Seven Dimensions of Portability for Language Documentation and Description

Steven Bird, Gary Simons

Language, Volume 79, Number 3, September 2003, pp. 557-582 (Article)

Published by Linguistic Society of America
DOI: 10.1353/lan.2003.0149

SEVEN DIMENSIONS OF PORTABILITY FOR LANGUAGE DOCUMENTATION AND DESCRIPTION

STEVEN BIRD
University of Pennsylvania and University of Melbourne

GARY SIMONS
SIL International

The process of documenting and describing the world's languages is undergoing radical transformation with the rapid uptake of new digital technologies for capture, storage, annotation, and dissemination. While these technologies greatly enhance our ability to create digital data, their uncritical adoption has compromised our ability to preserve this data. Consequently, new digital language resources of all kinds—lexicons, interlinear texts, grammars, language maps, field notes, recordings—are proving difficult to reuse and less portable than the conventional printed resources they replace. This article is concerned with the portability of digital language resources, specifically with their ability to transcend computer environments, scholarly communities, domains of application, and the passage of time. We review existing software tools and digital technologies for language documentation and description, and analyze portability problems in the seven areas of CONTENT, FORMAT, DISCOVERY, ACCESS, CITATION, PRESERVATION, and RIGHTS. We articulate the values that underlie our intuitions about good and bad practices, and lay out an extensive set of recommendations to serve as a starting point for the community-wide discussion that we envisage.*

1. INTRODUCTION. LANGUAGE DOCUMENTATION provides a record of the linguistic practices of a speech community, such as a collection of recorded and transcribed texts. LANGUAGE DESCRIPTION, on the other hand, presents a systematic account of the observed practices in terms of linguistic generalizations and abstractions, such as in a grammar or analytical lexicon.[1] It is now easy to collect vast quantities of language documentation and description and store them in digital form. It is easy to transcribe the material using appropriate scripts, to organize it into databases, and to link it to linguistic descriptions. It is also easy to disseminate rich language resources on the internet. Yet how can we ensure that this digital language documentation and description can be reused by others, both now and in the future?

Today's linguists can access printed and handwritten documentation that is hundreds (sometimes thousands) of years old. However, much digital language documentation and description becomes inaccessible within a decade of its creation. Linguists who have been quick to embrace new technologies, create digital materials, and publish them on the web soon find themselves in technological quicksand. Funded documentation projects are usually tied to software versions, file formats, and system configurations having a lifespan of three to five years. Once this infrastructure is no longer tended, the language documentation is quickly mired in obsolete technology. The issue is acute for endangered languages. 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. When the technologies die, unique heritage is either lost or encrypted. Fortunately, linguists can follow BEST PRACTICES in digital language documentation and description, greatly increasing the likelihood that their work will survive in the long term.

* This research was supported by NSF Grant No. 9983258 'Linguistic Exploration' and Grant No. 9910603 'International Standards in Language Engineering (ISLE)'. We are grateful to Dafydd Gibbon, David Nathan, Nicholas Ostler, and the *Language* editors and anonymous referees for comments on earlier versions of this article.

[1] For a lucid discussion of the terms 'language documentation' and 'language description' we refer the reader to Himmelmann 1998.

If digital language documentation and description should transcend time, they should also be reusable in other respects: across different software and hardware platforms, across different scholarly communities (e.g. field linguistics, language pedagogy, language technology), and across different purposes (e.g. research, teaching, development). In this article we address all these facets of the problem under the heading of PORTABILITY. Portability is usually viewed as an issue for software, but here our focus is on data. By 'data' we mean any information that documents or describes a language, such as a published monograph, a computer data file, or even a shoebox full of handwritten index cards. The information could range in content from unanalyzed sound recordings to fully transcribed and annotated texts to a complete descriptive grammar.

This article addresses seven dimensions of portability for digital language documentation and description, identifying problems, establishing core values, and proposing best practices. The article begins with a survey of existing tools and technologies, leading to a discussion of the problems that arise with the resources created using these tools and technologies. We identify seven kinds of portability problems under the headings of CONTENT, FORMAT, DISCOVERY, ACCESS, CITATION, PRESERVATION, and RIGHTS. Next we give statements about core values in digital language documentation and description, leading to a series of VALUE STATEMENTS that serve as requirements for best practices. Finally, we discuss OLAC, the Open Language Archives Community, which provides a process for identifying community-agreed best practices, and lay out an extensive set of recommendations to serve as a starting point for the community-based effort that we envision.

The structure of the article is designed to build consensus. Readers who take issue with a best practice recommendation in §6 are encouraged to review the corresponding statement of values in §4 and either suggest a different practice that better implements the values, or else propose a more appropriate value statement. The reader could turn further back to the corresponding problem statement in §3 and offer a critique of the analysis of the problems. In this manner, any disagreement about recommendations will lead to deeper understanding of the problems with current practice in the community, and to greater clarity about the community's values.

2. TOOLS AND TECHNOLOGIES FOR LANGUAGE DOCUMENTATION AND DESCRIPTION. Language documentation projects are relying more and more on new digital technologies and software tools. This section surveys a broad range of current practices, covering general-purpose software, specialized tools, and digital technologies. This snapshot of how digital language documentation and description are created and managed in practice provides a backdrop for our later analysis of data portability problems.

2.1. GENERAL PURPOSE TOOLS. Most computer-based language documentation work uses conventional office software. This software is readily available, often preinstalled, and familiar. Word processors have often been used in creating large dictionaries, such as a Yoruba lexicon with 30,000 entries split across twenty files (Yiwola Awoyale, p.c. 1998). Frequently cited benefits are WYSIWYG editing (i.e. 'what you see is what you get'), the find/replace function, the possibility of cut-and-paste to create sublexicons, and the ease of publishing. On the down side, a large fraction of the linguist's time is spent on maintaining consistency of both content and format or on finding ways to work around the lack of consistency. Word processors have also been used for interlinear text, with three main approaches: fixed-width fonts with hard spacing, manual setting of tab stops, and tables.[2] All methods require manual line-breaking, and

[2] http://www.linguistics.ucsb.edu/faculty/cu mming/WordForLinguists/Interlinear.htm

significant additional labor on presentation if line width or point size are ever changed. Another kind of office software, the spreadsheet, is often used for wordlists or paradigms.

Language documentation created using office software is normally stored in a proprietary format that is unsupported within five to ten years. While other export formats are supported, they may lose some of the structure. For instance, part of speech may be distinguished in a lexical entry through the use of italics, and this information may be lost when the data is exported to a nonproprietary plain-text format. Also, the portability of export formats may be compromised by being laden with presentational markup.

A second category of general-purpose software is hypertext processors. Perhaps the first well-known application to language documentation was the Macintosh hypercard stack that appeared in the late 1980s for *Sounds of the world's languages*, later published on the web[3] and on CD-ROM (Ladefoged 2000). More recently, the HTML standard coupled with universal, free browsers has encouraged the creation of large amounts of hypertext for a variety of documentation types. For instance, we have interlinear text with HTML tables (e.g. Peter Austin's Jiwarli fieldwork[4]), interlinear text with HTML frames (e.g. M. Eleanor Culley's presentation of Apache texts[5]), HTML markup for lexicons with hyperlinks from glossed examples and a thesaurus (e.g. Peter Austin and David Nathan's Gamilaraay lexicon[6]), gifs for representing IPA transcriptions (e.g. Steven Bird's Dschang tone paradigms[7]), and Javascript for image annotations (e.g. Bill Poser's annotated photographs of gravestones engraved with Déné syllabics[8]). In all these cases, HTML is used as the primary storage format, not simply as a view on an underlying database. The intertwining of content and format clearly makes this kind of language documentation difficult to maintain and reuse.[9]

The third category of general-purpose software is database packages. In the simplest case, the creator shares the database with others by requiring them to purchase the same package, and by shipping them a full dump of the database (e.g. the StressTyp database, which requires users to buy a copy of '4th Dimension'[10]). In other cases the dump is provided in a portable format, such as tab-delimited files or a set of SQL commands. A more popular approach is to host the database on a web-server and create a forms-based interface that allows remote users to search the database without installing any software (e.g. the Comparative Bantu Online Lexical Database[11] and the Maliseet-Passamaquoddy Dictionary[12]). Some databases support updates via the web (e.g. the Berkeley Interlinear Text Collector[13] and the Rosetta Project's site for uploading texts, wordlists, and descriptions[14]).

[3] http://hctv.humnet.ucla.edu/departments/li nguistics/VowelsandConsonants/

[4] http://www.linguistics.unimelb.edu.au/resear ch/projects/jiwarli/gloss.html

[5] http://etext.lib.virginia.edu/apache/ChiMe sc2.html

[6] http://coombs.anu.edu.au/WWWVLPages/Aborig Pages/LANG/GAMDICT/GAMDICT.HTM

[7] http://www.ldc.upenn.edu/sb/home/papers; shLDC2003S02

[8] http://www.ydli.org/dakinfo/dulktop.htm

[9] Our purpose in citing specific examples is not to single them out for criticism, but to show how serious work by conscientious scholars has grappled with a host of technical problems in the course of exploring a large space of imperfect solutions.

