Diversity is "all the ways we are different"[1]
-Hemphill and Haines

A growing awareness of possibilities created by digital technologies and heightened concern about endangered languages has led to an increase in language-focused archive initiatives. The Open Language Archive Community (OLAC) is one such early and continuing effort that has been made possible by the proliferation of digital technologies in the age of the Internet. The "language archive," a term which suggests a fuzzy monolithic space where languages can be safely stored away for future use, is in reality a dynamic network with a high degree of heterogeneity in content and level of development.

Jung-ran Park (2004) claims that OLAC can be divided into three rather different domains: (1) preserved primary data from endangered languages and cultures, (2) open-source software for manipulating human language, and (3) "archives of documentation of over 8000 languages across the world and of linguistic and ESL (English as Second Language) studies…" (p. 7). OLAC's offerings, prepared using a range of recipes, can be sorted according to a variety of tastes and appetites. While this essay does not claim to review the many instances of Open Language Archive-like entities across the globe, it does attempt to broaden the focus beyond Park's intended audience of academic librarians in order to sensitize information professionals in particular and users in general to the diverse nature of digital language archives. As with most early-stage information phenomena, when language archives were few in number, there was no great need to sort them out systematically. Now the steadily increasing number of sites has reached the point where information professionals must become cognizant of the different genres of digital language archives. They must also recognize for whom and by whom the archives are designed in order to help map the territory for a variety of users, who might otherwise be alienated when they wander into the strange conceptual universes or ontologies of language that such archives embody.[2]

The first part of this essay will attempt to parse some of the different kinds of language archive projects currently under way and projected for the future. In the second part, a review of a field study of Ban Khor Sign, an endangered indigenous Thai sign language, will serve as a rich example of the potential complexity of the data of language documentation and the serious challenge of managing this data, particularly when information professionals try to reconcile their commitment to social diversity with the need to promote standards that foster interoperable information systems. Park's (2004) observation that the information professions have stayed in the background of the language archive movement deserves serious consideration, especially since language archive activities "are parallel to ones of information professionals to the extent of

collection, resource organization by utilizing human language technology and standardization, distribution and provision of access, preservation of language and culture related resources" (p. 8).

What's in a Word, or How Do You Say "Computer" in Anishinaabemowin?

Language endangerment has become a high-profile and high-stakes international issue. It has made headlines in *Scientific American* and on the BBC. A special meeting of the UNESCO Programme of Safeguarding Endangered Languages in March 2003 explicitly linked the issue of language endangerment with those of cultural heritage and linguistic diversity by quoting the 2002 UNESCO Istanbul instrument on intangible cultural heritage:

> *Article 1*: The multiple expressions of intangible cultural heritage constitute some of the fundamental sources of the **cultural identity** of the peoples and communities as well as a wealth common to the whole of humanity. Deeply rooted in local history and natural environment and embodied, among others, by a great variety of languages that translate as many world visions, they are an essential factor in the preservation of cultural diversity, in line with the UNESCO Universal Declaration on Cultural Diversity (2001). (Cited in UNESCO, 2003, emphasis in original)

Translating this statement into action is quite another matter. It makes cultural diversity the primary value, conceiving language as a subset of "expressions of intangible cultural heritage" constitutive of cultural identity. The first obvious question to ask is where these languages are to be maintained. The most obvious response is that languages should be safeguarded in the communities that use them; that is, in a twist on the Istanbul instrument's statement that cultural heritage / identity are embodied in languages, languages are to be embodied in people. Ironically, the fact that each language translates a world vision makes "preserving" the language an ambitious and ambiguous goal, especially with respect to managing the various stakeholders' agendas. Matters are made even more complex by the fact that preservation is often paired with the term "revitalization" in endangered language discourse.

"Revitalization" may mean scooping up the remains of a dead language and reanimating them for contemporary use. Proponents of this optimistic procedure frequently cite Hebrew as an example of a language that has undergone this process. In the cases of moribund languages, language revitalization projects attempt to reverse a language's movement toward extinction; they can take many forms depending on the linguistic features of the language, the size of the

population that still "speaks" that language, and the type and quantity of resources available to the effort. However, the decision to revitalize does not suffice to define the specific goal of the operation. Is a particular moment of the language's history to be privileged as a museum-like space where speakers put on the clothing of times gone by and demonstrate for us "how it was?" Or does revitalization focus more on the "how it is" of contemporary society?

Patrick Eisenlohr (2004) reminds us that language revitalization projects are often political, ideologically-driven activities that can posit very different revitalization goals, ranging from perfect imitation of old language forms, to appropriation of old forms into a new vernacular, to an assertion of identity formation, having very little to do with the language itself, that can be created through the networked contacts developing around a revitalization project. Disagreement over the aims of revitalization can form around competing ideologies within the language community. The success of revitalization projects is therefore relative to the expectations of the participants. Walsh (2005) refers to Thierberger's contention that the ideal of intergenerational transmission of language "…is often in conflict with the actual preferences of the people group who identifies with the language in question" (Walsh, p. 304). To put the matter neatly into focus: An Amerindian language site explores the development and the legitimacy of using a word that could never have been in the lexicon of the "authentic" language. How do you say "computer" in Anishinaabemowin?[3]

