Reconceiving metadata: language documentation through thick and thin

David Nathan and Peter K. Austin

1. Introduction

Metadata can be described as 'data about data'. As a result of recent activities and discussion regarding documentation of endangered languages through projects such as OLAC (Open Language Archives Community) (Simons and Bird 2003) and IMDI (ISLE Meta Data Initiative), metadata within language documentation is now coming to be understood as information that is attached to a file or document for cataloguing purposes (see Johnson, this volume). We call this focus on cataloguing metadata 'thin metadata'. It runs the risk of not only being a simplistic view of the role of metadata in language documentation, but also, in the longer term, is likely to limit the accomplishments of the field. Thin metadata takes no account of information structuring within documents, and does not encourage state-of-the-art encoding or quality of documentary practice. Arguments that thin metadata is only intended to support resource discovery make it vulnerable to being made redundant by web search engines, and apply a consumerist approach to the complex and urgent problem of supporting endangered languages.

Over the past five years a new field of 'language documentation' has been emerging (see Himmelmann 1998, Woodbury 2003, DoBeS project specifications at http://www.mpi.nl/DOBES and also the Hans Rausing Endangered Languages Project materials at http://www.hrelp.org/documentation/). There are several features that distinguish language documentation from language description but a key one is that documentation embraces information and communication technologies to create digital sound and video recordings and to integrate them with text and other explanatory or analytical material. A component of this integration is metadata, the structured information describing characteristics of events and recordings and properties of data files. But there is far more to metadata than this. Much of the activity of traditional language description can be understood as creating metadata, 'data about data', that can potentially provide indexing, access, annotation, and classification for all data types, including recordings. Applying this descriptive and analytical material to recordings of linguistic performances is a primary driver of language documentation because linguistic analysis (transcription, translation and other higher levels of description) provides access keys to otherwise unsearchable recordings; in turn, recordings provide evidence for analysis and make the descriptive and analytical processes transparent and accountable. Thus, a richer, "thick metadata" approach that operates at all levels of linguistic analysis should be central to our field.

Endangered languages materials are characterised by diversity at several levels, so a "top down" imposition of standard, minimalist schemas does not provide the best

David Nathan and Peter K. Austin (2004) Reconceiving metadata: language documentation through thick and thin. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 179-187. London: SOAS.

path for describing them. There must be a complementary "bottom up" process that can explore and implement the kinds of metadata that linguists (and others, e.g. speech community members, language teachers, members of other academic disciplines) design and use.

What has emerged is a "metadata gap"; on the one hand we find minimalist cataloguing schemas promoted for the endangered languages field, and on the other are the rich descriptions that fieldworking linguists write as they create and analyse their data. This gap is also manifest in a lack of communication between the OWLs ('ordinary working linguists') and the computational linguistics community in terms of developing an agenda for better encoding practices as well as software for interaction with and mobilisation of endangered languages data.

2. Cataloguing metadata and resource discovery

In recent publications and discussion about endangered languages and their documentation, metadata has played a greater and greater role, while being presented purely as cataloguing data whose value lies in 'resource discovery', i.e. being able to identify and locate particular resources in internet catalogues (see Aristar-Dry 2003, Bird and Simons 2003, Good 2002 on this topic). Thus the E-MELD "School of Best Practice" states:

> "Metadata is information about resources. In this case, it is information about language resources: lexicons, audiotapes, transcribed texts, language descriptions, etc. It is similar to card catalogue information about library resources — it enables discovery and retrieval of resources through standardized information. Metadata should have a structured, unified and regular format so that it can be easily retrieved by mechanical, internet-based search engines like the OLAC harvester."

And again:

> "Metadata should include information about the language resource — information that will help others find the resource and assess its relevance to their own research. Examples of this type of information include the name of the linguist who created the resource; the subject (including the subject language); the language it is written in; the format (for example, audio or text); and so forth."

Equally clearly, Good 2002 writes:

> "One of the most important uses for metadata is to locate a resource. Thus, a book reference is designed to give enough information to allow someone to find that book. The other primary use of metadata is resource discovery — that is, finding resources relevant to one's research but which one is unaware of."

