elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 2*. Editor: Peter K. Austin

Language documentation and archiving, or how to build a better corpus

HEIDI JOHNSON

Cite this article: Heidi Johnson (2004). Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.) *Language Documentation and Description, vol 2*. London: SOAS. pp. 140-153

Link to this article: http://www.elpublishing.org/PID/026

This electronic version first published: July 2014

Language documentation and archiving, or how to build a better corpus

Heidi Johnson

1. Introduction

Archives have historically played a central role in the description of endangered languages. This is not surprising, since there is little sense in collecting data on languages that are disappearing if there is no plan for preserving that data. Archiving materials for the already nearly extinct languages of North America was an essential goal of the pioneers of Americanist linguistics: Franz Boas, Edward Sapir, and their intellectual descendants. They diligently deposited all their fieldnotes (and later, audio recordings) in archives and museums such as the Smithsonian Institution.

These archived materials have since formed the basis for decades of linguistic research. Archives facilitate collaboration across generations of researchers and have enabled the production of some of the greatest contributions to the field. One excellent example is the Onondaga-English/English-Onondaga Dictionary (Woodbury 2003) which was based on both the author's own fieldwork and on archived texts and earlier dictionaries.

Archives also support the maintenance and revitalization of endangered languages, by making materials from earlier periods, when the language was more widely spoken and a greater range of forms and genres were still alive, available to the speakers and their descendants. An example of this is the J.P. Harrington Database Project at the University of California, Davis, which is digitizing and publishing on the web the descriptive data he collected in the early 1900's (Macri, Golla, and Woodward 2004). These newly-available recordings and texts are being used in "the monthly 'language lessons' that are being held by members of the Juaneño Band of Mission Indians at San Juan Capistrano. Rather than the usual vocabulary drills and tutoring in a practical orthography, the Juaneños gather to listen to tape recordings of the last fluent speaker of their language, Anastacia Majel, dubbed from aluminium discs that Harrington and his nephew, Arthur, made in the mid-1930s" (Golla 1996).

Similar stories can be told around the world. In the nineteenth and early twentieth centuries, language materials consisted entirely of written text data. Transcriptions taken as direct dictation, notes of elicitation sessions, translations, field notes, and analyses, were all produced on paper. Linguists and anthropologists were careful to preserve these painstakingly produced materials by depositing them in archives, which were well-equipped to preserve such collections. These texts were accessible to researchers who were able to travel to the archive and work with the original, often

Heidi Johnson (2004) Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 140-153. London: SOAS.

hand-written, materials. As technological developments progressed, recordings of speech were made, and these were also duly deposited in archives (see De Graaf and Shiraishi, this volume, for a brief history of similar developments in Russia). Unfortunately, recordings on wax cylinders, vinyl disks, and open-reel magnetic tapes are not so easily accessed after a few decades. They are also difficult to copy, placing them at risk of destruction by natural forces such as mould and oxidation.

As recording technologies improved, making recordings in the field became easier and easier, but traditional archives are not well equipped to manage collections of recordings. They have neither the means to preserve them for the long term nor to make them accessible to researchers and speakers. There are exceptions, such as the Indiana University Archives of Traditional Music[1], whose mission is precisely the long-term preservation of recorded materials, on their original media. They make copies on cassette tapes on request.

There are few such repositories for analogue recording media in the world, however, and somehow during the middle decades of the twentieth century linguists stopped trying to deposit their language documentation in archives and museums (except in Australia where the Australian Institute of Aboriginal and Torres Strait Islander Studies has had an active tape archiving policy since 1964). There must be thousands, if not tens of thousands, of recordings of speech in endangered languages on open-reel and cassette tapes squirreled away in the attics and offices of linguists and anthropologists around the world[2]. These primary data have not been publishable, and so perhaps have been less valued by the field as a whole. The only "documentation" that has been available to the world at large, which includes speakers of endangered languages, has been the highly refined distillations of languages that are published as grammars, dictionaries, and scholarly articles.