The tools provided by scholarly research and media, no matter how innocuously they have been conceived, are perforce implicated in these conflicts. Eisenlohr (2004) tells us that "Generally speaking, there is a striking gap between expert discourses seeking to mobilize Western public awareness of widespread language loss across the world today and the concerns motivating users or former users of a linguistic variety to engage in practices of linguistic revitalization" (p. 21). The disparity in the resources and power positions of the parties can in fact work against the very goal of cultural identity in the name of cultural diversity. How much effect expert interventions have on language preservation goals is something that remains to be studied; it is difficult to separate the effects of the interpersonal dynamics of the process from the feasibility of maintaining the language proper, since, as the UNESCO instrument tells us, language is the embodiment of diversity. Some judgments are rather harsh. Newman, as quoted in Walsh (2005), for example, has criticized the "endangered language issue as hopeless" and revitalization efforts as "linguistic social work" whose benefits he clearly discounts (p. 302).

In some cases, revitalization does not enter the digital realm at all and remains a person-to-person activity that often emphasizes intergenerational relationships. In this case, the information professional's role might be assisting in research, gathering useful learning materials, or facilitating use of a community

room in the local library. Often, though, language revitalization may take a hybrid form, with the digitization of some materials and the exploitation of the multimedia capacities of the digital "archive." Sometimes, these efforts are expressions of intent or wishful thinking, as when a small lexicon or the translation of a poem is posted on the internet.

If differing motivations lead to different philosophies of how to combat language loss, the technologies used also exert an influence on how the language in question is experienced. Although Eisenlohr (2004) speaks specifically to mass media and endangered languages, the problematic effects of electronic mediation are also present with internet use and the additional sociocultural practices it engenders, shapes, or prevents. The effects are not only centripetal but centrifugal. Walsh (2005) observes that distributed technologies can help language communities with the "tyranny of distance" that separates indigenous languages, thereby allowing them to reconstitute the entire language community through networks. However, just as present-day groups of speakers will not have the same relationships with each other as their counterparts of a century ago, far-flung digitally-constituted communities will not be the equivalent of geographically unified communities.

Politics aside, information professionals must recognize the presence of many "user-published" web sites that serve as places where people put endangered languages. The actual quantity of linguistic material residing on the sites varies greatly; these virtual archives may contain anything from elaborate narratives and poems and lexicons of hundreds of terms to a simple vocabulary list or a place-holder for one. Language resource sites may point to non-digital materials at a specific geographical location, but at this stage of proliferation, many sites simply point the way to other language sites, which may themselves be pointers. While some users might find this digital hopscotch akin to running on a treadmill, these indexical chains also arguably constitute a kind of social network. In such cases, what the language expressions lack in informational content could be said to be recuperated and extended in the symbolic nature of the gesture as an assertion of cultural identity.

Finding One's Way: Toward a Typology of Language Archives

Grass-roots efforts to preserve cultural diversity by slowing if not reversing the movements of some languages toward extinction have coincided with a wide range of institutionally-based "documentation" projects, such as the Formosan Language Archive, the Oxford Text Archive, the Archive of the Indigenous Languages of Latin America (AILLA), the Alaska Native Language

Center Archives (ANLC), the Open Language Archives Community (OLAC), the Rosetta Project, the Hans Rausing Endangered Languages Project (HRELP), the Documentation of Endangered Languages Project (DoBeS) at the Max Planck Institute for Psycholinguistics, and the Berkeley Language Center, to name just a few.[4]

These archives can be sorted according to several expressed and implicit notions of *documentation.* From the perspective of the information specialist, they tend toward a different kind of pointing or linking than grass-roots archives; often their links are exclusively internal. Thus they often function as a kind of archival *finding aid*, a description of the contents of a particular repository. These descriptions can range from the very general to the extremely detailed. The language data, digitized or otherwise, may be in a specific geographical location, the archive's web site serving to indicate where a given item may be found. For example, the great majority of the Alaska Native Language Center's collection is in non-digital form at the University of Alaska Fairbanks. In other archives, the language data may be digitized and in principle available to authorized users online. The Rosetta Project, which describes itself as a digital library, promises an array of materials at its site, which has categories for various linguistic document types: detailed description, maps, orthography, phonology, grammar, Swadesh Word List, numbers, a Genesis translation, Glossed Vernacular Text, and Audio Files. In addition, the Rosetta Project is also beginning to prepare a word list database for all languages documented on the site. However, while there are place-markers for all of these categories for all languages having an Ethnologue code, the *de facto* standard for referencing languages, numerous languages have no data other than basic metadata associated with them.

Whereas the Rosetta Project grew out of an NSF grant, it has devised a clever mechanism for future funding: you can sponsor your favorite language in the same way that you can sponsor clean-up of a section of freeway. There is no central list of sponsors, so it is difficult to determine to what degree this ploy has been successful. The project has indeed been successful, however, in developing relationships with other major digital language archives, and it provides links to their sites, enhancing the idea of a growing networked language resource community.

