Archiving and linguistic resources *or*
How to keep your data from becoming endangered

Over the last several decades, the rise of new technologies has drastically changed the nature of linguistic resources. Whereas formerly primary documentary evidence for a language consisted of notebooks and analog recordings, today linguists doing documentary work have a bewildering array of choices of media, tools, and standards to choose from when creating materials.

The production of textual materials, for example, is complicated by the plethora of choices in computer software. Typical word processors are designed for the business world where documents do not generally need to be preserved for centuries, as is the case with endangered language documentation. Furthermore, much linguistic documentation does not consist of unstructured prose, which is what word processors are designed for, but, rather, highly structured information, like entries in a lexicon or morphological paradigms. In the ideal case, such data would be entered into a database where its structure could be properly encoded. There are a variety of database programs which can be used for this purpose, but most of them have the same limitations as word processors—they produce proprietary, short-lived formats that are difficult to migrate forward as technologies and methodologies evolve.

Another layer of complexity is that different digital resources need to be properly "linked" in order for them to have maximal value. Documentation may begin with audio or video recordings, but often a transcription and text analysis will also be created to accompany a recording. Ideally, recordings should not be inextricably linked to any particular analysis. However, any researcher making use of a recording would almost certainly want to be aware of an accompanying transcription or analysis. Dealing with the problem of indicating relationships among related resources is a problem in the creation of metadata—that is, archival data about resources which facilitates their access.

A final issue engendered by the rise of new technologies for documentation is how to manage "legacy" data—that is, data produced before the rise of digital tools—in order to make it accessible and to preserve it for future generations. Some important questions in this area include: What is the best way to digitize field notes and analog recordings? And, who has the rights to material produced before the academic community became more sensitive to the relationship between communities of speakers and data from their languages?

At present, it is difficult for linguists to locate specific recommendations for creating archivable resources. The purpose of this tutorial is to create a general forum wherein linguists who have created, or are planning to create, documentary linguistic resources can hear a range of talks on current accepted standards of best practice for resource production and conservation. The tutorial will aim for breadth, rather than depth, of coverage in order to address the needs of as many individuals as possible. By hearing from a number of experts from different areas, attendees of the tutorial will be able to identify appropriate individuals whom they can contact in order to get answers to their specific questions in the future.

Word count: 495

Invited participants and short abstracts

Gary Simons, *SIL International, Co-administrator of the Open Language Archives Community (OLAC), 7500 W. Camp Wisdom Road, Dallas, TX 75236, gary_simons@sil.org*

The Open Language Archives Community: Building a worldwide library of digital language resources

New ways of documenting and describing language via digital media coupled with new ways of distributing the results via the World-Wide Web offer a degree of access to language resources that is unparalleled in history. At the same time, the proliferation of approaches to using these new technologies is causing serious problems relating to resource discovery and resource creation. The Open Language Archives Community (OLAC) is an international partnership of almost 30 projects and institutions who are addressing these issues by (1) developing consensus on best current practice for the digital archiving of language resources, and (2) developing a network of interoperating repositories and services for housing and accessing such resources.

This talk presents the OLAC vision for creating a virtual library of the language resources that are housed all over the world by its member archives. It then describes the infrastructure that has been built in order to achieve this objective. Special attention is given to explaining the various mechanisms that make it possible for a project or institution to become a participating archive and to demonstrating the global search portal that allows any Web user to present a single search query to all participating archives at once.

Helen Aristar-Dry, *Eastern Michigan University and LinguistList, Co-Principal Investigator of the Electronic Metastructure for Endangered Languages Data initiative (EMELD), Dept. of English Language and Literature 612, Pray-Harrold Hall, Ypsilanti, MI 48197, hdry@linguistlist.org*

The E-MELD School of Best Practices

The Electronic Metastructure for Endangered Languages Data (E-MELD) Project, is a five-year collaborative project designed to build digital infrastructure for the long term preservation of linguistic documentation in "best practice" format. Best practice recommendations are designed to ensure that digital language resources will remain accessible and intelligible by future generations. A goal of the E-MELD project is to create a comprehensive but user-friendly website which offers information about creating such resources; this is the E-MELD School of Best Practices in Digital Language Documentation (http://emeld.org/school/), which will be demonstrated at this symposium. The site includes:

- A Classroom comprising links to information and tutorials on, for example, metadata creation, audio and video recording, and file storage
- Case Studies, which describe how legacy data was processed to create the documentation available in the Exhibit Hall
- The Exhibit Hall, which displays actual language documentation digitized according to recommended practices
- A Reading Room, which includes links to useful background material

• A Work Room, where users can work on their own documentation using online facilities resident on the LINGUIST List servers
• A Tool Room comprising descriptions of and links to hardware and software facilitating the implementation of best practice

