Language archiving and data ecology

Presented by *Hugh Paterson III*

Bio:

Hugh is an information designer, archivist, and linguist who has worked with SIL International's Language and Culture Archive for the last five years. His work includes:

- Establishing audio and text digitization workflows,
- Collecting metadata for submitted, and digitized works
- Working with various archive clients who are submitting one of a kind collections of language documentation and indigenous language pedagogy materials.

He works extensively with information design analysis for a customized version of DSpace (like the U of O's ScholarSpace) and with an auxiliary tool, called RAMP[1], which is used to upload digital objects and their metadata to DSpace.

Presentation Abstract:

Archives and language artifacts play an increasingly important role in linguistics, folklore studies, ethnic studies, language documentation, ethnomusicology, (minority language focused) multi-lingual education, and anthropology. Funding agencies like the NSF, NEH, DEL, CIRC, SOAS/ELDP, DoBeS, and private foundations require grantees to archive their content. However, as producers of language material artifacts, do archiving requirements come to us begrudgingly or do we embrace them? Building on previously presented work (Paterson & Nordmoe 2013) about the practice of Archiving at the SIL Language and Culture Archive, Hugh presents an overview of linguistic archiving and a discussion of his recent survey on the practice of archiving lexical databases. This is the first known global survey that attempts to assess the archiving habits of linguists across disciplines and institutional affiliations.

Bibliography

Nordmoe, Jeremy. 2011. Introducing RAMP: an application for packaging metadata and resources offline for submission to an institutional repository. Paper presented at Workshop on Language Documentation & Archiving, SOAS, London, 18 November 2011. <Accessed: 18 February 2013>. http://www.sil.org/acpub/repository/LDLT3_Nordmoe_preprint.pdf

Paterson, Hugh III & Jeremy Nordmoe. 2013. Challenges of implementing a tool to extract metadata from linguists: the use case of RAMP. Paper presented at 3rd International Conference on Language Documentation and Conservation (ICLDC), Honolulu, Hawai'i, (2 March 2013). <Accessed: 08. May 2014>. http://hdl.handle.net/10125/26178

---

[1] RAMP was first presented by Nordmoe in 2011 at a Workshop held at SOAS.

Crafters of Language Artifacts

Language documentation vs. language description

Linguistics is chiefly concerned with the identification and analysis of language use patterns; in contrast language documentation is chiefly interested in the creation and (re)use of language artifacts. (Himmelmann 1998, 2012, Woodbury 2003)

In a way, to follow in the footsteps of Ansel Adams, Language Documenters create displaced/frozen/stolen moments. There are quite a few similarities between making visual artifacts and creating language artifacts. Both are data. Both affect our perception of reality. Both have the ability to call us to action and create social responses.

> Photography is a reality so subtle that it becomes more real than reality.
> ~ Alfred Stieglitz
>
> A photograph is usually looked at — seldom looked into.
> ~ Ansel Adams

> Actually, I'm not all that interested in the subject of photography. Once the picture is in the box, I'm not all that interested in what happens next. Hunters, after all, aren't cooks.
> ~ Henri Cartier-Bresson

It is this last perspective as described by Cartier-Bresson that Nathan *(2010)* says he finds among linguists, and encourages a different view on the production of language artifacts.

> Often, however, these field recordings were poor in quality as a result of three factors:
> (a) equipment choices (such as using inbuilt microphones of recorders);
> (b) recording methodology (microphones placed far from language speakers, or not suitably aimed);
> (c) an elicitation genre neither attractive to listen to nor containing much content suitable for using in teaching. (Nathan 2010: 262)

What other disciplines make and use language artifacts?

Language artifacts are important because they serve as the basic unit of data in a variety of academic disciplines: folklore studies, ethnic studies, language documentation, ethnomusicology, multi-lingual education (including minority language focused education), and anthropology, radio/communications, and now the self promoter via iPhones (though probably not a academic discipline yet).

In linguistics, as part of the sciences, it is increasingly more common to cite data, rather than just present some data in the context of supporting theoretical arguments.

Data Ecology (Ecology of Data)

Data Ecology, as a term is newish (since the 1950s - google ngram). The term represents a growing concern in the Sciences. This concern revolves around the evidence base and discovery methods. Globally, we are beginning to see some concerted discussion centering around how scientist interact with data (including data sharing). Two recent events include:

- International Symposium Towards an Ecology of Data. Political and Scientific Issues of Digital Data. February 14th, 2013
- National Academy of Sciences discussion in May of 2013: Public Access to Federally-Supported Research and Development Data and Publications: Data

An accepted data lifecycle model in the hard sciences is presented by Altman (2013).

