is not hard to imagine with the licensing of recordings of music and cultural performances.

Several language archives routinely interact with and champion the causes of the language community. The Native American Languages (NAL) archive at the University of Oklahoma is actively involved in language revitalization efforts, grant writing for language communities, and is committed to prosecute those who profit improperly from archived materials. A case at PARADISEC illustrates a solution regarding potential revenues generated by archived materials; the case concerns Dye's work among the Bahinemo of Papua New Guinea. Though the language documentation gathered from the Bahinemo was unlikely to generate revenue, and Dye stressed to the village in which he worked that neither he nor the archive were going to make any money from having their materials archived, the possibility that authors or film producers or some other archive user might materially benefit from the archived recordings remained a point of concern for the community. Thus PARADISEC provided a way to share any potential income as part of its firm commitment to be fair to descendants of language group members. The archiving agreement specifies a local agency to which any royalties resulting from use of

archived materials should be paid. Additionally, Dye liberally shared with the community from funds he had been awarded to do the language documentation work.

4. Ongoing Relationship

In-depth questions: What types of interaction do I anticipate needing to take place between the language community and the archive? Will the archive support these? Is potential revenue sharing an issue for my deposit? If so, will the archive offer this service?

4.3 Access

Access refers to the accessibility and usage of the data and corresponding metadata once materials are deposited. The TAPS Checklist addresses Access in questions five through eight concerning discoverability, fixed identifiers, reach, and access restrictions.

4.3.1 Item 5: Discoverability

Are the metadata for materials deposited at the archive searchable online? I.e. posted on the web or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g., Google, Yahoo!, Bing, etc.)?

Once digital material is submitted and archived, it is advantageous for the materials to be discoverable over the Internet to reach the widest possible audience. Generally speaking, an archive's holdings do not have to be available as complete digital files online, but the catalog of what they contain, that is, the descriptive metadata, should

be online so that it is searchable and discoverable with ordinary Internet search engines; see section 2.2.4. If reaching the widest possible audience is deemed undesirable, the essential question is, “Does the archive provide adequate resource discovery opportunities to the designated community?”

Before materials can be cataloged and found on the Internet, sufficient descriptive metadata needs to be provided by the depositor. An archive should have guidelines and standards for this metadata. The quality of the metadata to be searched is another indicator of trustworthiness of the archive.

One way for an archive to ensure that search results for their holdings show up on the “first page” in a routine search on the Internet is to be a member of the Open Language Archives Community (OLAC), which aggregates metadata from all the participating archives into a combined catalog at the OLAC website. This aids the discoverability of materials deposited in both prominent and less prominent archives.

5. Discoverability

In-depth questions: Does the archive have the necessary guidelines and standards to help me, as the depositor, provide quality descriptive metadata? Is the metadata posted on the Internet? If so, are these records easily found on the Internet? Is the metadata aggregated through a service such as OLAC? Does my desired designated community have adequate resource discovery opportunities through the archive’s approach to descriptive metadata?

4.3.2 Item 6: Fixed Identifiers

Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?

The archive should have a system to assign externally visible and standardized persistent identifiers to materials in order to enable reliable referencing in academic citations and to ensure that they can be found in the distant future (NESTOR 2006:23, CCSDS 2002:2-6, Bird and Simons 2003:65-66). Each persistent identifier should be assigned permanently and remain unique within the system. The persistent identifier should not be based on any changeable attribute of the material being referenced. For instance, in the Digital Object Identifier (DOI)¹ system used in commercial publishing, a “dumb number” that is not based on any pattern avoids misleading assumptions or loss of meaning over time or across linguistic or cultural barriers. Another example is the Handle System² that is used in open-source repository systems like DSpace to assign persistent identifiers.

4.3.3 Item 7: Reach

Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?

The archive should communicate in advance and in a transparent manner its conditions of access and any costs that may arise. The linguist in turn needs to determine what constitutes reasonable access for the designated community and how that matches

¹ [<http://doi.org/>]

² [<http://www.handle.net/>]

up with the archive's policies on how materials will be accessed. Accessing the materials may include, but is not limited to, any or all of the following (NESTOR 2006:9):

- Accessing the materials at a given access point (*e.g.*, the access station pictured in figure 4.1)
- Creating or supplying an analog copy (*e.g.*, a print-out or a print-on-demand service)
- Creating or supplying a digital copy (*e.g.*, e-mail delivery or download by the user)
- Creating interfaces to permit online exploration or query of the materials.

7. Reach

In-depth questions: Will members of the designated communities be expected to have access to the Internet? Will members of the designated communities need to maintain an e-mail address? Will the metadata be available in English only, or will it be available in another language that is more accessible to the designated community? Will the archive charge fees for copies of data on media that are usable by members of the designated communities? Even if the fees are "at cost," will they be affordable for those communities?



Figure 4.1: An Access Station at NAL. The Division of Native American Languages (NAL) at the Sam Noble Oklahoma Museum of Natural History has designated spaces where persons may access language materials.

4.3.4 Item 8: Access and Use Restrictions

Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?

The policies and procedures of the archive should articulate usage rights and conditions and their enforcement. The policies should make clear what options are available for open access and for restricted access, and then the linguist needs to ensure that one of those options matches the current and anticipated future needs. The issues surrounding ownership, copyrights, and conditions on the use of deposited materials should also be weighed carefully. The needs of the designated community should be evaluated for compatibility with what the archive will and will not do once the materials are part of the archive's collection.

8. Access and Use Restrictions

In-depth questions: How does the archive deal with copyright? Does the archive require transfer of ownership? Does the archive allow materials to be deposited with restrictions on access? If so, what restrictions are possible and how are requests for access handled? Does the archive allow materials that are closed to access to be deposited? How long will periods of closed access last? What are the archive's conditions of use policies?

When materials are deposited in an archive, a contract that governs the use of the material should be signed by the institution and the depositor. This contract, or deposit agreement, typically takes into account issues of ownership and copyright, access restrictions, and conditions on the use of deposited materials. These issues are described in greater depth in sections 4.3.4.1 through 4.3.4.3.

4.3.4.1 Copyright and Transfer of Ownership

Depositing materials at an archive that makes its holdings freely available on the Internet is essentially publishing them through the archive (*e.g.*, PARADISEC, AILLA). Because of the inherent copyright in performances, one cannot actually do this without the informed consent of the performers to allow this kind of distribution. Once materials are deposited, unless other arrangements are made, the archive decides over questions of access to and use of the materials. These rights are not necessarily exclusive, however

(*e.g.*, AILLA³). Depositing in an archive at minimum involves grant of license to reproduce (a necessary condition for many preservation functions) and distribute. At the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) archive, copyright and responsibility for archived material may be retained by the depositor or be completely conferred upon the archive (see section 2.2.2).⁴ Ultimately, as AIATSIS and PARADISEC make clear in their deposit forms,^{5, 6} the archive is responsible to adhere to procedures that “safeguard the interests and sensitivities of relevant indigenous people”.⁷

It is debatable whether the traditional knowledge and stories of a people can be copyrighted, or whether such works should be treated as facts or ideas, which cannot be copyrighted.⁸ In either case, it is a given person’s performance of a work that is protected by copyright. However, under international and U.S. copyright laws, no work is protected in perpetuity though a language community may feel differently about their ownership of their traditional cultural heritage (SEM 2001:16-17). In AIATSIS and PARADISEC’s deposit forms, the depositor is required to list “relevant individual(s) and their community(ies) and/or other funding organizations” that may have rights to the material being deposited. Furthermore, in the “explanatory notes” attached to these two Australian archives’ deposit forms, similar statements carefully note that “the term ownership refers to ownership of the physical copy of the material being lodged with [the

³ [<http://www.ailla.utexas.org/site/ipr.html>]

⁴ [http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf],
 [http://www.aiatsis.gov.au/collections/docs/AVA_transfer.pdf]

⁵ The AIATSIS deposit form is available at: [http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf]

⁶ The PARADISEC deposit form is available at: [<http://www.paradisec.org.au/PDSCdeposit.pdf>]

⁷ [http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf],
 [<http://www.paradisec.org.au/PDSCdeposit.pdf>]

⁸ [<http://www.ailla.utexas.org/site/ipr.html>]

archive]. It is not a wider claim to the intellectual property or ownership of any traditional knowledge.”⁹ PARADISEC’s notes go on to say:

If the material was written, photographed, drawn, recorded or filmed by you, then you are the creator and owner of the physical copy of the material, or if you have collected, found or inherited the material you are the owner of the physical copy of the material and therefore you or your delegate are in a legal position to enter this agreement.

Much has been written about copyright issues, and each country has its own copyright laws for which the linguist is responsible. See sections 2.2.3 through 2.2.5 for a more in-depth discussion. More information may be found at AILLA’s webpage on Intellectual Property Rights¹⁰ and the included links, or the Society of Ethnomusicology’s *Manual for Documentation and Fieldwork & Preservation, Chapter 2, Ethical and Legal Considerations* (SEM 2000) for details concerning copyright law.

4.3.4.2 Access Restrictions

Varying levels of access to viewing and listening to materials is desirable in some cases to preserve privacy and confidentiality. Though AILLA encourages all depositors to make their resources freely available, its Graded Access System is an example of providing flexibility and leaving it up to depositors to specify restrictions on use.¹¹

AILLA provides the choice of four access levels to depositors who can choose to assign

⁹ [<http://www.paradisec.org.au/PDSCdeposit.pdf>],
[http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf]

¹⁰ [<http://www.ailla.utexas.org/site/ipr.html>]

¹¹ [<http://www.ailla.utexas.org/site/gas.html>]

any level to their entire collection or to any part of their collection: Level 1, access is open; Level 2, access is protected by password; Level 3, access is protected by a time limit, and Level 4, the depositor (or someone else) controls access to the resource.¹²

With AIATSI and PARADISEC, inquiry is made in their deposit forms regarding the depositor's "understanding of the attitude of the [language community towards] ... this material being made accessible to other people," and "whether any special conditions should be considered when handling this material, for example, ceremonial or gender restricted material, sensitive genealogical material, photographs or recordings of deceased people" so that the archive may act accordingly. Both archives further offer depositor specified conditions on access, but AIATSI reserves the right to refuse material which has unreasonable conditions, and PARADISEC will not hold material on permanent closed access.

NAL has a restricted-use policy which allows open access to most materials (*i.e.*, access is not based on tribal membership), but will permanently deny public access to portions of materials containing "injurious gossip." Materials pertaining to formal societies (*e.g.*, Kiowa Black Leggings) and chief societies (*i.e.*, men who are born into a chief family line) can also be restricted, though such sacred or sensitive content is officially "on loan" to the museum according to NAL's *Restrictions on Use Policy*.¹³ AILLA recommends that materials of extremely sensitive nature not be deposited.¹⁴

¹² [http://www.ailla.utexas.org/site/forms/ailla_depositor_packet.pdf]

¹³ [<http://www.snomnh.ou.edu/collections-research/cr-sub/nal/restrictions%20policy.pdf>]

¹⁴ [http://www.ailla.utexas.org/site/five_con.html]

4.3.4.3 Conditions of Use

Conditions of use have to do with what users of archived materials are allowed to do with them. NAL's *Restrictions on Use Policy* prohibits commercial or for-profit use of collected materials, and states the archive's commitment to prosecute for the improper use of deposited materials. AIATSIS provides its depositors open-ended choices, which include a choice for the depositor to be contacted each time material is copied. PARADISEC and AILLA outline responsibilities of and limitations on the user in their "Conditions of Access" and "Conditions for Use of Archive Resources" agreements.¹⁵ Users are not authorized to access the archives until they have read and signed these agreements.

4.4 Preservation

Preservation refers to the overall system and technical structures of the archive that ensure materials will be managed in ways that make them available and usable, with their authenticity and integrity intact, far into the future. The TAPS Checklist addresses Preservation in questions 9 through 12 concerning evidence of long-term planning, preservation strategies, integrity, and authenticity.

4.4.1 Item 9: Evidence of Long-Term Planning

Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g., the archive has written standards for

¹⁵ [<http://www.paradisec.org.au/PDSCaccess.pdf>], [http://www.ailla.utexas.org/site/use_conditions.html]

implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?

At the heart of any archive is the plan and implementation of a defined archival process that is sustainable over time (NESTOR 2006:20). At the highest level of planning for the long-term preservation of digital materials, the archive should plan to take into account legal and social changes, the needs and expectations of the designated communities, and technological developments relevant to the preservation and appropriate use of the deposited materials (NESTOR 2006:14). The day-to-day operations of the archive include the definition of digital objects “packaged” in a defined structure for long-term preservation (*i.e.*, content data in a suitable archival format, information needed to interpret the content data, and the relevant metadata). For digital language archives, the structure of complex objects, such as multimedia materials, needs to be adequately described so that they can be reconstructed and used as intended (NESTOR 2006:25). The procedures for creating and maintaining these “archival information packages” should be documented with written policies and procedures (NESTOR 2006:20, CCSDS 2002:1-7). The responsibility for each process may be assigned to particular individuals (NESTOR 2006:18) or to outsourced entities (NESTOR 2006:13). Trustworthy archives have a demonstrable commitment to the archival storage of digital materials to defined specifications, and will regularly review the appropriateness of those specifications over time (DANS 2010:11, NESTOR 2006:12).

9. Evidence of Long-term Planning

In-depth questions: Does the archive have written procedures for the tasks involved in implementing their defined archival process? Do these procedures specify deadlines for completing upcoming tasks as they pertain to the creation and maintenance of archival information packages? Is responsibility for each process clearly assigned to specific individuals or outsourced entities? Is the archive explicitly monitoring substantial changes, whether technical, organizational, or community-based? Will the archive change its procedures as needed?

4.4.2 Item 10: Preservation Strategies

Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?

A trustworthy archive has an overall strategy for preserving digital materials within their collection. The monitoring of technical developments noted above in section 4.4.1 includes the development and standardization of new file formats and new storage techniques and the phasing out of existing technologies as needed. The archive should keep pace with ongoing technical developments (such as changes to data carriers, data formats, and user demands), but even in the absence of such changes, the archive must have a plan to deal with the deterioration of the media on which the data is recorded, sometimes called “bit rot” or “digital decay,” in which the bits of data themselves are subject to corruption over time.

In order to carry out such responsibilities, the digital repository must identify which characteristics of the digital objects are significant for information preservation (NESTOR 2006:19). The process should be defined to determine for each item archived whether a maintenance measure must be undertaken to ensure long-term preservation, and when needed, the corresponding measure should be carried out and any changes to the digital object documented (see section 4.4.4 below on Authenticity) (NESTOR 2006:21, 25). Two long-term preservation measures are:

- Refreshing: the transfer of data from one medium to the same type of medium without any alteration to the data at the bit-level. Refreshing guards against the deterioration of physical media, but not obsolescence.
- Data migration: transferring files to a newer format (for example in 2001, from JPEG to JPEG 2000), when software or hardware required to read the data is no longer available. Data migration can be a time-consuming process, involving alterations to the data, and often sacrificing an element of the ‘look and feel’ of the original material.

Even when the migration strategy involves changing format, bit-level preservation of originals seems to be emerging as a best practice in the digital archiving community (Caplan 2004:6). Keeping the original file as it was submitted aids in demonstrating the integrity (see section 4.4.3 below) and authenticity (see section 4.4.4 below) of the digital objects. Additionally, there is always the possibility that a better migration algorithm may arise; if the original file are always retained, a “do-over” of the migration is always possible (Caplan 2004:6). No archive that I interviewed had conducted a full-scale

migration as the need had not arisen. The FDA had conducted proof-of-concept migrations, however.

10. Preservation Strategies

In-depth questions: On what medium will the archive store the materials I submit? What is their schedule for refreshing data on that medium? What will the archive do with the data as the medium approaches obsolescence? What will the archive do if the format in which the data are stored becomes obsolete? When was the last time the archive completed a migration from an obsolete format to a newer format?

4.4.3 Item 11: Integrity

Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g., a checksum, or digital signature, etc.)?

The archive should ensure the integrity of the digital materials and metadata throughout their lifecycle within the archive (as they are processed, stored, copied, and used). Here, integrity refers to (1) the completeness of the digital object, including metadata, and (2) the exclusion of unintended modifications as defined in the preservation rules. Integrity is measured in terms of the characteristics of the particular digital material being preserved (NESTOR 2006:41). Inappropriate modifications may be caused by human error (deliberate or accidental), imperfections in media, or damage to the technical infrastructure. The archive should take both organizational and technical precautions to secure the integrity of objects within their custody; that is, the archive

should operate a data management system that is able to ensure integrity of digital materials (NESTOR 2006:15). Best practice is to use fixity metadata, like checksums and digital signatures, to ensure the integrity of copies. The use of fixity metadata reflects an archive's institutional commitment to the integrity of digital materials, and is also an indicator of the quality of its archival implementation.

A checksum is created using an algorithm that adds all the bytes or words in an arbitrary block of data to create a value that is stored as part of the fixity metadata of the digital object. When data is transmitted or copied, the checksum is recomputed and compared to the checksum value stored in the metadata in order to detect an error. If the checksums match, it is unlikely that there was an error in transmission or copying, though it is possible that some pattern of altered bits in a message can result in an erroneously matching checksum value (Maxino 2006:1). A good checksum algorithm will yield a different result with high probability when the data is accidentally corrupted; thus, if the checksums match, the data is very likely to be free of accidental errors.¹⁶ There are tradeoffs, however, between the computing power used on the checksum calculation, the size of the block of data checked, and the probability of such undetected errors (Maxino 2006:1).