[10] http://www.let.leidenuniv.nl/ulcl/pil/stresstyp/

[11] http://www.cbold.ddl.ish-lyon.cnrs.fr/

[12] http://ultratext.hil.unb.ca/Texts/Malis eet/dictionary/index.html

[13] http://ingush.berkeley.edu:7012/BITC.html

[14] http://www.rosettaproject.org:8080/live/

2.2. SPECIALIZED TOOLS. Over the last two decades, several dozen tools with specialized support for language documentation and description have been developed; a representative sample is listed here.[15] Tools for linguistic data management include Shoebox[16] and the Fieldworks Data Notebook.[17] Speech analysis tools include Praat[18] and SpeechAnalyzer.[19] Many specialized signal annotation tools have been developed, including CLAN,[20] EMU,[21] and the Annotation Graph Toolkit[22] (including TableTrans, InterTrans and TreeTrans). There are many orthographic transcription tools, including Transcriber[23] and MultiTrans.[24] There are morphological analysis tools, such as the Xerox Finite State toolkit[25] and SIL's PC-Parse tools.[26] There are a wealth of concordance tools. Finally, some integrated multifunction systems have been created, such as LinguaLinks Linguistics Workshop.[27] The interested reader is referred to Antworth & Valentine 1998 for a full-length article on this topic.

In order to do their specialized linguistic processing, each of these tools depends on some model of linguistic information. All kinds of linguistic information—for example, time-aligned transcriptions, interlinear texts, syntax trees, lexicons—require suitable data structures and file formats. Given that most of these specialized tools have been developed in isolation, the models and formats are typically incompatible. For example, data created with an interlinear text tool cannot be subsequently annotated with syntactic information without losing the interlinear annotations. When interfaces and formats are open and documented, it is occasionally possible to cobble the tools together in support of a more complex need. However, the result is a series of increasingly baroque and decreasingly portable approximations to the desired solution. In consequence, specialized computational support for language documentation and description is in a state of disarray.

2.3. DIGITAL TECHNOLOGIES. A variety of digital technologies are used in language documentation owing to sharply declining hardware costs. These include technologies for digital signal capture (audio, video, physiological) and signal storage (hard disk, CD-R, DVD-R, DAT, minidisc).

Software technologies are also playing an influential role as new standards are agreed. At the micro level we have the simple hyperlink, which can connect linguistic descriptions to underlying documentation, for example, relating an analytical transcription to a recording. Hyperlinks streamline the descriptive process. Transcriptions can be checked with mouse clicks instead of unearthing an old tape or finding a speaker of the language. Hyperlinks help to organize the documentation, bringing temporally and

[15] Further examples may be found on SIL's page on *Linguistic computing resources* (http://www.sil.org/linguistics/computing.html) on the *Linguistic exploration* page (http://www.ldc.upenn.edu/exploration/), and on the *Linguistic annotation* page (http://www.ldc.upenn.edu/annotation/).

[16] http://www.sil.org/computing/shoebox/

[17] http://fieldworks.sil.org/

[18] http://fonsg3.hum.uva.nl/praat/

[19] http://www.sil.org/computing/speechtool s/speechanalyzer.htm

[20] http://childes.psy.cmu.edu/

[21] http://www.shlrc.mq.edu.au/emu/

[22] http://sf.net/projects/agtk/

[23] http://www.etca.fr/CTA/gip/Projets/ Transcriber/

[24] http://sf.net/projects/agtk/

[25] http://www.xrce.xerox.com/competencies/ content-analysis/fst/

[26] http://www.sil.org/computing/catalog/ pc-parse.html

[27] http://www.sil.org/LinguaLinks/LingWksh .html

spatially separated documentation together, and permitting a single artifact to play a role in multiple descriptions. This continual rearrangement of evidence is an important part of the analytic process.

At the macro level, software technologies and standards have given rise to the internet, which facilitates collaboration in the construction of language resources and low-cost dissemination of the results. Notably, it is portability problems that prevent the basic digital technologies from having their full impact. Thus, while the internet makes it easy to download a language resource, the would-be user may still face a daunting amount of set-up work before being able to derive the full benefits of that resource. The following download instructions for the Sumerian lexicon[28] illustrate the complexities (hyperlinks are underlined):

> Download the Sumerian Lexicon as a Word for Windows 6.0 file in a self-extracting WinZip archive.
> Download the same contents in a non-executable zip file.
> Includes version 2 of the Sumerian True Type font for displaying transliterated Sumerian. Add the font to your installed Windows fonts at Start, Settings, Control Panel, Fonts. To add the Sumerian font to your installed Windows fonts, you select File and Add New Font. Afterwards, make sure that when you scroll down in the Fonts listbox, it lists the Sumerian font. When you open the SUMERIAN.DOC file, ensure that at File, Templates, or at Tools, Templates and Add-Ins, there is a valid path to the enclosed SUMERIAN.DOT template file. If you do not have Microsoft's Word for Windows, you can download a free Word for Windows viewer at Microsoft's Web Site.
> Download Macintosh utility UnZip2.0.1 to uncompress IBM ZIP files. To download and save this file, you should have Netscape set in Options, General Preferences, Helpers to handle hqx files as Save to Disk. Decode this compressed file using Stuffit Expander.
> Download Macintosh utility TTconverter to convert the IBM format SUMERIAN.TTF TrueType font to a System 7 TrueType font. Decode this compressed file using Stuffit. Microsoft Word for the Macintosh can read a Word for Windows 6.0 document file. There is no free Word for Macintosh viewer, however.

The complexities illustrated in these download instructions are often encountered. Moreover, the ability of a technically savvy user to handle such complexities offers no guarantee that the software will actually work in that user's environment. For instance, the user could have a hardware or system software configuration that is substantially different than the one on which the resource was developed. Clearly our technologies for storing and delivering language resources fall far short of our need for easy reuse.

2.4. DIGITAL ARCHIVES. Recently several new digital archives of language documentation and description have been established, such as the Archive of the Indigenous Languages of Latin America,[29] and the Rosetta Project's Archive of 1000 Languages.[30] These exist alongside older archives that are in various stages of digitizing their holdings, for instance: the Archive of the Alaska Native Language Center,[31] the LACITO Linguistic Data Archive,[32] and the US National Anthropological Archives.[33] These archives and many others are surveyed on the *Language Archives* page.[34] Under the aegis of OLAC, the *Open Language Archives Community*,[35] the notion of language

[28] http://www.sumerian.org/

[29] http://www.ailla.org/

[30] http://www.rosettaproject.org/

[31] http://www.uaf.edu/anlc/

[32] http://lacito.vjf.cnrs.fr/archivage/

[33] http://www.nmnh.si.edu/naa/

[34] http://www.ldc.upenn.edu/exploration/ar chives.html

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

archive has been broadened to include corpus publications by organizations like the Linguistic Data Consortium[36] and archives of linguistic software like the Natural Language Software Registry.[37]

Conventional language archives face many challenges, the most significant being the unfortunate reality that data preservation is not as attractive to sponsors as data creation. Other challenges may include: identifying, adapting, and deploying digital archiving standards; setting up key operational functions such as processing digital submissions, offsite backup, and migration to new digital formats and media over time; supporting new access modes (e.g. search facilities) and delivery formats (e.g. streaming media); and obtaining the long-term backing of an established institution that can credibly commit to providing preservation and access over the long term.

This survey, brief and incomplete as it is, makes clear that there is an abundance of tools and technologies for language documentation and description, and that the community is impressively adept at creating digital data. Yet the snapshot also reveals an embarrassing level of digital detritus. Expensive data cannot be reused, or else it requires a major recycling effort to salvage the valuable pieces.

Computers are not to blame for all problems of portability in language documentation and description, however; many portability problems predate the digital era. No earlier generation of linguists was able to be confident of discovering, accessing, and interpreting all relevant language resources. While the digital revolution has exacerbated some portability problems, particularly in such areas as format, citation, and preservation, it has simultaneously provided new, promising solutions to these older open problems, along with efficient processes for geographically dispersed communities to reach consensus about best practice. In the next section we consider an extensive set of portability problems under seven headings encompassing both digital and nondigital practices.

3. SEVEN PROBLEM AREAS FOR PORTABILITY. During the rapid uptake of new digital technologies described in §2, many creators of language documentation and description have turned a blind eye to the issue of portability. Unfortunately, as a direct consequence of this, the fruits of their labors are likely to be unusable within five to ten years. In this section we identify seven problem areas for the portability of language documentation and description. While the tone of the discussion is negative, a full and frank assessment is necessary before we can articulate the core values that are being compromised by current digital and nondigital practices.

