Ethical issues in legacy language resources

Carolyn O'Meara
Department of Linguistics
University at Buffalo
609 Baldy Hall
Buffalo, NY 14260
USA
Phone: 716-645-2177
Fax: 716-645-3825
ckomeara@buffalo.edu

Jeff Good
Department of Linguistics
University at Buffalo
609 Baldy Hall
Buffalo, NY 14260
USA
Phone: 716-645-2177
Fax: 716-645-3825
jcgood@buffalo.edu

Abstract:
Recently, there has been extensive work in linguistics to develop recommendations for digitizing legacy language materials. However, relatively little work has been done on the social and legal concerns regarding rights and access to these materials, most of which were created when concerns surrounding intellectual property were less sensitive than they are today. We discuss four issues related to establishing rights and access to legacy language materials: (i) determining what "community" they should be associated with, (ii) establishing rights retroactively, (iii) establishing rights and access to "orphan" works, and (iv) assessing the sensitivities associated with different genres.

Keywords: ethics, legacy materials, digitization, archiving, endangered languages

Article outline:

5. The role of the researcher as gatekeeper
6. Prospects

1. Introduction

One of the products of linguistic research that took place in the latter half of the twentieth century is a large quantity of language material in the form of analog audio recordings, paper transcripts of those recordings, and associated field notes. Recently, there has been extensive work in linguistics to develop recommendations for digitizing materials like these for a variety of uses which range from long-term preservation to web-based dissemination (see, for example, Bird and Simons, 2003). However, relatively little work has been done on the social and legal concerns regarding rights and access raised by these so-called "legacy" materials, most of which were created when concerns surrounding intellectual property were less sensitive than they are today. In many respects, these issues are significantly more complicated than the technical concerns surrounding digitization because the questions that are raised can be very specific to particular researcher-community configurations. Nevertheless, addressing these concerns is a prerequisite for being able to both archive such materials and determine what modes, if any, are appropriate for their dissemination. Furthermore, we should immediately acknowledge that, more often than not, older legacy materials are in the possession of researchers, not speaker communities. Therefore, the latter group is more likely to be negatively impacted if the rights and access to these materials are not clarified.

In this paper, we discuss some of the issues which surround the process of establishing rights and access to legacy language materials. In particular, we introduce four issues we believe need to be addressed both in general terms and in the context of particular digitization and archiving projects: (i) devising principles to determine what we mean by "community", (ii) establishing rights to a resource retroactively, (iii) establishing rights and access to what we will refer to as "orphan" works, whose creators cannot be readily established, and (iv) assessing how the sensitivities associated with different genres may affect the archiving and dissemination process. We will not be able to provide definitive answers to the questions these issues raise. In fact, we doubt that there will ever be definitive answers to these questions. What is important, we believe, is to initiate a discussion which will allow each research project to consider these issues in more informed ways. In addition to these four issues, we will also briefly address how the researcher's role as a likely "gatekeeper" to their materials relates to the larger themes of this paper.

Within the broader context of ethical issues in fieldwork, one goal of this paper will be to highlight that a key component of ethical fieldwork is to consider the range of ethical issues surrounding the resources we collect. This paper, therefore, builds on work like that of Thieberger and Musgrave (2007) and addresses similar issues (though from a different perspective) as Dorian (this issue), Innes (this issue) and Macri and Sarmento (this issue). Furthermore, it is intended to complement more general work like Dwyer (2006) and Rice (2006) by exploring, in more detail, topics which they were only able to discuss more briefly (see, for example, Dwyer, 2006, pp. 41–49 and Rice, 2006, pp. 145–147). While the focus of this paper is on legacy materials, many of the issues that will be raised are also relevant for materials being created today, and we hope that this focus on the "past" will assist people in planning for the future.

In the discussion to follow, we will try to address two distinct points of view: that of the linguist and that of a language archive, reflecting the fact that our experiences in

this area derive largely from work on a pilot project to establish a new digital language archive, discussed below in section 2. In addition, we will often speak of entities like the "archive", "the researcher", and "the speaker". We are aware that such labels represent potentially problematic simplifications. (For example, what if the researcher is also a speaker?) Nevertheless, we adopt them here for the sake of moving the discussion forward. We further assume, in general, a scenario where the individual who deposits materials in an archive is a researcher since our primary audience here is academic linguists. However, that, too, represents a simplification from reality. We will also treat a "recording" as the prototypical linguistic resource, though most of the discussion would apply equally as well, in principle, to field notes or other data types.

2. Background on the Northeastern North American Indigenous Languages Archive (NNAILA)

Our context for this discussion is work done on a pilot project to determine the feasibility of establishing a digital language archive which we are provisionally calling the Northeastern North American Indigenous Languages Archive (NNAILA).[1] The aims of NNAILA, if successful, will be to serve as a trusted repository of digitized and "born-digital" materials on the Native American languages of northeastern North America (encompassing both the United States and Canada). Historical and political concerns surrounding linguistic work conducted on these languages required that a project like NNAILA proceed carefully in order to ensure that the concerns of all stakeholders in the relevant language materials are appropriately safeguarded, which, we believe, makes our experiences working on this project generally relevant. Here, we give some general background on the project so that the context of subsequent discussion can be made clearer.