Helen Agüera, *National Endowment for the Humanities (NEH), Acting Deputy Director, Division of Preservation and Access, Room 411, 1100 Pennsylvania Avenue, NW, Washington, D.C. 20506, haguera@neh.gov*

Archival projects and the NEH

Helen Agüera will discuss NEH's support for projects related to linguistic archives. She will describe the range of preservation and access activities funded by the Endowment through the Division of Preservation and Access. These activities include: the arrangement and description of a collection of linguistic materials that needs to be brought under intellectual control; the digital reformatting of sound and moving image collections for preserving and enhancing access to linguistic materials; and the creation of online archives that integrate multiple collections from widely dispersed sources or repositories to facilitate comparative studies and broad educational use.

Agüera will report on the characteristics of successful projects. She will touch on what aspects of a proposal NEH evaluators consider essential for endorsing a project—from information about the language or languages represented in a project to details about the proposed methodology and adherence to (or departure from) established standard and best practices. She will give special attention to questions concerning the long-term preservation of digital objects.

Finally, Agüera will discuss NEH's partnership with the National Science Foundation, "Documenting Endangered Languages," and the role linguistic archives can play in this effort to develop and advance knowledge concerning endangered languages.

Mark Kaiser, *Berkeley Language Center (BLC) (University of California, Berkeley), Associate Director, 29 Dwinelle Hall, Berkeley, CA 94720, mkaiser@socrates.berkeley.edu*

Digitizing the Audio Archive of Linguistic Field Work

The Berkeley Language Center manages three main and several minor archives of audio recordings. This presentation focuses on our efforts to digitally preserve and provide access to the Audio Archive of Linguistic Field Work, which consists of nearly 1,400 hours of field recordings of Native American languages. We discuss legal issues regarding copyright and the rights of consultants and Native American communities, as well as ethical issues surrounding the preservation and distribution of materials deposited at the BLC long before use of the Internet. We also address technical issues of archiving and delivery (bit depth and sampling rates, file formats, backup), and finally, our efforts to anticipate and comply with metadata standards.

Peter Wittenburg, *the Max Planck Institute for Psycholinguistics, Technical Director, Wundtlaan, PB 310, 6500 AH Nijmegen, The Netherlands, peter.wittenberg@mpi.nl*

From recordings to an organized language archive

While it is increasingly common for digital audio and video recordings to serve as the foundational resources for the linguistic documentation, these resources need to be properly associated with at least two kinds of information in order to be accessible to academic and native-speaker communities. First, they need to be annotated in ways that describe their content. Such annotation could include grammatical, ethnographic, or even musicological information. Annotation of digital recordings requires the use of a specialized tool, and one such tool, the ELAN tool, will be demonstrated.

Audio and video resources also need to be indicated as being associated with related resources—for example, an audio recording may be associated with a field notebook or even a lexicon. Properly indicating relationships between resources requires the use of an accepted metadata standard and tools for creating, editing, and viewing metadata. One metadata standard, the IMDI standard, and its associated editing and browsing system will be discussed.

Taken together, annotation tools and metadata tools can help accomplish two important goals of making the content of recordings more accessible and allowing a cluster of related resources to be linked together.

Heidi Johnson, *Archive of the Indigenous Languages of Latin America (AILLA) (University of Texas, Austin), Project Manager, Department of Anthropology, EPS 1.130, 1 University Station C3200, Austin, Texas 78712-1086, hjohnson@mail.utexas.edu*

Preparing documentary materials for archiving

The Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas at Austin is a digital repository of multimedia resources. Our primary mission is the digitization and preservation of "legacy" materals; that is, recordings and texts produced in analog media over the past half-century. We have received a wide variety of materials, with and without metadata (catalog information). This diversity is unavoidable when dealing with materials produced long ago, often by someone other than the person who sends us the package, but it is not particularly desirable.

In the course of her duties as manager of AILLA, and from her own experiences as a field linguist, Johnson has developed a set of guidelines for corpus management that she hopes will be useful for documentary linguists in the creation of an orderly, archive-ready, language documentation corpus. This talk will present those guidelines with examples from AILLA's materials. In the brief time allowed she will only touch on the essential elements: documenting consent, labelling, digital formats, and metadata. More detailed information on all of these topics and more is available on the web; the handout will include sites to which linguists can refer for further guidance.

Jeff Good, *the Max Planck Institute for Evolutionary Anthropology, Department of Linguistics, Deutscher Platz 6, 04103 Leipzig, Germany, good@eva.mpg.de*

Databases and archiving

Many types of linguistic data are highly amenable to being stored in databases, including, for example, lexical and typological data. Unfortunately, many commonly-used database programs produce, as a default, resources in proprietary formats that are not suitable for archiving.