Digital signatures are typically used to verify authenticity, which is the process of determining if a user or entity is who he, she, or it claims to be (OWASP 2002), but they also serve the purpose of simultaneously providing integrity over the signed data. This is a consequence of a necessary property of cryptographic hash algorithms and signature

¹⁶ [<http://en.wikipedia.org/wiki/Checksum>]

algorithms as any change in the input data leads to a large, unpredictable change in the output with very high probability. In other words, if the data has changed, the signature will fail to verify, and the loss of integrity will be obvious. If, on the other hand, the signature verifies, the digital object is likely unaltered (Adams and Lloyd 1999).

4.4.4 Item 12: Authenticity

Does the archive ensure that digital materials contain what they claim to contain (e.g., by verifying that digital objects are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital objects that occur while they are in the care of the archive)?

The archive should ensure the authenticity of digital materials and metadata throughout their lifecycle within the archive (as they are processed, stored, and used). Authentic here means that a digital object actually contains what the metadata claims that it contains. When authenticity cannot be demonstrated for a particular holding, the archive should document this fact in the metadata.



Figure 4.2: Materials Being Checked for Authenticity at NAL. OU graduate student, Amber Neely, listens to Kiowa language materials at NAL, checking the authenticity and noting any discrepancies.

After authenticity is verified in the initial deposit, it can be preserved through permanently associating adequate metadata so that the match between deposited materials and associated metadata can be verified, and using provenance metadata to document the origins and all changes to the materials and metadata (NESTOR 2006:17, 25). In language archives, the depositor may be solely responsible for vouching for the authenticity of deposited materials.

Provenance metadata should contain information about how the digital objects came about, and careful records of the outcome of preservation processes. In cases where material is migrated to new formats, users must understand which versions of a particular digital resource are available for access, and how the resources have been changed as a consequence of preservation (Lavoie and Dempsey 2004).

4.5 Sustainability

Sustainability refers to the demonstrated organizational robustness of the archive, lending long-range viability to the functions that it performs. The TAPS Checklist addresses sustainability in questions 13 through 16 concerning adequate infrastructure, financial sustainability, disaster preparedness, and succession planning.

4.5.1 Item 13: Adequate Infrastructure

Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g., quality media, environmentally-controlled storage, access-controlled storage area)?

Adequate infrastructure addresses two aspects of the archive: the staff and the technical infrastructure.

Staff: The qualifications and training of the staff should be adequate to the defined processes and mission of the archive (NESTOR 2006:12). Staff numbers should be sufficient to fully complete the tasks of the archive (OCLC and CRL 2007:11). Additionally, there should be programs to ensure adequate professional development of staff over the long-term.

Technical Infrastructure: The technical infrastructure of the archive should ensure the continuing maintenance and security of its digital objects. This infrastructure includes good overall computing practices described by international management standards, for example, ISO 27002, formerly ISO 17799 (OCLC and CRL 2007:43). It is

recognized that “without a secure and trusted infrastructure, the functions carried out on [archived materials] cannot be trusted,” and such an archive would be “built on a house of cards” (TRAC 2007:43).¹⁷ Responsible back-up procedures are included in section 4.5.3, Disaster Preparedness, but since it is likely beyond the scope of an informal interview to check the computing practices of a given archive, the TAPS question lists more tangible indicators of the quality of technical infrastructure as possible items to check. “Quality media” refers to the physical media on which data is stored; for example, hard disks are more durable and less prone to failure than CDs or DVDs. “Environmentally-controlled storage” refers to the physical environment in which physical copies of materials are stored, including temperature, humidity and pest controls. And “access-controlled storage area” indicates that digital and physical copies of materials are protected from misuse or theft by virtue of the security in the facilities in which they are kept.

4.5.2 Item 14: Financial Sustainability

Does the archive appear to have secured sources of long-term funding?

The archive should be able to demonstrate its financial sustainability. Though an archive may not be a for-profit business, it should adhere to good business practices and should have a plan for how it will “stay in business.” The business plan comprises a set of documents that lays out the past, present, and future of the repository and its activities, and which takes into account the financial implications related to development and

¹⁷ The requirements for an adequate technical infrastructure as it applies to digital archives are laid out in Section C of the *TRAC Criteria and Checklist*.

normal production activities, and may note factors that would affect operations. It is recommended that the business plan and financial fitness be reviewed at least annually (TRAC 2007:16). The digital repository should be able to demonstrate that the proposed services can be financed, both in the short and long term, whether it is on the basis of guaranteed funding or on the basis of charging for use of its services (NESTOR 2006:11).

4.5.3 Item 15: Disaster Preparedness

Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g., disaster recovery plan, offsite storage of backups)?

The archive should ensure that it has adequate hardware and software support for backup functionality that is sufficient for the services it provides and for its digital holdings. The following can demonstrate the adequacy of the processes, hardware, and software of an archive's backup systems: documentation of what is being backed up and how often; audit log of backups; validation of completed backups; "firedrills"—testing of backups; support contracts for hardware and software for backup mechanisms (TRAC 2007:44-45). Another important requirement is that backups be stored in a different physical location than the archive itself in order to mitigate the risk of fire, flood, tornado, and other disasters that could befall the building that houses an archive. The existence of (and long distances between) "mirror" sites also lend confidence to an archive's disaster preparedness.

In conjunction with responsible backup practices, the archive should have a written plan regarding what happens in specific types of disaster (fire, flood, earthquake, explosion, system compromise, *etc.*), and who has responsibility for which actions (TRAC 2007:49). Disaster with respect to digital archives is defined as “any event that threatens or interrupts the operation of the repository and that, without corrective action, threatens the long-term preservation of its holdings” (TRAC 2007:81). The level of detail in a disaster plan, and the specific risks addressed, are determined by the location and expected services of the archive. The disaster plan should also deal with specific consequences arising from unspecified situations, such as lack of access to a building or prolonged network outages. The archive should keep written disaster preparedness and recovery plans, including at least one off-site backup of all preserved information together with an off-site copy of the recovery plans (TRAC 2007:49).

15. Disaster Preparedness

In-depth questions: Is there a formally documented procedure for regular backups? Is compliance with the procedures audited? Where are the backups of archived material kept? If the building or location housing the archive is destroyed, how will materials be recovered? Is there a written disaster recovery plan, located off-site, that makes explicit what to do if a disaster occurs?

4.5.4 Item 16: Succession Plan

Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?

The archive should ensure the continuation of the preservation tasks if the archive itself ceases to exist. In order to avoid irreparable loss, consideration needs to be given to this responsibility while the archive and its holdings are viable, not when a crisis is occurring (TRAC 2007:10). To this end, the archive ideally should have a formal succession plan that includes trusted inheritors (TRAC 2007:10). Succession plans should describe processes that will enable the preservation work to continue within an alternative organizational framework, thereby ensuring that the requirements can continue to be completed; where this is not possible, any restrictions should be documented (NESTOR 2006:12).

If a formal succession plan is not in place, the archive should at minimum be able to identify the basis of a plan, for example, partners, commitment statements, likely heirs, and so forth. Succession plans do not need to transfer the entire collection to a single organization if this is not feasible. Multiple inheritors are acceptable as long as the data remains accessible (TRAC 2007:10).

It should be noted that, organizationally, the materials in an archive can be at risk whether the archive is run by a commercial organization or a government entity (*e.g.*, national library or archives):

At government-managed repositories and archives, a change in government that significantly alters the funding, mission, collecting scope, or staffing of the institution may put the data at risk. These risks are similar to those faced by commercial and research based repositories and

should minimally be addressed by succession plans for significant collections within the greater repository (TRAC 2007:10).

4.6 Limitations of the TAPS Checklist

This checklist is not a comprehensive tool and is not intended to be used to perform an outside audit of a given archive. Instead, a high degree of trust is placed on the self-reporting of the archives on their practices with regard to the items pinpointed in the checklist. The criteria contained in the TAPS Checklist are not exhaustive at sixteen items,¹⁸ but they are essential to the trustworthiness of digital language archives and concerns of linguists and language communities. Table 4.1 shows how the sixteen items of the TAPS Checklist align with “ten basic characteristics of digital preservation repositories”¹⁹ identified by four preservation organizations that convened in 2007 in Chicago under the auspices of the Center for Research Libraries. Note that some TAPS items are listed more than once in the table. Item 4, ongoing relationship, is the only item that does not appear as it pertains to the rights of language communities, which are not addressed by general digital archiving standards. The preservation organizations were: the Digital Curation Center (U.K.) and Digital Preservation Europe which created DRAMBORA, NESTOR (Germany) which created the *Catalogue of Criteria*, and the CRL (international consortium based in North America) which created TRAC.

¹⁸ TRAC, the most extensive of the auditing tools described in chapter 3 with 114 checklist items, refers to itself as a “starting point” and not an all-inclusive checklist.

¹⁹ [<http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re>]

Table 4.1: Distribution of TAPS Checklist Items among Ten Basic Characteristics of Digital Preservation Repositories
(material from [<http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re>] and appendix A)

Ten Basic Characteristics of Digital Preservation Repositories	TAPS Checklist
1. The repository commits to continuing maintenance of digital objects for identified community/communities.	1. Mission Statement 3. Designated Communities
2. Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfill its commitment.	14. Financial Sustainability 13. Adequate Infrastructure 9. Evidence of Long-Term Planning 10. Preservation Strategies
3. Acquires and maintains requisite contractual and legal rights and fulfills responsibilities.	8. Access and Use Restrictions
4. Has an effective and efficient policy framework.	9. Evidence of Long-Term Planning 15. Disaster Preparedness 16. Succession Plan
5. Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.	2. Submission Criteria 9. Evidence of Long-Term Planning
6. Maintains/ensures the (a) integrity, (b) authenticity and (c) usability of digital objects it holds over time.	11. Integrity 12. Authenticity 10. Preservation Strategies
7. Creates and maintains requisite metadata about (a) actions taken on digital objects during preservation as well as about (b) the relevant production, access support, and usage process contexts before preservation.	9. Evidence of Long-Term Planning 11. Integrity
8. Fulfills requisite dissemination requirements.	5. Discoverability 6. Fixed Identifiers 7. Reach 8. Access Restrictions
9. Has a strategic program for preservation planning and action.	9. Evidence of Long-Term Planning 10. Preservation Strategies
10. Has technical infrastructure adequate to continuing maintenance and security of its digital objects.	13. Adequate Infrastructure

Chapter 5: Results and Conclusions

In the preceding chapters, I reviewed the recommended best practices concerning digital archives and applied the findings in the context of language resource archiving to develop a new tool, the TAPS Checklist, to aid linguists and other depositors in choosing an archival home for their language materials. This chapter reports the results of using the TAPS Checklist. The repositories evaluated include six digital language archives, two special collections within university libraries, and one high-quality “dark” archive (see section 2.1.6) that did not specialize in digital language materials. Section 5.1 introduces the nine archives evaluated in this study. Section 5.2 contains a table summarizing the overall and average scores for each archive in the four sections of TAPS (Target, Access, Preservation, and Sustainability), and tables showing the total and average scores as they pertain to each section and item of TAPS. Section 5.3 lists the strengths and weaknesses evident among digital language archives, and a brief statistical analysis. Section 5.4 observes patterns concerning archives that are global or regional in scope. Section 5.5 summarizes the findings of the preceding sections. Sections 5.6 and 5.7 close with recommendations for further research and concluding remarks.

5.1 Archives Evaluated with TAPS

The nine participating archives in this study are listed below in alphabetical order. The means of evaluation with the TAPS Checklist is noted in parentheses:

- Archive of Indigenous Languages of Latin America (AILLA): a joint project of the Departments of Anthropology and Linguistics, and the Digital Library Services Division of the General Libraries at the University of Texas at Austin. (Site visit in Austin, Texas, U.S.)
- Endangered Languages Archive (ELAR): the digital archive of the Hans Rausing Endangered Languages Project (HRELP) at the School of Oriental and African Studies (SOAS) within the University of London, England. The work of Endangered Languages Documentation Programme (ELDP) grantees and others are deposited here. (Review of materials at website and Skype interview with archive representative)
- Florida Digital Archives (FDA): operated by the Florida Center for Library Automation (FCLA), which is a dark archive serving the libraries of the public universities of Florida. (Site visit in Gainesville, Florida, U.S.)
- Kaipuleohone: the University of Hawai'i at Mānoa's digital ethnographic archive within the Department of Linguistics, specializing in materials related to small and endangered languages. (Review of materials at website and phone interview with archive representative)
- The Division of Native American Languages (NAL) at SNOMNH, OU (Sam Noble Oklahoma Museum of Natural History, University of Oklahoma): a collection and resource center for researchers, educators, and language advocates of Native American languages. (Site visit in Norman, Oklahoma, U.S.)

- Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC): an archive operated by a consortium of three universities—University of Sydney, University of Melbourne, and Australian National University (Canberra). (Review of materials at website, e-mail communication with linguist-depositor, Wayne Dye, and phone interview with archive representative)
- SIL Language and Culture Archives: the archives of SIL International, a non-profit, faith-based non-governmental organization with 75 years experience in serving the world's ethnolinguistic minority language groups. (Site visit in Dallas, Texas, U.S.)
- University of California, San Diego (UCSD) Melanesian Archive: a joint project between the UCSD Department of Anthropology and the UCSD Libraries (Mandeville Special Collections), includes a wide variety of non-circulating physical materials, including sound recordings and field notes, with an emphasis on primary source materials that support selected UCSD instructional and research programs. An area of particular strength is Melanesian anthropology.²⁰ Some materials are digitized. (Separate review by linguist-depositor, Bob Conrad)
- University of Virginia (UVA) Albert and Shirley Small Special Collections Library: administers vast holdings of a wide variety of

²⁰ [<http://libraries.ucsd.edu/locations/mscl/overview/index.html>]

physical materials, including audio recordings;²¹ some materials have been digitized.²² (Separate review by linguist-depositor, Bob Conrad)

Table 5.1 shows the classification of the eight participating digital language archives according to the dimensions of scope, submitter restrictions, and focus as developed in chapter 2. No tribal archives were interviewed, but all other dimensions of scope (global, regional), and submitter restrictions (few restrictions, restricted with exceptions, restricted) are represented. All archives of regional scope aligned with few submitter restrictions, while global archives either were restricted or restricted with exceptions with respect to submitter restrictions. With the exception of the SIL Language and Culture archives, all the archives represented had university affiliations, though in wide-ranging capacities.

²¹ [http://www2.lib.virginia.edu/small/about/about_small_lib.html]

²² [<http://www2.lib.virginia.edu/small/using/documents/DigitalCameraPolicy.html>]

Table 5.1: Typology of Participating Language Archives Showing Focus

Scope	Submitter Restrictions	Name of Archive	Focus	
			Content	Type
Global	Restricted with exceptions	UVA Albert and Shirley Small Special Collections Library	Manuscripts, rare books, maps, broadsides, photographs and small prints, reels of microfilm, audio recordings, motion picture films, and ephemera (UVA 2002)	Analog and digital
		ELAR	Materials that relate to endangered languages	Digital
		SIL Language and Culture Archives	Books, journal articles, dissertations, and academic papers about languages and cultures; references for materials written in minority languages	Digital and analog
	Restricted	Kaipuleohone	Mostly digitized text; also audio and video recordings, photographs, notes, dictionaries, transcriptions	Digital
Regional	Few restrictions	AILLA	Audio, video, image, and text materials; all kinds of materials in and about the indigenous languages of Latin America	Digital
		NAL	Audio and video recordings, manuscripts, books, and teaching curriculum, lesson plans and materials, concentrating on Native languages of Oklahoma and incorporating Native languages of North America and endangered languages world-wide.	Digital and analog
		PARADISEC	Mainly digitized audio tapes; also textual materials, dictionaries, grammars, articles and other digital objects	Digital
		UCSD Melanesian Archive (Mandeville Special Collections Library)	Unpublished documentation pertaining to the peoples, cultures, languages, and history of Melanesia (UCSD 2009)	Analog and digital

5.2 TAPS Checklist Scores

Table 5.2 below summarizes the scores for each archive in each category of TAPS (Target, Access, Preservation, and Sustainability), the overall total scores, and the average categorical and overall scores of the eight language archives. FDA scores are listed but were not used in the calculation of the average scores since the FDA is not a digital language archive. Rather, it is a high-quality dark archive that serves as a case of known and attainable best practice with regard to the Preservation and Sustainability components of TAPS.

Table 5.2: Overall and Average Scores for Target, Access, Preservation, and Sustainability

Name of Archive	Target (out of 12)	Access (out of 12)	Preservation (out of 12)	Sustainability (out of 12)	TOTAL SCORE (out of 48)
AILLA	10	10	8	10	38
ELAR	11	10	9	9	39
FDA ²³	10	9	12	12	43
Kaipuleohone	9	10	7	6	32
NAL	11	10	10	10	41
PARADISEC	12	11	10	7	40
SIL Language and Culture Archives	10	11	8	7	35
UCSD Melanesian Archive (Mandeville Special Collections Library)	11	10	9	10	40
UVA Small Special Collections Library	11	10	9	11	41
AVERAGE SCORES	10.50	10.25	8.75	8.75	38.25

²³ The FDA is a dark archive that does not specialize in archiving language materials; FDA scores were not used to calculate the average scores.