Before doing so, however, we should make sure to distinguish three notions that will be important throughout the rest of the paper. The first of these is *digitization*, the process through which non-digital materials (e.g., traditional audio cassettes or paper documents) are analyzed by a machine in ways which allow for the creation of digital representations of their content (e.g., by scanning a document and producing a file that is readable on a computer). For materials created before digital storage technologies were the norm, digitization is a prerequisite to *digital archiving*, which is the process through which digital resources become housed in a *digital preservation archive*, an institution with a commitment to preserve materials in perpetuity. (In this paper, we will generally use the simpler term *archive* to refer to a *digital preservation archive*.) A related concept is *digital dissemination*, the process through which digital materials are copied and distributed. This is most commonly done, these days, via the internet. Digital preservation archives may offer digital dissemination as a service, though strictly speaking it is a secondary function for them, since their primary function is to preserve.

During the pilot phase of NNAILA, the focus has been on digitizing, archiving, and providing appropriate means of dissemination for materials documenting Onondaga [ono], an Iroquoian language.[2] Onondaga was chosen largely due to practical considerations. Most importantly, a member of the Onondaga Nation, Dr. Percy Abrams received his Ph.D. in Linguistics from the University at Buffalo (Abrams, 2006), giving us a point of contact to the community who had a clear understanding both of the culture

of academic linguistics and how the work of linguists could positively (or not) impact the Onondaga community. In addition, there were also substantial recordings on legacy media collected by various linguists outside the community. By way of background, Onondaga is spoken in parts of central New York (at Onondaga Nation) and in the area near Brantford, Ontario (at Six Nations Reserve). There are less than a hundred fluent speakers remaining (see Mithun, 1999, p. 423). The Onondaga community is shifting toward using English only and, because of this, the Onondaga community has established language teaching programs to ensure that the language is learned by a new generation of members of Onondaga Nation. One hope of NNAILA is that its activities will allow recordings of the language to be more efficiently disseminated to the community for use in Onondaga language curricula (for more information on the role that technology can play in language revitalization efforts see Galla, 2009).

The first stages of the pilot project were relatively straightforward in nature, relying largely on technological solutions to problems of preservation which had already been fairly well-explored within the linguistics community, for example, in the context of the Electronic Metastructure for Endangered Languages Data project (E-MELD) (see, e.g., Boynton et al., 2006 for discussion of E-MELD and Austin, 2006 for more general background). Analog recordings (mostly from reel-to-reel cassettes) were digitized to an archival file format and associated with basic metadata provided by the researcher. It is after this point in the archiving process where the issues at the heart of this paper become relevant. Once the resources had been transferred to a digital format which was not only archival, but also allowed for cheap duplication and dissemination (via the internet), questions of rights and access which had previously been largely theoretical suddenly required real answers. (See also Thieberger and Musgrave, 2007, pp. 29–30 for a similar point.)

In the rest of the paper, we will not focus on the work of NNAILA in particular but, rather, talk about the relevant issues in more general terms since the issues we have encountered, we believe, are not specific to these Onondaga materials but are of much wider significance.

3. Rights and access

Two crucial concepts around which the key issues raised in this paper revolve are *rights* and *access*. These concepts, of course, are of much broader relevance than the domain of language resources, and we do not intend to explore their possible senses in detail here. Rather, we use each term relatively informally as follows. By *rights*, we refer to the permissions or entitlements a given person has with respect to the use of a particular resource, including both legal rights and moral rights. The latter type of rights are those emanating from some notion of ethics— whether professional, community, or of some other kind—about appropriate uses of resources, whether or not they happen to be enshrined in law. In the present context, the crucial domains of interest with respect to rights are rights to specific language recordings as well as rights to the intellectual and artistic content found within those recordings. By *access*, we refer to the permissions granted to individuals not identified as having specific rights to a resource regarding the extent to which they should be allowed to inspect or make copies of a particular resource

or a representation of the content of that resource (e.g. in the form of a transcription or translation).

While the questions surrounding rights and access are both legal and ethical, we will give greater discussion to the ethical side of the issues rather than the legal ones.[3] One reason for this is, of course, the fact that neither author of this paper is qualified to interpret the finer points of copyright and other relevant areas of the law even within the United States, let alone the rest of the world.[4] Another reason is that, in some sense, legal issues surrounding rights and access are, from the perspective of the linguist, more straightforward than ethical ones. Laws are determined outside of the research context, and one must simply come to an understanding of how to follow them in the particular circumstances that one is working in. Ethical issues, on the other hand, are more difficult to deal with because we have an important role in shaping our responses to them, and it is our responsibility, both at the level of the profession and at the level of the individual, to make the complex determination both of one's general ethical obligations as a linguist and of how those obligations apply to specific recordings whose creation one has been a part of, with full recognition that each recording may raise its own distinct ethical concerns.[5]

In comparison with work done on developing digital standards for language materials (as exemplified, for example, by the E-MELD project), relatively little work has been done on issues relating to rights and access to digital language materials (though Thieberger and Musgrave (2007) is a noteworthy exception). This is almost certainly in part due to the difficulties that arise in arriving at generalizations in an area where details of personal relationships play a prominent role, but there are other issues involved as well. For example, linguists typically are not especially concerned with obtaining permission to access their own materials because, being in possession of them, this is rarely an immediate concern.[6] In contrast, they are more likely to be concerned about what digital formats they employ because making use of the right format is likely to facilitate their own research while making use of the wrong format may render their recordings unusable in a few years time.