OLAC's (Simon and Bird, 2000) founding statement of purpose clearly specifies the goal of collaboration to create a network of interoperating repositories and agreement on current best practice for language resource digital archiving. However, it is clear that for many participating archives, the train hasn't yet left the station, or as Bird and Simmons tell us "[i]n reality the user can't always get there from here" (2003, p. 377) even if OLAC points the way. The idea of standardizing digital language documentation practices via distributed computing to make available far-flung repositories' collections is extremely

appealing. However, in many cases, the repository presence online is unreliable, or it merely offers digital finding aids for its own collections. OLAC is not alone in being put at a disadvantage by this partial realization of a greater objective. The (re)searcher may mistakenly equate the common term *online archive* with a complete digital presence; the digital storefront's advertising may eventually go unheeded by users who visit multiple times only to find that the archive is still not open for business.[5]

Just what counts as appropriate data for a language archive is largely a matter of the disciplinary philosophies that determine where the boundaries of language lie. For one OLAC archive, the ACL Anthology, the contents are simply and unapologetically papers dealing with computational linguistics. At the other end of the scale, projects can aim for very dense and detailed description of language documentation. The Hans Rausing Endangered Languages Project (HRELP) web site tells us that:

> *Language documentation* (or *Linguistic documentation*) is one response to the common situation of language endangerment. It is often said to have been catalysed by Nikolaus P. Himmelmann, who wrote in 1999: 'The aim of a language documentation is to provide a comprehensive record of the linguistic practices characteristic of a given speech community… This … differs fundamentally from … language description [which] aims at the record of a language … as a system of abstract elements, constructions, and rules'… (HRELP).[6]

Looking at the information management required for this "new discipline within linguistics," as HRELP calls it, allows us to understand better the range of possible interpretations of the "documents" to be preserved in the process of language documentation, which cannot be reduced merely to collecting the writings produced by a community of language speakers. Rather, at its fullest, it is a multimedia attempt to capture the language *in situ* as both language and culture. The Ban Khor Sign example will show us many kinds of possible documents: oral histories, still photographs, sound recordings, videos, field notes, and drawings where villagers described their own community. The content of these materials is not the only information source. Since language documentation adheres to a principle of documenting a language ecology and the relationship of the researcher(s) to the documentation gathered, we also require detailed information about the document itself: format, medium, date of production, producer, subjects, and so forth.

By focusing on documents rather than uncertain future generations of native speakers to preserve the language, language documentation forces the issue not just of *where* the documents must be housed but *how* they will be kept available over time. In the world of digital artifacts, preservation means keeping

data in a machine readable format as well as keeping a machine that can read that format, a requirement that seems simple until we think of our 8-track tape collections. Moreover, the data must be housed in proper environments, handled with appropriate techniques, and migrated periodically. Larger institutions with a stable future and a strong commitment to this kind of preservation seem to be the most likely candidates for attention. These institutions are increasingly important given the fact that many grass-roots sites have already become defunct. The Rosetta Project proposes an unusual (and necessarily untested) solution for communities who store their digital data at the Rosetta site:

> The Rosetta digital library is archived at 5 year intervals on an extreme longevity, micro-etched nickel disk. This contemporary "Rosetta Stone" will soon be available to individuals, institutions and others who care to keep one. We hope the process of creating a new global Rosetta, as well as the imaginative power of having "all" languages in a single, aesthetically suggestive object, will help draw attention to the tragedy of language extinction as well as speed the work to preserve what we have left of this critical manifestation of the human intellect.[7]

Even if contentious issues within language communities engaged in revitalization projects might become less obvious in a documentary setting, questions of power and conflict cannot be avoided. The language documentation projects themselves are necessarily expert projects, even when linguists, anthropologists and other documentary specialists embed themselves within a language community for a considerable period. The very act of studying someone else's language can be perceived as imperialistic, and is at the very least a sensitive issue of which information specialists need to be aware. We must also recognize that all documentation projects are not equal in the degree to which they aspire or manage to capture a contextualized view of the language in question. Opinions vary as to whether it is more important to document the most salient features of as many endangered languages as possible rather than to explore the nuances of a single language. Differences in disciplinary philosophies between linguists and anthropologists as well as the distribution of institutional, political and economic clout will affect the representation of a particular language.

As documentation becomes more detailed and the impetus for preservation shifts from the speakers to the experts, the question of access to corpora becomes stickier. If a linguistic anthropologist has permission to videotape an individual speaker, to what degree does the scholar have the right to share the data? Such sharing is always problematic; it is especially so where the population of speakers is concerned, for the protocols that protect the rights of the informants can also restrict the access of "ordinary" members of the community to their own data.[8] The access and security problem was somewhat attenuated in the past by the

necessity to visit an archive in person, sometimes at a great distance, or to request a particular document for loan; today, these are issues that all information professionals must face. Archivists are already familiar with the nuances of access to any sensitive collection, and they have become particularly versed in recent years in the integrity and security of electronic data; now local librarians as well must increasingly cope with electronic indexing services and periodicals in order to serve their patrons.

