Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 3*. Editor: Peter K. Austin

The digital skills of language documentation

ROBERT MUNRO

The digital skills of language documentation

Robert Munro

1. Introduction[1]

Information technology (IT) plays an important role in language documentation (Woodbury, 2003). As language documentation is a multidisciplinary domain, it is not always easy to identify which parties need to know which IT skills. For example, the use of software supporting XML is widely recommended, but who needs to know how XML works: is it the documenter, the archivist, a software developer, a video maker, or all/none of them? As language documentation is an emerging field, it is timely to examine the nature of the IT skills required for language documentation, and to see how these might differ from related fields. There have been a number of recent papers looking at IT in language documentation from the perspective of digital archiving and data analysis, focusing on format standards, information system technologies and searching capabilities (Bird and Simons, 2003; Johnson 2004; Wittenburg and Broeder, 2002). This paper draws on them, but its primary objective is to complement them by giving an IT-informed account of the documentation itself, and so this paper is specifically intended for an audience of researchers planning or undertaking language documentation. It is assumed that the reader already possesses the linguistic knowledge and skills necessary to undertake a documentation project.

Section 2 gives an outline of the different types of IT professionals a documenter might include in a documentation project, arguing that IT skills such as systems analysis and design and data modeling will typically be more important to a language documentation project than IT skills such as programming and software development. Section 3 defines the three specific sets of IT skills needed for language documentation: consultation and elicitation, media management and data management. Sections 4 and 5 then compare these skills to those of language description and digital archiving, showing how exercising these skills will influence the utility and archivability of the documentation materials.

2. IT disciplines and language documentation

For a researcher planning and/or managing a documentation project it is important that they possess a clear idea of the IT skills and knowledge required by that project. IT knowledge is not a single scale of technological competence, but a blanket term for a collection of very different disciplines, and an IT professional is typically an expert in

[1] Much of this paper was presented at the ELAP workshop on *Multidisciplinary Approaches to Language Documentation* (Munro 2004). It also draws on the content and feedback from seminars on data management seminars given at ELDP grantee training workshops in September 2004 and July 2005. Thanks go to Peter Austin and David Nathan for their feedback and suggestions on a more recent version of this paper, and to comments of the anonymous reviewers. Any errors or omissions are my own.

Robert Munro (2005) The digital skills of language documentation In Peter K. Austin (ed.) *Language Documentation and Description Vol 3*. 141-156 London: SOAS

only a small subset of them. After identifying the IT skills required by a documentation project, a researcher must be able to acquire any missing skills, or be in a position to appoint and manage an appropriate person who has them.

The term ‘IT’ is still occasionally conflated with the term ‘computers’, or even more narrowly ‘programming’. However, the scope of IT in practice extends beyond the supporting technology and, in fact, programming knowledge is probably superfluous to most language documentation projects. For proficient language documentation a very different set of IT skills is required.

2.1 IT disciplines

There is no consensus for dividing IT into its sub-disciplines, and shifting technologies mean that this is unlikely to change, but IT can currently be divided into three areas with relatively little overlap. These are (1) software development, (2) network management, and (3) consultancy / systems analysis and design.

Software developers or programmers have comprised most IT staff in past research projects. Research institutions have long favoured the pairing of domain-experts with programmers for application development. Software developers and programmers will be skilled at the design and implementation of the software design of an application, but not the functional architecture, that is, they are builders, not architects. The domain-expert/programmer pairing is a model now used less and less in industry. For language documentation, where there is no single domain and therefore no domain-expert with complete expertise, this is rarely a suitable model. Nonetheless, for research projects there is more often more emphasis on processing the data than structuring it. Therefore, as language documentation is increasingly taking advantage of advances in computational linguistics for language description, the need for data processing skills is increasing (Bender et al, 2004).

Network management encompasses the duties performed by the IT support staff that maintain computer networks in many institutions, who often combine these duties with response to IT ‘helpdesk’ requests. These IT professionals are also known as ‘networks and systems programmers’ and are typically trained in installing and maintaining the infrastructure that supports end users, and in the programming of systems that support networked data and applications. These skills relate to language documentation only in that they support the infrastructure used by documenters, and of the three IT sub-disciplines described here they are the least directly related to the processes of documentation projects. The exception is when there is a need to transfer large volumes of digital information between working computers and/or recording equipment, which requires some knowledge of networking.