Furthermore, for many linguists—particularly those who began working decades ago when the possibilities for dissemination afforded by the internet were barely imaginable—issues of access and rights can be quite uncomfortable, not because they failed to behave responsibly in any way, but because they fear that raising questions regarding rights and access to their materials in the age of the internet may upset a manageable status quo.[7] Under such circumstances, they may have decided that the best option is to simply "keep quiet", even if this means that important questions about the long-term use of their materials are left open, perhaps until after their death, after which point archiving the resources becomes much more difficult because the original creator is no longer available for consultation regarding resource provenance or content.[8]

In principle, of course, many of the rights and access issues which we will be discussing below could apply to analog resources just as well as to digital ones, as exemplified in Debenport (this issue). For example, determining who represents the "community" is not a specifically "digital" question. In practice, however, the rise of digital media and the internet are of enormous consequence here. Previously, the barriers to dissemination of language resources were so high that one could avoid devising general schemes of rights and access and, instead, rely on ad hoc procedures on a request-

by-request basis. In particular, in the past, materials dissemination required specific human intervention (for example, the linguist may have had to make a duplicate copy of a recording using their own equipment). This human component to dissemination facilitated highly individuated judgments regarding rights and access. Now, however, if we want to take advantage of new technologies facilitating cheap and timely access to language materials, we require rights and access schemes that can be generalized to the point where they can be implemented on machines, for example in the form of a website giving graded access to different classes of users.[9]

Issues like these must be viewed against the backdrop of a discipline that, for the most part offers linguists no workable framework for conceptualizing, developing, and implementing appropriate rights and access statements for their materials in the first place. In some cases, there may be particular institutions, for example preservation archives, which can facilitate this process.[10] But many parts of the world lack a dedicated preservation archive (one of these being northeastern North America, which is one motivation for the development of NNAILA). Furthermore, the primary role of a preservation archive is merely to implement rights and access restrictions, and to neither formulate them, which is more properly the responsibility of the depositor, nor to devise a framework for establishing them in general terms, which is more properly the responsibility of the profession of linguistics itself in cooperation with other relevant stakeholders to language materials.

4. Four issues related to establishing rights and access to legacy materials

4.1. *The notion of "community"*

Discussions of linguistic fieldwork, in particular fieldwork on endangered languages, frequently invoke an opposition between the "researcher" and the "community". This opposition comes through most clearly, perhaps, in work discussing collaborative fieldwork models where the research agenda is shaped both by the linguist and by the group the linguist is working with (see, e.g., Czaykowska-Higgins, 2009; Dwyer, 2006; Mithun, 2001; Penfield et al., 2008; and Yamada, 2007, among others). However, understanding just what the "community" is in a given research project is not necessarily straightforward. Thus, for example, the Linguistic Society of America's statement of ethics can only offer relatively uninformative generalities on the topic: "While acknowledging that what constitutes the relevant community is a complex issue, we urge linguists to consider how their research affects not only individual research participants, but also the wider community."[11] Even trying to delimit the relevant notion of community to something like "speaker community"—that is, to try to define it in terms of who actually has competence in speaking a given language—may still be of little help since a given community's notion of "speaker" may be distinct from the linguist's notion (see, e.g., Evans, 2001, for discussion in an Australian context of the complexities of determining who the "last" speakers of a language may be).

In fact, the word "community" can be applied sufficiently broadly to render its usefulness in discussion of rights and access to language resources questionable. To give an example of the kinds of ambiguities one may find, on the website for the Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas at Austin[12], one of the three missions of the archive is listed as "community support". In

this section of the page, the word "community" is used in two different ways, namely, in the context of "fostering the community of speakers and scholars" with an interest in the archive's language materials, as well as the more usual sense of community in a linguistic context to mean something like the community of "speakers" of a language (keeping in mind the difficulties associated with the notion of "speaker" just discussed above) or, in the case of a moribund or extinct language, people who consider the language to constitute part of their heritage. While we borrow this example from the AILLA website, we do not mean to single out AILLA's vague use of the word community as representing an exceptional case. Rather, we use it as a fairly clear illustration of how "friendly ambiguities of language" (Sapir, 1949 [1932], p. 516) can lead to situations where an apparently obvious principle like "respect the wishes of the community" may obscure the presence of a complex issue which has no obvious general—or even case-specific—resolution.

In addition to the problem of establishing just what the "community" is, there is a further concern regarding who has the authority to speak for community interests when deciding on rights and access restrictions to materials. At one extreme, one has languages like English, where the notion that anyone can speak for "the community of English speakers" is not only impractical but also inconsistent with a cultural value associated with the most prominent English-speaking communities where individuals are taken to have clear rights to their "property" (intellectual and otherwise). At the other extreme, there are situations like that described in Wilkins' (1992) examination of his fieldwork in Australia where explicit agreements exist regarding precisely what administrative body has the power to review and control the work of the linguist. It is clear that there cannot be a "one-size-fits-all" solution to this problem. However, at this point, as a field, we even lack a set of principles through which one can determine what the legitimate authority might be to speak for a community. Again, Wilkins' situation is instructive. After having established a research relationship with one community, his assistance was requested to help with language development in another community where speakers of the primary language of his research, Mparntwe Arrernte, were to be found (Wilkins, 1982, pp. 182–183). His initial reaction to this request—that it would be a natural extension of his services to the *speaker* community—turned out to be rejected by the council which had official control over his work and which represented a community distinct from (though overlapping with) "vaguer notions such as the Arrernte...speaking community" (Wilkins, 1992, p. 183).