The Documentation of Endangered Languages Program (DoBeS) at the Max Planck Institute for Psycholinguistics in the Netherlands would seem to come closest to the realization of the linguistic documentation ideal, since it incorporates a holistic understanding into its system design. For example, rather than focusing on either preserving the data or making available a relatively useless if not linguistically misleading snippet of information, DoBeS approaches the problem of "chopping up" the data by "orienting the research less according to the positions of researchers in their contemporary intellectual landscape and more according to the positions of the speakers in *their* social landscape" (Widlok, 2004, p. 4). In other words, focusing on making available the context rather than merely abstracting out formal linguistic categories leads to a better system of access. Thomas Widlok (2004) claims that "ethnography helps to reduce the arbitrariness in data collections by drawing on the cultural context of speakers and it helps to make the language documentation materials meta-theoretical enough to be suitable for a long-term archive" (p. 4). This approach has the advantage of allowing the user to cut across pre-conceived categories as well as give "a voice" to the informants participating in the study. The ethnography allows users to view informants, settings, and activities beyond the bounds of a single session. As Widlok says, "ethnography counts because it is more than just another domain. Ethnography helps to re-connect what has been archived as separate sessions" (p. 5).

The DoBeS project has also made some headway in the next step of corralling enormous and numerous language corpora by dividing the sessions into smaller segments that can be searched, chunked, and reordered for various kinds of cross-linguistic comparisons. While this can be a matter for highly trained linguists, it is also of concern and interest to information professionals, for language archives are nodes of activity in an incipient semantic web which may change how we all look for and think about information.

These DoBeS sessions can be segmented to allow for various kinds of groupings only when they are in digital format, so even materials such as drawings must be re-represented digitally in order to become part of the corpus. The DoBeS planned workflow has digitization of original data as a high priority. The assumption is that copies of the data will be returned to the researcher who will help to segment and identify important elements of the segments. The

DoBeS archiving team then attempts to organize these data according to flexible hierarchies.

DoBeS provides tools with which researchers can manipulate their data. Their desire to include more data is based on the theoretical principle of rich contextuality and on the notion that more data variety will allow for a richer matrix of associative possibilities. The goal is to frontload as much of the metadata as possible in order to make these associations possible to users. Widlok's principal message to us is, "*[d]o not be judgmental about the diversity of space/time/person/group categories but make use of this diversity to facilitate access to the database"* (Widlock's emphasis, p. 6).

Sign(s) in Context: The Ban Khor Sign Project

Information professionals—librarians, archivists, and informaticists—are increasingly recognizing the need to examine their own institutional cultures in the light of diversity. The American Library Association, for example, counts diversity as one of five primary "action areas," and the Society of American Archivists' statement on diversity (1999) holds that the SAA "…is committed to integrating diversity concerns and perspectives into all aspects of its activities and into the fabric of the profession as a whole…" (¶ 1). Coping with differences, however, is not alien to the information professions whose stock and trade is dealing with dynamic heterogeneity and "sorting things out."

Diversity is a particularly sensitive question in the realm of language archives since they not merely represent different communities but inherently embody their speakers' cultures. Organization of information about the existence and locations of language data is one level for information professionals to consider, but another is the internal organization of the language—or the perception of its internal organization—that must to some degree condition how the information about the content itself is represented to users. Information professionals must be aware that the field of language preservation is pockmarked with political and intellectual land mines that remain invisible until they are stepped on.

The Ban Khor Sign project illustrates the complexity of the multiple issues that surround the collection and archiving of endangered language-related data, for in order for Ban Khor Sign to be documented, it must be recognized as a legitimate language with equal rights among other endangered languages. Sadly, even though linguists, anthropologists, archivists, and native speakers themselves have mobilized to document the most endangered of the world's 6000 to 7000 languages, they are addressing but half of the problem, for in fact, if sign

languages were included in the tally, the number of languages in the world would double (Branson and Miller, 2000; Jokinen, 2000). In spite of over 35 years of efforts on the part of linguists to demonstrate that sign languages are languages, Branson and Miller say that "[m]ost hearing people are still unaware of the existence of sign languages and even many of those who have contact with the Deaf, including teachers of the Deaf, still do not accept that sign languages are languages like any others" (p. 30). As Angela Nonaka (2004a) insists, "Nowhere is the impact of continued marginalization of Deaf people as linguistic minorities more evident than in contemporary discussions of language rights as human rights" (p.740).

The general monolithic view of sign language surely causes confusion about sign language as a phenomenon. Not all sign languages are "created" equal. There are natural sign languages generated spontaneously by deaf people that function independently of any spoken or written language. In contrast, manually coded sign languages are "parasitic" on a pre-existent spoken or written language that does not necessarily coincide with ethnicity or culture of the sign language's users. A country's national sign language may have very little to do with its national spoken language. The national Thai Sign Language (TSL), for example, resembles American Sign Language more than it does Thai, as a result of Gallaudet College's role in establishing the first school for the Deaf in Thailand in the 1950s (Nonaka, personal communication). At best, a national sign language may be a natural indigenous language that comes to dominate other sign languages. In contrast, a non-national sign language is doubly minoritized, first by orally spoken language and secondly by the sign language that has achieved dominance within its own nation-state. When a sign language is recognized or studied at all, it is likely to be the national sign language (Nonaka, 2004a).

While many areas of the world are characterized by great linguistic variety, Thailand is a particularly interesting case in that it did not experience the artificial imposition of a colonial language upon its populace. Rather, there was considerable influx of many ethnic-linguistic groups over porous borders that added to an already considerable linguistic diversity, with the result that there are over eighty languages spoken in Thailand (Smalley, 1994).