3.1. CONTENT. By CONTENT we mean the information content of the resource. The area of content involves three key concepts: the breadth and depth of COVERAGE, ACCOUNTABILITY for conclusions reached in description, and the TERMINOLOGY used in description.

COVERAGE. The language documentation community has been active since the nineteenth century (even earlier in some cases[38]), collecting wordlists and texts, and writing descriptive grammars. With the arrival of the digital era, we can transfer the endeavor from paper to word processor and carry on as before. However, new technologies provide opportunities to create new kinds of language resources. We can make digital multimedia recordings of rich linguistic events, documenting endangered languages

[36] http://www.ldc.upenn.edu/

[37] http://registry.dfki.de/

[38] Celebrated early grammarians include Pāṇinī (5th century BC), Dionysius of Thrace (2nd century BC), and Hesychius of Alexandria (5th century AD).

and genres, and fortuitously capturing items that turn out to be crucial for later analysis. However, even when extensive multimedia recordings are made, they may be of low quality (e.g. poor microphone placement, bad lighting), or they may not represent a balanced collection (e.g. twenty recordings of the same genre). Each of these weaknesses in coverage limits our ability to interpret the content. Many senses, collocations, and constructions will be missed or else unique, and we will not have a corpus from which we can draw reliable conclusions.

ACCOUNTABILITY. The content of a description is difficult to verify when it cannot be checked against the language documentation on which it is based. For example, if the reported phonetic transcription of a word contradicts the known phonotactic properties of the language, could this be a typographical error, a difference in transcription practice, or a bona fide exception? Similarly, incompatible descriptions cannot be reconciled when the documentation is unavailable. Without accountability, problems of interpretation may only be resolved by contacting the author or by locating speakers of the same speech variety (and that only when the point in question does not derive from the idiosyncratic performance of the original source), and these problems may present significant obstacles to the reuse of the language description. Accountability is also an issue for documentation: heavy editing of recorded materials may give an artificial or even misleading impression of the original linguistic event.

TERMINOLOGY. Many potential users of language data are interested in assimilating multiple descriptions of a single language to gain an understanding of the language that is as comprehensive as possible. Many users are interested in comparing the descriptions of different languages in order to apply insights from one analysis to the analysis of another language, or to test a typological generalization. However, two descriptions may be difficult to compare or assimilate because they have used terminology differently.

Language documentation and description of all types depend critically on technical notation and vocabulary, and ambiguous or unknown terms compromise portability. For instance, the symbols used in phonetic transcription have variable interpretation depending on the descriptive tradition: 'it is crucial to be aware of the background of the writer when interpreting an unexplained occurrence of [y]' (Pullum & Ladusaw 1986:168). In morphosyntax, the term 'absolutive' can refer to one of the cases in an ergative language, or to the unpossessed form of a noun as in the Uto-Aztecan tradition (Lewis et al. 2001:151), and a correct interpretation of the term depends on an understanding of the linguistic context.

The existence of variable or unknown terms leads to problems for retrieval. Suppose that a linguistic typologist wanted to search the full-text content of a large collection of data from many languages in order to discover which languages have a particular trait. Since the terms are not standardized, the user will discover irrelevant documents (low precision) and will fail to discover relevant documents (low recall). In order to carry out a comprehensive search, the user must know all the ways in which a particular phenomenon is described. Even once a set of descriptions is retrieved, users will generally not be able to make reliable comparisons between the descriptions of different languages without studying them in detail. We will return to this topic when we discuss the problem of discovery.

3.2. FORMAT. By FORMAT we mean the manner in which the information is represented electronically. The area of format involves four key concepts: the OPENNESS of the format, the ENCODING of characters within textual information, the MARKUP of structure in the information, and the RENDERING of information in human-readable displays.

OPENNESS. Language data frequently ends up in a secret proprietary format. To use such data one must typically purchase commercial software from the company that developed the format, then install it on the same hardware and under the same operating system as used by the person who created the data. By contrast, an open format is one for which the specifications are open to the public, and thus software is available from multiple sources (including noncommercial ones) and on multiple platforms.

ENCODING. Encoding is the property of textual data that has to do with how the characters are represented as numerical codes in storage (as opposed to how they are keyboarded, or how they are rendered on the screen [Becker 1984]). This has been a perennial problem for linguists who need to encode characters that are not part of the standard character sets that are supported by common software, whether these be 'special' characters that occur in the orthographies of little-studied languages or symbols that are used in phonetic transcription. In the void left by the lack of standards, linguists have devised a variety of ingenious solutions, including using combinations of available characters to transliterate unavailable ones and devising new character sets that assign the needed characters to specific numerical codes. The portability of such solutions depends critically on the transmission of documentation that explains the encoding schemes. The emergence of Unicode[39] as a character encoding standard for all the major orthographic systems of the world (including the International Phonetic Alphabet) holds much promise. But even Unicode has portability problems when the characters that linguists need are not covered by the standard and they are forced to use the Private Use Area to encode custom characters.

MARKUP. Markup is the property of textual data that has to do with how the information above the character strings themselves is represented. For instance, in a dictionary the markup has to do with identifying the various parts of the dictionary entries. The purpose of markup is to support format conversion, database storage, and query. In a word processor, a linguist might switch fonts (such as from normal face to bold face) to indicate a particular part of the entry (such as the part of speech), as shown in 1a. This is the least portable markup of all, since such binary formatting can easily be lost when the file format is converted. Another approach to markup using a conventional word processor is for the linguist to use punctuation marks in a disciplined way (e.g. putting square brackets around the part of speech in a lexical entry), as shown in 1b. However, when maintaining complex entries it is easy to introduce a formatting error (e.g. omitting a closing bracket), with unpredictable consequences for the software used for converting, storing, or querying that data.

(1) a. *chien* **n** dog.
b. chien: [n] dog.

A more robust approach to markup is to introduce special strings of characters (called MARKERS or TAGS) into the stream of text. For instance, the Shoebox program uses markers that begin with a backslash to mark the beginnings of information elements, as shown in 2.

(2) \ent chien
\pos n
\def dog

An even more robust approach to markup uses balancing tags to mark both the beginning and end of each information element. Two examples are shown in 3. These follow

[39] http://www.unicode.org/

the markup convention first established in SGML, the Standard Generalized Markup Language,[40] of placing tags in angle brackets and using a slash within the tag to indicate the balancing end tag for the start tag of the same name.

(3) a. ⟨p⟩⟨font size = + 1⟩⟨i⟩chien⟨/i⟩⟨/font⟩
⟨b⟩n⟨/b⟩ ⟨font color = blue⟩dog.⟨/font⟩⟨/p⟩

b. ⟨entry⟩
⟨headword⟩chien⟨/headword⟩
⟨pos⟩n⟨/pos⟩
⟨definition⟩dog⟨/definition⟩
⟨/entry⟩

In markup systems there is a basic dichotomy between PRESENTATIONAL versus DESCRIPTIVE markup (Coombs et al. 1987). In presentational markup, the markup tags document what the information is supposed to look like (e.g. an entry is formatted as a paragraph with the headword in italics one font size larger, with the part of speech in bold, and with the definition in blue), as shown in 3a. This example uses HTML, Hypertext Markup Language, which is the most widely used system of presentational markup. It is portable with respect to preserving the appearance of information for human readers, but is not portable for the purpose of enabling computer systems to read the information and manipulate it consistently. For this, descriptive markup is needed in which the markup tags identify the pieces of information with respect to their function (e.g. an entry contains a headword, a part of speech, and a definition), as shown in 3b. This example illustrates XML,[41] Extensible Markup Language, which is now the most widely used system for implementing descriptive markup. The portability of descriptive markup may be limited when the system of markup is not documented. XML addresses this by supporting the formal definition of the markup scheme by means of a Document Type Definition (DTD) or an XML Schema (Bradley 2002).

RENDERING. It is a basic requirement of language resources that they should be presented to human readers in conventionally formatted displays (Simons 1998:§6). Both encoding and markup may lead to problems for rendering. Character encoding (the representation of characters in digital storage) causes problems for rendering when the fonts needed to view the textual information are not available. This problem is exacerbated when custom fonts are developed to support custom character sets. This is because the fonts themselves are a special kind of resource and are subject to a wide range of portability problems.

Markup may also cause problems for rendering. As we have seen, resources employ descriptive markup to maximize portability across computer systems and potential uses. However, such resources fail to cross the gap from computer to human if there is no meaningful way to display them.

3.3. DISCOVERY. By DISCOVERY we mean the problem of finding digital resources in the first place. The area of discovery involves two key concepts: discovering the EXISTENCE of a resource, and then judging the RELEVANCE of a discovered resource.