Before digital technologies made the copying and dissemination of language materials relatively trivial, precisely delineating who belonged to a given community using operationalizable criteria would have been helpful, but not necessarily essential. The *technological* barriers to access of materials produced a *social* setting conducive to ad hoc case-by-case decisions. However, if we want to realize the promise of digital technologies for allowing individuals to easily access materials which they have a legitimate interest in, the process of determining who is a community member needs to be at least somewhat depersonalized. A digital language archive, in particular, will not be in a position to create an appropriate definition of "community" for all of the communities represented in its materials. Rather, its job is merely to enforce access restrictions, requiring that the groups to which those restrictions pertain must already be well-defined. When something as basic as community membership is difficult to define,

how is the archive supposed to facilitate access to the language materials it houses and has a duty to protect, but at the same time provide equitable access? This is ultimately not a question for archives but for linguists and the communities they work with as in the case described in Macri and Sarmento (this issue).

For the purposes of this paper, let us discuss how the notion of "community" relates to NNAILA. Given that the archive is centered around language materials, an initial criterion for "community" is that it consists of individuals united, in some sense, by a common language. With a focus on languages of northeastern North America, in some cases, of course, a given language may no longer have any native speakers or only a few but nevertheless be associated with a group that sees the language as a key component of their cultural heritage. Therefore, for NNAILA, defining "community" as "speaker community" would not be appropriate. In the case of Onondaga, as discussed in section 2, there are relatively few remaining fluent speakers. However, there is a language program attempting to revitalize the language (Abrams, 2006, p. 3). We have begun consultation with individuals at Onondaga Nation, who are being trained as language teachers, during the pilot phase, and we hope that this group will help us develop consensus regarding how to define the Onondaga community appropriately in the future. We, thus, have avoided arriving at a general definition of the "Onondaga community" at this stage by restricting our work to a sub-community of clear importance to the project which is also much easier to concretely define.

Ultimately, for the purposes of internet-based dissemination, any consensus that emerges regarding what the "community" is in a given context is will have to be operationalized—that is, reduced to criteria that can be straightforwardly applied by an archive. While certain obvious parameters for operationalization present themselves, none that we are aware of is at all ideal. For example, one could employ a geo-political distinction in order to determine community membership, based on residence in a certain village or town. This seems like a rather appealing and convenient way to determine community membership, but, in the Onondaga case, for example, it would problematically exclude individuals with strong community connections but who happen to live outside of official Onondaga areas.[13] Similar problems would arise if the community were defined in terms of a notion like nationality, since it might exclude access to materials by individuals who could reasonably claim that some set of language materials represented an important part of their heritage but who may not have official status as members of the relevant nation.[14] The goal should be how to achieve a balance between what is most appropriate socially and what can be straightforwardly operationalized. To the best of our knowledge, though, the conversations required to help us understand how to reach such a goal for any given case have yet to be seriously embarked upon.

While we have focused solely on the "community" up until this point, we should keep in mind here the important role of another party: the researcher who collected language materials. We will come back to this topic below in section 5, but here, we should briefly point out that the community that a given researcher may have worked with will not necessarily be the same as the community with a valid stake in the resources collected. As mentioned above, this scenario was described in Wilkins (1992). Assuming a given researcher is still living and in sufficiently good standing with the relevant community, they may have important input on just how the community should

be conceived with respect to the resources they collected. This is just a specific (though important) instance of a general principle that input from all valid stakeholders should be taken into account when considering an important issue of delineating what a given community can call its "own".

4.2. *Establishing rights to resources retroactively*

In the case of NNAILA or any other language archive which seeks to preserve analog materials via digitization, it is necessary to consider how to work with resources that predate the internet, and the possibilities it offers for widespread dissemination. Moreover, it is possible that these resources were created by the researcher without discussing with consultants what restrictions, if any, they would like placed on the recording if it were to get deposited into an archive or if it could even be archived. Under such circumstances, it is not necessarily the case that the researcher failed to ask these questions out of a lack of care. Rather, the wider range of uses of materials that are now possible may simply not have been anticipated at the time the recordings were made. So, what is the process for establishing rights and access to such materials after they were created in order to allow them to be appropriately archived?

In the ideal case the researcher can go back and discuss the relevant issues with the consultants who participated in the creation of the resources, and collectively they can decide on what access restrictions might be needed and work out a mutually agreeable system of rights. Under such a scenario, an archive like NNAILA is actually at an advantage compared to some other archives, to the extent that the speakers involved in the research will have enough access to contemporary technology that they can generally be expected to have a reasonable understanding of notions like digitization, digital archiving, and internet-based dissemination once they are explained to them by the researcher or archive (but see Robinson (this issue) for a more detailed and nuanced discussion of this issue). This may be less true for archives focusing on parts of the world where there is less access to digital technologies.

However, frequently, the process is not as easy as just described. Speakers may have passed away prior to the archiving of materials, or become otherwise unreachable. Even if the speakers are still alive and are easily contacted, the process can be not only time consuming, but also psychologically daunting to the researcher. What if speakers, years later, want to limit even the original researcher's access to the materials? Even if this possibility is highly unlikely, it may still make the researcher hesitate before trying to retroactively work out a system of rights and access explicit enough for the age of the internet.