Consultants and systems analysts are specialists in data management strategies. An information system architecture is any formal architecture describing the structures and flow of information, for example, a database architecture or markup ontology and the processes utilizing them. A systems analyst is able to undertake the tasks necessary to inform the design of an information system architecture or software application; a design process often known as ‘data modeling’. An information system is arrived upon through formal analysis of collected information, and in selecting the correct methods and sources for collection, and is typically an iterative process.

2.2 Seeking IT advice

Language documenters are performing tasks very similar to those carried out by a systems analyst in their consultation with speakers of endangered languages. An analyst will elicit information from organisations with the goal of building a comprehensive model of business processes, while a language documenter will elicit language from speakers with the goal of building a comprehensive model of language and speech processes. Therefore, even though software development deserves a privileged position within linguistics, language documenters are in more need of consultation and systems analysis skills.

Depending on the project, the IT skills necessary for a given language documentation project may be within the capability of linguists trained in language description, but including a professional in a project may still be desirable. For example, if a project required the development of new complex software for language elicitation, then a person with qualifications in computer science (encompassing software development and programming) would be an invaluable member of the project team. However, if a project requires a digital representation of the complicated relationships between a number of materials and the cultural context (see Harrison, this volume) then a person with qualifications in information systems (encompassing consultancy / systems analysis and design) would be a more valuable member of the project team.

For language documentation, IT practice does not start when a recording device is first switched on; it begins with a documentation project’s first contact with the speakers. More broadly, IT truly begins with the project’s conception, and so a team undertaking language documentation needs at least one member with the appropriate skills to be involved from the start.

Research institutions tend to be populated with IT professionals *other* than consultants and systems analysts and so it is not surprising that documenters report that good advice can be hard to find. IT support staff are accustomed to supplying solutions to problems outside their area of expertise, so if a documentation project is relying on an overstretched support staff member for advice, the head of that project needs to be clear about the IT person’s level of expertise in the specific area(s).

3. IT skills for language documenters

While the application of IT skills will depend on the goals of a documentation project, there are three sets of IT skills that will be required by all language documentation projects:

- consultation and elicitation
- media management
- data management

The first and third of these are skills of consultancy and systems analysis and design, the second of these falls closer to that of an 'expert user' rather than a specific IT discipline. These skills are described in further detail in the following sections and are summarized in Figure 1. The descriptions here focus on audio recordings, but the same principles apply to any collected or recorded documentation material. There is a slight chronological ordering in the application of these skills, with consultation being the most important skill at the beginning of a project and data management being the most important skill in the delivery of materials. However, the planning and application of all three will span the length of a documentation project.

Surprisingly, language documenters are currently much better at the more complicated IT skills (consultation and elicitation) than the simpler ones (data management). This is because the simpler ones rely on more recent technological developments, and they are less likely to be carried over from other fields of linguistics. The three sets of IT skills are more than static knowledge of the formats and standards required for creating publishable materials. For all three a documenter will need to apply the skills in a changing and unpredictable environment.

The time it takes to acquire the three skill sets differs significantly. Within data management most people take a day or two to become skilled in using XML or basic relational modeling. For media management it will usually take about a week of training to become familiar with audio recording techniques and quality (although video will take longer and vary according to the intended use) and a couple of months of irregular use to become familiar with linguistic software. At the end of the scale, consultation and elicitation skills will be refined over many years.

Figure 1: IT skills for language documentation

Skill set	**Summary of skills**	**Time needed to acquire skill**
Consultation and elicitation	Project management. Eliciting information. Consciousness of personal influence on recordings. Application of ethical knowledge Skills transfer	Many years
Media management	Recording techniques Transferring data between storage mediums Use of linguistic software	Audio recording: a week Video recording: weeks to months Software tools: a few months
Data management	Storing data and relationships as explicit structures Ensuring data integrity and portability	XML or relational modeling: a few days XML software and database software: days to weeks

3.1 Consultation and elicitation

Consultation and elicitation is how a documenter obtains knowledge about a language and the communities in which it is spoken. Combining consultation with sophisticated media and data management facilitates a better feedback loop, as the results of documentation can be viewed and negotiated with the language consultants and wider community while the documenter is still undertaking fieldwork.

