How can we ethically put language on the web?

D. H. Whalen, President, Endangered Language Fund.


The internet is a great tool for spreading information around the globe at minimal cost. As sound gets better integrated into this world-wide web, it becomes easier to include material from endangered languages. Since the majority of the world's language do not have an agreed upon writing system, putting them on the web in spoken form makes a lot of sense. But can we do it ethically? This was the question raised at a recent workshop at the University of Texas in Austin.

The workshop was part of the SALSA meeting-the Symposium about Language and Society, Austin, which is in its ninth year. The need for explicit agreement about ethical issues was brought into focus by the launch of UT's AILLA project. This is the Archive for the Indigenous Languages of Latin America (http://www.ailla.org). As with North America, virtually all of these indigenous languages in Latin America are endangered. For some of these language communities, the prestige of being put on the web is very advantageous, both for pride within the community and for political leverage with the majority language of the country. But there are ethical decisions that have to be made at every step of the process. The intent of the workshop was to bring some of these considerations to light, both to make others aware of them and to bring in other perspectives so that a consensus can begin to be formed.

The eight speakers had a variety of issues to raise, and certainly none of them were settled. In fact, most of the talks explicitly raised more questions than they answered. As a workshop to bring awareness to issues that are easy to ignore when only the technical challenges seem difficult, however, this workshop was quite a success.

The clearest consensus was reached on these points:
- "Publishing" on the web is different from publishing on paper. It has different consequences for the authors of the texts and should be treated differently.
- Agreements to be recorded are often made with an individual based on trust with that one person. Putting that same recording on the web brings the whole world into the picture and should not be assumed as part of the original agreement.
- This last point is especially difficult to accommodate when the recordings were made decades ago, before there was an internet to think about.
- The rights to linguistic material fall under the general issue of intellectual property rights. These are currently being debated in relation to indigenous culture for music, dance, and visual arts as well as for language material. In these domains as well, there are many unanswered questions.
- While problems will always arise, it is imperative that web archivists have an explicit ethics policy in place. It needs to cover the known issues but also be flexible enough to accommodate the developments that we know are going to take place in the realm of indigenous intellectual property rights.
- Indigenous peoples often assume that others are making money off of their products. With language material on the web, this is typically not true, but if it does become true, it is apparent that some method of sharing that income with the indigenous group is necessary.

The following is a more complete description of the workshop.

Joel Sherzer (Anthropology, U. Texas, Austin) introduced the session. We first agreed to tape the session so that it could be put on the web (our first ethical decision). We then agreed that two of our presentations would be in Spanish. Then he told us a bit about how he got into this issue. Some years ago, Sherzer recorded a Kuna speaker in Panama and got his agreement about the use of the tape. Permission was granted to transcribe and publish the text in any of Kuna, English and Spanish. But the internet did not even exist

Continued on page 2
Ethics - Continued from page 1

English and Spanish. But the internet did not even exist then, and so discussion of its use did not, of course, arise. Does he then have permission to put the material on the web or not? Clearly, the web brings up new issues, ethical, political and other.

The first speaker was Neyla Pardo (National University, Bogotá), speaking on "The State of Linguistic Studies in Colombia." Colombia is multi-cultural, -linguistic and -ethnic. Spanish is the national language, but there are many indigenous languages as well as creoles, many as yet unclassified. But there is discrimination against indigenous languages. Two large projects are currently underway, one on indigenous languages and one on Spanish. Because of the history of South America, the most isolated groups are in the most danger (including both disappearing through assimilation and direct threats due to violence). It is difficult to study the languages in this area.

Lucia Golluscio, of the University of Buenos Aires, described the situation in Argentina. Not until the constitutional amendment of 1994 were indigenous people in Argentina recognized as "preexistent peoples." The constitution of 1853 explicitly said that the conversion of indigenous people to Catholicism was the primary goal in dealing with these groups, while the amendment recognizes the preexistent cultures and guarantees bilingual education. Their cultures are seen as "profound rivers" which have survived more than five centuries thanks precisely to secrecy and hiding—should this be given up? There has been a history of mistrust, so building up trust in this new system will require collaboration and participation of the indigenous peoples.