In cases where the speakers have passed away, who gets the right to say what happens to the deceased person's recordings? The law may recognize special rights to a speaker's heirs, particularly in the domain of copyright (see section 4.3 for further discussion of issues relating to copyright). At the same time, for a severely endangered language, it would also seem possible to make a case that the "community" (though, see, section 4.1) should have a say in who can access the materials, especially if they have special value for revitalization. It could also be the case that the researcher's ideas of what the native speaker consultant would have wanted in terms of rights and access differ from those of the native speaker consultant's family members or of the community at large (see for example Dorian (this issue)). Finally, if the native speaker consultant is no

longer living, who decides what parts of the recording constitute sensitive material with respect to that speaker? We will come back to some of these issues in 4.3 and 4.4.

Some overall principles seem clear enough: If all individuals involved in the creation of a recording are alive, then they should play primary roles regarding issues of rights and access. Beyond this, the field of linguistics will need to open up a discussion on what principles should govern this determination when any of the parties involved in the creation of the resource have died. While we focused here on cases where the speaker has died—following our general framing of the issues here where the researcher is treated as the archive depositor—of course, the researcher, too, may die before materials have been properly archived and then appropriate individuals will need to be found to speak for their interests as well. Here, too, the tension between heirs and the "community" arises, though, here, the relevant community is that of academic linguists, since very often one's professional colleagues may have a better sense of the significance of language resources than one's family.

4.3. *Establishing rights and access to "orphan" works*

As a technical term in the context of copyrighted material, works are deemed to be *orphan* when a copyright owner of a given work cannot be identified (see Newman, 2007, section 7). In the present context, we extend the notion "orphan" work to resources where the main stakeholders, namely, the speakers, the researcher, or the community, may no longer be able to make decisions about access and restrictions to the materials they would otherwise have a stake in and where there is no obvious—or at least economically viable—way to determine who can speak for these parties. In the eyes of an archive, coming into possession of "orphan" language materials may be considered lucky insofar as these types of resources could have just as easily remained inaccessible and unpreserved, allowing them to deteriorate and potentially be lost forever. The most obvious scenario through which this might occur would be if one researcher obtains access to the materials of another researcher who has passed away (see Newman, 2007, Q13). Orphan language materials create problems for the archive since it is unclear how the archive manager is to determine access to these resources due to the fact that they are unlikely to be accompanied by the information needed to determine their rights and access.

With respect to copyright, when the original copyright owner passes away, it is quite likely that it is held by heirs, not whoever possesses the materials (see Newman, 2007, Q13).[15] Assuming the heirs can be located, the archive manager, for instance, can then ask them to transfer copyright to another linguist, an archive, a community representative, etc. If they cannot be located, the work is orphaned from a copyright perspective. However, this still leaves open the broader question of "moral ownership" from the perspective of the researcher and the community.[16] Which individuals, if any, should have unrestricted access to the materials? Is there someone in the community who has rights over the materials (which, again, brings up the issue of just who the "community" is—see 4.1)? And, perhaps, most importantly, if an orphaned work is deemed of great cultural interest but of little commercial value, how do we balance the ethical tensions of this situation where ignoring copyright concerns and, thereby, possibly breaking the law, may result in tremendous positive impact for research or for the community?

For endangered or moribund languages or languages with no speakers, it seems crucial to avoid the scenario that important works simply lie unused because no one knows who can use them. But how can this be handled in a way that is equitable to all stakeholders? And how does the archive navigate between the customs and needs of the community and those of the larger legal system in which it operates?

In this context, it seems worth pointing out that, at least in the United States, copyright law has changed in ways which make this problem considerably more acute than before. Whereas formerly, works fell out of copyright after twenty-eight years if copyright was not explicitly renewed, under current law copyright extends from seventy to 120 years (Newman, 2007, p. 29) without a need for explicit renewal. Before, it may have simply been possible to "wait out the clock" on older resources until the twenty-eight year barrier was passed, but now copyright restrictions are approximately as long (and even longer) than a lifetime, in some cases putting them in legal limbo for generations while speakers who could make use of those resources pass on before getting a chance to legally use them. In this case, at least, a fairly clear, if difficult to achieve, solution presents itself: Linguists and communities could lobby to change copyright law. Of course, this would be a major undertaking, but at least a resolution here is much more straightforward than, for example, the issue of defining "community".

4.4. *Sensitivities in resource content and how they may affect archiving and dissemination*

Different genres of resources are likely to be associated with inherently different levels of sensitivity regarding rights and access. For example, there is less likely to be conflict over the ownership of the intellectual content of a recording of a lexical elicitation session than an original song—to take just one dimension of the problem, the former is not subject to copyright, unlike the latter. However, if the recording of the lexical elicitation contains gossip in between elicitation, the recording might be just as or even more sensitive than an original song. Understanding these sensitivities will facilitate making an initial determination regarding which recordings and transcripts of a collection are likely to require the most discussion, as well as how to treat orphan works, for example.

Some complications derive from the differing requirements of copyright law and community notions of intellectual property. For example, copyright does not cover folktales (Newman, 2007, Q9). So, a community wishing to retain ownership of folktales may require very tight access restrictions on any recordings of them to prevent unwanted distribution of their content since, at least in the United States, they may have no legal recourse against those who use the content of a folktale against their wishes. Does this mean that, in general, folktales should be treated as a "sensitive" genre? Or are they only sensitive if there is a good chance they will have commercial value?