At the end of the twentieth century, this gloomy picture was transformed by the development of digital media for audio and video recordings, and by the Internet, which facilitates the global dissemination of digital text and media. Now, digital archives make it possible to preserve language documentation permanently and disseminate it widely. The emergence of documentary linguistics (see Himmelmann 1998, Woodbury 2003), accompanied by publicity about endangered languages (such as Webster 2003), and the efforts of international projects such as the Hans Rausing Endangered Languages Project (HRELP) and the DoBeS project of the Volkswagen Foundation. Documentary linguistics is characterised by integration with information and communications technology, which enables researchers to capture, store, and utilize enormous amounts of information (Woodbury 2003, Bird and Simons 2003, Nathan, this volume, Thieberger, this volume).

[1] http://www.indiana.edu/~libarchm/

[2] Dietrich Schüler, Austrian Sound Archive, estimates that 80% of recordings are in private hands (pers. comm.) — Editor.

Fortunately, this explosion of interest and capabilities has been accompanied by developments in the creation of digital archives. There are already several digital archives for endangered language materials ready to receive the documentation being produced today, and to digitize and archive legacy materials from previous decades. The Digital Endangered Languages and Musics Archive Network (DELAMAN[3]) has been formed to co-ordinate efforts and thus improve service to the field. The workshop at which this paper was originally presented was one result of DELAMAN's collaboration.

A list of current DELAMAN members is maintained on the DELAMAN website. Researchers are encouraged to contact any of these archives for information and assistance preferably at an early stage of their language documentation project.

2. Archiving whys and wherefores

This section attempts to answer the basic questions about archiving: who should archive, and where, why, when, and how one should archive.

2.1 Who should archive?

Any researcher who accepts funding from public sources, such as universities and private foundations like HRELP that have public application procedures, has an obligation to produce a public good. Archived materials are public goods, even if access and use is restricted to protect the rights or wishes of the speakers whose words are recorded therein. The resources are still preserved for future generations. Any researcher who works on an endangered language, and thus with an endangered language community, has an obligation to produce materials that can be used by that community well into the foreseeable future and beyond. In other words, all documentary linguists should archive at least a substantial portion of the documentation materials that they produce.

2.2 Where should you archive?

An archive is a trusted repository created and maintained by an institution with a demonstrated commitment to permanence and the long-term preservation of archived resources. A ***collection***, or ***corpus***, is the body of documentary materials created by linguists and native speakers in the course of their research. Note that digitization alone does not constitute archiving. Digital media are actually more vulnerable to loss and obsolescence than are analogue media. The open-reel tapes stacked in museum basements are far easier to retrieve and convert to digital form than a digital recording

[3] http://www.delaman.org

stored on a DVD-RAM double-sided storage disk, the drives for which were made for only one year.

Documentary linguists should seek help from the archive that serves their funding agency, (e.g., HRELP or DoBeS[4]), or the region or language area in which they work (e.g., ANLC[5], AILLA[6], or PARADISEC[7]). Consult the list of DELAMAN archives and feel free to write to any member for advice if you do not find an appropriate archive there.

2.3 Why should you archive?

Documentary linguists should archive their language documentation in order to:

- preserve recordings of threatened languages for future generations;
- facilitate re-use of primary materials (e.g. recordings and fieldnotes) for:
    - language maintenance and revitalization programs;
    - typological, historical, comparative studies;
    - any kind of linguistic, anthropological, or other study that you won't do;
- foster development of both oral and written literatures for endangered languages;
- make known what documentation there is for which languages.

You should also archive your language documentation to further your own career. Archiving can be considered a form of publishing: even if the materials themselves are archived with highly restricted access conditions, the metadata (see section 3.4) is published in the archive's catalogue. You should list all materials that you have archived on your curriculum vitae, so that future employers will know how much work you have done.