Tables 5.3 through 5.6 show the nine archives' detailed results for the categories of Target, Access, Preservation, and Sustainability, respectively. Each table is further broken out into the sixteen total items (four items per category) of the TAPS Checklist, and shows the ratings for each item, archive totals for the category, average score for each item, and overall average score for the category. The archives are listed in alphabetical order in each table. The FDA is a dark archive that does not specialize in archiving language materials. FDA scores are shown in a shaded row in each of the following four tables, but were not used to calculate the average scores.

Table 5.3: Results by Items in Target

TARGET					
Name of Archive	1. Mission Statement	2. Submission Criteria	3. Designated Communities	4. Ongoing Relationship	TOTAL SCORE (out of 12)
AILLA	3	3	3	1	10
ELAR	3	3	3	2	11
FDA	3	3	3	1	10
Kaipuleohone	3	2	3	1	9
NAL	3	2	3	3	11
PARADISEC	3	3	3	3	12
SIL Language and Culture Archives	3	2	3	2	9
UCSD Melanesian Archive (Mandeville Special Collections Library)	3	3	3	2	11
UVA Small Special Collections Library	3	3	3	2	11
AVERAGE SCORES	3.00	2.625	3.00	2.00	10.50

With regard to item 3, the designated communities for the two university library collections (UCSD and UVA) were specified as “linguistic researchers”; this designation was a good fit for these repositories. There were no specific designated communities in the other six archive evaluations. Thus, all archives were given the full score for item 1, designated communities. Though the FDA is a dark archive, serving only the libraries of public universities in Florida, it also received the full score as a control.

Table 5.4: Results by Items in Access

ACCESS					
Name of Archive	5. Discoverability	6. Fixed Identifiers	7. Reach	8. Access and Use Restrictions	TOTAL SCORE (out of 12)
AILLA	3	1	3	3	10
ELAR	2	3	2	3	10
FDA	2	3	1	3	9
Kaipuleohone	3	3	1	3	10
NAL	2	3	2	3	10
PARADISEC	3	3	2	3	11
SIL Language and Culture Archives	3	3	2	3	11
UCSD Melanesian Archive (Mandeville Special Collections Library)	3	2	3	2	10
UVA Small Special Collections Library	3	2	3	2	10
AVERAGE SCORES	2.75	2.5	2.25	2.75	10.25

Table 5.5: Results by Items in Preservation

PRESERVATION					
Name of Archive	9. Evidence of Long-term Planning	10. Preservation Strategies	11. Integrity	12. Authenticity	TOTAL SCORE (out of 12)
AILLA	2	2	1	3	8
ELAR	2	2	3	2	9
FDA	3	3	3	3	12
Kaipuleohone	2	1	3	1	7
NAL	3	3	1	3	10
PARADISEC	2	3	3	2	10
SIL Language and Culture Archives	2	2	2	2	8
UCSD Melanesian Archive (Mandeville Special Collections Library)	2	2	2	3	9
UVA Small Special Collections Library	3	2	1	3	9
AVERAGE SCORES	2.25	2.13	2	2.38	8.75

Table 5.6: Results by Items in Sustainability

SUSTAINABILITY					
Name of Archive	13. Adequate Infrastructure	14. Financial Sustainability	15. Disaster Preparedness	16. Succession Plan	TOTAL SCORE (out of 12)
AILLA	2	3	3	2	10
ELAR	3	2	2	2	9
FDA	3	3	3	3	12
Kaipuleohone	1	1	2	2	6
NAL	2	3	3	2	10
PARADISEC	2	1	3	1	7
SIL Language and Culture Archives	2	2	2	1	7
UCSD Melanesian Archive (Mandeville Special Collections Library)	3	3	2	2	10
UVA Small Special Collections Library	3	3	3	2	11
AVERAGE SCORES	2.25	2.25	2.5	1.75	8.75

5.3 Relative Strengths and Weaknesses of Digital Language Archives

Table 5.7 lists the average scores, from highest to lowest, for each item in the TAPS Checklist for the digital language archives evaluated. Digital language archives were strongest in items pertaining to the TAPS categories of Target (mission statement, submission criteria) and Access (access and use restrictions, discoverability, and fixed identifiers). Archives were weakest in items pertaining to the categories of Preservation (integrity, preservation strategies, evidence of long-term planning) and Sustainability (succession plan, adequate infrastructure, financial sustainability).

Table 5.7: Relative Strengths and Weaknesses of Digital Language Archives

TAPS CATEGORY	ITEM	AVERAGE SCORE
Target	1. Mission Statement	3.00
Target	3. Designated Communities	3.00
Access	5. Discoverability	2.75
Access	8. Access and Use Restrictions	2.75
Target	2. Submission Criteria	2.625
Access	6. Fixed Identifiers	2.50
Sustainability	15. Disaster Preparedness	2.50
Preservation	12. Authenticity	2.38
Access	7. Reach	2.25
Preservation	9. Evidence of Long-term Planning	2.25
Sustainability	13. Adequate Infrastructure	2.25
Sustainability	14. Financial Sustainability	2.25
Preservation	10. Preservation Strategies	2.125
Target	4. Ongoing Relationship	2.00
Preservation	11. Integrity	2.00
Sustainability	16. Succession Plan	1.75

The scores for the first and second categories of TAPS (Target and Access or T/A) and the third and fourth categories of TAPS (Preservation and Sustainability or P/A) were compared using the Mann-Whitney-Wilcoxon two-sample rank-sum test (also known as a Wilcoxon Rank Sum).²⁴ The medians of individual item scores for T/A and P/S were 3 and 2 respectively. The test for significance excluded all scores from the FDA, and scores for item 3, designated communities. The data points compared are summarized in table 5.8 below. The two sets of scores differed significantly (Mann-Whitney $U = 2317.0$, $n_1 = 56$, $n_2 = 64$, $P < 0.01$ two-tailed) (Avery 2004). We may therefore conclude that the digital language archives in this study are more effectively addressing items in Target and Access than items in Preservation and Sustainability.

²⁴ [http://en.wikipedia.org/wiki/Mann-Whitney_U_test]

Table 5.8: Summary of Data Points Used in Mann-Whitney-Wilcoxon Two-Sample Rank-Sum Test

TAPS Categories		TARGET and ACCESS (T/A), $n_1 = 56$							PRESERVATION and SUSTAINABILITY (P/S), $n_2 = 64$																						
Name of Archive	TAPS Items	1. Mission Statement		2. Submission Criteria		4. Ongoing Relationship		5. Discoverability		6. Fixed Identifiers		7. Reach		8. Access and Use Restrictions		9. Evidence of Long-term Planning		10. Preservation Strategies		11. Integrity		12. Authenticity		13. Adequate Infrastructure		14. Financial Sustainability		15. Disaster Preparedness		16. Succession Plan	
	TARGET Scores	ACCESS Scores						PRESERVATION Scores				SUSTAINABILITY Scores																			
AILLA		3	3	1	3	1	3	3	2	2	1	3	2	3	3	2	2	3	3	2	2	3	3	2	2	3	3	2	2		
NAL		3	2	3	2	3	2	3	3	3	1	3	2	3	3	2	2	3	3	2	2	3	3	2	2	3	3	2	2		
PARADISEC		3	3	3	3	3	2	3	2	3	3	2	2	3	3	2	2	1	3	1	2	2	1	3	1	2	3	2	2		
UCSD Melanesian Archive (Mandeville Special Collections Library)		3	3	2	3	2	3	2	2	2	2	3	3	3	2	2	2	3	3	3	2	2	3	3	2	2	3	3	2	2	
SIL Language and Culture Archives		3	2	2	3	3	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	
ELAR		3	3	2	2	3	2	3	2	2	3	2	2	2	3	2	3	2	2	2	2	3	2	2	2	2	2	2	2	2	
UVA Small Special Collections Library		3	3	2	3	2	3	2	3	2	1	3	3	3	3	2	3	3	3	3	2	3	3	3	2	3	3	3	2	2	
Kaipuleohone		3	2	1	3	3	1	3	2	1	3	1	1	3	1	1	1	1	2	2	2	1	1	2	2	2	2	2	2	2	

5.4 Relative Strengths and Weaknesses of Global versus Regional Archives

. The preliminary site visits and interviews are suggestive of a relationship between strengths and weaknesses of global versus regional archives (see section 2.1.3 on scope of archives' collections). Table 5.9 plots each archive by scope (global or regional) against each item of the TAPS checklist by category (Target, Access, Preservation, and Sustainability). The scores for each item and category of the TAPS Checklist were

summed for each archive and group of archives to compare relative strengths and weaknesses of global and regional archives. The sums suggest that regional archives are more likely to be stronger in three of the four categories. In every category except Access, in which the total categorical scores were the same, the total scores for regional archives were higher than for global archives. In every item of TAPS, except for item 6, fixed identifiers (Access), and item 11, integrity (Preservation), the scores for regional archives were the same or higher than for global archives. Regional archives may be better at maintaining ongoing relationships with language communities (item 4, Target), provide materials in more accessible ways (item 7, reach), have more robust preservation strategies (item 10, Preservation), and have better ways to check the authenticity of materials (item 12, Preservation). Regional archives may also be more sustainable, with higher scores for item 14, financial sustainability, and item 15, disaster preparedness with the category of Sustainability. On the other hand, global archives appear to be stronger in providing the technologies needed to assign fixed identifiers (item 6, Access), and maintain the integrity of digital materials (item 11, Preservation). These conclusions are tentative given the small sample of archives.

Table 5.9: Archives Sorted by Global versus Regional Scope

TAPS Categories	TAPS Checklist Items	Scope		Global					Regional				
		Archives	ELAR	Kaipuleohone	SIL Language and Culture Archives	UVA Small Special Collections Library	TOTALS by Category	AILLA	NAL	PARADISEC	UCSD Melanesian Archive	TOTALS by Category	
Target	1. Mission Statement		3	3	3	3	12	3	3	3	3	12	
	2. Submission Criteria		3	2	2	3	10	3	2	3	3	11	
	3. Designated Communities		3	3	3	3	12	3	3	3	3	12	
	4. Ongoing Relationship		2	1	2	2	7	1	3	3	2	9	
	TOTALS by Archive		11	9	9	11	40	10	11	12	11	44	
Access	5. Discoverability		2	3	3	3	11	3	2	3	3	11	
	6. Fixed Identifiers		3	3	3	2	11	1	3	3	2	9	
	7. Reach		2	1	2	3	8	3	2	2	3	10	
	8. Access and Use Restrictions		3	3	3	2	11	3	3	3	2	11	
	TOTALS by Archive		10	10	11	10	41	10	10	11	10	41	
Preservation	9. Evidence of Long-Term Planning		2	2	2	3	9	2	3	2	2	9	
	10. Preservation Strategies		2	1	2	2	7	2	3	3	2	10	
	11. Integrity		3	3	2	1	9	1	1	3	2	7	
	12. Authenticity		2	1	2	3	8	3	3	2	3	11	
	TOTALS by Archive		9	7	8	9	33	8	10	10	9	37	
Sustainability	13. Adequate Infrastructure		3	1	2	3	9	2	2	2	3	9	
	14. Financial Sustainability		2	1	2	3	8	3	3	1	3	10	
	15. Disaster Preparedness		2	2	2	3	9	3	3	3	2	11	
	16. Succession Plan		2	2	1	2	7	2	2	1	2	7	
	TOTALS by Archive		9	6	7	11	33	10	10	7	10	37	

5.5 Findings and Conclusions

The TAPS Checklist offers linguists and other potential depositors a simple way to “spot check” the practices of digital archives along the lines of sixteen items arranged in four categories (Target, Access, Preservation, and Sustainability). The digital language archives evaluated show that they were strongest on points addressing a specialized audience and issues surrounding access, which linguists are most concerned about, but were weakest on technical and organizational points that may impact the overall longevity of the archived material.

The archives scored the highest on the Target (mission statement, designated communities, and submission criteria) and Access (discoverability, access and use restrictions, and fixed identifiers) components. These results are as would be expected since the archives selected reflected a reasonably good fit for language materials. As specialized language archives, they are providing the kinds of services that linguists and language communities generally need and want. The archives also demonstrated an effective response to the particular challenges inherent to sensitive archived language materials, which require well-defined access policies. With the exception of the library-based archives (the UCSD Mandeville Special Collections Library (Melanesian Archive) and the UVA Albert and Shirley Small Special Collections Library), all digital language archives received the full score for access and use restrictions. Though the total scores for the libraries were among the highest scores, this slight disparity between libraries and language archives concerning access and use restrictions (item 8) may indicate that

specialized archives are more likely to provide the kinds of services that linguists and language communities require.

The archives scored the lowest on the Preservation (integrity, preservation strategies, and evidence of long-term planning) and Sustainability (succession plan, adequate infrastructure, and financial sustainability) components of the TAPS Checklist. These results indicate that the longevity of digital language archives is at risk. Linguists should steer clear of non-archives altogether when finding an archival home for their work, but even the archives themselves need to improve their implementation to ensure the preservation of digital data. Awareness should also be raised concerning the financial sustainability of digital language archives. The products created by the many language documentation projects that are now being funded, as well as older material from the pre-digital era, need good archival homes if they are to last. It does not stand to reason to skimp on the preservation of irreplaceable materials that have been collected at great cost and considerable effort. Most archives interviewed were fairly confident that, if they ceased operations, data would persist in the servers of their host institution, but could not guarantee access to them.

Another point of weakness common to most of the digital language archives was item 4 concerning an ongoing relationship with the language community.²⁵ While some examples of exemplary conduct were evident (NAL, PARADISEC), these relationships have not been fully explored by digital language archives. But if the heritage language

²⁵ Members of the language community may also be depositors, in which case many archives would continue to be in contact with at least one “insider” in the community. The commitment to an ongoing relationship with the community involves more than relating to depositors, however, as discussed in section 4.2.4.

communities' experiences in the U.S. are an indication, the language communities will potentially be the most invested "stakeholders" and the heaviest users of language archives.

An analysis of the scores suggests that regional and global archives have different strengths and weaknesses. Many global-scale archives (ELAR), and some large regional archives (AILLA) self-report that they are not resourced to interface with communities. Thus, larger archives may have slightly better technical processes, but they are not as good at directly relating to communities. In this small sampling of language archives, regional archives appear to be more sustainable, with higher scores for financial sustainability, and disaster preparedness. Regional archives are also sometimes located close to language communities, enhancing accessibility of resources.

It is postulated that regional archives may enjoy more secure funding because of a greater local commitment to local language materials (NAL at SNOMNH, OU), or a greater commitment with greater specialization (the Melanesian Archive within UCSD's Mandeville Special Collections Library). Additionally, staff sizes and resources for regional archives were generally comparable to global ones according to my site visits and interviews; that is, global archives are often trying to do more with similar resources. On the other hand, some large-scale regional archives with few submitter restrictions (AILLA, NAL) perceived a lack of adequate infrastructure given their backlog of materials needing digitization and proper accessioning.

Global language archives scored consistently higher on measures to ensure the integrity of data (Preservation) and more consistently used fixed identifiers (Access),

which indicate that global language archives may have a higher degree of technical sophistication than regional archives. However, lower scores on the “reach” (Access) of global archives suggest that they are weaker in the distribution of data. This may have implications for assumptions about economies of scale (see section 5.6 regarding recommendations for further research).

It is clear that even among digital language archives, archives have different specializations and focuses. For example, PARADISEC and ELAR illustrate how these differences arise by virtue of the way the archives were established. PARADISEC was initially funded for one year, and set-up within that year to justify further funding; it was fully operational within one and a half years, and knowing that funding would be intermittent, it set up self-sustaining structures. Even now, PARADISEC reportedly has “no resources,” has not been funded for three years, and its staff is mostly composed of a core group of dedicated volunteers. Initially conceived as a large regional archive, it has the technical infrastructure to be a global archive and has expanded to include non-text materials from the Kaipuleohone archive, which were too large for the current library system at UH. PARADISEC cannot accommodate all depositor requests, however, due to a lack of resources.

ELAR, in contrast, was granted funding for fifteen years by the Arcadia Foundation, which also funded its sister program, the Hans Rausing Endangered Languages Project (HRELP). Perpetual funding has not been guaranteed to ELAR, but funding was recently extended by five years (until 2016), and there is hope that it will become a permanent part of the linguistics department of SOAS at the University of

London. From the beginning, ELAR was conceived to be a global archive. ELAR systematically trains depositors, and advises them as to the most effective equipment and techniques, thereby improving the quality of language documentation projects. It is also leading the way in helping depositors interface with language communities so that depositors can make their own choices with regard to access conditions (using a tool modeled on “Facebook”). However, the development of these services may overshadow concerns about exactly what will happen when the linguist will no longer be there to make such decisions; the partial solution is to designate representatives if the original depositor is unavailable or deceased. Most archives appear to prefer open-access deposits, which do not require much further interaction with the depositor.