EXISTENCE. A given resource, even if it is of the highest quality, is of little practical value if the people who could benefit from it do not know that it exists. A large proportion of digital language resources (particularly those resulting from linguistic field work) are only to be found in the linguist's personal collection of computer files,

[40] http://xml.coverpages.org/sgml.html

[41] http://www.w3.org/XML/

and have no publicly available metadescription that would permit someone else to find them. When resources are entered into an institutional collection that is properly organized and cataloged, they remain virtually inaccessible if that catalog may only be consulted in person at the host institution. Even when a catalog is available electronically over the internet, the resources remain hidden unless the catalog is formatted in such a way that web search engines can appropriately retrieve its contents. In many of these cases, successful discovery of language resources depends on word-of-mouth and queries posted to electronic mailing lists.

RELEVANCE. Merely knowing that a resource exists is insufficient; the potential user must also be supplied with enough information in order to gauge the relevance of the resource. One may download a large resource only to discover that it is in an incompatible format. One may locate a binary file called dict.dat, then expend considerable effort to determine whether its content is relevant. Even where organized collections provide metadescription for subject language and linguistic type, they will typically use free text rather than a controlled vocabulary, reducing precision and recall in searching (cf. our discussion of terminology in §3.1).

3.4. ACCESS. By ACCESS we mean issues relating to the way in which the potential user of a resource gains access to it. Access involves three key concepts: the SCOPE of access that is granted, the PROCESS by which access is granted, and the EASE with which access is obtained.

SCOPE OF ACCESS. In the past, primary documentation was usually not disseminated. To listen to a field recording it was often necessary to visit the laboratory of the person who collected the materials, or to make special arrangements for the materials to be copied and posted. Digital publication on the web alleviates this problem, although projects usually refrain from full dissemination by limiting access to a restrictive search interface. This means that only selected portions of the documentation can be downloaded, and that all access must use categories predefined by the provider. Lack of full access means that materials are not fully portable.

PROCESS FOR ACCESS. It sometimes happens that an ostensibly available resource turns out not to be available after all, because there is no process whereby it may be obtained. One may discover the resource because its creator cited it in a manuscript or an annual research report. Commonly, researchers want to be recognized for the labor that went into creating primary language documentation, but do not want to make the materials available to others until they have derived maximum personal benefit. Despite its many guises, this problem has two distinguishing features: someone draws attention to a resource in order to derive credit for it—'parading their riches' as Mark Liberman (p.c., 2000) has aptly described it—and then applies undocumented or inconsistent restrictions to prevent access. The result may be frustration that a needed resource is withheld, leading to wasted effort or a frozen project, or to suspicion that the resource is defective and so must be protected by a smoke screen.

EASE OF ACCESS. Some resources are disseminated only on the web, making them difficult or impossible to access by people having a low-bandwidth connection or no connection at all. It may be particularly significant for communities that use endangered languages to have access to printed versions of language resources for use in efforts at language development and revitalization. In the case of multimedia resources, the absence of a low-bandwidth surrogate or a textual account of the content forces potential users to download and review the full resource in order to evaluate its suitability.

3.5. CITATION. By CITATION we mean the problems associated with making bibliographic citations of electronic language documentation and description. Citation in-

volves four key concepts: the ability to cite a resource in a BIBLIOGRAPHY, the PERSISTENCE of electronic resource identifiers, the IMMUTABILITY of materials that are cited, and the GRANULARITY of what may be cited.

BIBLIOGRAPHY. Research publications are normally required to provide full bibliographic citation of the materials used in conducting the research. Citation standards are usually high when citing conventional publications, but are much lower for citations of digital language resources. Many scholars do not know how to cite electronic resources; thus the latter are often incorrectly cited, or not cited at all.[42] When electronic sources are not properly cited, it is difficult to discover what resources were used in conducting the research or, following the linkage in the reverse direction, to consult a citation index to discover all the ways in which a given resource has been used.

PERSISTENCE. Often a language resource is available on the web, and it is convenient to identify the resource by means of its UNIFORM RESOURCE LOCATOR (URL) since this may offer the most convenient way to obtain the resource. However, URLs are notorious for their lack of persistence. They 'break' when the resource is moved or when some piece of the supporting infrastructure, such as a database server, ceases to work.

IMMUTABILITY. Even if a URL does not break, the item that it references may be mutable, changing over time. Language resources published on the web are usually not versioned, and a third-party description based on some resource may cease to be valid if that resource is changed. This problem can be solved by archiving each version and ensuring that citations reference a particular version. Publishing a digital artifact, such as a CD, with a unique identifier, such as an ISBN, also avoids this problem.

GRANULARITY. Citation goes beyond bibliographic citation of a complete item. We may want to cite some component of a resource, such as a specific narrative or lexical entry. However, the format of the resource may not support durable citations to internal components. For instance, if a lexical entry is cited by a URL that incorporates its lemma, and if the spelling of the lemma is altered, then the URL will not track the change. In sum, the portability of a language resource suffers when incoming and outgoing links to related materials are fragile.

3.6. PRESERVATION. By PRESERVATION we mean the problem of ensuring that digital resources remain accessible to future generations. Preservation involves three key concepts: the LONGEVITY of the format, the SAFETY of resources from catastrophic loss, and the ongoing migration of resources to current physical and digital MEDIA.

LONGEVITY. The digital technologies used in language documentation and description greatly enhance our ability to create data while simultaneously compromising our ability to preserve it. Compared to paper copy, which can survive for hundreds of years, and other media such as clay tablets, which have lasted for millenia, digitized materials are evanescent because they are based on binary formats. The problem is exacerbated when they use a proprietary format that becomes obsolete within a few years (e.g. Microsoft Word 3.0). Presentational markup with HTML and interactive content with CGI, Javascript, and specialized browser plugins require future browsers and servers to be backwards-compatible. Worse still, primary documentation may be embodied in the interactive behavior of the resource (e.g. the gloss of the text under the mouse may show up in the browser status line, using the Javascript 'mouseover' effect). Consequently, digital resources—especially dynamic or interactive ones—often have a short lifespan, and typically become unusable three to five years after they cease to be actively maintained.

[42] Incidentally, *The Columbia guide to online style* (Walker & Taylor 1998) is a good source on how to cite online resources.

SAFETY. Language resources are stored on some physical medium or device (the CARRIER), such as paper, magnetic tape, and various kinds of disk (e.g. floppy disk, hard drive, compact disk). Many undesirable eventualities may befall such physical artifacts; they may be degraded, damaged, lost, stolen, or destroyed. Such problems are usually greater in the field, where accidents may be more common (e.g. canoes capsizing), and where there may be less protection from extremes of climate. If the resource is digital it may be deleted, overwritten, or corrupted. While the individual guardian of the resource may exercise great care with it, mistakes nevertheless occur. Other agents also come into play: the people who share, manage, or repair the equipment; hostile third parties including thieves and computer viruses; political instability that may force sudden evacuation; elements of the environment such as dust, humidity, pests, mold, and power failure; catastrophes including fire, flood, lightning strike, and war; and natural disasters such as earthquakes, tornadoes, hurricanes, tsunamis, and volcanic eruptions. A resource may suddenly cease to exist if no steps are taken to mitigate these risks by ensuring that another copy is in a safe location.

MEDIA. Digital storage media may become inaccessible due to the absence of supporting hardware (e.g. 5.25'' floppy disks). While the problem of obsolete media predates the digital era (e.g. wax cylinder recordings), the problem has become more acute and is frequently noted in recent literature on digital archives: 'The lifespan of consumer physical digital media is estimated to be 5 years or less' (Cohen 2001); 'To date, none of the digital recording systems developed specifically for audio has achieved a proven stability in the market place, let alone in an archive' (International Association of Sound and Audiovisual Archives 2001). Magnetic media degrade in quality over time, with the loss of signal strength and, in the case of tapes, deformation of the backing, hydrolysis in the binder (St-Laurent 1996), and the imposition of bleedthrough.

3.7. RIGHTS. By RIGHTS we mean issues relating to what a potential user of a resource is permitted to do with the resource. The area of rights involves four key concepts: clarifying the TERMS OF USE for the resource, maximizing the public BENEFIT of the resource, protecting any SENSITIVITY that is inherent in the resource, and finding the proper BALANCE between public benefits and protecting sensitivities.

TERMS OF USE. A variety of individuals and institutions may have intellectual property vested in a language resource, and there is a complex terrain of legal, ethical, and policy issues involved (Liberman 2000). In spite of this, most digital language data is disseminated without identifying the copyright holder and without any license delimiting the range of acceptable uses of the material. Often people collect or redistribute materials or create derived works without securing the necessary permissions. While this is often benign (e.g. when the resources are used for research purposes only), the creator or user of the resource risks legal action, or having to restrict publication, or even having to destroy primary materials. To avoid any risk one must avoid using materials whose property rights are in doubt. In this way, the very lack of documented rights may restrict the portability of the language resource.