Archived materials should also be cited in scholarly and other publications, just as we cite any other published work. This enables those who read a work to locate the primary materials on which that work is based. It also ensures that the speakers whose knowledge and artistry are preserved in the documentation materials are given proper credit for their contributions.

DELAMAN and other organizations such as the Open Language Archives Community (OLAC[8]) are working to devise a standard format for citing archived materials. It will probably look something like the following:

[4] Dokumentation Bedrohter Sprachen, http://www.mpi.nl/DOBES

[5] Alaska Native Language Center, http://www.uaf.edu/anlc/

[6] Archive of the Indigenous Languages of Latin America, http://www.ailla.utexas.org

[7] Pacific and Regional Archive for Digital Sources in Endangered Cultures, http://www.paradisec.org.au

[8] http://www.language-archives.org

Sánchez Morales, Germán. (1994). "Satornino y los soldados." Heidi Johnson, (Researcher.) [online.] Archive of the Indigenous Languages of Latin America. http://www.ailla.utexas.org. ZOH001R010. Access=public.

Note that both the narrator (creator) of the text and the researcher who collected it are mentioned. This models the relationship between the author of a chapter of a book and the editors who put the book together and see it through to publication.

2.4 What should you archive?

Language documentation ideally consists of examples of the full spectrum of language forms and uses that the language community employs (see also Himmelmann, 1998). In general, documentary linguists should try to record, in audio and/or video, as much information as their means and their consultants allow. We should also make an effort to ensure that at least some part of what we record is amenable to open publication, so that some sort of introduction to the language and the culture of its speakers can be made visible to the rest of the world. For many of these peoples, obscurity is a grave historical wrong, pushing them to the margins of world events and facilitating their ultimate destruction.

Of course, we must always discuss the pros and cons of publication for each event or discourse that we record with the speakers, carefully documenting their wishes with respect to future uses of the recordings. In short: it is important to get permission before you start recording (recording includes taking photographs). There is a little more about permissions and intellectual property rights in section 3.2.

With that caveat, what kinds of things are good candidates for archival preservation? Here are some likely genres:

- public events: ceremonies, oratory, dances, chants;
- narratives: historical, traditional, myths, personal, children's stories;
- instructions: how to build a house, how to weave a mat, how to catch a fish;
- literature: oral or written, poetry, any creative work that people may offer;
- conversations: anything that's not gossip or too personal, e.g. conversations about a recent school event or holiday;
- transcriptions, translations, and annotations of recordings, in which anonymity is preserved if necessary;
- field notes, elicitation lists, orthographies - anything other people might find useful;
- datasets, databases, spreadsheets and other secondary (unpublishable) materials;
- sketches of all kinds: grammar, ethnography;

- photographs of speakers and public events.

There are also recordings that should not be archived, such as anything that would cause injury, arrest, or embarrassment to the speakers. One example is the collection of interviews conducted by Pamela Munro and her students with Zapotecs living in Los Angeles about their illegal border crossings (Pamela Munro, pers. comm.). Sacred texts that must not be heard or seen by outsiders are another example.

2.5 When should you archive?

In addition to making regular backups, ideally, you should archive everything you produce as soon as you return from the field, to make sure that nothing is lost. In practice, it usually takes a little time to prepare a field corpus for archiving (see section 3). Note that concerns about losing primary access to the research potential of your documentation should not prevent you from archiving as soon as possible. Students especially are encouraged to archive their corpora with password protection or some other restriction that allows them sole access, to give them time to finish their theses. It is expected that restrictions on access created to protect researchers' concerns will expire after an appropriate length of time (five years, in most cases).

2.6 How should you archive?

You should prepare your corpus according to the guidelines established by the archive where you will deposit it. Review the guidelines published on its website or write to the management for information. If there is no archive for your language or region, the general rules for corpus management given in section 3 will help you ensure that your language documentation is ready for archiving as soon as a suitable institution is established.