Additional sensitivities arise around recordings that contain personal information from or about speakers who may wish for access to such information to be restricted, if not forever then at least for an extended period of time. What should be done in cases where a recording contains personal information but the speaker has passed on? When can we assume that such a recording can be distributed and when can we assume it cannot? What about instances where other individuals are mentioned on a recording? Do those individuals, in addition to the speaker, need to be contacted before a recording can

be made accessible? In the case of recordings that contain a mix of personal and impersonal information—or, even worse, recordings whose content is unknown—who should be allowed to listen to them to determine if there may be sensitive information on them? Even if the archivist is deemed to professionally be in a position to perform this task, they are unlikely to understand the content of materials (for reasons of language or lack of context) in order to be able to make an appropriate assessment. For a highly endangered language, it may even be the case that the set of remaining fluent speakers who would be able to readily understand the recordings may be the worst group of people to hear the recordings from the personal perspective of the original creator. Who has the authority to determine a solution to such a problem? Is it possible to devise a set of general principles regarding sensitivities of different genres that apply across all communities or are the appropriate procedures going to differ on a community-by-community basis?

One procedure that may be of general use for cases like this can be borrowed from the world of "human subjects" research: data anonymization. This could be technically impossible when it comes to audio and video recordings, as attempts to anonymize them by manipulating voice quality could lead to the data being useless for linguistic purposes, at least with the technology that is presently easily available. However, for some data types, like transcriptions of recordings, this might be a feasible option to facilitate access. Corti et al. (2000) indicate that anonymization is not appropriate for many data sets and that "gate-keeping" (i.e., the use of variable access restrictions) is a good alternative to anonymization for securing appropriate access measures to data.

Anonymization brings up an interesting issue that frequently comes up when linguists in the United States apply for approval to work with human subjects through an institutional Internal Review Board (IRB). IRBs tend to, as a default, encourage researchers to maintain anonymous data and to destroy the data after a few years. However, this is at odds with general practice in linguistic work wherein consultants are acknowledged in ways consistent with their intellectual contribution to research. The same impulse that results in IRBs taking anonymization as a default stance is at work in the idea just discussed above that anonymization may be a reasonable route to deal with sensitivities of the content of language materials. However, even if anonymization is taken to be a valid option, one must keep in mind that actually destroying data in order to anonymize the participants in the creation of a resource runs counter both to the mission of language archives to preserve the contents of their deposited materials and linguists' position that their consultants are active intellectual contributors to their work. Accordingly, the default procedure for anonymization, if employed, should be one where an original non-anonymized copy of a resource is preserved even if the anonymized version is the only one made widely available.

As with the other questions we have raised here, there will be no simple solution to determining what kinds of content may be more sensitive than others. Still, however, this should not prevent the field from articulating some general principles in making the relevant assessments.

5. The role of the researcher as gatekeeper

Before being deposited in an archive, language materials may be held either by researchers or by speakers (or speaker communities). Here, we have focused on a scenario with the researcher as depositor since, as linguists ourselves, it is the scenario we feel most able to comment upon. In this sort of situation, it is important to keep in mind that, whether this job is desired or not, the researcher acts as the initial gatekeeper to materials they have collected if only because they are the ones who possess them—at least until the materials are archived.

This raises then an additional broad issue that does not center on the rights and access to resources which the researcher intends to deposit but, rather, on what obligations the researcher has to deposit collected materials in an archive at all (or make them otherwise widely available). Corti et al. (2000) discuss the skepticism they have heard from researchers who work with qualitative data surrounding archiving and potential re-use of their data. They indicate that the primary concerns of researchers tend to be along the following lines: a) once material is archived, the researcher could lose control of data in the archive, especially control of how the data is re-used; b) the researcher is unsure whether or not they even have permissions from project participants to archive the data; c) the researcher does not want to ask participants retroactively whether they have permission to archive the data (Corti et al., 2000).

In some cases, these concerns may stem from a lack of information regarding what archiving entails and what the procedure of archiving and access of materials is. More fundamentally, these concerns seem to be a result of the fact that up until fairly recently the role of archiving as it pertains to the social sciences has not generally been explicitly discussed (with some exceptions, for example, within oral history (Corti et al., 2000, paragraph 33)). We see that an important step to alleviating many of these concerns is to begin a dialogue with researchers in the social sciences generally, as well as within specific subfields like linguistics, regarding the extent of the control a researcher can assert over the data they collect. While this issue has not been addressed in any general way in linguistics, there are examples from other fields that can be examined. For example, within archaeology, there is a Register for Professional Archaeologists, with a set of standards for research performance which includes a stipulation that if, after ten years, an archaeologist has failed to produce a scholarly report for a project, then data emanating from the project, "should be made fully accessible to other archaeologists for analysis and publication."[17] This ten-year requirement is viewed as striking an adequate balance between a researcher's right to have first privilege of the use of the data they collect within their own scholarship and the right of other archaeologists to make use of important data for which there may be no other available source. No such system is in place for linguistics, though one could clearly argue the merits of a comparable kind of restriction, though not only emphasizing availability of data within the profession but also with the communities that linguists work with.