Sometimes resources are not made available on the web for fear that they will get into the wrong hands or be misused. However, this fear may be based on a confusion between dissemination medium and rights. The web supports secure data exchange between authenticated parties through data encryption. Copyright statements and user licenses can restrict uses. More sophisticated models for managing digital rights are emerging (Iannella 2001). The application of these techniques to language resources is unexplored, and today we have an all-or-nothing situation in which the existence of any restriction tends to prevent access across the board.

BENEFIT. Researchers typically want the results of their work to benefit human knowledge and experience as widely as possible. When permission is obtained for collecting primary language documentation, however, restrictions may be imposed on who is allowed to use it, how they are allowed to use it, and the time period of use. Such restrictions may originate from various sources including the language community, the government agency that provides the research permit, or an institutional review board. While researchers may wish the results of their work to benefit the public, they may discover too late that legitimate but unanticipated uses by unforeseen users are unintentionally jeopardized when permissions are tightly circumscribed.

SENSITIVITY. Many individuals and institutions are sensitive about the collection, dissemination, and uses of what linguists typically regard as neutral language documentation. The content of an oral discourse may contain sensitive personal, tribal, religious, or corporate information, or may be viewed as libel, breach of confidence, or even treason by others. There is a perceived risk of commercial exploitation of language documentation that, as with Western commercialization of indigenous music, may 'emphasize the exotic and the unexpected at the expense of the real substance' (Bebey 1997:1). Researchers may build a career on the data obtained from a language community without ever making the resources available in a form that benefits that community. Disregard for such sensitivities may compromise the standing or security of an individual or group, or may lead to the imposition of tighter access restrictions in the future (Wilkins 1992).

BALANCE. Access restrictions that protect a sensitive resource simultaneously limit the wider benefit that the resource may bring to human knowledge and experience. Researchers will typically want to maximize the wider benefit of the resource while protecting any sensitivities. The precise formulation of access restrictions, however, is often overgeneral, encompassing a greater timespan or a greater proportion of the resource than strictly necessary. It causes real problems when a sensitivity is stipulated without any time limit. An item that could never be accessed (including at no time in the future) would only be wasting space in an archive. The sensitivities inherent in a resource are often time-limited, for example, by the lifetime of the individuals involved in creating it, or the remaining lifetime of an endangered language. Sometimes, sensitivities that pertain to some part of the linguistic documentation are assigned scope over an entire collection. For instance, when a portion of a video recording contains some sensitive material this may constitute grounds for withholding the entire recording. The sensitivity may be generalized from a recording to the associated linguistic description, such as transcripts, even though the transcripts themselves may contain no sensitive material. In the reverse direction it is also possible for sensitivities about the linguistic description to be generalized to the underlying documentation. The researcher may not be prepared to release the primary documentation until satisfied with the transcriptions, on the grounds that his or her career will benefit more if he or she has sole access to the primary documentation while conducting the research.

3.8. SPECIAL CHALLENGES FOR LITTLE-STUDIED LANGUAGES. Many of these problems are exacerbated in the case of little-studied languages. The small amount of existing work on the language and the concomitant lack of established documentary practices and conventions may lead to especially diverse nomenclature. Inconsistencies within or between language descriptions may be harder to resolve because of the lack of significant documentation, the limited access to speakers of the language, and the limited understanding of dialect variation. Open questions in one area of description (e.g.

the inventory of vowel phonemes) may multiply the indeterminacies in another (e.g. the transcription and interpretation of texts). More fundamentally, existing documentation and description may be virtually impossible to discover and access, owing to its sparse or fragmentary nature.

The acuteness of these portability problems for little-studied languages can be highlighted by comparison with well-studied languages. In English, published dictionaries and grammars exist to suit all conceivable tastes, and it therefore matters little, relatively speaking, if some of these resources are not especially portable. However, when there is only one available dictionary for a little-studied language, it must be pressed into a great range of services, and so portability becomes a major concern.

Another issue that is more vexing in the case of endangered languages is access. Access may be prevented by the choice of inappropriate media for dissemination. For instance, an endangered language dictionary published only on the web will not be accessible to speakers of that language who live in a village without electricity. In the reverse direction, when a collection of recordings is transcribed in a little-studied language but not interpreted into a major language, then the content of those recordings is inaccessible to the outside world.

Sensitivity issues are often more acute for endangered languages. The wishes of the speech community (to control rather than disseminate their language) may conflict with the wishes of the linguists documenting the language (to disseminate rather than tie up the documentation). In balancing sensitivities it is often helpful to distinguish description from documentation; researchers *create* descriptions, while they only *collect* documentation. In the case of pure documentation, such as a video recording of a linguistic event in which the researcher has no creative input, the sensitivity of the participants takes precedence over any sensitivities of the researcher. In the case of pure description, such as a theoretical monograph on the language, the researcher's own sensitivities prevail. However, language resources such as grammars and analytical lexicons combine documentation and description. In such cases, resolving the conflicting sensitivities of the speech community and the linguists documenting the language will often depend on forging alliances and establishing shared goals.

This concludes our discussion of the portability problems in language documentation and description. The following sections respond to these problems by laying out the core values that constitute requirements for best practices (§4), describing how the the Open Language Archives Community supports the process of identifying community-agreed best practices (§5), and by providing a comprehensive set of best practice recommendations (§6).

4. VALUE STATEMENTS. Best practice recommendations amount to a decision about which of several possible practices is best. As anthropologist Henry Bagish points out in his critique of cultural relativism, indiscriminate tolerance of every possible practice is paralyzing (Bagish 1983). He proposes a formula that permits objective, crosscultural evaluation of competing practices, namely, 'If you value X, then A is better than B'. That is, before making a judgment as to which practice is better, one must clearly articulate the values that motivate the choice. If different parties can agree on the motivating values, then they should be able to come to agreement on the evaluation of competing practices.

In this section, we articulate the values that motivate the recommendations for best practice that are offered in §6. Our use of 'we' in the value statements is meant to include readers and members of the wider language resources community who share

these values. Note that these statements do not necessarily reflect an official position of the Linguistic Society of America.

4.1. CONTENT. COVERAGE. We value comprehensive documentation, especially for little-studied languages. Thus the best practice is one that establishes a record that is sufficiently broad in scope, rich in detail, and authentic in portrayal that future generations will be able to experience and study the language, even if no speakers remain.

ACCOUNTABILITY. We value the ability of researchers to verify language descriptions. Thus the best practice is one that provides the documentation that lies behind the description.

TERMINOLOGY. We value the ability of users to compare two resources by virtue of their terminology. Thus the best practice is one that makes it easy to identify the comparable aspects of unrelated resources.

4.2. FORMAT. OPENNESS. We value the ability of any potential user to make use of a language resource without needing to obtain unique or proprietary software. Thus the best practice is one that puts data into a format that is not proprietary.

ENCODING. We value the ability of users of a resource to understand the textual characters that are used in the resource, even in the absence of a font that can correctly render them. Thus the best practice is one that fully documents what the character codes in the resource represent.

MARKUP. We value the ability of users of a resource to be able to write programs that can process or present the information in novel ways. Thus the best practice is one that represents all of the information using a transparent descriptive markup, rather than in procedural code or in presentational markup.

RENDERING. We value the ability of users of a resource to be able to read the content of the information in a conventional presentation form. Thus the best practice is one that supplements the information resource with all the auxiliary software resources that are needed to render it for display.

4.3. DISCOVERY. EXISTENCE. We value the ability of any potential user of a language resource to learn of its existence. Thus the best practice is one that makes it easy for anyone to discover that a resource exists.

RELEVANCE. We value the ability of potential users of a language resource to judge its relevance without first having to obtain a copy. Thus the best practice is one that makes it easy for anyone to judge the relevance of a resource based on its description.

4.4. ACCESS. SCOPE OF ACCESS. We value the ability of any potential user of a language resource to access the complete resource, not just a limited portion of it or a limited interface to it. Thus the best practice is one that makes it easy for users to obtain a complete copy of the resource.

PROCESS FOR ACCESS. We value the ability of any potential user of a language resource to follow a well-defined procedure to obtain a copy of the resource. Thus the best practice is one in which there is a clearly documented procedure by which users may obtain a copy of the resource.

EASE OF ACCESS. We value the ability of potential users to access a version of a language resource from wherever they are located, even where the available computational infrastructure may be limited. Thus the best practice is one that makes such access possible.

4.5. CITATION. BIBLIOGRAPHY. We value the ability of users of a resource to give credit to its creators, as well as to learn the provenance of the sources on which it is

based. Thus the best practice is one that makes it easy for electronic language documentation and description to be cited.

PERSISTENCE. We value the ability of users of language resources to locate an instance of the resource, even though its actual location or filename might change. Thus the best practice is one that archives resources with identifiers that are independent of location or file name.