Linguists would do well to balance their concern for near-term services with a concern for the archive’s ability to preserve data in the long-term. It may be a good idea to follow suit with the Endangered Languages Documentation Programme (ELDP) requirement that its grantees deposit at their global archive, ELAR, plus another, local archive (perhaps one that is more likely to be readily accessible to the language community). The depositor and language community then conceivably get “the best of both worlds”: long-term preservation and near-term access. This solution points to the general principle espoused by LOCKSS or “Lots of Copies Keeps Stuff Safe,” a project at Stanford University that provides libraries with digital preservation tools.²⁶

The sensitivity of language materials to be deposited should be considered, however, before depositing them in diverse archives with varying levels of commitment

²⁶ [<http://lockss.stanford.edu/lockss/Home>]

to the security of their collections. It bears mentioning that in the digital age, it is all too easy to proliferate copies that can cause irreparable harm to a community, and digital information can conceivably live on indefinitely. Though NAL and AILLA are regional archives with university affiliations, their approaches to the security of their collections differ significantly. NAL, being located at a world-class museum, is very serious about security. Uniformed guards are stationed in an office at the entrance and handle specially issued name badges that authorize entrance into the building holding the collection; in the course of a few hours, the badges change to reveal stripes that invalidate it. Linn at NAL



Figure 5.1: Time-activated Security Badge. A souvenir from my visit to NAL, this badge was printed at the time of my arrival. No stripes were at first apparent. Stripes appeared the day after my visit, invalidating the badge.

noted that there is a tension between making the facilities friendly to visitors and maintaining an access-controlled environment. AILLA on the other hand, operates out of an office in the library of a large public university. Though it offers graded levels of access, AILLA makes no promises of the security of archived data and cautions

depositors to refrain from archiving highly sensitive material. In contrast, NAL was committed to prosecuting for the misuse of any archived material. One possible strategy is for a depositor to create two collections, an open collection that could be accessioned at multiple locations, and a sensitive collection that can be more tightly managed by an archive that has a firm commitment to maintaining access and use restrictions.

5.6 Recommendations for Further Research

A larger sampling of digital language archives using the TAPS Checklist would be desirable to show the true state of digital language archives. A survey of archives differing in institutional affiliation, such as those run by government and non-profit agencies, rather than almost exclusively academic institutions, would also give a clearer picture of the world's digital language archives, and make possible meaningful comparisons between and among archives with different kinds of institutional affiliations. In many instances, archives were continuing to improve and develop their services and capacities. Weaknesses may be remedied over time, though archives lacking in resources may regress in the quality of services and data preservation. Longitudinal studies could be conducted to explore the progress of digital language archives as the field is better defined.

It would be valuable to investigate archives that are tribal in scope since none were evaluated in the testing of the TAPS Checklist. I would expect the trend among global and regional archives to continue towards decreasing sophistication in Preservation issues with respect to even smaller tribal archives, but it is not clear if tribal

archives would be more sustainable or less sustainable than either global or regional archives.

The differences observed in global as opposed to regional archives could be due to the fact that global archives are generally larger than regional archives. Thus it would be good to study effectiveness in preserving data in comparison to the total size of the archive. Rather than measuring the size of digital collections in terabytes, which may not be very meaningful given the diversity of media and particular digitization choices made, collections could be measured in terms of time with respect to audio and video recordings, and perhaps some equivalent to linear feet used in the paper-based archiving world for digitized field notes. There may be a “critical mass” of efficiency for certain archives, and it may be that an “ideal” digital language archive is not too large, but is specialized to a degree.

Finally, the economics of digital preservation is an area of growing scholarship (BRTF-SDPA 2010). A focused study on the costs of archiving digital language materials, taking into account a given archive’s budget with respect to the volume of materials preserved, could inform the wider linguistic community as to how much digital language archiving costs. Knowing this, we could go forward with budget and funding proposals in building our “data economy,” envisioned as a system in which “those who care, those who pay, and those who preserve are working in coordination” (RPI 2010). These factors could be extrapolated for a variety of other kinds of digital archives. This may also serve to inform other memory institutions about a baseline for the costs of digital preservation. Such a study could take advantage of the diversity of digital

language archives, with various types of digital media, access rights, budget and funding models.

5.7 Concluding Remarks

Much work remains. While language shift is progressing all over the world, and is influenced by forces that seem much bigger than ourselves, it is striking that the curation of archived language resources of so many of the world's languages is concentrated in the hands of so few—just a handful of people comprise the entire staff at any one of the digital language archives of global scope. I hope that this thesis has been effective in shedding some light on some major issues involved in digital language archiving and that the TAPS Checklist will help anyone depositing language documentation to make good decisions about where they place such valuable materials. May this work help linguists to act wisely to save language documentation, language communities to value and recover their languages, and archivists to better preserve and provide access to digital language materials, both now and into the future.

Appendix A

TAPS (Target, Access, Preservation, and Sustainability): Checklist for Responsible Archiving of Digital Language Resources

Archive:	Date:	Reviewer:
TARGET	Yes ? No	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify: _____)?	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
3. Designated Communities: Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g., language community, academic community, etc.)?	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
4. Ongoing Relationship: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
ACCESS	Yes ? No	Comments
5. Discoverability: Are the descriptive metadata for materials deposited at the archive searchable online? That is, the metadata is posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g., Google, Yahoo!, Bing, etc.)?	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
6. Fixed Identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
7. Reach: Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
8. Access and Use Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	

Yes = best practice

? = in planning stage / partial practice / assumed done by others

No = not in scope of archive / unclear

PRESERVATION		Yes ?	No	Comments
9.	Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g., the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10.	Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11.	Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g., a checksum, or digital signature, etc.)?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12.	Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g., by verifying that digital materials are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital holdings)?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUSTAINABILITY		Yes ?	No	
13.	Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g., quality media, environmentally-controlled storage, access-controlled storage area)?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14.	Financial Sustainability: Does the archive appear to have secured sources of long-term funding?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15.	Disaster Preparedness: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g., disaster recovery plan, offsite storage of backups)?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16.	Succession Plan: Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Yes = best practice

? = in planning stage / partial practice / assumed done by others

No = not in scope of archive / unclear

In-Depth Questions for use with TAPS

TARGET

2. Submission Criteria

In-depth questions: What is the content of material I want to deposit? What formats are they in? Does my content fall within the collection policy of the archive? Do my materials fall within the preferred submission formats of the archive?

3. Designated Communities

In-depth questions: What are the desired user communities for the data I want to deposit? Do these fall within the designated communities of the archive? Particularly if you see the language community as an important user of the deposited materials, does the archive cater to that user community?

4. Ongoing Relationship

In-depth questions: What types of interaction do you anticipate needing to take place between the language community and the archive? Will the archive support these? Is potential revenue sharing an issue for your deposit? If so, will the archive offer this service?

ACCESS

5. Discoverability

In-depth questions: Does the archive have the necessary guidelines and standards to help the linguist provide quality descriptive metadata? Is the metadata posted on the Internet? If so, are these records easily found on the Internet? Is the metadata aggregated through a service such as OLAC? Does my desired designated community have adequate research opportunities through the archive's approach to descriptive metadata?

7. Reach

In-depth questions: Will members of the designated communities be expected to have access to the Internet? Will members of the designated communities need to maintain an e-mail address? Will the metadata be available in English only, or will it be available in another language that is more accessible to the designated community? Will the archive charge fees for copies of data on media that are usable by members of the designated communities? Even if the fees are "at cost," will they be affordable for those communities?

8. Access and Use Restrictions

In-depth questions: How does the archive deal with copyright? Does the archive require transfer of ownership? Does the archive allow materials to be deposited with restrictions on access? If so, what restrictions are possible and how are requests for access handled? Does the archive allow materials that are closed to access to be deposited? How long will periods of closed access last? What are the archive's conditions of use policies?

PRESERVATION**9. Evidence of Long-term Planning**

In-depth questions: Does the archive have written procedures for the tasks involved in implementing their defined archival process? Do these procedures specify deadlines for completing upcoming tasks as they pertain to the creation and maintenance of archival information packages? Is responsibility for each process clearly assigned to particular individuals or outsourced entities? Is the archive explicitly monitoring substantial changes, whether technical, organizational, or community-based? Will the archive change its procedures as needed?

10. Preservation Strategies

In-depth questions: On what medium will the archive store the materials you submit? What is their schedule for refreshing data on that medium? What will the archive do with the data as the medium approaches obsolescence? What will the archive do if the format in which the data are stored becomes obsolete? When was the last time a migration from an obsolete format to a newer format was completed?




SUSTAINABILITY**15. Disaster Preparedness**

In-depth questions: Is there a formally documented procedure for regular backups? Is compliance with the procedures audited? Where are the backups of archived material kept? If the building or location housing the archive is destroyed, how will materials be recovered? Is there a written disaster recovery plan, located off-site, that makes explicit what to do if a disaster occurs?

Appendix B

TAPS Checklist Evaluations

Key to Scoring

Yes	<p>Best practice</p> <p>The archive appears to follow best practices. The strongest indicator of this is that the staff is following a written policy or procedure that conforms to best practices.</p>	3 points	
?	<p>In planning stage / partial practice / assumed done by others</p> <p>The archive is in the planning stages of implementing best practices; or the archive partially follows best practices; or the archive is assuming that another entity to which they outsource follows best practice in implementing the functions indicated, but cannot document it.</p>	2 points	
No	<p>Not in scope of archive / unclear</p> <p>The item is not in the scope of the archive; or the degree to which the archive follows best practice is unclear.</p>	1 point	

Appendix B-1

Archive of Indigenous Languages of Latin America (AILLA)




University of Texas (UT), Austin, Texas

AILLA is a joint project of the Departments of Anthropology and Linguistics, and the Digital Library Services Division of the General Libraries at the University of Texas at Austin.


Archive Representative: **Heidi Johnson, Program Coordinator**





Reviewer: **Debbie Chang**

Date: **22 October 2009**





TARGET	Rating	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?		A three-part mission statement is found on the AILLA's homepage: http://www.ailla.utexas.org/site/welcome.html . "Mission #1: Preservation" reflects a long-term commitment to the preservation of digital information. Some concern was expressed that the mission statement was too "verbose" and that there were parts of the website that were outdated at the time of the interview.
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify:_____)?		The submission criteria can be found at the website. All legitimate materials in or about an indigenous language of Latin America are accepted, even "laundry lists." Materials in an indigenous language must have originated from a native speaker belonging to that language group. Borderline cases involve Spanish dialects within an indigenous language group. The submission criteria do not address formats.
3. Designated Communities: Is my desired audience (specify:_____) a good match for the groups of users the archive targets (e.g. language community, academic community, etc.)		The archive serves Latin American language communities. Materials, including metadata, are available in contact languages, Spanish and English, but not Portuguese. This reflects the degree of fluency the program coordinator has in English and Spanish. Some metadata has been made available in the language of specific language communities, but this effort depends on the availability of graduate research staff persons who are familiar with those communities.

Archive of Indigenous Languages of Latin America (AILLA)





4. **Ongoing Relationship:** Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)
-  AILLA is not involved in tribal politics. The physical distance of the archive and the vastness of its holdings were reasons cited for making interfacing with the language community impractical. The archive does not charge for use of or pay for materials.

ACCESS	Rating	Comments
5. Discoverability: Are the metadata for materials deposited at the archive posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g. Google, Yahoo!, Bing, etc.)?		AILLA is a participant in the OLAC community, and is also a well-established presence on the Internet.
6. Fixed Identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?		Identifiers used in AILLA are not strictly fixed according to best practice. A complicating fact is that identifiers contain meaning-based elements that may change over time (e.g. normalizing language codes that are contained in identifiers at a later date).
7. Reach: Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?		Users in language communities must have access to a computer and an Internet connection in order access materials in AILLA. For some, this would mean paying to go to an Internet café. Materials are readily accessible to the academic community for whom reliable Internet connections can be assumed.
8. Access and Use Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?		AILLA uses formal access restriction forms and agreements that clearly define conditions and disclaimers.

Archive of Indigenous Languages of Latin America (AILLA)

PRESERVATION	Rating	Comments
<p>9. Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g. the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?</p>		<p>It is assumed that UT Digital Library Sciences follows written policies and procedures.</p>
<p>10. Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?</p>		<p>It is assumed that UT Digital Library Sciences refreshes and updates digital materials as needed.</p>
<p>11. Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g. a checksum, or digital signature, etc.)?</p>		<p>Unknown.</p>
<p>12. Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g. by verifying that digital objects are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital objects that occur while they are in the care of the archive)?</p>		<p>Some digital materials can be traced back to tapes that have been split up into several digital files; tapes are kept in the Benson Latin American Collection within the UT library system. The authenticity of certain audio materials is verified with listening. Migrated formats are also noted, though there is concern that the archive is "not as detailed as [it] could be."</p>

Archive of Indigenous Languages of Latin America (AILLA)

SUSTAINABILITY	Rating	Comments
<p>13. Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g. quality media, environmentally-controlled storage, access-controlled storage area)?</p>		<p><u>Staff:</u> AILLA has a director and associate director who have policy roles in the archive, but do not serve as staff day-to-day in the archive. Heidi Johnson is the Program Coordinator; her time is split half-time in the university libraries, and half-time in the College of Liberal Arts. Two half-time graduate students work and are trained under Johnson. The digitization of materials done by graduate students who are learning on the job is recognized as "expensive," taking an estimated three to four times more time in comparison to professionals. However, learning takes precedence over efficiency as grants support the graduate students' work and reflect that the archive is part of a larger teaching institution. <u>Technical Infrastructure:</u> It is assumed that the digital library services division of UT maintains state-of-the-art equipment.</p>
<p>14. Financial Sustainability: Does the archive appear to have secured sources of long-term funding?</p>		<p>AILLA is funded by NEH and NSF grants as well as UT, which is funded by the state of Texas. Johnson is UT staff and her role in the archive is not contingent on grants. UT covers Johnson's salary, office, and servers---items that would not be first in line for budget cuts.</p>
<p>15. Disaster Preparedness: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g. disaster recovery plan, offsite storage of backups)?</p>		<p>Digital materials are backed up daily and weekly in a central location (the "tower") and are safely stored in an abandoned salt mine.</p>
<p>16. Succession Plan: Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?</p>		<p>No formal succession plan was evident. Materials would most likely stay with the UT library server. Preservation but not access would be assured.</p>

Appendix B-2

Endangered Languages Archive (ELAR)

School of Oriental and African Studies (SOAS), London, England



The digital archive of the Hans Rausing Endangered Languages Project (HRELP) at the SOAS within the University of London, England. The work of Endangered Languages Documentation Programme (ELDP) grantees and others are deposited here.

Archive Representative: **David Nathan, Director**

Reviewer: **Debbie Chang**



Date/s: **16 March 2010** (review of website)

31 March 2010 (phone interview)




TARGET	Rating	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?		The first of four aims of the archive is to "provide a safe long-term repository of language materials" (http://www.hrelp.org/archive/).
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify:_____)?		With regards to content, ELAR specializes in deposits of materials that relate to endangered languages as the archive to the Hans Rausing Endangered Languages Project, a "co-sibling" with the granting body, Arcadia. The archive accepts materials mainly from three categories of depositors: (1) Endangered Languages Documentation Programme (ELDP) grantees, (2) Endangered Languages Academic Programme (ELAP) students, (3) people who have deposited material at ELAR previously. Many depositors fall into the first category, ELDP grantees (http://www.hrelp.org/archive/depositors/), but the archive also welcomes deposits from other researchers who have worked on endangered languages (http://www.hrelp.org/languages/help/#6). The archive is thus open to all depositors, but gives priority to grantees. To date, a handful of deposits have been made in ELAR from people who have not been affiliated as funding recipients. There is preference for materials that are open access. The archive takes a liberal interpretation of what constitutes an endangered language, and reserves the right to evaluate the quality and extent of data to be deposited.

Endangered Languages Archive (ELAR)

With regards to type, the archive specifies preferred deposit formats and offers guidelines (<http://www.hrelp.org/archive/depositors/formats.html>). Most of the materials deposited in the archive are already in digital form, though analog conversion solutions are available on a case by case basis; i.e. digitization services are not offered by default. (<http://www.hrelp.org/archive/depositors/>).