Fieldwork is often called an art (Wolcott 1995). However, as documenters manage their interaction with the speakers in order to record speech practices (Himmelmann, 1998) and need to document the field methods used (Bird and Simons, 2003) fieldwork for language documentation is better described as a science.

Approaching language documentation as a science is a 'soft systems methodology' (Checkland and Scholes 1990). A soft systems methodology is a system of inquiry where the researcher/analyst perceives complex and potentially confusing systems, and organizes the exploration and learning of these systems. This is distinct from a 'hard systems methodology' where the perceived systems are understood almost immediately by the observer and can be modelled/recorded with much less interaction, although the difference between the two is not always a hard boundary. For language documentation, the 'systems' are any that might be documented or influence the documentation process, ranging from the rules and constraints governing phonological alternation to the personal relationships between individuals within a speech community. As should be evident, a soft systems methodology has the researcher taking a more active role in the analysis than in a hard systems methodology, giving them a more direct influence on the nature of the systems modelled.

An example can be seen in the scale of the 'naturalness' of speech acts (Himmelmann, 1998), where the documenters' self-awareness of their interaction is a fundamental parameter in documentation planning. At the least natural end of the scale, well-known formal elicitation methods can be easily adapted to most documentation projects. For recording staged, observed or natural communicative events the interaction of the documenter can have a varying personal and/or cultural influence on the content and quality of the recordings. Therefore, the nature of event, the recording environment, ethical considerations, intellectual properties issues, and the speakers' cultural and personal characteristics are all factors that need to be planned for and documented.

An important step in establishing a relationship with speakers of endangered languages is formalizing the professional relationships with language consultants and members of their communities. Not surprisingly, there is usually a strong correlation between the amount of community involvement in a project and the quality of the documentation (Grinevald, 2003; Woodbury and England, 2004), especially when the outcomes include developing specific resources for use by speakers (Csato and Nathan, 2003). The exact skills that are required will depend on the nature of the communities and participants, the scope of the project, and the time and resources available. In addition, any strategies for consultation and elicitation will necessarily develop along with the project.

If fieldwork should be increasingly undertaken by speakers, and the arguments for it are strong (Grinevald, 2003; Woodbury and England, 2004), then the relationships between documenters and speakers will become more formalised and managing these will become an important part of the documentation process. Therefore, skills transfer plays an important role in documentation. In order to involve the language speakers in the documentation, documenters will usually need to train consultants in techniques for transcription and possibly annotation. This could range from simply teaching the conventions for hand-written aligned transcriptions to the use of transcription and annotation software, and even the formulation of formal data management strategies for representing language-specific or culture-specific phenomena. If a documentation project is undertaken by only one person, then it will often be a good idea to train a language consultant or related person in the operation of recording equipment, especially for the more natural communicative events, allowing the documenter to focus on the elicitation and interaction without compromising the recording quality. People from most backgrounds will usually welcome the opportunity to be employed to acquire the skills necessary to operate video and audio recording equipment, so this is also a way to immediately benefit the speaker community.

The consultation and elicitation skills described above are often grouped under the banner of 'project management' (Duncan, 1996) describing how the outcomes of any project are determined by its planning and ongoing strategies of execution. This extends to the conscious management of the roles of people involved in a project. For example, beyond making a person's duties clear, they should also undertake work with a clear understanding of the possible future uses of their contributions and the objectives of the greater project. For example, for work that can sometimes be a little uninteresting, like transcription, this would include planning explicit motivation strategies (even for yourself). These project management skills are probably already familiar to documenters, as they are all current practices of people performing language documentation and description. Because they are undertaken with a view to creating digital recordings, they are also sophisticated IT skills.

3.2 Media management

Media management is how a language documenter records speech acts and the participant's knowledge of a language. This encompasses the technical aspects of conducting fieldwork, from audio and video recording to transcription and annotation.

Recording techniques are a well described area and are not covered here, although the relative importance of a good microphone is often overlooked (Nathan, 2004, Barwick, this volume). The relative quality of different audio recorders and sound formats has also been described elsewhere, but it is worth repeating the value of recording to an open, uncompressed format (Bradley 2004).

Typically, the ability to transfer data between storage mediums is also necessary, as current PC's and notebooks do not contain hardware of sufficient quality to record directly to them. Techniques need to be learned for lossless transfer in the case of digital-to-digital transfer and minimal-loss in the case of analogue-to-digital capture.