Information lifecycle
- Design/Creation/Collection
- Storage/Ingest
- Processing
- Internal Sharing
- Analysis
- External dissemination/publication
- Re-use
- Long-term Access

Stakeholders
- Data source / Subjects
- Service & Infrastructure Providers
- Researchers
- Research Organizations
- Research Sponsors
- Scholarly Publishers
- Consumers
- Data Archives/Publishers

Language related example of data ecology:

Audio Data:

Some linguists have suggested that it always pays off to email other reachers on languages that are being investigated.

What is an Archive?

Sometimes they are corpora, which are call "archives" by their compilers. Sometimes these are typological "projects" or interactive "datasets" which are hosted on a computer somewhere.

TAPS (Target, Access, Preservation, and Sustainability): Checklist for Responsible Archiving of Digital Language Resources (Chang 2010: 136-7)

Target

Mission Statement: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?

Submission Criteria: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type?

Designated Communities: Is my desired audience a good match for the groups of users the archive targets (e.g., language community, academic community, etc.)?

Ongoing Relationship: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)

Access

Discoverability: Are the descriptive metadata for materials deposited at the archive searchable online? That is, the metadata is posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g., Google, Yahoo!, Bing, etc.)? Fixed Identifiers: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?

Reach: Will the audience that I wish to reach be able to access the materials once they are deposited in the archive? Access and Use Restrictions: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?

Preservation

Evidence of Long-Term Planning: Does the archive adhere to written policies and procedures for the long- term preservation of digital materials (e.g., the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?

Preservation Strategies: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?

Integrity: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g., a checksum, or digital signature, etc.)?

Authenticity: Does the archive ensure that digital materials contain what they claim to contain (e.g., by verifying that digital materials are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital holdings)?

Sustainability

Adequate Infrastructure: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g., quality media, environmentally-controlled storage, access-controlled storage area)?

Financial Sustainability: Does the archive appear to have secured sources of long-term funding?

Disaster Preparedness: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g., disaster recovery plan, offsite storage of backups)?

Succession Plan: Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?

Nathan, David. 2008. Digital archives: essential elements in the workflow for endangered languages documentation. In Peter K. Austin (ed.), Language Documentation and Description, vol. 5, 103-19. London: The Hans Rausing Endangered Languages Project, School of Oriental and African Studies.

Why Archive?

Long Tradition

Robinett (1954, 1955) published listings of languages that have language artifacts held at the Indiana University language archive.

> "...this report is concerned with listing acquisitions in the Archives of the Languages of the World, together with certain information regarding each acquisition." (1955)

Other archives have audio collections which go back to 1910

Resource discovery

Focused funding dollars, and research efforts (Holton 2011, Johnson 2004)

Ethical Reasons

Audio formats to use for historical preservation (IASA Technical Committee 2005)

The role of archiving as part of protecting cultural heritage (UNESCO 2003)

Community Access

The language community

Traditionally linguists have accepted and rallied around the concerns, goals and needs of the language community. The language community is certainly a stakeholder and a concerned party in the creation and archiving of language artifacts. Various authors have written about the impact that being able to access archived materials has had on their language revitalization (language teaching) programs. Yurok, California (Garrett 2011); Blackfeet, Montana (Kipp 2007)

In March (2011) Peter Austin asked several archives about their users. In reply Gary Holton writes:

> Our numbers vary considerably year to year, and so far we are not keeping track of online usage. In-person visits average about 200 distinct visits per year. Length of visit varies considerably, from a few minutes to a few weeks. This number may under-report actual usage because many visitors from Native communities represent a project or village council and bring back materials to that project or village. In other words, a single representative may bring materials which are eventually used by many more people. About 5% of visitors are linguists. But … the few linguists who do use the archive tend to use it fairly intensely, often over a period of days or weeks. For usage, there are several common agendas: (1) to acquire materials in their language; (2) to acquire pedagogical materials; (3) oral history, usually focused on particular person/village; (4) songs. We don't keep track of secondary products, but we should.