In 2003, linguistic anthropologist Angela Nonaka undertook an in-depth ethnographic field study to document an indigenous sign language in Ban Khor. As she describes it,

> Ban Khor is a small village in northeastern Thailand. By official accounts, it is an unremarkable area—poor, agricultural, and remote. From a linguistic anthropological perspective, however, Ban Khor is a very special place. Ban Khor has an unusually large number of deaf residents, and in response to the high incidence of deafness in the population, villagers invented an indigenous sign language that is used fluently by many members—hearing and deaf—of the

> community. In such an environment, deafness is not an impediment to communication, and deaf people are well integrated into the mainstream of village life. Increasingly, however, the local sign language and the delicate sociolinguistic ecology associated with it are threatened (2004b, p. 1).

Such a community perfectly illustrates sign language linguistics pioneer William Stokoe's (1980) advocacy of studying deaf speech communities in sociological terms, "not just as minorities within dominant cultures but as models of human organization and reasonably complete cultural systems operating in a four-sense world" (p. 387).

The Ban Khor Sign project was "holistically conceived to include historical, geographical, structural, ideological, and interactional dimensions"; it employed a variety of anthropological methods such as oral histories, sociolinguistic interviews, surveys, mapping, kinship diagramming, demography, ethnographic focal-follows, and participant-observation (Nonaka, 2004b, p. 2). Nonaka's methodologies were not simply linguistically motivated but geared to provide documentation on the speech community as a whole, for as she indicates, there is a paucity of local written records.

Ban Khor sign is an important site for research and documentation for a multiplicity of reasons. As a small, isolated village, Ban Khor is well suited to creating a speech community with reasonably well-defined parameters. The sign language, a spontaneous response to an incidence of deafness that exceeds normal expectations by 1.5 to 3 times and even by 5 to 10 times in certain sectors (Nonaka, 2004a), will have had a short life cycle. Its history can be traced from some 60 years ago to its likely demise within the current generation as contact with the official National Thai Sign Language grows, both through residents (only in their twenties) who have returned from a regional school for the deaf and through the education of the new generation of deaf children at this school.

The situation is typically paradoxical. The deaf villagers will eventually have more access to programs geared in their favor and to better occupational opportunities if they become fluent in TSL. They will also gain in status, for the already status-conscious Thai society applies the same stratified values to its sign languages as it does to the many spoken languages. They will be able to integrate into a much larger Deaf community, yet the community configuration where deaf and hearing mix easily will disappear. This is a clear case for non-intervention on the part of the researcher, but it is also an imperative for careful and thorough documentation.

The Ban Khor Sign project, while invaluable as the only documentation of a moribund indigenous sign language, also provides collateral documentation of a society in which five different oral languages are spoken: Nyo(h), So(e), Phutai, Lao, and Thai-ka-dai. (Nonaka, personal communication). Data collection efforts

were not limited to elicitation lists; Nonaka was also able to capture myriad special and everyday cultural practices in her field notes and tapes. While even limited release of parts of records such as that of a funeral ceremony and a Deaf Thai woman giving birth merit careful consideration, others such as the Loy Krathong Festival, the importance of the water buffalo in rice cultivation and as a sporadic source of meat, and extensively filmed socialization of young children would be valuable resources for further study by others whose mission is not primarily to document the linguistic ecology *per se*.

Nonaka's description of the data corpus under consideration reveals a wide variety of materials:

> Video recordings (approximately 225+ hours) are of two general types: linguistic and anthropological. The former, supplemented by audio recordings, document formal linguistic elicitation sessions designed to learn more about both the local sign language, *pasa bai* (language deaf/mute) and the spoken language, Nyoh. The anthropological footage includes semi-formal interviews, ethnographic recordings of language use in everyday life, and focal-follow case studies of children acquiring language(s) and being socialized into the local speech/sign community. Combining elicited and spontaneous conversational data, the corpus documents multiple language genres as well as language use within and across various linguistic and social contexts, activities, and interactions (2004b, p. 2).

The first step in making this data useful to a non-expert community would naturally be the creation of a Ban Khor Sign dictionary. *Thai Sign Language Dictionary* is an example of existing documentation for the same corner of the globe and is predicated on use primarily by the deaf person, ordering the entries on the basis of their morphology rather than on written Thai or English words. A notational system indicates hand position and shape as well as views from above, movements (including direction and speed), and line drawings of real people and their expressions.

Nevertheless, the print dictionary is first of all just a lexicon and not a grammar; secondly, it can represent only partially the space involved in the entirety of the enunciation; and thirdly, it cannot capture the simultaneity that may characterize Thai signs and larger locutions. The representation is in every sense flat, but creating a digitally available sign language dictionary is an extremely difficult undertaking. Many online attempts represent a simple transfer of the paper rendering to the digital world or, as in a more complex online ASL dictionary, short video sequences demonstrating a given sign are glossed by an English "translation," which perpetuates the unfortunate misunderstanding that full signs are merely iconic representations and do not belong to a system that can accommodate a richness of expression beyond nouns and adjectives. By analogy,

a student in a foreign language class can learn perfectly a list of 20 words and be incapable of speaking, as opposed to mimicking sound form.