The discussion began with the question of how indigenous people can effectively participate in the language revitalization projects that might originate in a different setting than their own. Some of the suggestions were: serving on advisory boards, participating in the Common Ethics discussions, being fieldworkers and linguistic consultants for the projects that do arise, and becoming interns in digital technology.

The next presentation was by Lev Michael (UT, Austin) speaking on "Technical implications of ethical issues in web-based archiving." The most obvious need in electronic databases of indigenous texts is graded access, in which the "stakeholders" of the text can limit access for more sensitive items. The agreements over recording are usually reached between the recorder and the recorder, and it is paramount that such explicit agreements be abided by. But there are many recorded materials for which no further understanding exists, and we must interpret what would be the rights and wishes of the recorder in the digital domain. The graded access system need to be flexible; address the legitimate concerns of the native language community; but also maximize access. Mountains of linguistic data were collected before the internet even existed, so it is not possible for an explicit agreement to have been made at the time of the recording.

Continued on page 3
Obtaining permission at the current time may be impractical or impossible. So, what is the best practice in these cases where the agreement is non-ideal? Should the recorder be allowed to make a decision? Are consultations with the native community necessary? No obvious answers are available.

The next speaker was Chris Beier (UT Austin), describing some of the procedures used in the AILLA project. This is the Archive of the Indigenous Languages of Latin America database at the university (http://www.aiila.org). One of the goals of this archive is to maintain the dataset through changes in technology, to have plans to migrate the archive to new systems as the old ones become obsolete.

There are some difficult ethical issues even for datasets that have an original agreement. What happens if a speaker gives permission for a text to be put into the online archive but later wants it removed? Will it be removed even though archiving was one of the original goals of the recording session? It is probably best to be explicit about this at the outset.

There are also potential cases where a previously uninvolved party becomes involved. Joel Sherzer posed the possibility of having a recording to put into the archive, but then a grandson of the speaker might object, saying that the recording now belonged to him. Should the archivists attempt to mediate these situations? What should be done in the interim (which can be a very long time)? It would be best for the archive as an organization to have made decisions about these cases ahead of time and as explicitly as possible.

Patrick McConvell, of AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies), discussed some of the difficulties of one of the earliest online archives, ASEDA (the Aboriginal Studies Electronic Data Archive). Most of Australia’s 250 aboriginal languages are gone, and ASEDA was created to make material on them available. It has just gone through a period of crisis and dormancy. It is not directly web-based, since documents have to be sent. In Australia, not only the indigenous people but the linguists who work with them were aghast at the thought of putting language material directly on the web.

One issue that comes up is: What are aboriginals getting out of this? A way of thinking about new projects is called "gamma", based on a word for a lagoon where fresh and salt water mix. This resonates with the word "gamma," a space where people perform when many different groups are together in public. What aboriginals want from the internet is a gamma/garma, a place where there can be two-way interaction that will benefit the aboriginal communities.

This raises the question of who the correct representatives are. Is it the community at large? How are individual intellectual property rights to be dealt with? And what if the gatekeepers lock the gate?

Tony Woodbury (UT, Austin) then spoke on “preparing for archive dormancy.” What happens when the money runs out? Unlike books, which can be maintained for relatively low cost, computer archives are expensive to keep running. An archive is a thing, but, more importantly, it is a service. If other groups are approached for support or maintenance of the archive, it is necessary to have a means of enforcing the original agreements about access to the archive. Is it better to have the archive in one place (for control) or many (for access and survivability)?