Can a readiness to associate the the term 'community' with the 'language speaking community', keep us from identifying other communities through time and space?

What if we were to re-conceptualize "community" to be both the "broad community" and the "deep community"?

The broad community

Today's communities are interdisciplinary and multifaceted. There are multiple academic audiences who are interested in language artifacts. These audiences are not bound by geographical limits. (Holton 2012) Language speakers certainly fit within this category.

The deep community

Communities don't just exist for a single moment but rather exist and change shape over time. So, concepts involving community and community interactions should consider time depth and the evolutionary nature of the "broad community".

Narcissism

Its good for publishing.

> Archiving can be considered a form of publishing: even if the materials themselves are archived with highly restricted access conditions, the metadata … 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. (Johnson 2004: 143)

Getting data out there and seen early is good for future citations (Brody & Harnad 2005). Acedemia.edu is based on these ideas.

Funding connections:

> The Project Description should discuss plans for archiving recordings, field notes, and processed documentary materials in a stable environment. Simply placing materials on a CD or a Web site will not in and of itself guarantee sustainable archiving. In discussing methods to be employed in recording, documenting, and archiving the endangered language(s), include reference to current statements of best practices (e.g. Bird and Simons, 2003; E-MELD; "Methodology and Standards" statements of the NEH Preservation and Access Division). http://www.nsf.gov/pubs/2011/nsf11554/nsf11554.htm (National Science Foundation 2011: Section A)

What is OLAC, why is it important?

Wouldn't it be great if there was an aggregated listing of resources available about languages? (Simons & Bird 2011, Simons, Bird & Spanne 2008)

Archives as the centers of engagement not just data bins.

Archives become the center of interaction around content and have a stabilizing effect on society. (Hartley, Lucy & Briggs 2013, Lothian 2013)

Do we believe in Archiving? (A case Study)

* "we" being users of FLEx and Toolbox who have responded

Positive (Let's Archive!):

> Even though my Toolbox file is a bit of a mess, I'd rather people in the future have a useful mess than nothing at all!
>
> I have archived my audio files at LACITO's Pangloss; I'm busy archiving my fieldnotes; but oddly enough, never my lexical Toolbox files. Probably because (a)_my dictionaries are work in progress, and unfinished, so the files never feel ready to be archived; (b)_no archive ever showed interest towards archiving this type of data (vs WAV or MPEG files); (c)_I'm not sure what files should be best archived: the raw Toolbox files? an XML version? a LexiquePro output?
>
> I was not aware of archiving. Thanks for the infor. I have some contact with PARADISEC

Not willing:

> Have you Archived: No (never took the time to find out how to do it)
>
> Yes it's been produced (dictionary), no it is not published. And no, I do not share copies of it. It's unfortunate that I lack the funding and time to complete it, but it won't see the light of day until then.
>
> The dictionary is still a "work-in-progress" and not ready to be archived.
>
> I using version 7.2.6 in an effort to compare several languages of South Bougainville. I have nothing ready for any archive. I am still trying to figure out how to make the glossing work with my assumptions about the languages. Ex: if a morpheme ends in velar nasal is positioned just before a bilabial stop. I just discovered that I can just use //ang// in the analysis line and leave the /amp/ in

the baseline. I had been unnecessarily declaring a morpheme variant in the lexicon. I have no urge to archive that level of my ignorance.

My FLEx database includes some vulgar terms that would not likely be in a public dictionary, so I would want to delete those before uploading it to a public archive.

Location:

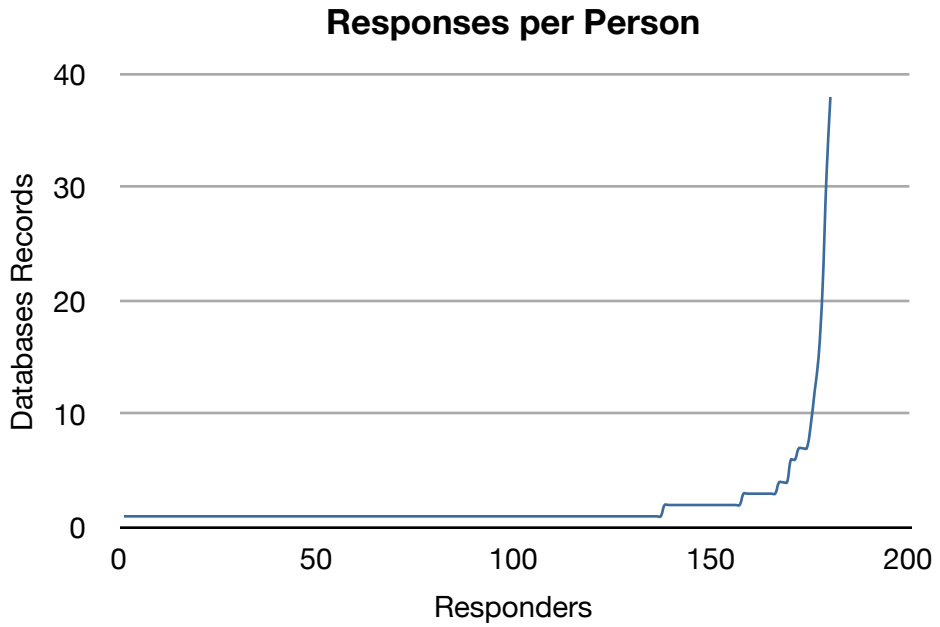
Have you archived your date?: It's in a git repository at github. - I have tried the move from Shoebox to Lex once and got bogged down in cross references and sub-entries.

Process:

Since it was submitted to SIL for e-publication and is online with Webonary.org, I assumed that it was de facto archived. If not, what should I do?

Distribution of responses

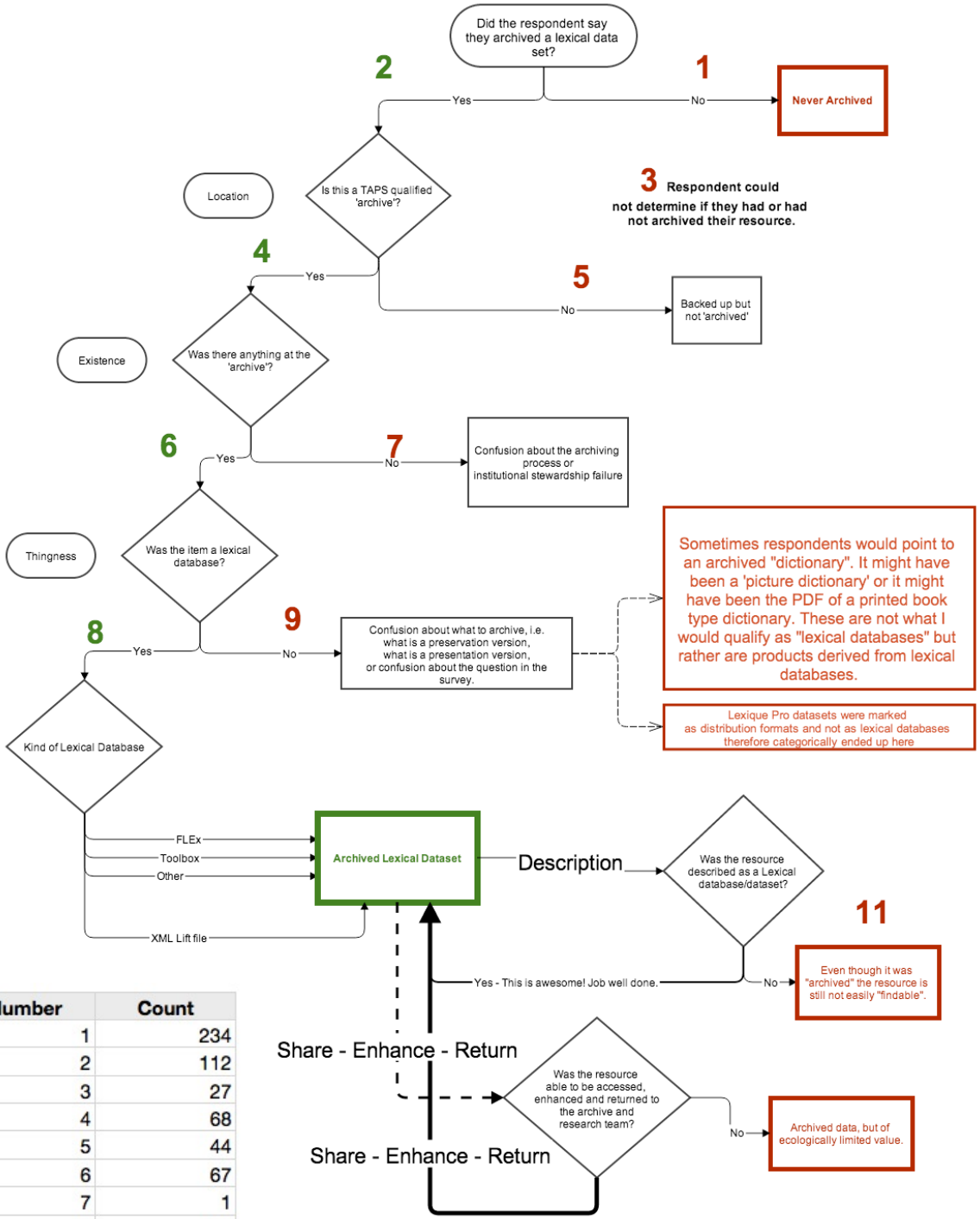
- About 179 Respondents
- About 373 lexical datasets are represented
- About 311 languages are represented in the questionnaire results
- The questionnaire has been open for responses since 6 November 2013