-
3. **Designated Communities:** Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g. language community, academic community, etc.)
- 
- Archived materials are for potential use by the language community, researchers, and others. The archive can accommodate requests to restrict access (<http://www.hrelp.org/archive/depositors/index.html>). Because ELAR recognizes that (1) endangered language materials are diverse, and (2) sensitive to all kinds of access restrictions and through time, the archive is further developing a platform for the depositors to directly negotiate access to materials from users. The “default” setting will be for materials to be open to community members, but direct contact, anticipated to be a foreground rather than background case, can cross-cut that in a Facebook-like model. Delegates could be appointed by depositors to handle negotiations over the long-term (i.e. when original depositors are gone).
-
4. **Ongoing Relationship:** Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)
- 
- The archive takes seriously its obligation to help provide data in usable ways to language communities, but as an international archive with more than 200 provider communities, it is not realistic to interface with each one. Revenue sharing is not explicitly in its funding model. In principle the archive would not be opposed, but it is not in a position to generate revenue. ELAR is developing better ways for communities to interact with linguistic multimedia, however (see item #7).
-

Endangered Languages Archive (ELAR)

ACCESS	Rating	Comments
<p>5. Discoverability: Are the metadata for materials deposited at the archive posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g. Google, Yahoo!, Bing, etc.)?</p>		<p>Deposits of endangered languages documentation materials are described in the ELAR catalogue. At this time, the catalogue appears to be a first-release beta version that supports views of summary information about deposits (http://elar.soas.ac.uk/catalogue). The archive has search facility and catalog is exposed to search engines, but it is still under development concerning privacy issues. Metadata established with deposit form and is "crawled" by Google on the web. Results vary: the archive reports that they were encouraged with the results from a simple Google search, but a Google search done by reviewer for a language with materials deposited at the archive yielded no "first page" results.</p>
<p>6. Fixed Identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?</p>		<p>Yes. The archive's local system ingestion software assigns random numbers as fixed identifiers (to a file or group of files). The archive also has a plan to have a handle- system persistent urls for these identifiers.</p>
<p>7. Reach: Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?</p>		<p>"ELAR does not normally charge for depositing, storing or accessing materials (in some cases people requesting data may pay the cost of delivering them, such as disks and postage)" (<i>Endangered Languages Archive Deposit Form v0.91PR</i> 2007:5, http://www.hrelp.org/archive/depositors/depositform/ELAR_Deposit_09PR.pdf). The cost of CDs plus shipping are below the cost of production; accessibility for reasons of these nominal costs have not been an issue. Participants in ELDP can use grant money to publish to community. The grant stipulates that documentation must publish to another archive besides ELAR that is accessible to the community, so most depositors will be depositing somewhere else.</p>

Endangered Languages Archive (ELAR)




The archive is doing research in order to improve accessibility to language communities, for example, developing a “people database” of speakers and consultants that will be more accessible to language communities than searching and interacting with materials according to standards designed for/by language descriptivists. The archive is also exploring how to provide metadata in interface languages.

-
8. **Access Restrictions:** Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?






ELAR fully respects and implements requests to restrict access if the depositor or the relevant language community do not wish to publish the materials, but nevertheless recommends archiving for “preservation and cataloguing purposes” (<http://www.hrelp.org/archive/depositors/index.html>) . ELAR pledges to support the preservation of deposited materials while protecting the depositor’s interests and is providing “leadership and innovation” in this area. The archive makes the materials available to the depositor, provides facilities for the depositor to manage them, and allows access to them consistent with the depositor’s wishes. (Endangered Languages Archive Deposit Form v0.91PR 2007:5, http://www.hrelp.org/archive/depositors/depositform/ELAR_Deposit_09PR.pdf). If permissions are not updated after three years, access becomes less restricted (Endangered Languages Archive Deposit Form v0.91PR 2007:5, http://www.hrelp.org/archive/depositors/depositform/ELAR_Deposit_09PR.pdf).

Endangered Languages Archive (ELAR)



PRESERVATION	Rating	Comments
<p>9. Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g. the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?</p>		<p>ELAR's commitment to long-term planning is "not in policy documents as such," but does carry out this type of planning. "ELAR stores data using high quality equipment and manages its collection according to recommended practice in the languages archiving community.... Data systems may be changed where necessary to meet changes in legislation or new legislation. ELAR collaborates with other leading UK and international archives to ensure that materials are preserved well into the future" (<i>Endangered Languages Archive Deposit Form v0.91PR 2007:5</i>, http://www.hrelp.org/archive/depositors/depositform/ELAR_Deposit_09PR.pdf).</p>
<p>10. Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?</p>		<p>The archive intends to refresh and update digital materials, but it is "not a priority" at this time and is prepared to live with "legacy formats" for a while. The archive has lots of material in un-ideal formats with no stated policy to deal with them. A few practice migrations from Word to html have been done, but original formats have been kept in consideration of depositors who want access to materials in the same format in which they were submitted.</p>
<p>11. Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g. a checksum, or digital signature, etc.)?</p>		<p>Checksums are in use.</p>

Endangered Languages Archive (ELAR)

12. **Authenticity:** Does the archive ensure that digital materials contain what they claim to contain (e.g. by verifying that digital materials are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital holdings)?
-  ELAR staff do not have time to check the authenticity of all materials, but they are “fairly careful” about tracking metadata, tracking relationships of some materials, and listening to some representative samples listened to during accessioning. Roughly 40% of depositors, many of whom are systematically trained by ELDP and ELAR, take the opportunity to send audio and metadata samples to the archive for quality control.

SUSTAINABILITY	Rating	Comments
<p>13. Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g. quality media, environmentally-controlled storage, access-controlled storage area)?</p>		<p><u>Staff:</u> ELAR is adequately staffed and includes a full-time archivist (Nathan), a full-time software developer, a half-time digital technician who manages digital integrity and tape back-up, occasional access to another technician, and grad students who help curate data, find problems, respond to depositors, and normalize metadata. A shared individual from the local university IT department does server maintenance, “systems programmer” (same person). <u>Infrastructure:</u> the archive is equipped with “all of the above.”</p>
<p>14. Financial Sustainability: Does the archive appear to have secured sources of long-term funding?</p>		<p>The Endangered Languages Documentation Programme (ELDP) is funded by Arcadia (previously known as the Lisbet Rausing Charitable Fund), which provided £20 million pounds for the Project, including £17 million for grant funding" (http://www.hrelp.org/grants/index.html). The original budget for the archive budget was £1 million, and funding has been extended 5 years (~£1.5 million total through 2016). Archival responsibilities could then be handed over to another repository, but ELAR is making case for its centrality to SOAS. There is secured commitment to data “should the worst happen.”</p>

Endangered Languages Archive (ELAR)

-
15. **Disaster Preparedness:** Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g. disaster recovery plan, offsite storage of backups)?
- 
- The archive has a basic policy regarding back-up tapes, but has no written disaster recovery plan. With regards to storage offsite, the backed-up files are currently stored in a different, fireproof building; this data will also be stored in another town sometime this year. Due to some bad equipment, the archive has done some disaster recovery successfully.
-
16. **Succession Plan:** Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?
- 
- The archive has a succession plan that would guarantee preservation, but not accessibility since ELAR is “very media focused,” and requires specific access interfaces.
-

Appendix B-3

Florida Digital Archives (FDA)




Florida Center for Library Automation (FCLA), Gainesville, Florida

The FDA is operated by the Florida Center for Library Automation (FCLA), which serves the libraries of the public universities of Florida.


Archive Representative: **Lydia Motyka, M.L.S., Manager**





Reviewer: **Debbie Chang**

Date: **13 October 2009, revised 19 February 2010**





TARGET	Rating	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?		The mission statement is available online and in brochure form. The mission statement is also documented in the FDA's Policy Guide, first published in January of 2006. At the date of the interview, Version 2.4 of the Florida Digital Archive (FDA) Policy Guide, published in August 2007, was most current: http://www.fcla.edu/digitalArchive/pdfs/DigitalArchivePolicyGuide .pdf . It has since been updated to version 2.5, published April 2009.
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify: _____)?		The FDA takes all materials submitted by affiliates (Florida's public university libraries). A relationship would need to be established with an affiliate before materials can be deposited in the FDA; a linguist must go through administrative and technical contacts within the library system to have authorization to deposit materials. Not all formats are equal, however. Available online is a table that lists all formats that have full preservation support, and indicates the confidence level in preserving other formats: http://www.fcla.edu/digitalArchive/formatInfo.htm
3. Designated Communities: Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g. language community, academic community, etc.)		The FDA serves a very limited designated community as a state-funded entity and as documented in the Policy Guide. The designated community is clearly defined as FDA affiliates which are public university libraries in the state of Florida. Each of these state university libraries have formal agreements concerning archiving with the Florida Center for Library Automation (FCLA).

Florida Digital Archives (FDA)




<p>4. Ongoing Relationship: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)</p>		<p>Any "ongoing relationships" are conducted outside the domain of the FDA. Individual universities' policies and affiliates' internal structures, rather than the archive, would determine the level of interaction with the language community. It is unclear that affiliates would be committed to ongoing relationships.</p>
---	---	--

ACCESS	Rating	Comments
<p>5. Discoverability: Are the metadata for materials deposited at the archive posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g. Google, Yahoo!, Bing, etc.)?</p>		<p>Holdings in the FDA are not discoverable by the public, but the same materials and relevant metadata are likely discoverable through affiliates (individual state university libraries).</p>
<p>6. Fixed identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?</p>		<p>A unique and persistent identifier, an Intellectual Entity ID (IEID), is assigned to each Archival Information Package (AIP) in the FDA repository. Each file within the AIP also has a unique identifier, a Data File ID (DFID). The process is detailed in the Archive Services Reports (http://www.fcla.edu/digitalArchive/dalInfo.htm).</p>
<p>7. Reach: Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?</p>		<p>The FDA is a dark, or closed, archive where public access is not part of the mission. The accessibility of materials depends on a particular affiliate's policies; a relationship would need to be established with an affiliate (e.g. a Florida university library) before materials can be deposited in the FDA.</p>
<p>8. Access and Use Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to or use of the materials will be honored?</p>		<p>The FDA enforces access restrictions strictly by allowing only contractual contacts indicated by affiliates to access their holdings. Rights information is included in the metadata. Libraries separately enforce access restrictions.</p>

Florida Digital Archives (FDA)

PRESERVATION	Rating	Comments
<p>9. Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g. the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?</p>		<p>Written policies and procedures are found in the Policy Guide. Included are format information and action plans for formats receiving full support. Action plans are reviewed and updated at regular intervals.</p>
<p>10. Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?</p>		<p>There are defined action plans to refresh and migrate data in fully supported formats. Not all formats are supportable for forward migration, however. A formats specialist on staff determines the confidence level of different formats. (FDA affiliates use FDA guidelines and have their own submission guidelines.) FDA hardware is up-to-date and there is regular review of action plans. No large-scale migration has actually taken place, but this has been done as a proof-of-concept.</p>
<p>11. Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g. a checksum, or digital signature, etc.)?</p>		<p>Information packages contain MD5 (a Message-Digest algorithm) and checksum within the integrity metadata; these are checked at every stage of archival processes.</p>
<p>12. Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g. by verifying that digital objects are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital objects that occur while they are in the care of the archive)?</p>		<p>"Yes to all." Valid Submission Information Packages (SIPs) must contain a METS (Metadata Encoding and Transmission Standard) XML (Extensible Markup Language) descriptor file, and at least one content file. The descriptor acts as a "packing list" for the content files/digital objects. Originals of all files submitted as part of the SIP are archived as a permanent part of the AIP, regardless of any file transformations (normalization, migration) that may be performed on them by the archive. The archive's database maintains records of all actions performed on the AIP and its content files.</p>

Florida Digital Archives (FDA)

SUSTAINABILITY	Rating	Comments
<p>13. Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g. quality media, environmentally-controlled storage, access-controlled storage area)?</p>		<p><u>Staff:</u> The FDA is the "brainchild" of Priscilla Caplan, who continues to serve as Assistant Director for Digital Library Services and was on premises during my interview. Lydia Motyka has been the Manager of the archive since 2008 and is a back-up Operations Technician. A full-time Operations Technician runs daily production for the FDA. The archive also has access to the services of FCLA Systems Administrators in the administration of our hardware and network access. Both Ms. Motyka and Ms. Caplan have MLS degrees (Master of Library Science). A total of four programmers, including a formats specialist, are also on staff. <u>Technical Infrastructure:</u> The FDA is part of FCLA, which is committed to the continuing existence of the FDA and providing the resources it needs. The software used by the archive is written locally under DAITSS (pronounced "dates," Dark Archives in the Sunshine State). The servers were offsite and can be assumed to be state-of-the-art.</p>
<p>14. Financial Sustainability: Does the archive appear to have secured sources of long-term funding?</p>		<p>FDA's budget is intertwined with the FCLA, which is funded by an act of state legislature rather than by grants. A renewed annual financial commitment, and even funding in perpetuity, can be reasonably assumed. The FDA is not-for-profit and streamlines many of its processes to be more efficient in terms of each information package to make good use of its ample resources. It would "take a lot" for the state of Florida to disband the FCLA, but if this ever happens, the FDA may survive as an institution by possibly billing for its services with a prior six-month "warning" period.</p>
<p>15. Disaster Preparedness: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g. disaster recovery plan, offsite storage of backups)?</p>		<p>The archive engages in responsible backup practices. There is a continuity of operations plans in the FCLA. Offsite storage is in Jacksonville, Florida. The FDA participates in the TIPR (pronounced "tipper," Towards Interoperable Preservation Repositories) project with NYU and Cornell; this project, if successful, would have positive outcomes for the preservation of digital materials across diverse archives.</p>

Florida Digital Archives (FDA)

16. **Succession Plan:** Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?



A succession plan is defined in the FDA Policy Guide. Affiliates have two options in the event that the FDA ceases operations: either (1) the return of archived content to the affiliate, in effect since the inception of the FDA as a working repository, or (2) the sending of archived content to another archive of the affiliate's choice in an acceptable exchange format, once a standard exchange format has been established within the international preservation community. The FDA is working with other archives on the grant-funded TIPR (Towards Interoperable Preservation Repositories) project, to be completed September 2010, to ensure interoperability of content. The FDA's metadata is PREMIS- (Preservation Metadata: Implementation Strategies-) based, and is compliant with a "core" preservation metadata element set applicable to diverse archives. An affiliate may choose either option for all archived materials, or different options for defined (by project coding) subsets of materials.

Appendix B-4

Kaipuleohone, University of Hawai'i at Mānoa




Kaipuleohone is the University of Hawai'i's digital ethnographic archive within the Department of Linguistics, and specializes in materials related to small and endangered languages.

Archive Representative: **Nick Thieberger, assistant professor**


Reviewer: **Debbie Chang**



Date: **17 February 2010** (review of website)

18 February 2010 (Skype interview)



TARGET	Rating	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?		Kaipuleohone was established "to ensure that priceless and unique research recordings will be digitized, described and safely housed in the long[-]term" (http://www.ling.hawaii.edu/langdoc/archive.html).
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify: _____)?		The submission criteria are not as fully-formed as it could be. Kaipuleohone accepts only materials from those who are associated with the University of Hawaii (UH), and non-text files that are "too large" for the UH Library's EPrint system are being sent to an alternate archive, PARADISEC. The archive accepts audio and video recordings as well as photographs, notes, dictionaries, transcriptions, and other materials related to small and endangered languages (http://www.ling.hawaii.edu/langdoc/archive.html). "Though the archive is housed in the Department of Linguistics, it is meant to hold a wide range of ethnographic materials" (http://hdl.handle.net/10125/4422 , Albarillo and Thieberger 2009:8).
3. Designated Communities: Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g. language community, academic community, etc.)		The order of priority with regards to designated communities is: speakers, descendants, researchers, and general public. A guiding principle of the archive is that "[language] documentation takes seriously the notion that the material we record should be accessible for others, including the speakers of the language. This means that we have to locate the recordings in a suitable archive, and any text that we annotate can then be referenced to the original media on which it was recorded" (http://www.ling.hawaii.edu/langdoc/index.html).

Kaipuleohone, University of Hawai'i at Mānoa




4. **Ongoing Relationship:** Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)
-  There are no ongoing relationships between Kaipuleohone and a language community. The UH Library maintains conservative policies, especially where the ownership rights are unclear. Publicizing materials to interested groups could pose problems.

ACCESS	Rating	Comments
5. Discoverability: Are the metadata for materials deposited at the archive posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g. Google, Yahoo!, Bing, etc.)?		The current catalog can be searched at (1) ScholarSpace, UH's DSpace-based digital repository (http://scholarspace.manoa.hawaii.edu/handle/10125/4250), which "can be browsed or searched in a number of different ways, including via any Open Archives Initiative search engine (and Google)" (http://hdl.handle.net/10125/4422 , Albarillo and Thieberger 2009:6-7). (2) Kaipuleohone's own catalog database, which sends metadata to ScholarSpace. The archive catalog accommodates specific linguistic metadata better than the general structure of the ScholarSpace catalog allows, and has a data entry screen more suitable for linguistic information, using drop-down menus to enforce consistency of data entry using controlled vocabularies. Or (3) via OLAC (http://www.language-archives.org/archive/scholarspace.manoa.hawaii.edu) (http://hdl.handle.net/10125/4422 , Albarillo and Thieberger 2009:7).
6. Fixed identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?		Fixed identifiers are assigned with ScholarSpace. A system was devised by a programmer Daniel Ishimitsu to assign handles to items by obtaining a block of random handles and giving a handle to each item as it was created.