However, if appropriate measures are taken, it is possible to use almost any kind of database software and still create resources which can be archived over the long term. To do this, it is important to maintain a distinction between at least two different formats which the data in a database can take on: archival and working.

An archival format for a database is one which is expected to be readable in the long term. Typically, it will take on the form of a text file, perhaps one which uses XML to annotate the data. A working format is typically optimized for data entry and searching. There is nothing intrinsically wrong with making use of a working format as long it is regularly exported to an archival format.

In addition to covering basic distinctions in database formats, this talk will discuss the advantages and disadvantages of using common database software with respect to creating archivable resources.

Gary Holton, *Alaska Native Language Center (ANLC) (University of Alaska, Fairbanks), P.O. Box 757680, Fairbanks, AK 99775-7680, gary.holton@uaf.edu*

Ethical practices in language documentation and archiving

This paper presents some ethical guidelines for language documentation and archiving, drawing on experiences at the Alaska Native Language Center archive and other primary language archives. A clearly defined approach to intellectual property rights is crucial in order for a language archive to meet its dual obligations of preservation of and access to language documentation materials. This point is perhaps most obvious with respect to access: proper access cannot be achieved unless legal intellectual property responsibilities are met. But ethical issues are also crucial to preservation efforts. This is because a lack of clear ethical guidelines may actually impede or inhibit the collection of documentary material, leading to the potential loss of irreplaceable data. Creators/authors of endangered language material are reluctant to deposit materials with a language archive without assurances as to the maintenance of intellectual property rights. On the other hand, archives have traditionally been reluctant to accept materials without full legal rights or ownership. Here we suggest ethical guidelines by which language archives can work in collaboration with creators of documentary materials to ensure preservation of materials while respecting restrictions on access to materials imposed by the creators and by language communities.

Nick Thieberger, *the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), Project Manager, Department of Linguistics and Applied Linguistics The University of Melbourne, Vic 3010 Australia, Nicholas.Thieberger@paradisec.org.au*

Archiving and the work flow of field work

Archiving is not something we do at the end of our fieldwork, it is part of everyday work. Recent technological advances have pointed to the importance of planning data management and workflow for ethnographic recording. Recordings should always be of high quality, but it is in the context of small and endangered cultures and languages that the quality of recording takes on new significance (quality here refers both to the content and the form of the recording). If we are the only recorders of the last remaining speakers or performers then we are providing historical documents that will be of use not only to other researchers, but primarily to those recorded and their descendants. So, right from the moment of recording, we must be concerned with making good documents which will be placed into a suitable repository for storage and discovery.

In this session we discuss a workflow that builds in development of archival data. We show that making the initial recordings and their digital representation citable by means of a persistent identifier allows further work to be located with reference to that primary data. Further description of the data with standard metadata terms allows its discovery in the long term.

Timetable

We aim for the tutorial to be four hours in length. A preliminary timetable (using 12:00 as a starting point for convenience) is given below. The slots given include time for both presentation and discussion. However, we also include some slots for general discussion.

12:00–12:20	Gary Simons	*The Open Language Archives Community: Building a worldwide library of digital language resources*
12:20–12:40	Helen Aristar-Dry	*The E-MELD School of Best Practices*
12:40–1:00	Helen Agüera	*Funding archival projects at the NEH*
1:00–1:15	**General discussion**	
1:15–1:35	Mark Kaiser	*Digitizing the Audio Archive of Linguistic Field Work*
1:35–1:55	Gary Holton	*Ethical practices in language documentation and archiving*
1:55–2:15	Peter Wittenburg	*From recordings to an organized language archive*
2:15–2:30	**General discussion**	
2:30–2:45	**Break**	
2:45–3:05	Heidi Johnson	*Preparing documentary materials for archiving*
3:05–3:25	Jeff Good	*Databases and archiving*
3:25–3:45	Nick Thieberger	*Archiving and the work flow of field work*
3:45–4:00	**General discussion**	

Note: It is the tutorial organizers' understanding that another organized session entitled *Unicode for Linguists* is being proposed for the 2005 conference. Since this covers topics in character encoding which are relevant to documentation and archiving, we anticipate that the potential audiences for the two sessions could greatly overlap. Should both sessions be approved for the conference, we would greatly appreciate it if they were not both scheduled at the same time.

Tutorial Organizers

Jeff Good
Max Planck Institute for Evolutionary Anthropology
Department of Linguistics
Deutscher Platz 6
04103 Leipzig
Germany
good@eva.mpg.de
+49 (341) 3550-319

Heidi Johnson
The University of Texas at Austin and AILLA
Department of Anthropology, EPS 1.130
1 University Station C3200
Austin, Texas 78712-1086
hjohnson@mail.utexas.edu
(512) 495-4604