In the spirit of holistic language documentation, a Ban Khor Sign Dictionary would make the language accessible to both experts and non-experts. A grammar, albeit much more difficult to create, would allow for better understanding of the language and help to combat the persistent belief that sign is not a language, a fact that Stokoe observes even among professional linguists with no sign language exposure. However, this would be a partial and imperfect representation of the nature of sign language.

As Stokoe (1979) points out, it is difficult to analyze sign language "words" precisely because of a sign's extremely complex structure. In his *Field Guide to Sign Language Research*, Stokoe indicates that "[t]he sign should also be described according to these five categories at least: (a) the sign's spatial position, (b) the specific configuration of the executing extremity (arm, hand, fingers), (c) points where the hand(s) touch the body or one another, (d) movements of the body or its parts, and (e) the facial expression" (p. 16). Other relevant body parts may include the head, shoulder, forearm, elbow, hand, wrist, fingers, face, forehead, eyebrows, eyes, nose, mouth, lips, teeth, trunk, legs, and feet.

According to Stokoe, sign language exhibits a significantly different syntactic dimensionality from spoken language. That is to say, sign language fully exploits three dimensions of space and one of time (1980), and it forces serious rethinking of the flat Saussurian sign composed of *signifiant* and *signifié*. As a result, presentation of a sign language dictionary with no other form of documentation to an audience not familiar with the nature of sign language risks serving only to reinforce the stereotype of sign language as a choppy and unexpressive series of mimicking gestures. Most will not understand that, for example, "English lexicon and modulated ASL signs do not match well…" (1980, p. 372). Translation issues are magnified several times over when one of the languages in question is a sign language. For Stokoe, untranslatable concepts are evidence of language's embeddedness in culture and that "each culture is unique coming to terms with internal and external reality" (Stokoe and Kuschel, 1979, p. 11), but this untranslatability is likely to go unperceived in a sign language dictionary. The implications of making this sort of documentation available are dramatic for many disciplines, for reconceptualization or recontextualization of Saussurean linguistics and its relationship to Charles Peirce's theories would instigate a reevaluation of some 80 years of structuralist and post-structuralist linguistic theory and its products. The fact that sign language has been unheard has obfuscated an entire universe of analysis.

Ideally, one would take a step further and make available small videos (currently in the form of iMovies of 3 to 4 minutes in length) that show the

language functioning in daily life activities as well as on special or unique occasions. The oral histories conducted by Nonaka could potentially contain many diamonds in the rough for people in other disciplines. She indicates in passing, for example, that she recorded interesting footage of elderly villagers describing the day sometime after World War II when a man from the government went through the village asking who "owned" what property. He then prepared and distributed copies of documents corresponding to the property descriptions, and institutional codification of property through recordkeeping was born in a society that had not previously known or cared to know this practice.

The partial availability of Nonaka's ethnographic field notes would add immeasurably to the contextual richness of the language material presented. The notes could be made available for access as a parallel text to the videos or still photos that they describe. Fabian (2002) makes the point that presentation of material on the internet allows for the addition of commentary that is simply impossible in the economically defined world of academic publishing. He posits that we could be witnessing the birth of a new genre. While it has been countered that this may mean nothing more than providing a dustbin for "additional junk," it is perhaps precisely this junk that may be of interest to future users.

A Last Word: Transdisciplinarity

The DoBeS team also makes clear the need for collaborative efforts. As media has grown more complicated over the years and as the team has attempted to integrate media signals and annotations, they have realized that they had to solve the problems of organization and infrastructure at the team level: "[i]t was understood that the individual researcher was not capable of carrying out this task in a way that could guarantee re-usage even within the institute. The corpus infrastructure generated by about 30 researchers was collapsing into chaos" (Wittenburg, Brugman, and Broeder, 2000, p. 1).

The notion of transdisciplinarity might be a useful thought paradigm for overcoming the ethical dilemmas surrounding accessible yet diverse information management, for such a perspective tends to shift the focus from people as objects of documentation for the archive to people as partners in a given documentation process. An overview of the entire language archive problem also stresses the commonality of the concerns of various information professionals and minimizes differences that have become naturalized and entrenched over time. While it is true that the linguistic documentation effort is a particularly nice example of transdisciplinary cooperation among archivists, sociolinguists, linguists, informants, interpreters, educators, and even funding foundations and institutional archives, coordination of these problems of mutual concern to a multiplicity of

groups might well require the specialized skills of information professionals. As we have seen in the DoBeS example, the interaction among members of the various documentation teams, as they are called, can make use of everyone's expertise to create a resource and record that is far more accessible and useful than anything produced heretofore by either party going it alone. Information specialists are particularly well-suited to match networks of interested users with networked information.

Nonaka (2004a) predicts that the rare language of Ban Khor Sign and its unique, fragile ecology will disappear in a fairly short time. The pathos of the situation lies in the fact that language documentation cannot be exhaustive; all languages resemble Ban Khor Sign in that their feel and texture to their users today cannot possibly be preserved whole for posterity. Ironically, the fragile digital media containing valuable information are more vulnerable to the ravages of time than Franz Boas' paper transcriptions of Kwakiutl tales. The urgency of the need to keep its data alive makes Ban Khor exemplary of a larger world whose linguistic representations we attempt to preserve by digital means. Risky though the digital gamble may be, the possibilities for documenting Ban Khor Sign are of an unprecedented informational richness. Because information professionals will play an increasingly important role in organizing, preserving, and making available such digital representations, we must get to work sooner rather than later.