Kaipuleohone, University of Hawai'i at Mānoa




-
7. **Reach:** Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive? 
- No attempt is currently being made to provide elaborated interfaces to the data (<http://hdl.handle.net/10125/4422>, Albarillo and Thieberger 2009:3). Theoretically there can be differentiated access through ScholarSpace, but currently, everything is being mediated by Thieberger, who implemented the repository and communicates with the UH Library about access to materials on a case by case basis. The archive hopes to address [discovery and reach] simultaneously by "creating multiple access points that will facilitate discovery and make it easy to obtain digital files while respecting any access limitations requested by the depositor" (<http://hdl.handle.net/10125/4422>, Albarillo and Thieberger 2009:6). "Ideally an item discovered in ScholarSpace will [in the future]...link to the actual digital file that can be downloaded to the user's computer (of course, access to any item in the collection is subject to deposit conditions)" (<http://hdl.handle.net/10125/4422>, Albarillo and Thieberger 2009:6-7).
-
8. **Access and Use**
- Restrictions:** Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to and use of the materials will be honored? 
- Every item in the collection has access conditions specified by the depositor on the deposit form (<http://www.ling.hawaii.edu/langdoc/archive.html>, <http://www.ling.hawaii.edu/langdoc/UHKaipuleohoneDeposit.pdf>). When requesting copies of such material, users are required to submit signed access forms, which will then allow ScholarSpace to provide access to just the items requested. It is hoped that this process can be automated to some extent, with a clickable agreement form providing access to all items whose deposit conditions allow such access" (<http://hdl.handle.net/10125/4422>, Albarillo and Thieberger 2009:7). There are not many requests at present with the rationale that materials are being preserved for posterity.
-

Kaipuleohone, University of Hawai'i at Mānoa

PRESERVATION	Rating	Comments
<p>9. Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g. the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?</p>		<p>Everything is “on hold” at this time. Structures for long-term planning are assumed to be built-into ScholarSpace as a DSpace repository. Kaipuleohone conforms to international archiving standards for digital archives. Audio files are stored at high resolution and the metadata conforms to the Open Language Archives Community, Open Archives Initiative and Dublin Core. Workflow documents are available for digitizing cassettes, reel to reel tapes, and for imaging fieldnotes. All digital files are curated by the Library system at the University of Hawai'i's D-Space repository, ScholarSpace (http://www.ling.hawaii.edu/langdoc/archive.html). Ultimate responsibility for preserving the digital objects lie outside the Linguistics Department. An agreement was negotiated with UH's Hamilton Library to deposit material in ScholarSpace (http://hdl.handle.net/10125/4422, Albarillo and Thieberger 2009:6).</p>
<p>10. Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?</p>		<p>Preservation activities have been planned, but have not been needed since digitization was originally done to archival standards without many obsolescing formats. The archive can access objects and do updates as necessary.</p>
<p>11. Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g. a checksum, or digital signature, etc.)?</p>		<p>ScholarSpace reliably maintains the integrity of digital materials.</p>

Kaipuleohone, University of Hawai'i at Mānoa

- | | | |
|---|---|---|
| <p>12. Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g. by verifying that digital objects are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital objects that occur while they are in the care of the archive)?</p> |  | <p>The archive does not effectively maintain authenticity of materials.</p> |
|---|---|---|

SUSTAINABILITY	Rating	Comments
<p>13. Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g. quality media, environmentally-controlled storage, access-controlled storage area)?</p>		<p><u>Staff:</u> The archive has no staff persons at present. Digitization activities have ceased without staff. <u>Technical Infrastructure:</u> The archive is well-equipped, but is currently a static repository. In terms of equipment, the archive has playback machines for audio cassettes, DAT and minidisk (analog out only), as well as a reel-to-reel player arriving soon (http://www.ling.hawaii.edu/langdoc/archive.html).</p>
<p>14. Financial Sustainability: Does the archive appear to have secured sources of long-term funding?</p>		<p>"[T]he Kaipuleohone archive does not have guaranteed long-term funding, nor the resources required to build and maintain a digital repository" (http://hdl.handle.net/10125/4422, Albarillo and Thieberger 2009:6).</p>
<p>15. Disaster Preparedness: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g. disaster recovery plan, offsite storage of backups)?</p>		<p>It is assumed that ScholarSpace has more than one kind of built-in backup for the whole collection, but this is not true of the entire archival process (digitization, etc.).</p>

Kaipuleohone, University of Hawai'i at Mānoa

16. **Succession Plan:** Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?



There is no formal succession plan in place. However, "[a]n institutional repository, like ScholarSpace, also provides continuity and guarantees preservation of the collection if the archiving project itself ceases to function (for example due to a lack of funding or the retirement of key personnel)" (<http://hdl.handle.net/10125/4422>, Albarillo and Thieberger 2009:8).

Appendix B-5



The Division of Native American Languages (NAL) Sam Noble Oklahoma Museum of Natural History (SNOMNH) at the University of Oklahoma (OU), Norman, Oklahoma

NAL is a collection at the SNOMNH is a resource center for researchers, educators, and language advocates of Native American languages. The SNOMNH is a research division of the University of Oklahoma.



Archive Representative: **Mary Linn, Associate Curator**



Reviewer: **Debbie Chang**

Date: **6 November 2009**

TARGET	Rating	Comments
<p>1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?</p>		<p>A four-part mission statement is found on the NAL main page within the SNOMNH website (http://www.snomnh.ou.edu/collections-research/nal.htm). In the third part of the statement, "archiving and migrating materials" is listed among the services that NAL provides to Native American communities. Additionally, the museum names "stewardship of the earth and its peoples" as part of the museum's overall vision (http://www.snomnh.ou.edu/information.shtml#mission); in the context of a natural history museum, the documentation and preservation of languages takes on the significance of "language ecology."</p>
<p>2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify: _____)?</p>		<p>Submission criteria is "not easily found," though Linn stated "I've never had to turn anything down." Verbally, Linn could broadly state that NAL contains language resources (1) located in Oklahoma and (2) are Native American. The archive also accepts family-made tapes and teaching materials made by language community members in addition to formal documentation done by linguists and materials produced by "professionals." In the case of some African language materials that were clearly out of the scope of this archive, Linn opted to help the depositor find another archival home rather than accession these materials in the NAL collection.</p>

The Division of Native American Languages (NAL)

- | | | |
|---|---|---|
| <p>3. Designated Communities: Is my desired audience (specify:) a good match for the groups of users the archive targets (e.g. language community, academic community, etc.)?</p> |  | <p>The collection is open to "anyone who wants to use the collection," but primarily serves (in order of priority) (1) Native Americans (2) Academic researchers (3) Students. The metadata, however, is all in English, which may not be accessible to all members of a language community.</p> |
| <p>4. Ongoing Relationship: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)</p> |  | <p>NAL takes seriously its long-term commitments to communities and actively collaborates with donors, speakers, and organizations such as the Caddo Heritage Museum and the Kiowa Museum. Once NAL obtains a holding, main potential users in the language community are informed; these contacts include existing language programs, organizations, and individuals teaching or learning the language. NAL participates in revitalization efforts (e.g. Breath of Life). Members of the community also participate in NAL (e.g. digitization of Osh Nation materials were done by a native speaker). Though not opposed to revenue sharing, no examples of revenue sharing were cited. Linn indicated that she makes herself available for grant writing.</p> |

ACCESS	Rating	Comments
<p>5. Discoverability: Are the metadata for materials deposited at the archive posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g. Google, Yahoo!, Bing, etc.)?</p>		<p>Discoverability is "an obligation" and Linn would like to make it easier. However, as a sub-collection within the museum, the archive does not have control over a dedicated website. They are given a limited amount of Internet real estate by the museum's webmaster and museum director. Prior to my interview, Linn had taken steps for NAL to become a member of the OLAC community. At the time of my interview, the web interface to search for materials in archive on the SNOMNH website was not functional.</p>
<p>6. Fixed identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?</p>		<p>Identifiers start with a year number, but the rest of an identifier is a string of "completely arbitrary" numbers and never change. There are related fixed identifiers for a digital holding and its corresponding physical holding.</p>

The Division of Native American Languages (NAL)

7. **Reach:** Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?



NAL's central location in the state facilitates access by language groups located in Oklahoma. Visitors to the archive can request materials on premises and are afforded a spacious and comfortable research area much like a closed-stack library. However, there are some concerns about the need to interact with guards at the ground floor entrance to the building in order to gain access to the archive. There are also some concerns that the NAL website is "buried" within the museum's website.




8. **Access and Use**
Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?



NAL has adopted AILLA's access restrictions agreement which operates as a contract. It is preferable for donors not to put access restrictions on materials, as advised by Leanne Hinton of SCOIL. Material pertaining to formal societies (e.g. Kiowa Black Leggings) and chief societies (i.e. men who are born into a chief family line) can be restricted. However, NAL will not close access to certain families, but will help individuals digitize and archive materials to be restricted in such a manner "at home." The person or entity that deposits the materials with an approved restriction are expected to keep in contact; they must provide current contact information and names of people who have the authorization to make decisions over the materials if required by the restriction. No changes to access restrictions are permitted, barring a "very good reason" yet to be seen. "Licentious gossip" is taken out of user copies. When distributed, individual copies of materials are assigned a security code and download code.

NAL has a restrictive use policy, which is clearly laid out in a form that all users must sign before access to the materials is granted. The archive is also committed to prosecute those who profit from misuse of the archive's holdings.



The Division of Native American Languages (NAL)

PRESERVATION	Rating	Comments
<p>9. Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g. the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?</p>		<p>NAL follows best practices for its digital and physical holdings (e.g. digitization is done at the highest resolution available). The archive has a full-time collections manager, Terri Jordan.</p>
<p>11. Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g. a checksum, or digital signature, etc.)?</p>		<p>The IT staff of the museum does not specialize in maintaining the integrity of digital materials. No measures to ensure the integrity of materials is currently taking place within the archive.</p>
<p>12. Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g. by verifying that digital objects are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital objects that occur while they are in the care of the archive)?</p>		<p>Records that are kept to verify the authenticity of materials include: all copies made, what format, conservator reports (e.g. details of treatment of a reel-to-reel), for whom/for what/how many copies are made. Sometimes the authenticity relies on the expertise of the donor or a trained linguist (e.g. a Ph.D. graduate student is double-checking materials in the Kiowa language that she is working in); to date, verification of materials is not systematic.</p>

The Division of Native American Languages (NAL)

SUSTAINABILITY	Rating	Comments
<p>13. Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g. quality media, environmentally-controlled storage, access-controlled storage area)?</p>		<p><u>Staff:</u> In addition to the full-time collections manager and the half time audio/video technician on staff, Linn's time is split 59% as curator for NAL, and 41% as associate professor in Anthropology. She reportedly could spend all her time dealing with tribes, for whom she is a "language social worker," and also acts as a "friend to linguists." The collections manager, Ms. Jordan, is well qualified, with degrees in anthropology (BS), library science (MLS), and folklore with an emphasis in museum studies (masters). Funds are lacking, however, to hire people to transcribe all holdings. More technicians are needed to keep up with digitization requests, and NAL can no longer promise fast turnaround for digitizing materials. Specifically, the archive could use 2 full-time staff, one in audio/video editing, and one in digitization.</p>
<p>14. Financial Sustainability: Does the archive appear to have secured sources of long-term funding?</p>		<p><u>Technical Infrastructure:</u> NAL shares IPM (Integrated Pest Management) and an IT department with the museum. The collection is housed at the SNOMNH, which has state-of-the-art fire protection, and is F5 tornado-, earthquake-, and bomb-proofed. From all appearances, the museum and NAL has excellent facilities and a sound security system in place.</p> <p>The SNOMNH is one of the largest state museums of its kind, and its collections are thus relatively secure. The state is supported mostly by oil and gas revenues, and funding is not a foreseeable problem. An act of legislature recognized the natural history and languages of Oklahoma as a state treasure, so while the museum is subject to budget cuts, it is not as susceptible as some other museums. Some funding is also provided by grants.</p>

The Division of Native American Languages (NAL)

-
15. **Disaster Preparedness:** Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g. disaster recovery plan, offsite storage of backups)?
- 
- The museum's digital archives are backed up offsite (currently San Diego, California). A board of advisors consisting of Native American elders and language teachers (and which excludes elected tribal leaders) do not want holdings to leave Oklahoma, but NAL may choose to back-up data with the Max Planck Institute. The NAL staff are equipped with binders with instructions, and spend one day each year practicing what to do in the case of disaster. SNOMNH is part of the Oklahoma Museum Association Emergency Response Team that would help smaller museums.
-
16. **Succession Plan:** Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?
- 
- A formal succession plan has not been discussed with the director of the museum as the SNOMNH is thought to be the "end of the line." NAL in many cases acts as a back-up to tribal collections whose futures are subject to changing governance of any given tribe. Some healthy redundancy exists: e.g. the Caddo Nation Archives already has listening copies of everything at NAL as well as the full printed catalog of the Phil and Vynola Newkumet Collection, a large collection of Caddo language and music. Ideas for succession include giving copies of everything to the Tulsa Public Library as it is in-state, and possibly to UTA, SCOIL, or National Anthropological Archives (these options are less desirable due to their remote distance).
-

Appendix B-6 Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)



PARADISEC is a consortium of three universities: University of Sydney, University of Melbourne, and Australian National University (Canberra).

Archive Representative: **Nick Thieberger, Project Manager, University of Melbourne**



Reviewer/s: **Wayne Dye** (linguist/depositor), **Debbie Chang**

Date/s: **2 December 2009** (review of website)



18 February 2010 (phone interview)

TARGET	Rating	Comments
<p>1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?</p>		The purpose of PARADISEC is stated on its homepage (http://www.paradisec.org.au/home.html) and reflects a commitment to long-term preservation of digital materials.
<p>2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify: _____)?</p>		<p>PARADISEC accepts "endangered materials from the Pacific region, defined broadly to include Oceania and East and Southeast Asia," including digital audio and video files as stated on its homepage, although it has started to adopt a less restrictive submission policy since the infrastructure is in place to accept materials from other regions. As of December 2009 PARADISEC's collection contained 2520 hours of digital audio and video files on 4.43 TB of disk space; 614 languages from 60 countries were represented. PARADISEC does not currently have the resources to actively collect data and instead relies on the community of linguists, ethnomusicologists and ethnographers to deposit material in the archive for safe keeping and long-term accessibility (http://www.paradisec.org.au/deposit.html). Preferred deposit formats are described at: http://paradisec.org.au/deposit.html. The archive's audiovisual format standards may also be found online: http://paradisec.org.au/PARADISEC_digital_format_standards.pdf.</p>



Pacific and Regional Archive for Digital Sources in Endangered

-
3. **Designated Communities:** Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g. language community, academic community, etc.)
-  PARADISEC is committed to providing access to materials to interested communities (<http://www.paradisec.org.au/home.html>). The primary communities concerned are the performers/speakers and their descendants. Wayne Dye, a linguist and depositor at PARADISEC, has worked with the archive to develop phonemic and grammar sketches to give context and broad understandability to language data from the Bahinemo of Papua New Guinea.
-
4. **Ongoing Relationship:** Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)
-  A founding principle of PARADISEC is that small and endangered cultures need support for locating and reintroducing material that was recorded in the past (<http://www.paradisec.org.au/services.html>, Cultural renewal). Dye has indicated that the archive has been excellent at keeping the best interests of the Bahinemo community in mind, and helped to set up a revenue-sharing structure for the community. PARADISEC accepts responsibility but has no resources at present. Because many language communities are small and remote, the archive has partnerships with a local agency (e.g. a cultural studies center or museum) and make copies on CD available to these agencies, which include Institute of Papua New Guinea Studies, Tjibaou
-




Pacific and Regional Archive for Digital Sources in Endangered

ACCESS	Rating	Comments
<p>5. Discoverability: Are the metadata for materials deposited at the archive posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g. Google, Yahoo!, Bing, etc.)?</p>		<p>The archive is a participant in OLAC; the PARADISEC collection metadata can be searched via OLAC at http://www.language-archives.org/tools/search/search.php. Metadata accompanies all items in the collection. The collection is cataloged using descriptors based on Dublin Core and the Open Languages Archives Community (OLAC) recommendations, which also conform to the Open Archives Initiative guidelines. PARADISEC's current metadata set is available for download from its website (http://paradisec.org.au/downloads.html). The catalog also references items which have been assessed for eventual incorporation into the archive, but which are not currently in digital form; this permits discovery of otherwise undiscoverable resources. The goal is that any resource from the region be discoverable regardless of where it is located, and regardless of where the researcher is located. (http://www.paradisec.org.au/services.html, see Information discovery)</p>
<p>6. Fixed identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?</p>		<p>PARADISEC conforms to best practice with regards to fixed identifiers. Upon submission of materials, curators assign unique identifiers to all resources which are ingested. These identifiers are never reassigned and location independent to facilitate persistence. (http://paradisec.org.au/naming.html)</p>


Pacific and Regional Archive for Digital Sources in Endangered


-
7. **Reach:** Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?
-  Digital outputs from PARADISEC are available in various formats depending on the needs of the users. While audio files are archived at high resolution, they can be made available as MP3 or other formats for delivery on CD or over the web (<http://www.paradisec.org.au/services.html>, see Cultural renewal). Currently, deposits in PARADISEC are not freely available on the web (<http://www.paradisec.org.au/deposit.html>), though some special collections such as the scanned Capel papers have been made freely available (14K pages). There are plans, however, for the archive to become a clearinghouse for relevant content and support material (<http://www.paradisec.org.au/future.html>). A grant has been requested to build access modules via the Internet. These access modules would not currently be of great use to small, remote people groups, but could be in the future.
-
8. **Access Restrictions:** Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?
-  Though the metadata is fully searchable, access to materials requires permission (<http://www.paradisec.org.au/PDSCaccess.rtf>) which is specified for each item in the collection. Access is currently available to depositors only via password at: <http://paradisec.org.au/repository/login> (<http://www.paradisec.org.au/services.html>, see Information discovery). Normal copyright restrictions apply, and each item in the collection has its own access conditions (<http://www.paradisec.org.au/PDSCaccess.htm>), as specified by the depositor and performer using PARADISEC's deposit form (<http://www.paradisec.org.au/PDSCdeposit.rtf>). If an item is distributed, the moral rights of speakers and A notice on the publicly available fieldnotes at the archive's website (<http://www.paradisec.org.au/fieldnotes.html>) requests that the material be linked to rather than copied for further distribution so that PARADISEC's digitization work is acknowledged.
-