Let's look at metadata from this resource discovery point of view. Aside from side-stepping the question of what should be regarded as a "resource", a focus on resource discovery diverts attention from the fact that there are several types of resource-level metadata, including at least the following:

- *cataloguing* — title, speakers, collectors, time and place of recording, language name etc)
- *descriptive* — information about content, relationship to other resources etc
- *structural* — what structural devices and patterns exist in the document
- *technical* — performance and preservation information, description of formats etc
- *administrative* — work log, responsibilities, access protocol statements etc

Here is an example:

Cataloguing	Title: Sasak.dic Collector: PKA Speakers: YM, LH Language code: SAS
Descriptive	Trilingual Sasak-Indonesian-English dictionary, linked to finderlists, morpheme forms link to Sasak text collection
Structural	Dictionary entries with headword, part of speech, gloss in Bahasa Indonesia and English, cross-references for semantic relations; FOSF record format
Technical	Shoebox 5.0 ASCII text file
Administrative	Open access to all Last edited version dated 2004-06-25

Which of these types of metadata researchers choose to prepare depends on the type of materials under description, the usages and audiences that the materials are likely to have, and the metadata scheme adopted. However, there is no fixed boundary between cataloguing (resource discovery) metadata, and other resource-level metadata. The same metadata term could turn out to belong to different metadata types depending on the purpose and context of the materials; for example, the duration of a recording session might be cataloguing or resource-discovery metadata for a multimedia collection, but is more likely to be technical metadata for an analytically-oriented corpus. Similarly, the identity of the speaker may be useful for cataloguing, but also be relevant for administration of access protocols (which might require, eg. that some material is only available to relatives of the person in the recording).

The resource discovery metadata concept is useful and important for language documentation. However, it raises some serious questions. Who, for example, are the users most likely to benefit from it? — will it be ‘drive-by’ typologists who wish to rapidly assemble large amounts of disparate material, rather than those concerned with documenting and supporting endangered languages, or, indeed, the speaker communities of such languages?[1] Some descriptions of resource discovery projects are characterised by what we can call an ‘academic consumerist rhetoric’ (see Whalen 2003 for a clear example) that values linguistic resources for their ability to be exploited by those interested in topics such as relative phoneme frequencies or word order universals above their ability to provide support for language maintenance and description.

We might also ask to what extent resource discovery has been demonstrated to be a real and relevant bottleneck for endangered languages documentation; and what evidence is there that documentation efforts suffer as a result of not being able to locate resources? Or, even if it is a real issue for documentation, perhaps resource discovery may be already addressed, for example, by search tools such as Google and by other cataloguing efforts such as Ethnologue, Native Languages of the Americas[2], Aboriginal Studies Electronic Data Archive[3], Aboriginal Languages of Australia Virtual Library[4], all of which bring their own added value of applying assessment and skills to their publications, and have organisational commitment to quality and coverage.

Given these complementary efforts, and the limited resources available for documentation of the world’s endangered languages, it might be asked whether the scale of the resource discovery problem is proportionate to the amount of funding being devoted to it. In the light of these questions, it would be helpful to see empirical data establishing the resource discovery “problem”, and reporting the effectiveness of projects addressed to it.

More worrying, though, is an increasing belief among the language documentation community that (resource discovery) metadata is the primary determinant of language documentation. This has the effect of making documentation a narrow window that only looks out on ‘data’, in the same way that the current age has been characterised as shifting its perspective from ‘texts’ to ‘resources’, and from engaging with narrative to the manipulation of data. Paradoxically, however, resource discovery as currently envisaged cannot deliver any of the potential advantages of this shift in perspective, because its document- or resource-level focus places it firmly within the manuscript age — its concept of resource is no more modern than the technology of the printed book.

Documentation needs more emphasis on the coverage, quality and usability of the materials it is creating. Its greatest information and communication technology

[1] Interestingly, records of access to sound archives of indigenous languages which have been established for a generation, such as the Australian Institute of Aboriginal and Torres Strait Islander Studies, or the University of California Berkely Survey of California Indian languages show that 90% of usage is by descendants of the speakers recorded on the tapes in the archives, not academic researchers.

[2] http://www.native-languages.org/linguistics.htm.

[3] http://coombs.anu.edu.au/SpecialProj/ASEDA/ASEDA.html

[4] http://www.dnathan.com/VL/austLang.htm

needs lie in development of interfaces/software for effective delivery of language resources (not merely language data) to multiple audiences, including those involved in the maintenance and revival of languages (see Nathan and Csató 2005).