Acknowledgements

The author wishes to thank Angela M. Nonaka, a Ph.D. candidate in Anthropology at UCLA, for her boundless generosity in sharing her expertise and important original research. Her Ban Khor Sign project provides a unique insight into the problems of language archiving. Any errors contained herein are wholly my own.

Notes

1. Cited in Owens (2000): Hemphill, H. & Haines, R. (1997). *Discrimination, harassment, and the failure of diversity training: What to do now*. Westport, CT: Quorum Books.
2. See Phil Agre's "Designing Genres for New Media" (1995).
3. Dibaajimowin: From Birchbark Designs to Computers -- Looking at Anishinaabemowin Word-roots. http://www.kstrom.net/isk/stories/words.html

[4] Obviously there are many more, including archives of Australian, African, Oceanic, Native American, and a host of other languages. I visited the above archives on several occasions for one contingent reason or another. If I didn't stay long at an archive, it was because of server problems on their end or a general lack of accessibility. My list in no way constitutes a comment on hierarchical importance of these languages.

[5] For example, see the TRACTOR Test Archive base URL: http://www.language-archives.org/cgi-bin/gateway/gateway.cgi/ota.ahds.ac.uk/olac/tractor.xml, available from the OLAC TRACTOR Archive details page - http://www.language-archives.org/archive.php4?id=29

[6] See HRELP's reference to Himmelmann: "p. 166, 'Documentary and descriptive linguistics,' Himmelmann, N.P. (1998), *Linguistics* 36. 161-165. Berlin: de Gruyter."

[7] See http://www.rosettaproject.org/about-us/about-us

[8] For example, see Amith (2000).

References

ACL Anthology: A Digital Archive of Research Papers in Computational Linguistics (2005, January 8). Retrieved March 29, 2005 from http://acl.ldc.upenn.edu/.

Agre, P. (1995). Designing genres for new media. *The Network Observer*, *2*(7) and 2(11). Retrieved March 5, 2005, from http://polaris.gseis.ucla.edu/pagre/tno/november-1995.html#designing.

Alaska Native Language Center. Retrieved April 18, 2005, from http://www.uaf.edu/anlc/.

American Library Association. *Diversity. ALAAction*, *4*. Retrieved May 18, 2006, from http://www.ala.org/ala/ourassociation/governingdocs/keyactionareas/morediversity/diversitybrochure.htm.

Amith, J. (2000). Legal, ethical, and policy issues concerning the recording and publication of primary language materials. Paper presented at the workshop on Web-Based Language Documentation and Description, 12-15 December 2000, Philadelphia, USA. Retrieved March 29, 2005, from http://www.ldc.upenn.edu/exploration/expl2000/papers/amith/amith.htm.

Bird, S. & Simons, G. (2003, November). Extending Dublin Core metadata to support the description and discovery of language resources. *Computers and the Humanities,* 37(4), 375-388.

Branson, J. & Miller, D. (2000). Maintaining, developing and sharing the knowledge and potential embedded in all our languages and cultures: on

linguists as agents of epistemic violence. In R. Phillipson (Ed.), *Rights to Language: Equity, Power, and Education: Celebrating the 60th Birthday of Tove Skutnabb-Kangas* (pp. 28-32). Mahwah, NJ: Lawrence Erlbaum.

Hans Rausing Endangered Languages Project. What is language documentation? Retrieved March 12, 2005, from http://www.hrelp.org/documentation/whatisit/.

Eisenlohr, P. (2004). Language revitalization and new technologies: cultures of electronic mediation and the refiguring of communities. *Annual Review of Anthropology*, 33, 21-45. Retrieved April 15, 2006, from http://arjournals.annualreviews.org/doi/pdf/10.1146/annurev.anthro.33.070203.143900.

Ethnologue, languages of the world. Retrieved March 29, 2005 from http://www.ethnologue.com/.

Fabian, J. (2002). Virtual archives and ethnographic writing. *Current Anthropology 43*(5), 775-786.

Gibbs, W. W. (2002, August). Saving dying languages. *Scientific American*, *287*(2), 78-85.

Giese, P. (1997, January 11). *Dibaajimowin: From birchbark designs to computers — Looking at Anishinaabemowin word-roots*. Retrieved April 16, 2006, from http://www.kstrom.net/isk/stories/words.html.

Jokinen, M. (2000). The linguistic human rights of sign language users. In R. Phillipson (Ed.), *Rights to language: Equity, power, and education: Celebrating the 60th birthday of Tove Skutnabb-Kangas* (pp. 203-213). Mahwah, NJ: Lawrence Erlbaum.

Max Planck Institute for Psycholinguistics. *Documentation of Endangered Languages Project*. Retrieved May 20, 2006, from http://www.mpi.nl/DOBES/.

Nonaka, A.M. (2004a). The forgotten endangered languages: Lessons on the importance of remembering from Thailand's Ban Khor Sign Language. *Language in Society*. *33*(5), 737-767.