Pacific and Regional Archive for Digital Sources in Endangered



PRESERVATION	Rating	Comments
<p>9. Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g. the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?</p>		<p>PARADISEC personnel are subscribed to “digital preservation lists,” but overall long-term planning “depends on the efforts of a few.” A lot of attention went into how systems work, and the overall infrastructure is “good enough” such that it does not need constant maintenance. The archive has been able to successfully A more institutional framework was indicated as more desirable (e.g. such as a national data service running the archive). The archive adopts current best practice for preserving audio data by digitizing it at the highest quality available (http://www.paradisec.org.au/services.html, see Data preservation). The main focus of PARADISEC's current work is the digitization of audio files. Their Quadriga system uses the AudioCube workstation to digitize audio material at 24-bit, 96 kHz in Broadcast Wave Format (BWF); detailed work-flow charts may be found on their website (http://www.paradisec.org.au/services.html, see Technicalities). The PARADISEC ingestion workflow model can be found at: http://paradisec.org.au/pdsc-As of February 16, 2008, the archive's estimated backlog is at least 2,600 hours, which is the result of a brief initial survey, this is expected to increase as it investigates more collections of analog media (http://www.paradisec.org.au/future.html).</p>
<p>10. Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?</p>		<p>Preservation activities have been planned, but have not been needed since digitization was originally done to archival standards without many obsolescing formats. The archive can access objects and do updates as necessary.</p>
<p>11. Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g. a checksum, or digital signature, etc.)?</p>		<p>Checksums are in use. Weekly reporting on collections (length of file and checksums checked) makes sure that files sent at one end are same at other end.</p>

Pacific and Regional Archive for Digital Sources in Endangered


12. **Authenticity:** Does the archive ensure that digital materials contain what they claim to contain (e.g. by verifying that digital materials are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital holdings)?
-  Depositors are trusted to provide authentic materials and associated metadata. The archive does not have resources to do checking. Notes are made of changes to permanently associated metadata, e.g. a change in the language name or to record the return of analog materials. When a file is digitized, optional notes may be written concerning the original condition of a tape, though student researchers do not write as many notes as a professional has done in the past. Digitization is currently done without taking notes.

SUSTAINABILITY	Rating	Comments
<p>13. Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g. quality media, environmentally-controlled storage, access-controlled storage area)?</p>		<p><u>Staff:</u> PARADISEC appears to be well-run by qualified individuals. None of the core team of people are salaried, however; their work with the archive reflects their high degree of commitment. The Director, Linda Barwick, is an experienced field researcher and works with communities and linguists to produce well-documented published recordings of sung traditions. The Project Manager, Nick Thieberger, has also had significant field, digital archive, and computer programming experience. Other personnel include an Audio Preservation Officer (Aidan Wilson), Project Coordinator (Tom Honeyman), a one-day-a-week Project Liaison Officer (Amanda Harris), and a Research Fellow in Ethnomusicology (Aaron Corn) (http://paradisec.org.au/personnel.html). Digitization comprises a part-time job for one person, Aidan Wilson, in Sydney.</p> <p><u>Technical Infrastructure:</u> Dye describes a high level of confidence that the archive has state-of-the-art equipment and systems. He characterized PARADISEC as "small and human" with a "high level of sophistication" with regards to technology.</p>

Pacific and Regional Archive for Digital Sources in Endangered

-
14. **Financial Sustainability:** Does the archive appear to have secured sources of long-term funding?
-  PARADISEC got its start with a number of research grants (<http://www.paradisec.org.au/resgrants.html>). Currently, the archive has “no resources,” and has not been funded for three years though the archive anticipated intermittent funding from the beginning and set up self-sustaining structures wherever possible. Overall, the archive is not financially sustainable and is in need of grants. Two ways that PARADISEC is funded, aside from grants, are (1) charitable gifts (the archive is a registered “Deductible Gift Recipient,” which means that tax-deductible donations can be made to PARADISEC, <http://www.paradisec.org.au/funding.html>) and (2) charging “cost recovery fees” for its services to depositors who can afford to pay them. This includes digitization of audio material and training in ethnographic documentation techniques of recording, data management, and data linkage (<http://www.paradisec.org.au/services.html>). The fees for digitization services covers some of the costs. Anything left over gives “a margin to do a bit more,” but expenses are covered by digitization, which is a part-time job for one person. PARADISEC could be doing much more with more money. The archive suggests that future depositors build the costs of archiving into their grant applications since results of publicly funded research needs to be publicly available.
-
15. **Disaster Preparedness:** Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g. disaster recovery plan, offsite storage of backups)?
-  Several copies of digitized materials are stored in separate locations to mitigate the loss of an only copy in “in cyclones, fires or simply as a result of poor storage conditions” (<http://www.paradisec.org.au/services.html>, see Data Preservation). A backup version of all data is held offsite at the Australian Partnership for Advanced Computing (APAC, <http://nf.apac.edu.au/facilities/mdss/>) facility in Canberra, using the GrangeNet network to deliver the data from Sydney (<http://www.paradisec.org.au/services.html>, Technicalities). PARADISEC has successfully recovered material as needed. The archive has no mirror systems, so online access is not guaranteed if there is a catastrophic failure of the server in Sydney, but the data and metadata would be safely preserved.
-

Pacific and Regional Archive for Digital Sources in Endangered

16. **Succession Plan:** Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?
- 
- The archive has no formal succession plan. The PARADISEC Ingestion Workflow Model document is well enough described, however, that it someone else could take it and understand all the essential processes. The archive has strived to arrive at “clean, open, and simple” solutions. With three universities contributing resources, various partners could probably take it on. Presently, the operations of the archive are dependent on just a few people.
-

Appendix B-7

SIL Language and Culture Archives

SIL International, Dallas, Texas

SIL is a non-profit, faith-based organization with 75 years experience in serving the world's ethno-linguistic minority language groups.



Archive Representatives: **Jeremy Nordmoe, Director/Archivist**

Vurnell Cobbey, Archivist



Joan Spanne, Systems Administrator/Information Architect


Reviewer: **Debbie Chang**

Date: **16 September 2009**




TARGET	Rating	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?		The mission statement is found on SIL's corporate intranet, InSite, which "needs work." Only SIL members may access InSite. A public interface with the archive is available through the Bibliography posted online at the Ethnologue (http://www.ethnologue.com/biblio_docs/biblio_intro.asp), but no explicit mission statement was found at that site. The SIL archive and the Ethnologue will be undergoing significant changes in the next two years.
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify:_____)?		The submission criteria are found in the SIL Administrative Policy Manual, but needs to be updated to include unpublished materials. Submission of materials is open to SIL personnel and affiliates (e.g. organizations engaged in language development). The existence and availability of a wealth of materials dating from the inception of the organization in 1935 is made known through the Bibliography and is updated monthly. These materials include over 15,000 references to books, journal articles, book chapters, dissertations, and other academic papers about languages and cultures. The Bibliography also has about 9,500 references for materials written in minority languages (literacy books, instructional books on other basic education topics, story and folk tale books, and translated works). The mode of submission, currently either by hand, post, or e-mail, is "not as easy as it should be."

SIL Language and Culture Archives





- | | | |
|---|---|---|
| <p>3. Designated Communities: Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g. language community, academic community, etc.)</p> |  | <p>"End C," the third part of SIL's tri-fold board policy on corporate aims, prioritizes the sharing of knowledge gained through the work of the organization and identifies four communities to be benefited by this knowledge: the academic community, the church worldwide, governmental and other policy-making bodies, and the language communities. Preservation of certain materials is "mandatory" and is under corporate control. The language community may have "secondary ownership" of materials, which is currently minimal. The bibliography of holdings is available to the world.</p> |
| <p>4. Ongoing Relationship: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)</p> |  | <p>The archive presently does not make any materials available specifically to a language community, but there are plans to make materials belonging to a language community available on the Internet to that community. There is recognition that some communities will be more engaged than others in language development and language archiving, and SIL will treat each community on an individual basis. Thus far, the archive has not dealt directly with language communities but has depended on SIL personnel in the field as intermediaries, but there is an acknowledged need to look to the long-term when intermediaries are gone.</p> |

ACCESS	Rating	Comments
<p>5. Discoverability: Are the metadata for materials deposited at the archive posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g. Google, Yahoo!, Bing, etc.)?</p>		<p>The archive is a participant in OLAC.</p>



SIL Language and Culture Archives

-
- | | | |
|---|---|---|
| <p>6. Fixed identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?</p> |  | <p>The archive has assigned accession numbers as fixed identifiers since its inception. These are used as entry numbers in ethnologue.com records (a web and print reference work published by SIL that catalogues all known languages of the present-day world) and appended to the end of OAI (Open Archives Initiative) identifiers for the OLAC records. Fixed identifiers will continue to be assigned in an upcoming DSpace configuration, which will be a venue for electronic archiving and publishing.</p> |
| <hr/> | | |
| <p>7. Reach: Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?</p> |  | <p>Plans are in place to reach a global audience. Bibliography -- can obtain materials that are not available online already. Relevant resources already come up in Google searches via OLAC records and ethnologue.com records, but presently the vast majority of search results are only records that describe the existence of something. A digital library is planned and should make more items fully accessible over the Internet. The digital library project will scan many materials and accept submissions through DSpace.</p> |
| <hr/> | | |
| <p>8. Access Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?</p> |  | <p>Tiered restriction levels are in place. As part of the accessioning process, the cataloguer assigns the level according to the submitter's requests.</p> |
-



SIL Language and Culture Archives

PRESERVATION	Rating	Comments
<p>9. Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g. the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?</p>		<p>Well-documented procedures were not in evidence for digital materials. However, it was indicated that certain digital materials are marked for upgrades as the archive has a policy to convert Word documents to PDF/A, an ISO-approved archival format. There is a planned migration of digital materials to DSpace which embodies OAIS best practices.</p>
<p>10. Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?</p>		<p>The SIL archive is committed to preserving digital materials and is following industry recommendations, but formal procedures are not yet documented.</p>
<p>11. Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g. a checksum, or digital signature, etc.)?</p>		<p>No measures are currently being taken to ensure the integrity of digital materials, but using the system built into the planned installation of DSpace, checksums will be a matter of course by 2010.</p>
<p>12. Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g. by verifying that digital objects are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital objects that occur while they are in the care of the archive)?</p>		<p>The planned DSpace installation will account for actions taken on an object after ingest and will do a better job than is currently being done. The submitter will do some of the entry directly and the submitter will also upload directly. The submitter is often the field linguist or someone who has a closer relationship with a language project. A quality check by a professional is done in that case of non-linguist submissions. Currently, authentic copies of all textual materials can be verified, but this cannot be done with other kinds of media (audio, video, photographic). Physical copies of materials most often stay in the country of origin, while the SIL archives take care of the digitized versions.</p>

SIL Language and Culture Archives

SUSTAINABILITY	Rating	Comments
<p>13. Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g. quality media, environmentally-controlled storage, access-controlled storage area)?</p>		<p>The SIL archives are "getting the job done" but a lot of materials that should be submitted are not being submitted. <u>Staff:</u> There is a need for archivists in the field in different SIL entities, as well as one or two more in the SIL International office in Dallas. In August 2009 Jeremy Nordmoe, who has more than 12 years of experience working as an archivist for a major university, became the full-time director and archivist. From its inception ten years ago (October 1999), the archive was headed by Joan Spanne until October 2005 (now the systems administrator and information architect of the archive), and then by Vurnell Cobbey who continues to work at the archive. <u>Technical Infrastructure:</u> Considering the resources at hand, the SIL archives has a solid infrastructure. The decision to go to DSpace was cited as a sound one as it has a proven track record, and is open source software.</p>
<p>14. Financial Sustainability: Does the archive appear to have secured sources of long-term funding?</p>		<p>The commitment to archiving is there from the standpoint of SIL administration. The archive has had a budget allocation since its inception to maintain operations. It was established in FY 2000 as a separately functioning, budgeted department within the former Academic Affairs division of SIL (now Language Program Services), though some of its services and activities existed for roughly 20 years prior to that within the Academic Publications Department of Academic Affairs. Fundraising for the migration to DSpace has been assigned to SIL's Chief Information Officer; however, the future source and proportion of the corporate budget dedicated to archiving is yet to be determined.</p>

SIL Language and Culture Archives

-
15. **Disaster Preparedness:** Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g. disaster recovery plan, offsite storage of backups)?  There was some evidence of disaster preparedness, but no disaster recovery plan. Backup is done professionally by a vendor in greater Dallas, located in Irving which is 40 miles away from the archive. However, it is not certain if the servers containing the archival materials are backed up in Orlando, Florida, or Waxhaw, North Carolina. There is a mirror of the corporate intranet InSite; this would include a bibliography of holdings, though not the archival materials themselves.
-
16. **Succession Plan:** Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?  A succession plan for the SIL archives would be a corporate decision, and to date, the board of directors does not have a policy. Some redundancy in national archives (in the Americas, Philippines, and Australia) is known, though there is no formal tracking of materials (e.g. "last known copies"). It is uncertain if the archive in Dallas is capable of housing everything within the domain of SIL if it were given everything.
-

Appendix B-8

TAPS (Target, Access, Preservation, and Sustainability): Checklist for Responsible Archiving of Digital Language Resources

Archive: UCSD Melanesian Archive Date: 30 March 2010 Reviewer: Bob Conrad

TARGET	Yes ? No	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify: _____)?	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
3. Designated Communities: Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g., language community, academic community, etc.)?	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
4. Ongoing Relationship: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)	<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	
ACCESS	Yes ? No	Comments
5. Discoverability: Are the descriptive metadata for materials deposited at the archive searchable online? That is, the metadata is posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g., Google, Yahoo!, Bing, etc.)?	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
6. Fixed Identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?	<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	
7. Reach: Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?	<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
8. Access and Use Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?	<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	

Yes = best practice

? = in planning stage / partial practice / assumed done by others

No = not in scope of archive / unclear

PRESERVATION		Yes ?	No	Comments	
9.	Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g., the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
10.	Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
11.	Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g., a checksum, or digital signature, etc.)?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
12.	Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g., by verifying that digital materials are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital holdings)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
SUSTAINABILITY		Yes ?	No		
13.	Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g., quality media, environmentally-controlled storage, access-controlled storage area)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
14.	Financial Sustainability: Does the archive appear to have secured sources of long-term funding?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
15.	Disaster Preparedness: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g., disaster recovery plan, offsite storage of backups)?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
16.	Succession Plan: Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	

Yes = best practice

? = in planning stage / partial practice / assumed done by others

No = not in scope of archive / unclear

Appendix B-9

TAPS (Target, Access, Preservation, and Sustainability): Checklist for Responsible Archiving of Digital Language Resources

Archive: UVA Small Special Collections Library Date:30 March 2010 Reviewer:Bob Conrad

TARGET	Yes	?	No	Comments
1. Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
2. Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify: _____)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
3. Designated Communities: Is my desired audience (specify: _____) a good match for the groups of users the archive targets (e.g., language community, academic community, etc.)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
4. Ongoing Relationship: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
ACCESS	Yes	?	No	Comments
5. Discoverability: Are the descriptive metadata for materials deposited at the archive searchable online? That is, the metadata is posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g., Google, Yahoo!, Bing, etc.)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
6. Fixed Identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
7. Reach: Will the audience that I wish to reach (specify: _____) be able to access the materials once they are deposited in the archive?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
8. Access and Use Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	

Yes = best practice

? = in planning stage / partial practice / assumed done by others

No = not in scope of archive / unclear

PRESERVATION		Yes ?	No	Comments
9.	Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g., the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10.	Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
11.	Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g., a checksum, or digital signature, etc.)?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
12.	Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g., by verifying that digital materials are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital holdings)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
SUSTAINABILITY		Yes ?	No	
13.	Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g., quality media, environmentally-controlled storage, access-controlled storage area)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14.	Financial Sustainability: Does the archive appear to have secured sources of long-term funding?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15.	Disaster Preparedness: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g., disaster recovery plan, offsite storage of backups)?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16.	Succession Plan: Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Yes = best practice