3. Building a better corpus

Documentary linguists typically produce a plethora of materials in a wide range of formats, including audio and video recordings, digital and manuscript texts, spreadsheets, and databases. All of these things can be archived.

Steven Bird and Gary Simons identify seven factors affecting the portability of language documentation materials, where portability refers to the continuing usefulness of such materials across time, disciplines, and functions (Bird and Simons 2003). The seven 'pillars' are: content, format, discovery, access, citation, preservation, and rights. Content was discussed briefly in section 2.4. Access and preservation are considered primarily the responsibility of the archive in this guide. Discovery refers to the ability of other interested persons, in our case generally speakers and academic researchers, to locate and access the resource. Metadata, or catalogue information, is what makes

discovery possible[9]. It is also what makes proper citation of the resource possible, so that pillar will not be given its own section here. And since I am an archivist who has had to deal with all manner of legacy materials, I add an eighth fundamental pillar to the foundations for a preservable corpus: labelling.

The following crucial elements in building a better corpus will be discussed here:

- format;
- permissions (rights);
- labelling;
- metadata (discovery and citation).

3.1 Formats

Some data formats are more amenable to long-term preservation than others. These formats may not generally be the most convenient to work with or to use in presentations, publications, or computer-based displays. We distinguish three classes of contexts in language documentation: archival, presentation, and working (the last is where researchers manipulate, edit, annotate their data, etc.) The archival media materials should be uncompressed (this especially applies to digitization of an analogue original) and text data should be in eXtensible Markup Language (XML) structured files. Presentation and working formats can be derived from the archival format.

This is rather abstract, but given the proliferation of digital formats in recent years, it is worth expanding a bit to make clear. The following table of examples of each class of contexts should make these definitions concrete:

	a grammar	a recording	a film
archival context	XML	wav (at least 44.1Khz/16 bits)	MPEG2
presentation context	pdf, html	mp3	Quicktime
working context	MS Word	ATRAC (on minidisk)	proprietary digital camera formats

The general requirements for archival-quality (master copy) formats are that they be:

- non-proprietary; that is, their encoding is in the public domain;
- amenable to forward migration to new formats over time;
- portable, re-useable, repurposeable;
- the best possible reproduction of the original (if not the originals themselves.)

[9] For a different conception of the roles of 'metadata' see Nathan and Austin, this volume.

Legacy materials should be digitized. New materials should be recorded in archival formats. For example: new audio recordings should be created in PCM wav format at a sample rate of at least 44.1Khz with a bit depth of at least 16. DAT recorders, CD recorders, flash ram recorders, and high-density minidisk (Hi-MD) recorders all meet this requirement. Archive the original, and use your copy to produce mp3 files and cassettes for your consultants, make sound snippets for interactive multimedia dictionaries (see Nathan, this volume), and whatever other creative purposes you can devise.

3.2 Permissions

Define a policy concerning Intellectual Property Rights (IPR) and develop a consistent practice for obtaining consent, e.g. forms and/or recorded statements. Learn how to talk to your consultants about IPR. The best source for information on developing your policy will be other researchers who have worked in your region or language community, who are familiar with the customs and mores of the area, and your native-speaker consultants. Note the IPR status of each resource in its metadata (section 3.4).

As mentioned above, you should always get permission before recording anything. Getting permission means discussing the potential uses and abuses to which the recordings and other documentary materials that you and your consultants produce may be subject over time. Generally, we seek permission to publish language documentation for use only for academic, educational, and other non-commercial purposes.

If your consultants are familiar with forms, you could ask them to sign a licence agreement, such as the one on AILLA's website. You could expand this form to include every potential use that you can imagine, such as the following:

- archiving with the following access conditions:
  - open public access for non-commercial purposes;
  - access restricted by password;
  - access restricted for a certain length of time;
  - permission must be granted by a specific agency or individual;
  - special conditions (to be specified) apply.
- other publications, such as books or CD-ROM;
- excerpts published and/or used in classrooms.