Transcription and annotation also fall under media management. Since they are built on top of a recording these are value-adding exercises. From the point of the data, transcription and annotation are creating rich 'thick' metadata (Nathan and Austin, 2004). From the point of view of linguistics this is simply the process for recording the participants' knowledge of a recording. The skills needed here are the use of transcription and annotation software, including their installation.

3.3 Data management

Data management enables documenters to share the recordings and their knowledge. The way that data is structured is, in itself, information and so every relationship that is described between items adds to the richness of the documentation. Something that a documenter only considers in passing (for example, that a speaker is the sister of someone recorded yesterday), might become the most important piece of information that a person later discovers. If this information is stored in a structured format, then it will be easier for a later archive or digital publication to explicitly represent the links between the relevant items and to facilitate the development of rich searching and navigation capabilities. It will also aid in the development of multimedia and other materials that will be of more immediate interest to an endangered language community.

Although it is best practice to design the necessary data structures before a documentation project is undertaken, it is optimistic to think that the exact data structures needed to represent a certain language and its contexts can be defined before the project begins. Hopefully, the project and language will yield many unanticipated and interesting phenomena that a documenter will wish to record in a formalised structure. Therefore, documenters need to know how to create formal data structures in order to record the unanticipated phenomena in a structured format. These are outlined below.

3.3.1 Data management standards

The most well-known data storage structure is the directory structure ('folders') found on all PCs. Unfortunately, a directory structure is too simple for representing language documentation materials. Formal relationships between files can only be represented by the directory hierarchy. Within a directory structure, relationships can be represented through file-naming conventions, which is imperfect but is one way to represent non-hierarchical relationships between files. A further problem with directories is that they do not allow us to represent relationships between pieces of information within the files. For example, you might want to formally represent that one of the speakers in an audio file corresponds to the written profile of a person stored within a document elsewhere.

There are two well-known data management standards that have been developed for storing structured data with relationships of arbitrary complexity:

- Extensible Markup Language (XML)
- Relational format

Ideally, a language documenter should manage all collected data in one of these two formats. It should be emphasized that these are formats, not softwares or hardwares and neither requires specialised software: relational format can be maintained with any application that contains tables and XML with any application that handles text. Specialised software (such as a relational database) is simply a set of functions that wrap around these standard formats to ensure data integrity, portability and efficient, flexible access for management.

3.3.2 Data integrity and portability

Data management involves ensuring that the data is correct and able to be shared. These two objectives are respectively known as 'data integrity' and 'data portability'.

Data portability refers to the data’s reliance on specific software and computing environments. For example, portable data could be accessed by one application running on a Mac and also by a different application running on a PC, without any transformations needed to meet the specific requirements of those applications or operating systems. If data is stored as XML or relational format, then this is possible. More broadly, portability also refers to the data’s longevity and its ability to be used by non-linguists. For a much more detailed account of portability in language documentation and the surrounding issues, see Bird and Simons, 2003.

Data integrity refers to correctness of the data and the correctness of the references between pieces of information. Barring maliciousness or computer error, problems with the content of the data largely arise from human error. Referential integrity can be a much larger problem. Every recording, transcription etc needs a unique identifier so that it is possible to make unambiguous reference to an item. Referential integrity also requires that all references are correct. For example, referential integrity is compromised if a transcription points to the wrong sound file, a sound file that does not exist, or it is not clear which of many files it points to. In discussion of the ambiguities that arise when referential integrity is not maintained, Johnson calls the consistent labeling of objects the ‘eighth’ dimension of portability (Johnson, 2004).

Many documenters will be tempted to create their own formal or semi-formal methods of storing data. This is not advised as both XML and relational modeling can be learned in a couple of days and you will spend more time tweaking your personalized structures than you would have spent learning XML or relational modeling in the first place. Nonetheless, a few perfectly cross-referenced set of tables in a spreadsheet is better than broken XML. This is because converting between any two structured formats is a relatively simple task, but as soon as the cross-references in the spreadsheet contain errors or inconsistencies the relative effort in conversion will soar. This is why adopting one of these two standards is recommended: data management standards, and the software that uses them, are designed so that errors in data integrity are much less likely to occur, and when they do occur they are localized and easier to diagnose and correct.