3. Thin and thick metadata

We call the resource discovery metadata approach ‘thin metadata’[5]. It can, as seen earlier, be compared with library practice. Libraries do provide an example of good metadata management; to maintain their catalogues (their ‘resource discovery systems’) librarians expect to receive metadata describing a publication’s provenance and perhaps its structure, but they do not tell authors how to structure their books, name their chapters etc. This division of responsibility is possible because institutionalised print publishers act as both guarantors of quality and suppliers of metadata. Publishers inherit stable, clearly distinguished, standardised categories for describing their objects (e.g. author, title, publisher, date, ISBN) and they generate endorsed publications with accompanying metadata. For language documentations, however, there are currently **no** comparable institutions (in fact what distinguishes most documentations is that they are *not* published in the traditional sense). Decoupling gatekeeping institutions from metadata supply may open up diverse opportunities for dissemination, but it does not mean that the traditional boundary between publishing/cataloguing metadata and other types of metadata must remain.

Even if resource discovery could achieve for cataloguing of language documentation what librarians have achieved for books, we still wonder if something better should be aimed at. Language documentation, as a new field operating in a largely digital environment, can aspire to exploit all the capabilities of new technologies. Our ‘objects’ are not as simple as books: everything except the recorded signal itself could be regarded as some kind of metadata. Other fields, ranging from mathematics to mapping — and even comic books — are increasingly using rich mark-up schemes that provide knowledge representation at all levels, and at any granularity, of their domain. Such markup schemes allow not only discovery but also navigation, querying, and repurposing of materials. The DoBeS project has made a start in this direction by developing software that assists in producing descriptions according to the IMDI scheme prescribed for its archive deposits, allowing those deposits to be navigated in terms of the IMDI categories[6].

4. Thick metadata and time-based media

Another defining feature of language documentation is its emphasis on the collection of recordings of authentic linguistic events. It is sometimes argued that the most pressing duty of documenters is to obtain such recordings while it is still possible to

[5] We allude here not only to the lack of depth of this kind of metadata but also to the ‘thin versus thick description’ discussed by Clifford Geertz and others within anthropology.

[6] http://www.mpi.nl/DOBES/; http://www.mpi.nl/IMDI/

do so, noting only the situational metadata (location, speaker etc.), and leaving any analysis to be done later when time and resources permit. We believe that this is incorrect; it is methodologically better to create transcriptions, annotations, and other commentary and analysis — in other words, thick metadata — as early as possible. Initiatives such as the development of an annotation framework (Bird and Liberman 1999) have provided methodologies for a few specific types of such rich metadata[7].

Collection of thick metadata will benefit from taking place in the community setting at the same time as recordings are made, since researchers are more likely to have access to those who were recorded or who were around at the time. One of the advantages of basing linguistic outputs on recordings is to build in the participation of community members and thereby reduce the informational distance between information providers and eventual users of the materials (see Nathan, this volume, regarding the approach taken in the Paakantyi CD project). Integrating the linguist's contribution by facilitating the collection of thick metadata can enhance the ways that recordings can be accessed and used, leading to an increase of 'linguistic bandwidth' that can potentially be mobilised to support urgent language work (for an example of such increased linguistic bandwidth see Csató and Nathan 2003).

However, the most compelling reason for needing thick metadata is the nature of digital sound and video. These time-based media (and, to a lesser extent, images) are resource-hungry and intractable to access and manage. Without thick metadata — for example in the form of time-aligned annotations — we are, in terms of accessibility, plunged back to a time before books and libraries existed, while at the same time bearing the contemporary costs of creating and maintaining electronic data.

Problems of access to non-textual data are not new, and go back at least to Comenius in the 17th century who linked text to images via numbers in his *Orbis sesualium pictus* (published in 1657, and described and illustrated in McArthur 1986: 144-6). Such numbers are a form of metadata, similar to 'stand-off' metadata that is currently used for many purposes such as aligning transcriptions with sounds (see Thieberger, this volume).

Unless we insist on the collection of thick metadata, formulated by the relevant knowledge bearers (rather than cataloguers) to accompany recordings, we face a future of wading in digital quicksand — a rapidly expanding mass of digitised sound, image and video, with no way to get a foothold. The metadata-as-cataloguing approach builds signposts that will entice us to visit these quicksand deserts, and leave us stranded there.

5. Thick metadata and electronic publishing

While commentators worry that the Internet will be "flooded with ... texts that are not subject to the traditional 'gatekeeping' editorial functions" (e.g. Levinson 1998: 76),

[7] Though oriented, again, primarily to an academic linguistic audience.

linguists are increasingly making their data available electronically. Nodes of authority and ways to evaluate quality are required; without them, emerging fields such as language documentation may never achieve trusted status in the electronic world.

