The Archive of the Indigenous Languages of Latin America: Goals and Visions

Heidi Johnson

Department of Anthropology
EPS 1.130
The University of Texas at Austin
Austin, Texas 78712
hjohnson@mail.utexas.edu
www.ailla.org

Abstract

This paper presents an overview of the goals and visions of the Archive of the Indigenous Languages of Latin America at the University of Texas at Austin. The archive is a digital repository of multi-media language resources with a web-based interface. While the primary goal of the archive is to preserve recordings of natural discourse in endangered languages, we are also committed to serving our users by supporting activities that seek to improve the understanding and appreciation of these languages, and to promote their survival.

1. The AILLA community

The Archive of the Indigenous Languages of Latin America (AILLA) was founded at the University of Texas at Austin in 2000[1]. The principal investigators are Joel Sherzer (Anthropology), Anthony C. Woodbury (Linguistics), and Mark McFarland (Digital Library Services Division).

AILLA has two main sets of constituents: the global academic community of linguists, anthropologists, and other scholars, and the communities of speakers of Latin America's indigenous languages. The two groups overlap; there are many speakers of indigenous languages who are also scholars.

Scholars generally have good access to the Internet on a daily basis at their institutions, and are increasingly likely to be aware of the possibilities emerging in digital archives. They tend to be most interested in AILLA as a place to safely house their data collections and as a potential resource for their various research interests. Most of our current collection has come from members of the academic community, and we expect that they will continue to be our primary constituency for the near term, both as a source for new deposits and as consumers of the existing collection.

Speakers of indigenous languages have widely varying levels of access to the Internet, ranging from institutional access on a par with that of the academic community to intermittent access through Internet cafés to the most common case: no access nor any expected in the near future. However, a great many speakers are aware that even intermittent Internet access can solve a host of communication problems for their own projects. Speakers, especially speakers of the more robust languages, are primarily interested in language revitalization and education, cultural preservation and reclamation, and in developing literatures in their languages. Those with whom we have communicated so far have expressed enthusiastic support for our plans, recognizing that AILLA can be a way for them to publish their work to each other, as well as to a narrow academic community. We have recently received a large collection of recordings of cultural reminiscences and a series of primary school textbooks, both collections in a variety of Mayan languages, sent to us from speakers working on language and culture reclamation projects. Support for this constituency is an extremely important part of our mission, which means that we must continue to consult with representatives from as many groups as we can, and to consider their needs at every level of system design.

A third constituency is that part of the general public whose parents or grandparents were members of an indigenous group in Latin America. There are many people in this group living in the United States and elsewhere in the world. We occasionally receive mail from such people, typically seeking information about their heritage languages and cultures that they can pass on to their children. We would like to be able to help them.

2. Key tasks

The key tasks of an archive, according to the Reference Model for an Open Archival Information System (CCSDS, 2001), are acquisition, data management, long-term preservation, and provision of access (dissemination). I would add as an important secondary task the provision of tools and methods for exploiting archive resources to achieve users' goals. In the sections that follow, I will visit each of these key tasks in turn, discussing AILLA's plans, goals and visions for the future.

2.1. Acquisition

Ultimately, we would like to have every kind of resource imaginable for every indigenous language in Latin America: audio and video recordings of natural discourse in a wide range of contexts, with transcriptions and translations; literary works written by native speakers; a full slate of pedagogical materials; dictionaries,

[1] AILLA was started with seed money from the Dean of the College of Liberal Arts at UT Austin. It is currently supported by grants from the National Endowment for the Humanties and the National Science Foundation.

grammars, and scholarly articles; even radio and television broadcasts.

In practice, we can only process so much material a year, and so must focus our efforts with a schedule of priorities:

1. Data from extinct or moribund languages. AILLA's most important mission is the preservation of irreplaceable data from languages that are extinct or moribund.
2. Recordings made on obsolete media, like magnetic reel-to-reel tapes. Even for robust languages, it is vital to digitize these materials so that the data can be preserved and disseminated. This set includes data in obsolete or hard-to-duplicate formats, like field notes and extinct word processors.
3. Breadth of coverage. It is important to seek breadth for our collection so that no region or group feels left out, and to better support typological and comparative studies.
4. Support for the goals of speaker communities. The best way to preserve endangered languages that are not on the brink of extinction is to support efforts to ensure that they stay alive.
5. Depth of coverage. We would like to have the basic Boazian requirements for each language: a grammar, a dictionary, and a collection of texts.
6. Quality of supporting materials: metadata, transcriptions, and translations. Data from extinct or moribund languages must be archived even without supporting materials, but there will come a point when we are less willing to archive recordings without the information that makes them useable by non-speakers.

2.2. Data management

A number of the questions posed by the workshop organizers[2] have to do with data management, a broad category that spans archive functions from processing deposits to providing access tools that are interoperable with other language archives. This section will give AILLA's view on some of these questions.