Data management also extends to more basic practices such as maintaining back-ups of data, consistent labeling of physical objects, and ensuring correct references to physical objects from within the digital representations.

4. IT in language description

Many consultation and elicitation skills used in documentation can be carried over from language description. The major additional requirement is that more care needs to taken to ensure that intellectual property rights are clear, as, among other reasons, the stories, ceremonies and everyday gossip can become published materials. While good language description will also use of sophisticated data management, the invention of ad-hoc data management strategies and use software not supporting data integrity or portability has been the norm. This has often been the biggest problem in attempting to turn legacy recordings into materials suitable for archiving (Aristar and Dry 2001; Holton, 2003).

Documentation differs most significantly from description in the nature of the materials that need to be of publishable standard, as documentation aims to publish the primary materials. Himmelmann gives a breakdown of the differences, focusing in particular on data collection and data analysis (Himmelmann 1998), but the starkest difference between description and documentation in terms of IT skills is in data publication. A summary is given in Figure 2.

Figure 2: Materials for publication in language description and language documentation

	Language Description	**Language Documentation**
Materials:	academic papers, formal grammars, dictionaries(?) (secondary materials)	audio and video recordings, images, texts. (primary materials)
Quality defined by:	researchers (linguists)	researchers, community members, teachers and learners
Within the domain knowledge of the collector of materials?	yes	no

The broader sense of portability, its ability to be used by non-linguists, also separates documentation from description as the quality of a recording and clarity of transcription and annotation also need to meet the requirements of people outside the linguistic research community. For language description, an audio recording only needs to be understood by the person transcribing and analyzing that recording. The ‘publication standard’ here is found in academic articles, where care is taken to ensure a clear writing style using a set of standard well-known terms that can be understood by an audience of linguists. In language documentation, the same attention needs to be made in recording, managing and structuring the primary recordings. Here the ‘publication standard’ is that defined by any one of researchers, community members, learners or teachers (Austin, 2003) and the potential to improve quality by re-editing or re-recording is limited.

The most important distinction can be seen here: ‘quality’ in language description is defined by peer-review alone, ‘quality’ in language documentation is not. The result for documenters is that while they will have a good idea of the materials and quality of recording necessary for description, for language documentation the possible uses of the materials are outside their domain knowledge. The implications of this go broader than fieldwork. Conferences and workshops are tailored for researchers within a single domain, and are perhaps not the best forums for reaching cross-disciplinary consensus. As language documentation becomes a more established field, the definition of 'quality' will become clearer and will be moderated by peers, but this will only be a process of mediation, as the definition of quality will still need to be derived from a broad variety of stakeholders.

Within IT, the quality of recordings can be the weakest link in language documentation for endangered language communities. Audio/Video players, computers and televisions are much easier to obtain than the corresponding software and recordings of endangered languages. While technological infrastructure will mostly improve, often dramatically, the opportunity to make quality recordings will mostly decline. Documenters often have the opportunity to create the only digital materials of a language that are of a publishable standard. For fieldworkers, there is some convergence here that offers a good guideline. The materials that are of the highest value for documentation, high quality recordings of speech with heritage value, are also of the highest value for many endangered language communities. As all fieldworkers wish to contribute to the language speakers and language community, this gives a useful guideline for deciding upon the quality of recording: ideally, the quality of a recording, as determined by the recording equipment and recording environment, should reflect the respect we have for a language and its speakers.

5. IT in archiving

Archives have long been a central party in the documentation and description of endangered languages (Johnson, 2004). They ingest the information recorded by different language documenters and make it accessible to the large variety of stakeholders and interested parties. Here, the broad sense of 'accessible' is intended, with archives providing multiple and flexible modes of access to the data stored. An archive of endangered languages has much stricter technological constraints than language documenters: they need to ensure that the file formats and data structure formats are open and will have longevity.

This paper has deliberately avoided describing specific software, but it is worth grounding the scale here. On a scale of 1 to 6, OLAC/IMDI compliant Shoebox (Toolbox) output would score between 3 and 4, and OLAC/IMDI compliant ELAN output would score 5. From a data management point of view, this is because ELAN stores data in well-formed XML, but the structure of Shoebox's output can be ambiguous and difficult to transform without loss of information, meaning that Shoebox does not ensure data integrity or data portability. If a researcher is aware of the limitations of Shoebox but wishes to use it due to familiarity or functionality, then the application of good data management knowledge can improve Shoebox output to 5 (Austin, 2002).