It is possible that the respected cataloguing institutions that provide access to materials through resource discovery metadata can bestow a sufficient level of authority on materials, however their means of doing so — through verifying and holding a small amount of document-level metadata that bears little relation to the quality of the resource itself — seems inherently unreliable. On the other hand, the success of such institutions in becoming arbiters of quality for web-based linguistic materials would constitute a reinvention of the gatekeeping that some believe the Internet has set out to remove. The Internet environment allows materials to be located and evaluated in flexible ways that are facilitated by rich linguistic metadata. Resource-discovery schemas would not seem to offer a resolution to the Internet's productive tension between freedom from gatekeeping and the construction of authority and credibility.

6. Conclusion

We have described a polarised conception of metadata that pervades our field. On the one hand we find a plentiful and increasing amount of knowledge representation (e.g. the kind of richly structured interlinear text and lexical annotation that many linguists are now creating using tools such as Shoebox and Elan); on the other, we find simple, increasingly ubiquitous cataloguing metadata. However, as we have shown, metadata actually ranges along a continuum. It is as if a bridge is being built from both ends without thought for how they can meet in the middle: users can, for example, retrieve resources based on document-level properties, but are not encouraged to create and exploit resources in terms of their unlimited range of 'thick metadata' categories.

Current projects based around "ontologies" (eg. the E-MELD GOLD ontology) are one attempt to deal with this gap. These projects establish standard ways to express concepts and then encourage researchers to 'map' their local terminologies to standard terms in order to provide interoperability. We have looked briefly at the design of some of these projects and note that generally they investigate relatively well-defined structural domains such as lexicography and interlinearised data, and are typically focused on computational rather than linguistic outcomes, so, for language documenters and OWLs, the metadata gap is not being narrowed.

What is needed to support language documentation is a metadata methodology that provides flexible, richly articulated knowledge representation schemas to encode linguists' cascading layers of data and metadata. These schemas should be based on the evolving practices of working linguists and the communities of interest in the materials. In addition, software for creating and working with richly structured materials is required. Shifts of emphasis from knowledge to resources, and from expressive skills to access tools, may be trends of our time, but they do not provide the route to combating language loss. Perhaps future generations will not thank us for the thin gruel of cataloguing metadata that we leave for them to digest.

7. References

Aristar-Dry, Helen. (2004). Metadata. *Presentation at E-MELD Symposium on Endangered Data vs. Enduring Practice*, LSA, Boston, MA.

Bird, Steven and Gary Simons (2003). Seven dimensions of portability for language documentation and description. *Language* 93: 557-582

Bird, Steven and Mark Liberman (1999). A formal framework for linguistic annotation. *Technical Report MS-CIS-99-01*. Department of Computer and Information Science, University of Pennsylvania.

Csató, Eva A. and David Nathan (2003). Multimedia and documentation of endangered languages. *Language Documentation and Description, Vol 1*, 73-84.

Dauenhauer, N and R. (1998). Technical, emotional, and ideological issues in reversing language shift: examples from South East Alaska, in Lenore Grenoble and Lindsay Whaley (eds.) *Endangered Languages: Language loss and community response*. Cambridge: CUP

Good, Jeff. (2002). *A gentle introduction to metadata*. http://www.language-archives.org/documents/gentle-intro.html

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

Levinson, P. (1998). *The soft edge: a natural history and future of the information revolution*. London: Routledge

McArthur, Tom (1986). *Worlds of Reference: Lexicography, learning, and language from the clay tablet to the computer*. Cambridge University Press, Cambridge

Nathan, David and Eva A. Csató (2005). Multimedia: A Community-Oriented Information and Communication Technology. In Anju Saxena (ed.) *Minor Languages of South Asia*.

Simons, Gary (2002). The Electronic Encoding of Lexical Resources: A Roadmap to Best Practice. *Proceedings of EMELD Workshop on Digitizing Lexical Information*, Ypsilanti, MI. [HTML]

Simons, Gary (2003). The Electronic Encoding of Text Resources: A Roadmap to Best Practice. *Proceedings of EMELD Workshop on Digitizing and Annotating Texts and Field Recordings*, East Lansing, MI.

Simons, Gary and Steven Bird (2003). The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. *Literary and Linguistic Computing* 18/2: 117-127

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 Piet van Sterkenburg (ed.) *Linguistics Today*. Amsterdam: John Benjamins.

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

Websites:

http://emeld.org/school/index.html

http://www.hrelp.org/documentation/