IMMUTABILITY. We value the ability of users to cite a language resource without that resource changing and invalidating the citation. Thus the best practice is one that makes it possible for users to cite particular versions that never change.

GRANULARITY. We value the ability of potential users to cite the component parts of a language resource. Thus the best practice is one that ensures each subitem of a resource has a durable identifier.

4.6. PRESERVATION. LONGEVITY. We value ongoing access to language resources over the very long term. Thus the best practice is one that stores resources in formats that are likely to remain usable for generations to come.

SAFETY. We value ongoing access to language resources over the very long term. Thus the best practice is one that stores copies of resources in multiple locations so as to ensure against catastrophic damage to a single repository.

MEDIA. We value ongoing access to language resources beyond the life span of any particular storage medium. Thus the best practice is one that migrates resources to new physical and digital media before the ones they are stored in become unusable.

4.7. RIGHTS. TERMS OF USE. We value the ability of potential users of a language resource to understand any restrictions on its permissible use before they begin to use it. Thus the best practice is one that clearly states the terms of use as part of the resource package.

BENEFIT. We value the maximal application of language resources toward the benefit of human knowledge and experience. Thus the best practice is one that does not hinder the fair use of a language resource for scientific, educational, humanitarian, or other noncommercial uses.

SENSITIVITY. We value the rights of the contributors to a language resource. Thus the best practice is one that protects any sensitivities stipulated by the contributors.

BALANCE. We value the potential long-term benefits of a resource, even when sensitivities prevent its dissemination in the near term. Thus the best practice is one that clearly identifies the nature of a sensitivity and associates it with an explicit time frame.

These value statements lead us to propose the detailed best-practice recommendations listed in §6. Before proceeding to these recommendations we give a brief overview of OLAC, which provides structures to support the elaboration and implementation of such recommendations.

5. OLAC, THE OPEN LANGUAGE ARCHIVES COMMUNITY. While this article sketches a set of values and practices designed to enhance the portability of digital language documentation and description, it is ultimately the community that must work out the details and reach a consensus. A community that can fill this role has already begun to form.

In December 2000, an NSF-funded workshop, Web-Based Language Documentation and Description, was held in Philadelphia. The workshop brought together a group of nearly 100 language software developers, linguists, and archivists who are responsible for creating language resources in North America, South America, Europe, Africa, the

Middle East, Asia, and Australia (Bird & Simons 2000). The outcome of the workshop was the founding of the Open Language Archives Community (OLAC),[43] with the following purpose:

> OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.

Today OLAC has over twenty participating archives in seven countries, with over 30,000 records describing language resource holdings. The OLAC gateway at the LINGUIST List site[44] permits users to search the contents of all archives from a single location, before being directed to the website of the individual archive for information about how to obtain the resource. Anyone in the wider linguistics community can participate, not only by using the search facilities, but also by documenting their own resources, or by helping create and evaluate new best practice recommendations.

The technical infrastructure for OLAC is built on a framework developed within the digital libraries community by the Open Archives Initiative.[45] It has two components: a metadata standard (DCMI 1999) and a metadata harvesting protocol (Lagoze et al. 2002). These standards define how data providers—the institutions that want to make their resources known—publish metadata about their holdings, and how service providers—the institutions that want to provide value-added services for an entire community—can harvest the metadata and add it to the information pool on which they base their services. The OLAC versions of these standards, namely the OLAC Metadata standard and the OLAC Repositories standard, are designed to address the particular needs of language archiving (Bird & Simons 2003, Simons & Bird 2003a,b).

'Metadata' is structured data about data—descriptive information about a physical object or a digital resource. Library card catalogs represent a well-established type of metadata, and they have served as collection management and resource-discovery tools for decades. The OLAC Metadata standard (Simons & Bird 2002a) defines the elements to be used in metadata descriptions of language archive holdings, and how such descriptions are to be disseminated using XML descriptive markup for harvesting by service providers in the language-resources community. The OLAC metadata set contains the fifteen elements of the Dublin Core metadata set (DCMI 1999), plus several refined elements that capture information of special interest to the language-resources community. In order to improve recall and precision when searching for resources, the standard also defines a number of controlled vocabularies for descriptor terms. The most important of these is a standard for identifying languages (Simons 2000).

The OLAC Repositories standard (Simons & Bird 2002c) defines the protocol by which service providers query web-accessible repositories to harvest the metadata records they publish. Any other site may use the protocol to collect metadata records in order to provide a service, such as offering a union catalog of all archives or a specialized search service pertaining to a particular topic. To facilitate widespread discovery of the resources held in OLAC archives, all OLAC metadata is mapped to the more general-purpose Dublin Core metadata set and disseminated to the broader community of digital libraries; it is also mapped to an HTML format to facilitate indexing by web search engines. In the same way, more specialized metadata formats, such as the IMDI

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

[44] http://www.linguistlist.org/olac/

[45] http://www.openarchives.org/

format for fine-grained description of linguistic field recordings,[46] can be mapped to OLAC metadata for dissemination to the wider language-resources community.

In addition to this technical infrastructure, OLAC also provides simple infrastructure to support interaction among the human participants of the Open Language Archives Community. The OLAC Process standard (Simons & Bird 2002b) defines: (i) the governing ideas of OLAC, including a summary statement of its purpose, vision, and core values; (ii) the organization of OLAC, in terms of the groups of participants that play key roles: coordinators, advisory board, council, participating archives and services, working groups, and participating individuals; and (iii) the operation of OLAC, in terms of a document process that defines how documents are generated and how they progress from one status to the next along the five-phase life cycle of development, proposal, testing, adoption, and retirement.

This last aspect of the OLAC Process (i.e. the document process) is already leading to new standards and best practice recommendations. In the future, we envision best practices for a variety of players, including linguists, archivists, developers, and sponsors. By participating in the OLAC Process—setting up working groups, reviewing current practices, formulating best practice recommendations, and forging a consensus in the wider community through cycles of review and revision—the community that creates and uses digital language documentation and description will move forward to a new era of highly portable language resources.

Having described suitable community infrastructure for developing best practice recommendations, we now present our own recommendations. By presenting them here we do not intend to bypass the consensus-building process, but rather to stimulate widespread discussion leading to better, more carefully articulated recommendations.

6. BEST PRACTICE RECOMMENDATIONS. This section recommends best practices in support of the values set out in §4. These guidelines need to be fleshed out in more detail by the language-resources community. Note that these statements do not necessarily reflect an official position of the Linguistic Society of America.

6.1. CONTENT.

(1) COVERAGE.
   a. Make rich records of rich interactions, especially in the case of endangered languages or genres.
   b. Document the 'multimedia linguistic field methods' that were used.

(2) ACCOUNTABILITY.
   a. Provide the full documentation on which language descriptions are based. For instance, a grammar is based on a text corpus.
   b. When texts are transcribed, provide the primary recording (without segmenting it into clips).
   c. Transcriptions should be time-aligned to the underlying recording in order to facilitate verification.
   d. When recordings have been significantly edited, provide the original recordings to guarantee authenticity of the materials.

(3) TERMINOLOGY.
   a. Map the terminology and abbreviations used in description to a common ontology of linguistic terms.

[46] http://www.mpi.nl/world/ISLE/documents/draft/ISLE_MetaData_2.5.pdf

b. Map the element tags used in descriptive markup to a common ontology of linguistic terms.
c. Map the symbols used in transcription to phonological descriptors that are mapped to a common ontology of linguistic terms.

6.2. FORMAT.

(4) OPENNESS.
a. Store all language documentation and description in formats that are open (i.e. whose specifications are published and nonproprietary).
b. Prefer formats supported by software tools available from multiple suppliers.
c. Prefer formats with free tools over those with commercial tools only.
d. Prefer published proprietary formats, e.g. Adobe Portable Document Format (PDF) and MPEG-1 Audio Layer 3 (MP3), to secret proprietary formats, e.g. Microsoft formats.

(5) ENCODING.
a. Encode the characters with Unicode.
b. Avoid Private Use Area characters, but if they are used, document them fully.
c. Document any 8-bit character encodings.
d. Document any scheme used to transliterate characters.

(6) MARKUP.
a. Prefer descriptive markup over presentational markup.
b. Prefer XML (with an accompanying DTD or Schema) over other schemes of descriptive markup.
c. If the XML DTD or Schema is not a previously archived standard, archive it. Give each version a unique identifier.
d. If a descriptive markup scheme other than XML is used, prepare and archive a document that explains the markup scheme.
e. When a resource using descriptive markup is archived, reference the resource to the archived version of the definition of the associated markup format.
f. If punctuation and formatting are used to represent the structure of information, document how they are used.

(7) RENDERING.
a. If the fonts needed to appropriately render the resource are not commonly available, archive them and reference the resource to the archived version of the needed fonts.
b. Provide one or more human-readable versions of the material, using presentational markup (e.g. HTML) or other convenient formats. Proprietary formats are acceptable for delivery as long as the primary documentation is stored in a nonproprietary format.
c. If you have used stylesheets to render the resource, archive them as well.