The next speaker was Joan Spanne of SIL, presenting a case study in the intricacies of bringing a collaborative work to fruition. (Steve Eckert of SIL did much of the work for this case study.) The model that most linguists have is the isolated linguist working with a small group of consultants. But many projects require collaboration across researchers and institutions, especially when the goal is to have a large amount of content become available. This raises many complications, even within one institution, if it is as large and complex as SIL. Work produced in the course of regular employment with SIL belongs to SIL. But SIL does not own the work of non-SIL collaborators. So when beginning a research project, it is best to be explicit (and in writing) about ownership of various aspects of the project. One such project was the publishing of annotated texts in a language that had been worked on by SIL and non-SIL linguists. SIL wanted to produce the book, another U.S.-based researcher became interested, and the community was open to it. Negotiations for the various components involved a lot of give and take, but the final product was a good one.

There were four common misconceptions about intellectual property that this project highlighted: 1) The publisher automatically owns the copyright—this is not necessarily so. 2) The language community owns the copyright for traditional material—in Western law, this is not so, though it could be given to a legal persona. 3) The speaker owns the rights to a recorded text—translations are derivative works which are separately owned, but the publication of it still requires the speaker’s permission. 4) Owning the copyright to the collection means owning the copyright to the parts—no, since editing is an act in its own right, creating a unique work which is copyrighted independently of the copyright ownership of the individual elements.

Doug Whalen of the Endangered Language Fund (ELF), spoke about the ethical issues that have arisen from the ELF’s grant program. ELF has been giving out grants for the past five years for projects working on endangered language throughout the world. The projects are both traditional, as with dictionary making and text collection projects, and non-traditional, as in the production of a Choctaw videodrama and the support for a weekly Dakota radio program.

ELF does not own the material collected under these grants, though it does request the right to use excerpts on the internet. The reaction to this request has run the gamut from those groups who would like to see everything on the web to those who will not even allow the material to be sent to the ELF at all. The rationale for such positions can be compelling, even if it is difficult to iron out the differences between

Continued on page 4
Ethics- Continued from page 3
ELF's goal of greater access and the limitations requested by the native groups. Right now, ELF has taken the position that access has to be in place or we will not issue the grant; there are more worthwhile projects than ELF has the money for as it is, so this does not restrict the range of awards very much.

The time dimension is most forcefully brought out by a project of the Long Now Foundation, the Rosetta Project (http://rosettaproject.org). This is a collection of texts and language material for 1,000 languages that will be stored on nickel disks designed to last 2,000 to 3,000 years. While the amount of language material that will be preserved will only give a key into each language, the project is an interesting one and one which ELF is collaborating on, at least in the collection of the Swadesh word lists.

To be ethical, a digital archive of language material has to be accessible to the native community, even if they do not have internet access. While this sounds contradictory, it is quite true. It must be possible for language material to be taken from the web and put into a format that the native community can use if it is to fulfill its need to serve the native speakers as well as the linguistic community. This means that audio recordings should be easy to re-record onto cassettes or cds so that these more accessible formats can be made use of. Written material should be in such a form that it is easy to print out just the parts that the community needs.

Patrick McConvell endorsed looking for usable products from native communities. We could at least give suggestions for how to make tapes and transcripts (for budding native linguists) on the archival web site. We could also post ideas about maintenance techniques, intellectual property rights, copyright, etc.

Joel Sherzer pointed out that connections in Latin America are currently too slow to allow for the access of the audio in any event. This shows that internet access is a continuum, not an all or nothing event. It also points out the need for others outside the region to be able to transfer audio into other formats.

Lucia Golluscio wondered whether indigenous people should be on the advisory boards of these databases. Joel Sherzer said that not only should they be, they are on the board of the AILLA project. But how many such representatives is enough?

Doug Whalen reiterated that it seems immoral at this point not to put language material on the web. Joel Sherzer reminded us that there is a great deal of material that has already been collected that is out there in desk drawers, shoe boxes and attics. Some of this material is on languages that have since gone extinct. We should not allow them to go extinct again.

A slightly expanded version of this report can be found on our web page (http://www.ling.yale.edu/~elf/ethics.html).