If your consultants are unwilling or unable to sign a form, you could record on audio or video a statement of their agreement to specified uses of their works. This recording would then become a part of the archive's documentation for the work.

Although the legal and ethical issues are complicated, particularly when viewed from a global perspective, it is not really that hard to talk to the consultants we work with about potential uses of their work. It is incumbent upon us all to learn how to talk to speakers about intellectual property rights and publication, to take the time in the field for full discussion of all related issues, and to document in permanent form the resulting agreement between speakers and researchers, so that the archive can handle the documentation materials appropriately in the future[10]. That said, worries about property rights should never be used as an excuse for not archiving documentation for present and future generations.

3.3 Labelling

Nothing could possibly be more important than labelling every single item you produce — each track, tape, disc, notebook, digital file, photograph — with **RUTHLESS CONSISTENCY**.

Give this some serious thought. Your system must be infinitely extensible, ensure that related parts can be put back together, and facilitate sorting and general corpus management. You should be using this system from the very start, so figure it out before you begin your project. Think of it as your 'hit-by-a-bus' insurance: if something happens to you, another person will be able to make sense of your corpus, so that the speakers and others who are depending on you to do a good job are not disappointed (though of course, they will be grieved).

The first step is to decide what constitutes an archival object in your corpus. This is not necessarily the same thing as a digital file, and not necessarily the same thing as a unit of media, such as a CD. Consider the difference between a digital video casette, an MPEG2 file, and a documentary film: the file encodes the film which resides on the casette (along with, perhaps, other films). The file will go in the archive and be converted to whatever new format comes along in a decade (or less); the film will be described in the metadata, cited in articles, and 'repurposed' into alternative formats, such as CD-ROMs and BBC special broadcasts. The casette will probably end up in a landfill somewhere. But you still have to label it, so that the archivist, the BBC producer, and you can locate the file to view the film.

The useful 'object' over the long term is the content — the film, in our example above. This should generally be the basic object in your labelling scheme, if possible. In handling legacy materials, archivists often resort to considering the carrier (tape, casette, disk) as the basic object, simply because we can't understand the intellectual content it contains well enough to distinguish one story from another. But for new materials, this should not be a problem. Each individual story (song, interview, etc.)

[10] Note that IPR restrictions can be subsequently changed to be more or less restrictive and are not set in stone. It is important that consultants understand this flexibility.

that you record onto your high-density minidisc constitutes a separate archival object and should be labelled accordingly.

One other factor that must be considered is ensuring that related things are kept together by your labelling scheme. Language documentation materials often come in sets, or bundles, of related items. The prototypical example is an audio recording of a narrative with an annotation text that includes the transcription and translation of the recording. These two things may exist on different media in different physical locations — like a DAT tape in a storage box and an Shoebox file on your hard drive — for the duration of your project (although increasingly documenters use time-aligned annotations that link digital media and text — see Thieberger, this volume, for an example). Your labelling scheme sholud ensure that they can be properly paired by someone else and that they will be archived together. Long recordings may span several carriers, resulting in parts 1 and 2 (or more): these must be labelled so that people can listen to the whole recording in the proper order. Some people make both audio and video recordings of the same discourse event: be sure that your labels allow this relationship to be recovered. You may want to consider some member of the set as primary and the others as secondary. For example, an audio recording is primary, while transcriptions, translations, and other annotations are obviously derivative products, and thus secondary. A dataset that you construct during analysis may be regarded as a single object in itself and receive its own label.

I strongly recommend using a numeric labelling system (that may appear to be an opaque, and user unfriendly) and keeping track of all the details in an auxiliary database, spreadsheet, index cards, or some other sortable form. Numeric labels, which should be unique, are infinitely extensible and compact; this means you will be able to fit them on tiny media labels and use them as keys in your database. Do not use titles of stories: you have no idea how many versions of “El Tigre” you will ultimately end up recording. Always write labels on everything in good indelible black ink using clear, legible print. In the following subsections, I give you three examples of extensible labelling schemes.