We should emphasize that one way to avoid issues revolving around the researcher's role as (possibly unwilling) gatekeeper is for dialog among the researcher, the community, and the archive to happen sooner rather than later. The hard questions will need to be addressed at some point, and this is better done while people's memories of the circumstances around which a recording was made are still clear. Thus far, it seems that the responses to questions surrounding rights and access are more positive than anticipated by researchers. For instance, Corti et al. (2000) indicate that contrary to

the fears of researchers, participants, when contacted retroactively, do not oppose the archiving of the material they contributed to creating.

These things being said, we should also acknowledge that there may be cases when a researcher may decide to withhold materials from an archive for legitimate reasons. Corti et al. (2000, paragraph 31), for example, give as a general principle that data that holds particular risks to an individual if made public may be best handled by not archiving it all. While most of the content of language resources would likely not fall into this category, especially for speakers of languages where there are particular political sensitivities, it is possible to imagine instances where data they produce should not be permanently preserved as may be the case with some of the materials collected by Nambiar and Govindasamy's informants (this issue). As another example, Johnson (2004, p. 145) mentions a case where information was collected about illegal border crossings from speakers of Zapotec. The arguments against archiving such data are clear.

6. Prospects

This paper has raised many difficult questions and issues associated with rights and access of language materials in the context of digital language archiving. There are certainly many more questions that could, and probably should, be raised, especially in this era of increasing language endangerment. Currently in the field of linguistics, much emphasis has been placed on technical standards associated with making audio and video recordings, metadata formats, as well as specifications for digitizing legacy material. However, relatively less attention has been paid to the social and ethical issues that arise in association with the collection of language material, especially as it pertains to the archiving of these legacy materials. By pointing out some of the questions that need to be asked—even if they cannot be readily answered—we hope that we have made some progress towards resolving key issues regarding rights and access not only for legacy materials but also for materials being created today.

References

Abrams, P. W., 2006. Onondaga pronominal prefixes. Unpublished PhD dissertation, University at Buffalo, State University of New York.

Anderson, J. and Koch, G., 2003. The politics of context: Issues for law, researchers and the creation of databases. In: Barwick, L., Marett, A., Simpson, J., and Harris, A. (Eds.), Researchers, Communities, Institutions, Sound Recordings. University of Sydney, Sydney. http://hdl.handle.net/2123/1513

Austin, P. K., 2006. Data and language documentation. In: Gippert, J., Himmelmann, N. P., and Mosel, U. (Eds.), Essentials of Language Documentation. Mouton de Gruyter, Berlin, pp. 87–112.

Bird, S. and Simons, G., 2003. Seven dimensions of portability for language documentation and description. Language 79, 557–582.

Boynton, J., Moran, S., Aristar, A., and Aristar-Dry, H., 2006. E-MELD and the school of best practices: An ongoing community effort. In: Barwick, L. (Ed.), Sustainable Data From Digital Sources: From Creation to Archive and Back. Sydney University Press, Sydney. http://hdl.handle.net/2123/1296

Corti, L., Day, A., and Backhouse, G., 2000. Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 1, Art. 7. http://nbn-resolving.de/urn:nbn:de:0114-fqs000372

Czaykowska-Higgins, E., 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. Language Documentation and Conservation 3, 15–50.

Dwyer A., 2006. Ethics and practicalities of cooperative fieldwork and analysis. In: Gippert, J., Himmelmann, N. P., and Mosel, U. (Eds.), Essentials of Language Documentation. Mouton de Gruyter, Berlin, pp. 31–66.

Evans, N., 2001. The last speaker is dead—long live the last speaker! In: Newman, P. and Ratliff, M. (Eds.) Linguistic Fieldwork. Cambridge University Press, Cambridge, pp. 250–281.

Galla, C. K., 2009. Indigenous language revitalization and technology from traditional to contemporary domains. In: Reyhner, J. and Lockard, L. (Eds.). Indigenous Language Revitalization: Encouragement, Guidance and Lessons Learned. Northern Arizona University, Flagstaff, pp. 167-182.

Johnson, H., 2004. Language documentation and archiving, or how to build a better corpus. In: Austin, P. K. (Ed.), Language Documentation and Description, Volume 2. The Hans Rausing Endangered Language Institute, London, pp. 140–153.

Lewis, W. D., Farrar, S., and Langendoen, D. T., 2006. Linguistics in the internet age: Tools and fair use. In: Proceedings of E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art. http://emeld.org/workshop/2006/papers/lewis.pdf

Liberman, M., 2000. Legal, ethical, and policy issues concerning the recording and publication of primary language materials. In: Proceedings of the Workshop on Web-based Language Documentation and Description. http://www.ldc.upenn.edu/exploration/expl2000/papers/liberman/liberman.html

Michener, W. K., 2006. Meta-information concepts for ecological data management. Ecological Informatics 1, 3–7.

Mithun, M., 1999. The Languages of Native North America. Cambridge University Press, Cambridge.

Mithun, M., 2001. Who shapes the record: The speaker and the linguist. In: Newman, P. and Ratliff, M. (Eds.) Linguistic Fieldwork. Cambridge University Press, Cambridge, pp. 34–54.

Nason, J. D., 1997. Native American intellectual property rights: Issues in the control of esoteric knowledge. In: Ziff, B. and Rao, P. V. (Eds.), Borrowed Power: Essays on Cultural Appropriation. Rutgers University Press, New Brunswick, New Jersey, pp. 237–254.