Nonaka, A.M. (2004b). *Pasa Bai*: language socialization of an indigenous sign language in a northeastern Thai village. (Study #6983) Unpublished final report to the Thai-U.S. Educational Foundation (TUSEF) and IIE Fulbright, submitted January 2004.

Online Language Archives Community (OLAC). *TRACTOR test archive details*. Retrieved June 9, 2006, from http://www.language-archives.org/archive.php4?id=29.

Owens, I. (2000, April/May). Maintaining diversity in information agencies: Accountability, professionalism, job performance, policies, and standards. *Bulletin of the American Society for Information Science*, 26(4). Retrieved May 20, 2006, from http://www.asis.org/Bulletin/May-00/owens.html.

Park, J. F. (2004). Language-related open archives: impact on scholarly communities and academic librarianship. *Electronic Journal of Academic and Special Librarianship*, 5(2), 2-3. http://southernlibrarianship.icaap.org/content/v05n02/partk_j01.htm.

Phrasebase. *Thai language facts and information; Thai statistics*. Retrieved March17,2005,from http://www.phrasebase.com/languages/index.php?cat=222.

Rosetta Project. Retrieved April 10, 2005. http://www.rosettaproject.org/.

Simons, G. & G. Bird. (2000, December 7). *The seven pillars of open language archiving: A vision statement*. [draft] Retrieved March 23, 2005, from http://www.language-archives.org/docs/vision.html.

Smalley, W. A. (1994). *Linguistic Diversity and National Unity: Language Ecology in Thailand*. Chicago and London: University of Chicago Press.

Society of American Archivists. (1999, June 13). *Diversity statement*. Retrieved April1 8, 2005. http://www.archivists.org/statements/diversitystatement.asp.

Stokoe, W. C. (1980). Sign language structure. *Annual Review of Anthropology 9*, 365-390.

Stokoe, W. C. & R. Kuschel. (1979). *A Field Guide for Sign Language Research*. Silver Spring, Maryland: Linstok Press, Inc.

Thieberger, N. (2005). *Archiving and the work flow of field work: PARADISEC*. Annual meeting of the Linguistic Society of America, 2005, Oakland, California. Retrieved March 22, 3005 from http://www.language-archives.org/events/olac05/olac-lsa05-thiesberger.ppt.

TRACTOR Test Archive. Retrieved June 8, 2006, from http://www.language-archives.org/cgi-bin/gateway/gateway.cgi/ota.ahds.ac.uk/olac/tractor.xml.

UNESCO. (2003, October). *Charter on the preservation of the digital heritage. Adopted at the 32nd session of the General Conference of UNESCO, 17 October 2003*. Retrieved April 15, 2005, from http://portal.unesco.org/ci/en/ev.php-URL_ID=13366&URL_DO=DO_TOPIC&URL_SECTION=201.html.

UNESCO. (2002, September). *Istanbul Declaration. Final Communiqué of the Third Round Table of Ministers of Culture on "Intangible Cultural Heritage, mirror of cultural diversity," Istanbul, Turkey 16-17 September 2002*. Retrieved April 15, 2005, from http://portal.unesco.org/en/ev.php-URL_ID=6209&URL_DO=DO_TOPIC&URL_SECTION=201.html.

UNESCO. (2003, March). *International expert meeting on UNESCO Programme Safeguarding of Endangered Languages: Recommendation for Action Plans*. Retrieved March 30, 2005, from http://portal.unesco.org/culture/en/file_download.php/4ac2ba0017bb5f947232457dbae62b50recommendation_for_action_plans.pdf.

UNESCO. *Register of Good Practices in Language Preservation*. Retrieved March 30, 2005, from http://portal.unesco.org/culture/en/ev.php-URL_ID=23506&URL_DO=DO_TOPIC&URL_SECTION=201.html.

UNESCO. (2003, March). *Safeguarding of the Endangered Languages. International expert meeting on the UNESCO Programme. Paris, 10-12 March 2003*. Retrieved March 30, 2005, from http://www.unesco.org/culture/heritage/intangible/meetings/paris_march2003.shtml.

Walker, D. (2005, January 19). In defence of "lost" languages. *BBC News.* Retrieved March 29, 2005, from http://news.bbc.co.uk/go/pr/fr/-/1/hi/magazine/4172085.stm.

Walsh, M. (2005). Will indigenous languages survive? *Annual Review of Anthropology*, 34, pp.293-315. Retrieved January 23, 3006, from http://arjournals.annualreviews.org/doi/pdf/10.1146/annurev.anthro.34.081804.120629.

Widlok, T. (2004). Ethnography in language documentation. *Language Archive Newsletter* [of the Max Planck Institute for Psycholinguistics], *1*(3), 4-6. Retrieved April 23, 2005, from http://www.mpi.nl/LAN/vol_01_n03.pdf.

Wittenburg, P., H. Brugman, & D. Broeder. (2000). Annotations, formats, and data types in the DOBES project. Paper presented at the workshop on Web-Based Language Documentation and Description, 12-15 December 2000, Philadelphia, USA. pp. 1-6. Retrieved March 29, 2005, from http://www.ldc.upenn.edu/exploration/expl2000/papers/wittenburg/phil-workshop-dobes-paper.pdf.

Author

Stacey Meeker is a graduate student in the Department of Information Studies at the University of California, Los Angeles.