Participants	137	19	9	3	2	3	1	1	1	1	1	1
Database records contributed	1	2	3	4	6	7	9	12	15	21	31	38

by Hugh Paterson III

Analysis of responses



Number	Count
1	234
2	112
3	27
4	68
5	44
6	67
7	1
8	62
9	6
10	
11	6
10b	11

Key Terms

Ecology: is the scientific study of interactions among organisms and their environment, such as the interactions organisms have with each other and with their abiotic environment. Topics of interest to ecologists include the diversity, distribution, amount (biomass), number (population) of organisms, as well as competition between them within and among ecosystems. - definition for ecology shamelessly lifted from wikipedia

Data ecology: the system of exchange of data between persons, machines, and institutions. This is conceptualized as an organic network based on economic, technological, and relational factors.

Data: Stuff or knowledge, mostly the organized embodiment of knowledge, often data is considered to be digital stuff, but it could actually take a variety of forms. Examples often encountered in linguistic and language documentation work include: Language Artifacts (typetexts, audio recordings, video recordings,), statistics, geographical information, processing scripts (like PRAAT), annotations (like ELAN), descriptions, linguistic metadata.

Archive: an institution dedicated to the preservation and continued appropriate access of things stored within it

Language Artifact: is a recorded thing representing a linguistic performance. These might take the form of a typetext, a video recording, or and audio recording. It can be considered a sub-set of data that linguists commonly use.

Lexical Database: properties: is digital. Is in a structured format. Structured format is not based on visual stylesheet differentiation. Often the core data store of applications like FLEx, Toolbox, Shoebox, TshwaneLex, etc.

Dictionary: can be a derivative of a lexical database, but can be produced without a database. Is traditionally prepared for printed mode, though can be in digital formats like stardict (for interactive digital dictionaries) or .PDF for 'digital paper' dictionaries. Dictionaries might include grade-school and literacy primer materials which include a significant number of images.

Bibliography

Altman, Micah. 2013. Connecting Research, Publications, and Evidence: The Lifecycle and Institutional Ecology of Data. Paper presented at Public Access to Federally-Supported Research and Development Data and Publications: Data, Washington D.C., (16-17 May 2013). <Accessed: 21. May 2014>. http://vimeo.com/71358834

Austin, Peter K. 2011. Who uses digital language archives? In *Endangered Languages and Cultures*. 2011/04/29/. <Accessed: http://www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/

Brody, Tim & Stevan Harnad. 2005. Earlier Web Usage Statistics as Predictors of Later Citation Impact. <Accessed: 10. May 2014>. http://arxiv.org/abs/cs/0503020

Chang, Debbie. 2010. *TAPS: Checklist for Responsible Archiving of Digital Language Resources*. M.A. dissertation, Graduate Institute of Applied Linguistics, Dallas, Tx.