3.3.1 AILLA labels

At AILLA, we label every resource, which in our archive refers to a bundle of related files, and every file inside that bundle. The resource label is used to sort the collection and appears in citations of archive resources. If you used something like this, you would make the resource label the key in your supporting database, and write it on the CD, minidisc, notebook, diskette, or any other thing that includes a part of this resource’s bundle of related files.

Our labels work like this:

ZOH001R010	the 10th resource in the first deposit for language ZOH (Zoque of Oaxaca)

ZOH001R010I001.wav	the audio recording in wav format
ZOH001R010I001.txt	the Shoebox interlinearization in text format

We use the language code[11] as the first element so that all the materials in the archive for a given language will sort together. This is extremely helpful if you are working with more than one language. The deposit number helps us manage the archiving workflow. The zeros make sure that all the files in the archive will sort properly; they aren't necessary if you have fewer than 100 or so objects to manage.

3.3.2 Participant initials plus a media type code

A participant is a person who plays an important role in the creation of a resource. The central participants are the speaker who narrates a story, sings a song, or contributes to elicitation sessions, and the researcher who elicits all this verbal behaviour. You could use a labelling scheme based on the initials of your consultants; this would let you sort entries in your database so that all the materials created by or with a given consultant would fall together. If more than one consultant has the same initials, you'll have to add some letters to distinguish them. Examples from my work with Germán Sánchez Morales are:

gsm1_au1	audio recording part 1
gsm1_au2	audio recording part 2
gsm1_sb	Shoebox interlinearization of the audio
gsm1_tx1	text, notes
gsm1_ph1	photo of Germán

The next resource that you create with this consultant will be labelled gsm2_xx etc. Resources that you create by yourself, such as morphological paradigms, will be labelled with your own initials, numbered in the same fashion, and included in your corpus management database.

3.3.3 Label by media unit

This is a very straightforward way to manage recordings made on removable media such as CDs, minidisks or DAT tapes:

md1t1	minidisc 1, track 1

[11] The language codes used by all members of the Open Language Archives Community come from the Ethnologue language codes developed by the Summer Institute of Linguistics. This set encodes several thousand languages and is thus the most complete set of such codes available. For more information or to search for a code, visit the Ethnologue web site at http://www.ethnologue.com/.

md1t1_sb1 Shoebox database for that minidisk track

This method is not likely to be much help for materials produced on one large hard disk or flash memory card. However, as long as your labels are consistent and your materials described fully in your corpus management database, it really doesn't matter which scheme you employ.

3.4 Metadata

One of the reasons that labelling is so important is that it makes it possible to associate all sorts of useful information with each object in your corpus by means of a metadata record. This information is essential for portability, in the fullest sense of the word (Bird and Simons 2003). Metadata catalogue information is especially vital for digital materials, because they are not amenable to direct inspection, as is a book or other printed matter. Metadata facilitates discovery of archived resources, since it provides an assortment of terms for which researchers and speakers can search using interfaces such as the OLAC Search Engine (Hughes, Kamat, and Bird 2004). The metadata record for a resource also provides a place to maintain information about the intellectual property rights inherent in that resource, such as the full names of its creators (and copyright holders) and any special terms and conditions of use.

At an absolute minimum, the metadata for any resource must include:

- creators' full names: this is required for proper citation[12];
- name of the language: be specific! Zoque of San Miguel Chimalapa, Oaxaca, Mexico, not just Zoque;
- date of creation: use the primary (recording) date for all related items if you want, but be sure to note the date of each recording;
- place of creation: again, be specific;
- access restrictions: note any special conditions or restrictions on the use of the resource. Include a password, if necessary;
- genre keyword: this will be dependent on your choice of schema (see below). Keywords, such as narrative, dataset, word_list, make it easier for people to find the resources they are looking for.