Newman, P., 2007. Copyright essentials for linguists. Language Documentation and Conservation 1, 28–43. http://hdl.handle.net/10125/1724

Penfield, S. D., Serratos, A., Tucker, B. V., Flores, A., Harper, G., Hill, Jr., J., and Vasquez, N., 2008. Community collaborations: Best practices for North American indigenous language documentation. International Journal of the Sociology of Language 191, 187–202.

Rice, K., 2006. Ethical issues in linguistic fieldwork: An overview. Journal of Academic Ethics 4, 123–155.

Sapir, E., 1949 [1932]. Cultural anthropology and psychiatry. In: Mandelbaum, D. (Ed.), Selected Writings of Edward Sapir in Language, Culture, and Personality. University of California Press, Berkeley, pp. 509-521.

Story, A., Darch, C., and Halbert, D. (Eds.), 2006. The Copy/South Dossier: Issues in the Economics, Politics, and Ideology of Copyright in the Global South. Copy/South Research Group, Canterbury, UK.

Tatsch, S., 2004. Language revitalization in native North America: Issues of intellectual property rights and intellectual sovereignty. Collegium Antropologicum 28, supplement 1, 257–262.

Thieberger. N. and Musgrave, S., 2007. Documentary linguistics and ethical issues. In: Austin, P.K. (Ed.), Language Documentation and Description, Volume 4. The Hans Rausing Endangered Languages Institute, London, pp. 26–37.

Warner N., Luna, Q., Butler, L., and van Volkinburg, H., 2009. Revitalization in a scattered language community: Problems and methods from the perspective of Mutsun language revitalization. International Journal of the Sociology of Language 198, 135–148.

Wilkins, D., 1992. Linguistic research under aboriginal control: A personal account of fieldwork in Central Australia. Australian Journal of Linguistics 12, 171–200.

Yamada, R-M., 2006. Collaborative linguistic fieldwork: Practical application of the empowerment model. Language Documentation and Conservation 1, 257–282. http://hdl.handle.net/10125/1717

Zeitlyn, D., 2000. Archiving anthropology. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 1, Art. 17. http://nbn-resolving.de/urn:nbn:de:0114-fqs0003172

Notes

1 Financial support for the pilot phase of NNAILA was granted from the University at Buffalo College of Arts and Sciences, Libraries, Department of Linguistics, Digital Humanities Initiative, and Interdisciplinary Research Fund.

2 By "appropriate means of dissemination", we intend to cover methods both that make use of technologies which are readily accessible to the communities with a legitimate interest in NNAILA's resources and that allow for any access restrictions on the materials to be respected.

3 See Newman (2007) for an overview of the legal issues relating to copyright from the perspective of U.S. law and Anderson and Koch (2003) for a detailed case study of the legal issues surrounding intellectual property in field recordings and the impact of digital technologies on the application of the law to such materials in an Australian context. Liberman (2000) offers an informal discussion of a range of legal, regulatory, and ethical issues pertaining to language documentation, with an emphasis on how they apply in the United States. Story et al. (2006) offers a critique of how notions of copyright, as applied in relatively wealthy countries, impact poorer parts of the world.

4 Of course, as is clear from work like Nason (1997) and Tatsch (2004), we should not necessarily assume that the intellectual property regimes adopted by the government in a country like the United States should be considered valid for indigenous peoples who may live within its borders.

5 Some guidance on ethical matters can be found in ethics statements adopted by relevant professional societies, for example, the Linguistic Society of America adopted an ethics statement in 2009 (see http://www.lsadc.org/info/pdf_files/Ethics_Statement.pdf).

6 Zeitlyn (2000) makes a related point in discussing anthropologist's attitudes towards archives wherein they gladly use materials made publicly available but are more hesitant when it comes to granting access to the resources they create.

7 For obvious reasons, we cannot cite specific names here. However, the position we are describing here is not one we have constructed simply for the sake of argument,

but which we have encountered multiple times during informal discussions with linguists both within and outside of the context of NNAILA.

8 Though drawn from a different research domain, Michener's (2006, p. 4) schematization of the "natural" degradation of the information content of a resource is relevant here.

9 While our focus in this paper is on the consequences of these new technologies with respect to the relationship between language resources, linguists, and communities, it is clear that many of the concerns we raise are relevant to the domain of intellectual property more generally in the digital age. Lewis, Farrar, and Langendoen (2006), for example, attempt to formulate principles for regulating the appropriate use of online data within the community of academic linguistics itself.

10 For language resources, there is even a network of preservation archives dedicated to language and music materials, the Digital Endangered Languages and Musics Archive Network (DELAMAN; http://delaman.org).

11 See http://lsadc.org/info/pdf_files/Ethics_Statement.pdf.

12 http://www.ailla.utexas.org/site/welcome.html

13 More generally, there may be language communities which cannot be readily associated with any geographically or politically defined area. Warner et al. (2009) discuss one such case.

14 For instance, Abrams (2006, p. 5) discusses how the traditional Iroquois system of matrilineal descent means that his children do not have his Onondaga nationality but, rather, are members of their mother's nation, Tuscarora. Defining "community" in terms of nationality could mean his children have only limited access to Onondaga materials, which is not obviously an ideal outcome, especially given Abrams' own commitment to Onondaga language revitalization.

15 Of course, referring to the "original copyright owner" implies it is clear who that is. See Newman (2007, Q20).

16 We use the phrase "moral ownership" informally here to refer to an ethical, rather than legal, sense of ownership.

17 http://www.rpanet.org/displaycommon.cfm?an=1&subarticlenbr=4