Garrett, Andrew. 2011. An Online Dictionary with Texts and Pedagogical Tools: The Yurok Language Project at Berkeley. *International Journal of Lexicography* 24.4: 405-19. http://ijl.oxfordjournals.org/content/24/4/405.abstract

Hartley, John, Niall Lucy & Robert Briggs. 2013. DIY John Curtin: Uncertain futures for heritage and citizenship in the era of digital friends and foes. *International Journal of Cultural Studies* 16.6: 557-77. http://ics.sagepub.com/content/16/6/557.abstract

Himmelmann, Nikolaus P. 1998. Documentary and Descriptive Linguistics (full version). *Linguistics* 36.1: 161-95. <Accessed: 2 January 2011>. http://www.uni-muenster.de/imperia/md/content/allgemeine_sprachwissenschaft/dozenten-unterlagen/himmelmann/linguistics98.pdf

Himmelmann, Nikolaus P. 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation & Conservation* 6.1: 187-207. http://hdl.handle.net/10125/4503

Holton, Gary. 2011. "Unknown Unknowns" and the Retrieval Problem in Language Documentation and Archiving. *Language Documentation & Conservation* 5.1: 157-68. <Accessed: 26. April 2014>. http://hdl.handle.net/10125/4496

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart et al. (eds.), *Potentials of Language Documentation: Methods, Analyses, and Utilization* (LD&C Special Publication 3), 111-7. Honolulu, Hawai'i: University of Hawai'i Press.

IASA Technical Committee. 2005. The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy, Version 3. In *Standards, Recommended Practices and Strategies*, ed. Dietrich Schüller, p. 12: International Association of Sound and Audiovisual Archives. <Accessed:

Johnson, Heidi. 2004. Language Documentation and Archiving, or How to build a better Corpus. In Peter Austin (ed.), *Language Documentation and Description* (2), 140-53. London, UK: SOAS.

Kipp, Darrell. 2007. Swimming in words. *Cultural Survival Quarterly* 31.2: 36-43.

Lothian, Alexis. 2013. Archival anarchies: Online fandom, subcultural conservation, and the transformative work of digital ephemera. *International Journal of Cultural Studies* 16.6: 541-56. http://ics.sagepub.com/content/16/6/541.abstract

Nathan, David. 2010. Sound and unsound practices in documentary linguistics: towards an epistemology for audio. In Peter Austin (ed.), *Language Documentation and Description*, vol. 7, 262-84. London: SOAS.

National Science Foundation. 2011. Documenting Endangered Languages (DEL) data, infrastructure and computational methods. Program Solicitation NSF 11-554. <Accessed: 21. May 2014>. http://www.nsf.gov/pubs/2011/nsf11554/nsf11554.htm

Nordmoe, Jeremy. 2011. Introducing RAMP: an application for packaging metadata and resources offline for submission to an institutional repository. Paper presented at Workshop on Language Documentation & Archiving, SOAS, London, 18 November 2011. <Accessed: 18 February 2013>. http://www.sil.org/acpub/repository/LDLT3_Nordmoe_preprint.pdf

Paterson, Hugh III & Jeremy Nordmoe. 2013. Challenges of implementing a tool to extract metadata from linguists: the use case of RAMP . Paper presented at 3rd International Conference on Language Documentation and Conservation (ICLDC), Honolulu, Hawai'i, (2 March 2013). <Accessed: 08. May 2014>. http://hdl.handle.net/10125/26178

Robinett, Florence M. 1954. First Report on the Archives of Languages of the World. *International Journal of American Linguistics* 20.3: 241-7. http://www.jstor.org/stable/1263351

Robinett, Florence M. 1955. Second Report on the Archives of Languages of the World. 21.1: 83-8. http://www.jstor.org/stable/1263221

Simons, Gary F. & Steven Bird. 2011. OLAC: Accessing the World's Language Resources. Paper presented at Poster session on Metadata in Language Documentation and Description, Annual Meeting of the Linguistic Society of America, Pittsburgh, 6–9 January 2011. <Accessed: http://www.hrelp.org/events/external/lsa2011/assets/LSA2011_BirdSimons_poster.pdf

Simons, Gary F., Steven Bird & Joan Spanne. 2008. OLAC Metadata Usage Guidelines. <Accessed: 10 January 2011>. http://www.language-archives.org/NOTE/usage-20080711.html

UNESCO, Ad Hoc Expert Group on Endangered Languages. 2003. Language Vitality and Endangerment. Paper presented at International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages, Paris, France (10–12 March 2003). <Accessed: 7 January 2013>. http://www.unesco.org/culture/ich/doc/src/00120-EN.pdf

Woodbury, Anthony C. 2003. Defining Documentary Linguistics In Peter K. Austin (ed.), *Language Documentation and Description*, vol. 1, 35-51. London: SOAS.