There are two metadata schemas (sets of elements) that have been defined for use by the linguistic community. The OLAC schema is based on the metadata elements used by libraries and other disciplines[13]. The IMDI (International Standards for Language Engineering Metadata Initiative[14]) schema was developed by the Max Planck Institute

[12] In some cases, eg. where consultants request anonymity, you may wish to use abbreviations and store the full names in a password protected file.

[13] Dublin Core Metadata Initiative, http://www.dublincore.org

[14] http://www.mpi.nl/IMDI

for Psycholinguistics on behalf of the DoBeS project. It is specifically designed for cataloguing language documentation materials and bundling related items together properly.

You can choose either the IMDI or the OLAC schema for your corpus. If you already know which archive you will be depositing your corpus with, use the one they require. Detailed documentation of each schema can be found on the respective websites.

Label every metadata entry with the same label that you use for the resource. List every related item in the metadata. Add as many notes about the circumstances of the participants and the creation of the resource as you can while they are still fresh in your mind. The provenance, or history, of a documentation resource is often of great interest to future generations of community members. Always be thinking about your consultants' great-grandchildren when you work with an endangered language.

If you are using the IMDI schema, you can use the IMDI Corpus Browser, downloadable from the IMDI website, to manage your corpus. AILLA, which also uses the IMDI schema, will have a Shoebox 5.0 metadata template available from its website by the time this volume is published. We also offer paper forms that you can download in a variety of formats. You can create your own metadata editor easily enough, using any database or spreadsheet program that you happen to have handy.

4. Conclusion

Doug Whalen has written: "we are poised to see a revolution [in the field of linguistics] caused by an unprecedented level of access to the raw materials of our discipline, using tools that have only recently become available ... The vanguard of the revolution will be those who study endangered languages" (Whalen 2003:30). I hope that this brief guide to corpus management will help ensure that these unprecedented quantities of materials documenting endangered languages are indeed accessible for speakers and researchers for generations to come.

5. References

Bird, Steven and Gary Simons (2003). Seven dimensions of portability for language documentation and description, *Language* 79/3: 557-582.

Golla, Victor (1996). *Newsletter of the J.P. Harrington Conference*. Number 10: May 1996. [online] http://www.rock-art.com/jph/nl10.htm. Accessed 2004-06-04.

Himmelmann, Nikolaus P. (1998). Documentary and descriptive linguistics, *Linguistics* 36: 161-195.

Hughes, Baden, Amol Kamat, and Steven Bird (2004). The OLAC Search Engine, presented at *Workshop on Linguistic Databases and Best Practice*; EMELD Language Digitization Project, Detroit, Michigan, July 16-18, 2004. [online.] http://www.emeld.org/workshop/2004/hughes3-demo.html. Accessed 2004-07-09.

Macri, Martha J., Victor Golla, and Lisa Woodward (2004). *J.P. Harrington Database Project.* [online.] http://cougar.ucdavis.edu/nas/NALC/JPH.html. Accessed 2004-07-09.

Webster, Andy (2003). Digital race to save languages, *BBC News World Edition*, Thursday, 20 March, 2003, 09:02 GMT. [online.] http://news.bbc.co.uk/2/hi/technology/2857041.stm. Accessed 2004-07-09.

Whalen, Douglas H. (2003). How the study of endangered languages will revolutionize linguistics, XVII International Congress of Linguists, Prague, Czech Republic, July 24-29, 2003. To appear in *Linguistics Today*, Piet van Sterkenburg (ed.), Amsterdam: John Benjamins.

Woodbury, Hanni (2003). *Onondaga-English/English-Onondaga Dictionary*. Toronto: University of Toronto Press.

Woodbury, Anthony C. (2003). Defining documentary linguistics, in Peter K. Austin (ed.) *Language Documentation and Description, Vol 1*: 35-51. SOAS.