2.2.1. Diversity of resources

We envision an eclectic collection, representing the diversity of the languages and cultures of Latin America and the diverse purposes of our users. AILLA is like a library in that we intend to preserve our collection forever and make its contents available to a wide public. But it is not like a library in that we accept our resources directly from the creators, rather than from intermediaries, and we expect a far greater overlap between resource producers and resource consumers.

2.2.2. Dynamic resources

We also expect greater volatility in our collection than a library would, since we encourage people to deposit incomplete or unfinished materials[3] rather than risk an accidental loss. We want people to use the archive as a medium for collaboration, since it is more secure and convenient than sharing large resources by other means. This obliges us to be prepared to update resources on request, and to maintain different versions of materials when the differences are interesting (e.g. translations).

2.2.3. Standards, uniformity,and interoperability

Interoperability among similar digital archives is a highly desirable goal. It allows us to duplicate our collections, affording greater security and accessibility. It also makes it easier for users, because we can build common searching tools and strategies.

To achieve interoperability, we must have standards. The metadata that describes each archive resource should have a common core of descriptors, such as the language code (see Aristar & Dry, 2001), and some clue about content and provenance. AILLA has adopted the IMDI metadata standard (IMDI team, 2001), customized for our specific needs.

It would also be useful to develop common methods for packaging multi-media resources. Linguistic data comes in bundles - collections of files in various formats with essentially the same content. These bundles can vary greatly, from a single text file (a dictionary, poem, or textbook) to "metabundles" (a book about several recorded performances which are archived separately). The metadata should describe how items in a bundle are related, so the parts can be properly assembled by the user. Standards for the organization and management of such bundles would facilitate the sharing and searching processes.

2.2.4 Data types and strange formats

We accept data in any format, and produce audio files (both .wav and .mp3, with 1-minute samples of recordings over 10 minutes long), and text files (.pdf format for downloading, with the original version archived as well). We have not yet received any video or photographic materials, but we will process them similarly.

We have to accept any format to serve our mission, since the most valuable data according to our acquisition guidelines is the most likely to arrive in a difficult form. We have to convert these myriad formats into a small number of manageable ones, to facilitate dissemination and long-term preservation. Text formats are the most problematical, since there are so many of them, and we are averse to maintaining archive resources in proprietary formats. Our plan is to spend significant staff hours seeking a sound solution to the text format problem; our dream is that there is one to be found.

2.3. Preservation

Long term preservation is a requirement; in fact, it's the main objective of the whole enterprise. We intend that AILLA will exist as long as there is a library at the University of Texas. Achieving long range stability requires two kinds of effort: research on the forward migration of large data sets into new formats and media; and attraction of funds to keep the whole enterprise going. For the first effort, the plan is to stay abreast of new technologies and build well-managed systems that

[2] Peter Wittenburg, Peter Austin, and Helen Dry.

[3] No one is ever really finished processing their field data, even if they haven't looked at it in years.

facilitate periodic transformation. For the second effort, the vision is an endowment that will guarantee basic archive functionality, supplemented by grants for specific archive projects.

2.4. Dissemination

Making resources accessible to the user community is another central requirement of any archive. Several of the questions posed by the workshop organizations concern this task, and I have added one concerning useability and responsiveness to the various members of user community.

2.4.1. Is online availability necessary?

For AILLA, the answer is emphatically "Yes." Our users are widely dispersed geographically, and in many cases are more likely to have Internet access than reliable mail service or a library with significant language resources. All the metadata must be accessible online, so people can know what is available. We also plan to make most resources available for downloading (in compressed formats), given the appropriate access permissions (see 4.3).

However, since at present most people in Latin America connect to the Internet via telephone, and since multi-media resources are often too large to download at low speeds, we must also maintain a parallel system of offline access. To that end, we are adopting the suite of corpus management and metadata editing tools developed at the Max-Planck Institute for Psycholinguistcs (MPI) (Broeder, etal., 2001), and develop parallel online tools so that our users only have to learn one new trick. Then they can use the online tools to search the whole archive, order CDs with a subset of resources, and employ the offline tools to manage and exploit those resources locally.

We are also obliged to program for a wide range of platforms, testing especially carefully on older computers running early versions of web browsers and operating systems. We have to learn to use new technology in ways that remain compatible with old technologies, or we will frustrate the portion of our user community that we most desire to serve. To make this testing real, we will have to recruit volunteers in a range of situations in Latin America. For example, we have recently found one excellent reviewer and ally in Tulio Rojas, at the Universidad de Los Andes in Colombia, who was an engineer before he became a linguist, and is thus able to be explicit about what exactly does and does not work.

2.4.2. Accessibility vs. protection

The vast majority of our holdings will be publically accessible, but we have provided a graded access system that allows depositors to restrict access to sensitive resources (Johnson, 2001). Sensitivity comes in many flavors: a given resource might be sacred to some community, to be heard only by a certain set of people; it might be prone to mis-use, like music or medicinal lore; it might not be ready for publication; or it might be material that a researcher hasn't yet had time to exploit for her or his own purposes[4].