? = in planning stage / partial practice / assumed done by others

No = not in scope of archive / unclear

References

- Adams, Carlisle and Steve Lloyd. 1999. Core PKI Services: Authentication, Integrity, and Confidentiality. *Understanding Public Key Infrastructure: Concepts, Standards, and Deployment Considerations*. Indianapolis: Macmillan Technical Publishing. Online version [<http://technet.microsoft.com/en-us/library/cc700808.aspx#XSLTsection126121120120>]
- ANA. 2005. *Native Languages Archives Preservation: A Reference Guide for Establishing Archives and Repositories*. Washington, D.C.: Administration for Native Americans. [<http://www.aihec.org/resources/documents/NativeLanguagePreservationReferenceGuide.pdf>]
- Avery, Leon. 2004. Mann-Whitney U Test / Wilcoxon Rank Sum Test. *Leon Avery's Home Page*. Dallas: University of Texas Southwestern Medical Center. [<http://elegans.swmed.edu/~leon/stats/utest.html>]
- Beagrie, Neil, Julia Chruszcz, and Brian Lavoie. 2008. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. Salisbury, UK: Charles Beagrie Ltd. [<http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>]
- Beagrie, Neil, Brian Lavoie, and Matthew Woollard. 2010. *Keeping Research Data Safe 2*. Salisbury, UK: Charles Beagrie Ltd. [<http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf>]
- Bird, Steven and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79(3).557–582. Preprint available at [<http://www ldc.upenn.edu/sb/home/papers/0204020/0204020-revised.pdf>]
- BRTF-SDPA. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. La Jolla, California: Blue Ribbon Task Force on Sustainable Digital Preservation and Access. [http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf]
- Cahill, Mike. 2004. From Endangered to Less Endangered: Case Histories from Brazil and Papua New Guinea. *SIL Electronic Working Papers* 2004-004. Dallas: SIL International. [<http://www.sil.org/silewp/2004/silewp2004-004.htm>]

- Caplan, Priscilla. 2004. *How to Build Your Own Dark Archive (in your spare time): A Talk for the Cornell Digital Preservation Management Workshop*. ms. [<http://www.fcla.edu/digitalArchive/pdfs/Howtobuildyourowndarkarchive.pdf>]
- CCSDS. 2002. *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, Blue Book. Washington, DC: NASA Consultative Committee for Space Data Systems. [<http://public.ccsds.org/publications/archive/650x0b1.pdf>]
- Conathan, Lisa and Andrew Garrett. 2009. *Archives, Communities, and Linguists: Negotiating Access to Language Documentation*. LSA Annual Meeting. Online version [<http://linguistics.berkeley.edu/~garrett/OLAC-2009.pdf>]
- CRL. 2010. *CRL Report on Portico Audit Findings*. Chicago: The Center for Research Libraries.
- Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press.
- . 2004. *The Language Revolution*. Cambridge, U.K.: Polity Press.
- Dalabajan, Dante. 2001. The Healing of a Tagbanua Ancestral Homeland. *Hope Takes Root: Community-based Coastal Resource Management (CBCRM) Stories from Southeast Asia*, ed. by Elmer M. Ferrer, Lenora P. de la Cruz, and Gary Newkirk, 169–193. Quezon City, Philippines: CBCRM Resource Center and Coastal Resources Research Network. [<http://www.idrc.ca/uploads/user-S/11174853201Case10.pdf>]
- DANS. 2008. *Data Seal of Approval*. The Hague: Data Archiving and Networked Services.
- DANS. 2010. *Data Seal of Approval - Quality guidelines for digital research data in the Netherlands*, Second edition, ed. by Laurents Sesink, Rene van Horik, and Henk Harmsen. The Hague: Data Archiving and Networked Services. [http://www.datasealofapproval.org/sites/default/files/Data_Seal_of_Approval_1-4_3.pdf]
- Dauenhauer, Nora Marks and Richard Dauenhauer. 1998. Technical, Emotional, and Ideological Issues in Reversing Language Shift: Examples from Southeast Alaska. *Endangered Languages: Current Issues and Future Prospects*, ed. by Lenore A. Grenoble and Lindsay J. Whaley, 57–98. Cambridge, U.K.: Cambridge University Press.
- DCC and DPE. 2007. *DCC and DPE Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*, Version 1.0 Draft for Public Testing and Comment.

Edinburgh: Digital Curation Centre and Glasgow: Digital Preservation Europe.
Register to download at [<http://www.repositoryaudit.eu/download>]

DCC and DPE. 2009. *Drambora Interactive: User Guide*. Edinburgh: Digital Curation Centre and Glasgow: Digital Preservation Europe.
[http://www.dcc.ac.uk/docs/tools/DRAMBORA_Interactive_Manual.pdf]¹

Dobrin, Lise, Peter K. Austin and David Nathan. 2007. Dying to Be Counted: Commodification of Endangered Languages. *Proceedings of Conference on Language Documentation and Linguistic Theory*, ed. by Peter K. Austin, Oliver Bond and David Nathan, 59–68. London: SOAS. Preprint
[http://www.hrelp.org/publications/ldlt/papers/dobrin_austin_nathan.pdf]

FEL. 2002. Manifesto. *Endangered Languages and Their Literatures: Proceedings of the Sixth FEL Conference, Antigua, Guatemala*, ed. by R. McKenna Brown, 155-156. Bath, U.K.: Foundation for Endangered Languages. Online version
[<http://www.ogmios.org/manifesto/index.htm>]

Ferrari, Maurizio and Dave de Vera. 2004. A Choice for Indigenous Communities in the Philippines. *Human Rights Dialogue*, Series 2 No.11 (Spring 2004): Environmental Rights.
[http://www.cceia.org/resources/publications/dialogue/2_11/online_exclusive/4457.html]

First Archivists Circle. 2007. *Protocols for Native American Archival Materials*. Flagstaff, Arizona: First Archivists Circle. [<http://www2.nau.edu/libnap-p/PrintProtocols.pdf>]

Fishman, Joshua A. 1991. *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Clevedon, Avon: Multilingual Matters Ltd.

———. 2000. Three Success Stories (More or Less): Modern Hebrew, French in Quebec and Catalan in Spain. *Can Threatened Languages Be Saved? Reversing Language Shift, Revisited: A 21st Century Perspective*, ed. by Joshua Fishman, 287–336. Clevedon, Avon: Multilingual Matters Ltd.

Garrett, Andrew, Leanne Hinton, and Eric Kaiser. 2008. The Berkeley Linguistic Archives: Archives as Community Resources. *LSA Annual Meeting*. Columbus Ohio: Linguistic Society of America.
[<http://linguistics.berkeley.edu/~garrett/lisa2008.pdf>]

¹ Link is broken May 21, 2010.

- Gibbs, W. Wayt. 2002. Saving Dying Languages. *Scientific American*, August 2002. 287(2).78–85. [<http://www.language-archives.org/documents/sciam.pdf>]
- Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel (eds.). 2006. *Essentials of Language Documentation*. Berlin: Mouton de Gruyter.
- Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas: SIL International.
- Grinevald, Colette. 2003. Speakers and Documentation of Endangered Languages. *Language Documentation and Description* Volume 1, ed. by Peter K. Austin, 52–72. London: Hans Rausing Endangered Language Project. Preprint [http://www.hrelp.org/events/workshops/eldp2008_6/resources/grinevald.pdf]
- Hale, Kenneth. 1992. Language Endangerment and the Human Value of Linguistic Diversity. *Language* 68(1).35–42.
- Harrison, David K. 2007. *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*. New York: Oxford University Press.
- Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 165–191.
- Hinton, Leanne and Kenneth Hale. 2001. *The Green Book of Language Revitalization in Practice*. San Diego: Academic Press.
- Hinton, Leanne, Matt Vera, and Nancy Steele. 2002. *How to Keep Your Language Alive: A Commonsense Approach to One-on-One Language Learning*. Berkeley, California: Heyday Books.
- Hinton, Leanne. 2005. What to preserve: A viewpoint from linguistics. *Native Languages Archives Preservation: A Reference Guide for Establishing Archives and Repositories*, 24–26. Washington, D.C.: Administration for Native Americans. [<http://www.aihec.org/resources/documents/NativeLanguagePreservationReferenceGuide.pdf>]
- ISO. 2003. *Space Data and Information Transfer Systems—Open Archival Information System—Reference Model* ISO 14721:2003. Geneva: International Organization for Standards. [http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683]
- Kenney, Anne R. and Ellie Buckley. 2005. Developing Digital Preservation Programs: the Cornell Survey of Institutional Readiness, *RLG DigiNews*, August 15, 2005. [<http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file3006.html>]

- Kipp, Darrell R. General Guidelines. 2005. *Native Languages Archives Preservation: A Reference Guide for Establishing Archives and Repositories*, 23–24. Washington, D.C.: Administration for Native Americans.
[<http://www.aihec.org/resources/documents/NativeLanguagePreservationReferenceGuide.pdf>]
- Krauss, Michael E. 1992. The World's Languages in Crisis. *Language* 68(1).4–10.
- Landweer, Lynn. 2009. Indicators of Relative Ethnolinguistic Vitality. *SIL PNG Viability Workshop: Markers of Language Maintenance and Shift*. Dallas: SIL International.
- Lavoie, Brian. 2004. Of Mice and Memory: Economically Sustainable Preservation for the Twenty-first Century. *Access in the Future Tense*, 45–54. Washington, DC: Council on Library and Information Resources.
[www.clir.org/pubs/reports/pub126/pub126.pdf]
- Lavoie, Brian and Lorcan Dempsey. D-Lib Magazine. 2004. Thirteen Ways of Looking at...Digital Preservation. *D-Lib Magazine* 10(7/8), July/August 2004.
[<http://www.dlib.org/dlib/july04/lavoie/07lavoie.html>]
- Leggett, John J. 2005. *Preservation, New Media, Oral Cultures*. College Station, Texas: Texas A&M University.
[www.csdl.tamu.edu/~leggett/courses/dl/slides/Chapter9.ppt]
- Lewis, M. Paul. 2009. *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Texas: SIL International. Online version [<http://www.ethnologue.com>]
- Liberman, Mark. 2001. Panel 1: Legal, Ethical, and Policy Issues Concerning the recording and Publication of Primary Language Materials. *Exploration 2000 Workshop: Web-Based Language Documentation and Description* December 12, 2000, updated June 3, 2001.
[<http://www ldc.upenn.edu/exploration/exp12000/papers/liberman/liberman.html>]
- Lynch, Clifford A. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL BiMonthly Report* 226.
[<http://www.arl.org/resources/pubs/br/br226/br226ir.shtml>]
- Maxino, Theresa. 2006. Revisiting Fletcher and Adler Checksums. *DSN 2006 Student Forum*. Pittsburgh: Carnegie Mellon University.
[http://www.zlib.net/maxino06_fletcher-adler.pdf]

- Mifflin, Jeffrey. 2008. Saving a Language: a Rare Book in MIT's Archives Helps Linguists Revive a Long-unused Native American Language. *Technology Review*, May/June 2008.M16–M17. Online version: [<http://www.technologyreview.com/article/20629/>]
- Nathan, David. 2009. Archiving and Language Documentation: from Diskspace to MySpace. *3L Summer School*. London: School of Oriental and African Studies.
- . In press. Archives 2.0 for Endangered Languages: from Disk Space to MySpace. *International Journal of Humanities and Arts Computing*, Volume 4 (Special issue), 2010.
- NESTOR. 2006. *Catalogue of Criteria for Trusted Digital Repositories Version 1 (draft for public comment)*. Frankfurt am Main: Network of Expertise in long-term STORage and long-term availability of digital Resources. [<http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf>]
- NESTOR. 2008. *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive Version 2*. Frankfurt am Main: Network of Expertise in long-term STORage and long-term availability of digital Resources. [<http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf>]
- Niedzielski, Henry Z. 1992. The Hawaiian Model for the Revitalization of Native Minority Cultures and Languages. *Maintenance and Loss of Minority Languages*, ed. by Willem Fase, Jaspaert Koen, and Sjaak Kroon, 269–384. Volume 1, Maintenance and Loss of Minority Languages. Amsterdam: John Benjamins Publishing Co.
- OCLC and CRL. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist, Version 1.0*. Dublin, Ohio: Online Computer Library Center, Inc. and Chicago: The Center for Research Libraries. [http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf]
- OCLC. 2007. *Trustworthy Repositories Audit and Certification Checklist is Published*. Dublin, Ohio: Online Computer Library Center, Inc. [<http://www.oclc.org/research/news/2007-03-12.htm>]
- OWASP. 2002. Authentication. *A Guide to Building Secure Web Applications*. Open Web Application Security Project. [<http://www.cgisecurity.com/owasp/html/ch06.html>]
- Pandikar, Tan Sri Datuk Pandikar Amin Haji Mulia. 2003. Language Development and Revitalization in a South East Asian community: An Insider's Perspective. *Conference on Language Development, Language Revitalization and Multilingual*

Education in Minority Communities in Asia, 6-8 November 2003: Plenary Presentations, 1–11. Bangkok, Thailand: SIL International, the Institute of Language and Culture for Rural Development of Mahidol University, Salaya (Thailand), and UNESCO.

[http://www.sil.org/asia/lcdc/plenary_papers/tan_sri_datuk.pdf]

Rahman, Tariq. 2003. *Language Policy, Multilingualism and Language Vitality in Pakistan*. Dallas: SIL International.

[http://www.sil.org/asia/lcdc/parallel_papers/tariq_rahman.pdf]

RLG and OCLC. 2002. *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report*. Mountain View, California: Research Libraries Group and Dublin, Ohio: Online Computer Library Center, Inc.

[<http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf>]

RPI. 2010. *Blue Ribbon Task Force Report: Preserving Our Digital Knowledge Base Must be a Public Priority*. Troy, New York: Rensselaer Polytechnic Institute.

[http://news.rpi.edu/_update.do?artcenterkey=2691]

SEM. 2001. *A Manual for Documentation Fieldwork and Preservation for Ethnomusicologists*, Second Edition. Bloomington, Indiana: The Society for Ethnomusicology.

Sherzer, Joel. 2002. Archive of the Indigenous Languages of Latin America. *Endangered Languages and Their Literatures: Proceedings of the Sixth FEL Conference*, ed. by R. McKenna Brown, 7–11. Bath: Foundation for Endangered Languages.

Simons, Gary F. 2006. Ensuring That Digital Data Last: the Priority of Archival Form Over Working Form and Presentation Form. *SIL Working Papers* 2006-003.

[<http://www.sil.org/silewp/abstract.asp?ref=2006-003>]

Slaughter, Inee Yang. 2005. What are the priorities? Why prioritize? *Native Languages Archives Preservation: A Reference Guide for Establishing Archives and Repositories*, 26–27. Washington, D.C.: Administration for Native Americans.

[<http://www.aihec.org/resources/documents/NativeLanguagePreservationReferenceGuide.pdf>]

Smith, Abby. 2004. Mapping the Preservation Landscape. *Access in the Future Tense*, 1–8. Washington, DC: Council on Library and Information Resources.

[www.clir.org/pubs/reports/pub126/pub126.pdf]

Task Force on Archiving of Digital Information. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information Commissioned by the Commission on Preservation and Access and the Research Libraries Group*.

- Washington, DC: Commission on Preservation and Access.
[<http://www.clir.org/pubs/reports/pub63watersgarrett.pdf>]
- UCSD. 2009. Information for Potential Donors. San Diego: Regents of the University of California. [<http://libraries.ucsd.edu/locations/ssh/resources/featured-collections/melanesian-studies-resource-center/information-for-potential-donors.html>]
- UVA. September, 2002. Collection Development Policy. Charlottesville, Virginia: University of Virginia Library.
[<http://www2.lib.virginia.edu/small/collections/policy.html>]
- Witten, Ian H. and David Bainbridge. 2002. *How to Build a Digital Library*. San Francisco: Morgan Kaufmann Publishers.
- Woodbury, Tony. 2003. Defining Documentary Linguistics. *Language Documentation and Description* Volume 1, ed. by Peter K. Austin, 35–51. London: Hans Rausing Endangered Language Project.
- Wurm, Stephen A. 1991. Language Death and Disappearance: Causes and Circumstances. *Endangered Languages*, ed. by Robert H. Robins and Eugenius M. Uhlenbeck, 1–18. Oxford: Berg Publishers Limited.
- Yamamoto, Akira Y. 1998. Retrospect and Prospect on New Emerging Language Communities. *Endangered Languages: What Role for the Specialist?: Proceedings of the Second FEL Conference*, ed. by Nicholas Ostler, 113–120. Bath, U.K.: Foundation for Endangered Languages.

VITA

Debbie Chang was born in China. At the age of two, she immigrated with her family to the United States where she grew up. She graduated with honors from Gainesville High School in Gainesville, Florida. She holds a B.S. in Architectural Design (2000) from the Massachusetts Institute of Technology. After finishing her undergraduate studies, she worked as a project manager for five years at S+H Construction, Inc. in Cambridge, Massachusetts. She also taught English as a second language in Boston's Chinatown for American Chinese Christian Educational and Social Services, Inc. She left the Boston area in 2005 to teach undergraduate English majors in Beijing. She has been trained in applied linguistics at the SIL program at the University of North Dakota (SIL-UND) and the Graduate Institute of Applied Linguistics in Dallas, Texas. She served as a teaching assistant in Second Language Acquisition at SIL-UND in 2007, and was part of the Student Body Association at GIAL from 2006 to 2008. Beginning in 2007, she has been a graduate research assistant for the Open Language Archives Community project under the supervision of Gary F. Simons. She has been an active member of Cambridge Community Fellowship Church, Beijing International Christian Fellowship, and Gainesville First Church of the Nazarene in Florida. Since coming to Dallas in 2006, she has been a part of Grace Community Church in Arlington, Texas. She speaks English, and to lesser degrees the Shanghainese and Mandarin dialects of Chinese. She currently enjoys spending time with and learning from refugees in Fort Worth, Texas. She may be contacted at debbiechang at gmail dot com.