N.B. Format is a critical area for the definition of best practices. We propose that recommendations in this area be organized by type (e.g. audio, image, text), possibly following the inventory of types identified in the Dublin Core metadata set.[47]

[47] http://dublincore.org/documents/dcmi-type-vocabulary/

6.3. DISCOVERY.

(8) EXISTENCE.

a. List all language resources with an OLAC repository.

b. Any resource presented in HTML on the web should contain metadata with keywords and description for use by conventional search engines.

(9) RELEVANCE.

a. Follow the OLAC recommendations on best practice for describing language resources using metadata, especially concerning language identification and linguistic data type. This will ensure the highest possibility of discovery by interested users in the OLAC union catalog hosted on the LINGUIST List site.[48]

6.4. ACCESS.

(10) SCOPE OF ACCESS.

a. Publish complete primary documentation, providing a documented method by which anyone may obtain the documentation.

b. Publish documentation and description in such a way that users can gain access to the files to manipulate them in novel ways. (That is, do not just publish through a fixed user interface like a web search form, or a fixed presentation view like a PDF file.)

c. Transcribe all recordings in the orthography of the language (if one exists).

(11) PROCESS FOR ACCESS.

a. Document the process for access as part of the metadata, including any licenses and charges.

b. Document all restrictions on access as part of the metadata.

c. For resources not distributed over the web, document the expected delivery time.

d. For resources not distributed over the web, publish online surrogates that are easy for potential users to access and evaluate.

(12) EASE OF ACCESS.

a. Publish digital resources using appropriate delivery media, e.g. web for small resources, and CD or DVD for large resources.

b. Provide low-bandwidth surrogates for multimedia resources, e.g. publish MP3 files corresponding to large, uncompressed audio data.

c. Provide transcriptions for extended recordings to facilitate access to the relevant section.

d. For little-studied languages where the speech community has limited web access, publish print versions to facilitate access by the community, and provide a written account of any multimedia content using a major language.

6.5. CITATION.

(13) BIBLIOGRAPHY.

a. Furnish complete bibliographic data in the metadata for all language resources created.

b. Provide complete citations for all language resources used.

c. Provide instructions on how to cite an electronic resource from the collec-

[48] http://www.linguistlist.org/olac/

tion as part of the web site for a digital archive (e.g. see the instructions on SIL's Electronic Working Papers site[49]).

d. Use the metadata record of a language resource to document its relationship to other resources (e.g. in the OLAC context, use the RELATION element).

(14) PERSISTENCE.

a. Ensure that resources have a persistent identifier, such as an ISBN, an OAI identifier, or a Digital Object Identifier.[50]

b. Ensure that a persistent identifier resolves to an online instance of the resource, or else to detailed online information about how to obtain the resource.

(15) IMMUTABILITY.

a. Provide fixed versions of a resource, either by publishing it on a read-only medium, or by submitting it to an archive that ensures immutability.

b. Distinguish multiple versions with a version number or date, and assign a distinct identifier to each version.

(16) GRANULARITY.

a. Provide a formal means by which the components of a resource may be uniquely identified.

b. Take special care to avoid the possibility of ambiguity, such as arises when lemmas are used to identify lexical entries, and where multiple entries can have the same lemma.

6.6. PRESERVATION. Many organizations have published detailed recommendations concerning the archival preservation of paper, audio, video, and images. Readers are referred to: the Library of Congress Preservation Directorate[51] which has recommendations concerning paper and images (Library of Congress 1995, 2001); the UNESCO Archives Portal[52] which has a section on preservation and conservation, including a reader on audiovisual archives focusing on the practical needs of audiovisual archivists in developing countries (Harrison 1997); the International Association of Sound and Audiovisual Archives[53] which has published recommendations for audio preservation (International Association of Sound and Audiovisual Archives 2001); The Council on Library and Information Resources[54] which publishes a series of reports containing chapters on audio and video preservation (Brylawski 2002, Cohen 2001, Wactlar & Christel 2002); The Conservation Online (CoOL) website,[55] with the most comprehensive set of links to online resources for the preservation of audio materials,[56] and recommendations for the handling of media (St-Laurent 1996); the Preservation Metadata Working Group of the Online Computer Library Center,[57] the Research Libraries

[49] http://www.sil.org/silewp/citation.htm l

[50] http://www.doi.org/

[51] http://lcweb.loc.gov/preserv/

[52] http://www.unesco.org/webworld/portal_archives/pages/

[53] http://www.iasa-web.org/

[54] http://www.clir.org/

[55] http://palimpsest.stanford.edu/

[56] http://palimpsest.stanford.edu/bytopic/ audio/

[57] http://www.oclc.org/research/pmwg/

Group,[58] developing a standard to 'document and evaluate the processes that support the long-term retention and accessibility of digital content' (OCLC/RLG 2002); and the International Standards Organization, providing a standard concerning the structure and function of a digital archive in ISO 14721 *Reference Model for an Open Archival Information System*.[59]

The recommendations in this section touch on key themes from the literature we cite that are directly relevant to language archiving. However, readers are advised to consult the literature for full discussion and detailed recommendations.

(17) LONGEVITY.
   a. Commit all documentation and description to a digital archive that can credibly promise long-term preservation and access.
   b. Ensure that the archive satisfies the key requirements of a well-founded digital archive, for instance, that it implements digital archiving standards, provides offsite backup, migrates materials to new formats and media/devices over time, is committed to supporting new access modes and delivery formats, has long-term institutional support, and has an agreement with a national archive to take materials if the archive folds.
   c. Digitize analog recordings, to permit lossless copying in the future.
   d. Publish language documentation and description on the web using standard open formats so that they are fortuitously captured by internet archives (e.g. the Wayback Machine[60]).
   e. When digital language resources are stored offline, transfer them to new storage media before the existing media type becomes unsupported (for many media types this would be necessary every five years).
   f. Archive physical versions of the language documentation and description (e.g. printed versions of documents, any tapes from which online materials were created).
   g. Prefer the file formats—including markup and encoding—that have the best prospect for accessibility far into the future (e.g. use type 1 (scalable) fonts in preference to bitmap fonts in documents).

(18) SAFETY.
   a. Ensure that copies of archived documentation and description are kept at multiple locations (e.g. following the LOCKSS concept, 'Lots of copies keeps stuff safe'[61]).
   b. Create a disaster recovery plan, such as that developed by the Syracuse University Library (1995), containing procedures for salvaging archived resources in the event of a disaster.

(19) MEDIA.
   a. Whenever possible, maintain language resources on digital mass-storage systems, for easy backup and transfer to upgraded hardware.
   b. Refresh offline digital storage by transferring the data to new storage at regular intervals (e.g. 1–5 years). Choose intervals appropriate for the performance of the media and location (e.g. offline magnetic media suffer

[58] http://www.rlg.org/

[59] http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

[60] http://www.archive.org/

[61] http://lockss.stanford.edu/

from signal loss and bleedthrough and should normally be refreshed every 1–2 years; nonmagnetic media and media maintained in climate-controlled storage may only need refreshing after 5–10 years).

c. Language resources that are stored in a proprietary binary format should be migrated to new formats before the existing format becomes unsupported (for many formats this would be necessary every five years).

6.7. RIGHTS.

(20) TERMS OF USE.

a. Ensure that the intellectual property rights relating to the resource are fully documented.

b. Ensure that there is a terms-of-use statement that clearly states what a user may and may not do with the materials.

(21) BENEFIT.

a. Ensure that the resource may be used for research purposes.

b. Ensure that the use of primary documentation is not limited to the researcher, project, or agency responsible for collecting it.

(22) SENSITIVITY.

a. Ensure that the nature of any sensitivity is documented in detail. To aid interpretation in the distant future, include concrete examples of any eventualities that must be avoided.

(23) BALANCE.

a. Limit any stipulations of sensitivity to the sensitive sections of the resource, permitting nonsensitive sections to be disseminated more freely.

b. Associate each sensitivity with an expiry date or a review date. List objective criteria that can be applied to determine whether the sensitivity has expired.

c. When primary documentation is closed in order for a researcher to derive maximal personal benefit, the expiry date should be no later than five years after the recording date.

As stated at the outset, we have structured this article to build consensus. Readers who take issue with any of our best-practice recommendations are encouraged to join the OLAC community[62] and enter into the consensus-building process. We further recommend that they review the corresponding statements of problems (§3) and values (§4). Baseline consensus on the problems and values provides a secure foundation for constructive discussions about the community's best practices.