The graded access system will be supported by policies that advise archive users about their intellectual property rights, and about getting permission from creators to archive their works. We can't fully anticipate the range of possible misuses, but we can explore ways of ensuring that we can prove an item was appropriated from our archive (by means of digital watermarks, for example).

On the theme of archives and protection of sensitive data, however, it is important to note that not publishing data on endangered languages is also a disservice to those languages and to their communities of speakers. We must not let fears of potential abuses prevent us from building these collections.

2.4.3. Useability

Building web interfaces that run well on older platforms is actually the easiest part of the useability equation: we can identify a range of platforms, set them up or simulate them in our lab, recruit field testers, and fix bugs on a regular schedule. But user-friendliness implies more than just not crashing at random intervals; it implies ease of use and intuitive operations. Many of AILLA's users will have virtually no prior experience with computers. Many will have had only two or three years of primary school, and maybe some experience working with a linguist or anthropologist or missionary as a consultant. Our interfaces must be explicit, detailed, and jargon-free, with lots of help, and clear, consistent ways of getting that help.

This is a serious challenge, but not an impossible goal. We will have to devise ways of testing our interfaces thoroughly with a wide range of users. We will most likely find our best allies in this effort to be the language education specialists, especially those who are also speakers, since they will have the best understanding of their students' capabilities (and we can often reach them relatively easily.) The ideal would be for AILLA members to conduct tutorial sessions in various environments, such as a classroom with Internet connections (there are some), at a meeting of indigenous leaders in some capitol city, and an Internet café in a market city where there are many indigenous peoples. This would give us an opportunity to elicit direct feedback while people are using the system and give them a chance to express what they wish were possible as well as what might be wrong with what is currently there.

2.5. Exploitation

The workshop organizers asked "Should an archive add value to the archived corpus?" Although this is not a key task of an archival system, the answer should also be "Yes." The archive can at least provide tools that enable users to add value to archive resources, such as the viewing and annotation tools developed at MPI (Brugman & Wittenburg, 2001).

---

[4] We think it is fair for researchers to have first fruits of data that the collect, provided that it will become publically available after some reasonable interval.

We consider it a part of our job to identify, review, and possibly localize to Spanish any software that could help AILLA users exploit the archive's resources. We would like to see a full suite of tools for language documentation: digitization, transcription, interlinearization, and other kinds of annotation and analysis. We would also like to see a range of tools for developing pedagogical materials[5].

Finally, we will provide services for building a community of people interested in the indigenous languages of Latin America. We are planning a variety of bilingual communications facilities, including a newsletter, bulletin boards, and an on-line journal.

3. Conclusion

What is AILLA's long range vision? We want an endowment that allows us to operate as far into the future as anyone can imagine, and management tools and policies to keep the collection viable. We want to house every last scrap of information in and about the indigenous languages of Latin America, in easy-to-use formats that can be downloaded from anywhere in the world in minutes. We want every scholar who has hopes of a respectable career to recognize that archiving their data is an absolute requirement. We want speakers of indigenous languages to regard the archive as a permanent yet flexible resource that can accommodate all their projects and objectives while safeguarding their valuable writings and recordings, and respecting their wishes with respect to public accessibility and use. In short, we want to provide full-service support for all creative and scholarly work concerning these languages.

What is AILLA's short range plan? To do the best we can to follow the vision, taking small steps and solving the easy problems first.

4. References

Aristar, Anthony and Helen Dry. (2001). *The EMELD Project.* Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, 11-13 December 2001. pp. 11-16.

Broeder, Daan, Freddy Offenga, Don Willems, and Peter Wittenburg. (2001). *The IMDI Metadata Set, Its Tools and Accessible Linguistic Databases.* Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, 11-13 December 2001. pp. 48-55.

Brugman, Hennie and Peter Wittenburg. (2001). *The Application of Annotation Models for the Construction of Databases and Tools: Overview and Analysis of MPI Work since 1994.* Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, 11-13 December 2001. pp. 65-73.

Consultative Committee for Space Data Systems. (2001). *Reference model for an Open Archival Information System.* CCSDS 650.0-R-2. http://www.ccsds.org/RP9905/RP9905.html.

Johnson, Heidi. (2001). *The Archive of the Indigenous Languages of Latin America.* Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, 11-13 December 2001.

ISLE Team. (2001). *Metadata Elements for Session Descriptions.* ISLE Metadata Initiative, Draft proposal version 2.4 (7), May, 2001. http://www.mpi.nl/ISLE/documents/docs_frame.html

Wittenburg, P. and D. Broeder. (2000). *Metadata Discussions at Philadelphia Workshop.* Manuscript.

[5] E.g., the MaxAuthor software from the Computer Aided Language Instruction Group at the University of Arizona (http://cali.arizona.edu/)