7. CONCLUSION. Today, the community of scholars engaged in language documentation and description is in the midst of transition between the paper-based era and the digital era. We are still working out how to preserve knowledge that is stored in digital form. During this transition period, we observe unparalleled confusion in the management of digital language documentation and description. A substantial fraction of the resources being created can only be reused on the same software/hardware platform, within the same scholarly community, for the same purpose, and then only for a period of a few years. However, by adopting a range of best practices, this specter of chaos can be replaced with the promise of easy access to highly portable resources.

Using TOOLS as our starting point, we described a diverse range of practices and discussed their negative implications for DATA portability along seven dimensions, lead-

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

ing to a collection of ADVICE on creating portable resources. These three categories, tools, data, and advice, are three pillars of the infrastructure provided by OLAC, the Open Language Archives Community (Bird & Simons 2001). Our best-practice recommendations are preliminary, and we hope they will be fleshed out by the community using the OLAC Process.

We leave off where we began, namely with tools. It is the community's use of new tools that has led to data portability problems. And it is only newer tools—supporting the kinds of practices we advocate—that will address these problems. An archival format is useless unless there are tools for creating, managing, and browsing the content stored in that format. Needless to say, no single organization has the resources to create the necessary tools, and no third-party developer of general-purpose office software will address the specialized needs of the language documentation and description community. We need nothing short of an open source[63] revolution, leading to new open source tools based on agreed data models for all of the basic linguistic types, connected to portable data formats, with all data housed in a network of interoperating digital archives. On their own, technological solutions will be inadequate, as they have been in the past, only contributing further to the digital carnage we experience today. Instead, the technological solutions must be coupled with a sociological innovation, one that produces broad consensus about the design and operation of common digital infrastructure for the archiving of language documentation and description.

REFERENCES

ANTWORTH, EVAN, and J. RANDOLPH VALENTINE. 1998. Software for doing field linguistics. In Lawler & Aristar Dry, 170–96. Appendix online: http://www.sil.org/computing/routledge/antworth-valentine/.

BAGISH, HENRY H. 1983. Confessions of a former cultural relativist. Anthropology annual editions 83/84, ed. by Elvio Angeloni, 22–29. Guilford, CT: Dushkin Publishing Group.

BEBEY, FRANCIS. 1997. African music: A people's art. Trans. by Josephine Bennett. New York: Lawrence Hill and Company.

BECKER, JOSEPH D. 1984. Multilingual word processing. Scientific American 251.96–107.

BIRD, STEVEN, and GARY SIMONS (eds.) 2000. Proceedings of the workshop on web-based language documentation and description. Online: http://www.ldc.upenn.edu/exploration/expl2000/.

BIRD, STEVEN, and GARY SIMONS. 2001. The OLAC metadata set and controlled vocabularies. Proceedings of the ACL/EACL workshop on sharing tools and resources for research and education, compiled by Mike Rosner, 27–38. East Stroudsburg, PA: Association for Computational Linguistics. Online: http://arXiv.org/abs/cs/0105030.

BIRD, STEVEN, and GARY SIMONS. 2003. Extending Dublin Core metadata to support the description and discovery of language resources. Computers and the Humanities 37, to appear.

BRADLEY, NEIL. 2002. The XML companion. Harlow, UK: Addison Wesley.

BRYLAWSKI, SAMUEL. 2002. Preservation of digitally recorded sound. Building a national strategy for preservation: Issues in digital media archiving. Washington, DC: Council on Library and Information Resources and the Library of Congress. Online: http://www.clir.org/pubs/reports/pub106/sound.html.

COHEN, ELIZABETH. 2001. Preservation of audio. Folk heritage collections in crisis. Washington, DC: Council on Library and Information Resources. Online: http://www.clir.org/pubs/reports/pub96/preservation.html.

COOMBS, JAMES H.; ALLEN H. RENEAR; and STEVEN J. DEROSE. 1987. Markup systems and the future of scholarly text processing. Communications of the ACM 30.933–47.

DCMI. 1999. Dublin Core Metadata Element Set, version 1.1: Reference description. Online: http://dublincore.org/documents/1999/07/02/dces/.

[63] http://www.opensource.org/

HARRISON, HELEN P. 1997. Audiovisual archives: A practical reader. Paris: UNESCO. Online: http://unesdoc.unesco.org/images/0010/001096/109612eo.pdf.

HIMMELMANN, NIKOLAUS P. 1998. Documentary and descriptive linguistics. Linguistics 36.161–95.

IANNELLA, RENATO. 2001. Digital rights management (DRM) architectures. D-Lib Magazine 7.6. Online: http://www.dlib.org/dlib/june01/iannella/06iannella.html.

INTERNATIONAL ASSOCIATION OF SOUND AND AUDIOVISUAL ARCHIVES. 2001. The safeguarding of the audio heritage: Ethics, principles and preservation strategy. Online: http://www.iasa-web.org/iasa0013.htm.

LADEFOGED, PETER. 2000. Vowels and consonants: An introduction to the sounds of languages. Cambridge, MA: Blackwell.

LAGOZE, CARL; HERBERT VAN DE SOMPEL; MICHAEL NELSON; and SIMEON WARNER. 2002. The Open Archives Initiative Protocol for Metadata Harvesting. Online: http://www.openarchives.org/OAI/openarchivesprotocol.html.

LAWLER, JOHN M., and HELEN ARISTAR DRY (eds.) 1998. Using computers in linguistics: A practical guide. London and New York: Routledge.

LEWIS, WILLIAM; SCOTT FARRAR; and D. TERENCE LANGENDOEN. 2001. Building a knowledge base of morphosyntactic terminology. Proceedings of the IRCS workshop on linguistic databases, ed. by Steven Bird, Peter Buneman, and Mark Liberman. Online: http://www.ldc.upenn.edu/annotation/database/ (Click on 'Papers').

LIBERMAN, MARK. 2000. Legal, ethical, and policy issues concerning the recording and publication of primary language materials. In Bird & Simons 2000.

LIBRARY OF CONGRESS. 1995. Guidelines for electronic preservation of visual materials. Online: http://lcweb.loc.gov/preserv/guide/.

LIBRARY OF CONGRESS. 2001. The deterioration and preservation of paper: Some essential facts. Online: http://lcweb.loc.gov/preserv/deterioratebrochure.html.

OCLC/RLG. 2002. Preservation metadata and the OAIS information model: A metadata framework to support the preservation of digital objects. Online: http://www.oclc.org/research/pmwg/pm_framework.pdf.

PULLUM, GEOFFREY K., and WILLIAM A. LADUSAW. 1986. Phonetic symbol guide. Chicago: University of Chicago Press.

SIMONS, GARY. 1998. The nature of linguistic data and the requirements of a computing environment for linguistic research. In Lawler & Aristar Dry 10–25. Appendix online: http://www.sil.org/computing/routledge/simons/.

SIMONS, GARY. 2000. Language identification in metadata descriptions of language archive holdings. In Bird & Simons 2000.

SIMONS, GARY, and STEVEN BIRD. 2002a. OLAC metadata. Online: http://www.language-archives.org/OLAC/metadata.html.

SIMONS, GARY, and STEVEN BIRD. 2002b. OLAC process. Online: http://www.language-archives.org/OLAC/process.html.

SIMONS, GARY, and STEVEN BIRD. 2002c. OLAC repositories. Online: http://www.language-archives.org/OLAC/repositories.html.

SIMONS, GARY, and STEVEN BIRD. 2003a. Building an Open Language Archives Community on the OAI foundation. Library Hi Tech 21.2.210–18. Online: http://www.arxiv.org/abs/cs.CL/0302021.

SIMONS, GARY, and STEVEN BIRD. 2003b. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. Literary and Linguistic Computing 18.117–28.

ST-LAURENT, GILLES. 1996. The care and handling of recorded sound materials. Online: http://palimpsest.stanford.edu/byauth/st-laurent/care.html.

SYRACUSE UNIVERSITY LIBRARY. 1995. Procedures for recovering audio and sound recording materials. Online: http://libwww.syr.edu/information/preservation/audio.htm.

WACTLAR, HOWARD D., and MICHAEL G. CHRISTEL. 2002. Digital video archives: Managing through metadata. Building a national strategy for preservation: Issues in digital media archiving. Washington, DC: Council on Library and Information Resources and the Library of Congress. Online: http://www.clir.org/pubs/reports/pub106/video.html.

WALKER, JANICE R., and TODD TAYLOR. 1998. The Columbia guide to online style. New York: Columbia University Press. Companion site: http://www.columbia.edu/cu/cup/cgos/.

WILKINS, DAVID. 1992. Linguistic research under Aboriginal control: A personal account of fieldwork in central Australia. Australian Journal of Linguistics 12.171–200.

Bird
Department of Computer Science
University of Melbourne
Victoria 3010
Australia
[sb@cs.mu.oz.au]

Simons
SIL International
7500 W. Camp Wisdom Rd.
Dallas, TX 75236
[gary_simons@sil.org]

[Received 23 July 2002;
accepted 27 March 2003]