An archive will usually accept materials in a much broader range of formats than it will store them in, and so it will need to automate or semi-automate the conversion/migration of materials into formats and structures suited for long-term preservation and the needs of the archive's user communities. In order to avoid the loss of information in any conversions, an archivist needs a documenter to have exercised their data management skills (Holton, 2003; Wittenburg, 2003; Johnson, 2004). Figure 3 gives the scale of increasing richness in the nature of data deposited with an archive. Note that for simplicity, the scale assumes that the recordings, transcriptions and other materials are already high quality in terms of choice of content, audio quality, choice of formats and correctness of transcription and annotation. These properties could all form dimensions of archivability on a more complicated multi-dimensional scale. The scale in Figure 3 is a conflation of two dimensions of data management for language documentation: using a metadata standard and a data management standard. This means that there are combinations of properties not covered, for example, a documentation project may record all data in rich well-formed XML but not compliant to a particular metadata standard. As this is a very general model ranking, these can simply be approximated along the scale.

6. Conclusions

This paper has looked at the role of IT in the emerging field of language documentation. It has argued that while there is a need for software development and programming, documentation has a particular need for consultation and systems analysis skills.

The ultimate goal of language documentation as IT is to provide access to high quality materials, and language documenters require three specific sets of skills in order to do this: consultation and elicitation, media management, and data management. Consultation and elicitation is how a documenter obtains knowledge about an endangered language and the communities in which it is spoken, media management is how they record that knowledge, and data management is how documenters can share that knowledge with the immediate community and with people they will never meet.

Figure 3: The scale of archivability for language documentation materials

<table>
<tr><th>Scale</th><th>Description</th></tr>
<tr><td>1. Unlabeled recordings</td><td>Recordings with no independent information about the contents. For example, a tape, minidisk or WAV file with metadata such as speaker name and date only captured as spoken at the beginning of the recording.</td></tr>
<tr><td>2. Labeled recordings</td><td>Recordings with unique file names using an informative file-naming convention. For example, A recording entitled “rg_12_2.wav”, where “rg”, “12” and “2” are interpretable codes describing metadata about the recordings. For example, “rg” might represent the speaker’s name.</td></tr>
<tr><td>3. Information compliant to a metadata standard.</td><td>Recordings with unique file names and accompanying information that corresponds to a metadata standard such as OLAC or IMDI. For example, a recording with an accompanying text document describing the contents:

“File rg_12_2.wav is a recording of Roger Gasket, who is pictured in file rg_p.jpg, speaking Atinle. It was made on the 12th March 2005 at Mt Solitaire by Robert Munro. It is transcribed in file rg65.xml and annotated in fda.xml. The recording continues on file rg_12_3.wav…”*</td></tr>
<tr><td>4. Structured information compliant to a metadata standard.</td><td>Recordings with unique file names and <i>structured</i> accompanying information corresponding to a metadata standard. As 3, but with the information in a table in an application such as Filemaker or MS EXCEL:

<table>
<tr><th>File</th><th>Creator</th><th>Format</th><th>Contributor</th><th>Language</th><th>…*</th></tr>
<tr><td>rg_12_2.wav</td><td>R. Munro</td><td>WAV</td><td>R.Gasket</td><td>Atinle</td><td>…</td></tr>
<tr><td>rg65.xml</td><td>R. Munro</td><td>XML</td><td>R.Gasket</td><td>Atinle</td><td>…</td></tr>
<tr><td>rg_p.jpg</td><td>R. Munro</td><td>JPG</td><td>R.Gasket</td><td>-</td><td>…</td></tr>
<tr><td>rg_12_3.wav</td><td>R. Munro</td><td>WAV</td><td>R.Gasket</td><td>Atinle</td><td>…</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
</table></td></tr>
</table>

5. Information compliant to a metadata standard, and to a data management standard.	As 4, with the structure conforming to a data management standard such as XML or relational format:
	<olac xmlns="http://www.language-archives.org/OLAC/0.4" >
	    <creator>Robert Munro</creator>
	    <format>WAV</format>
	    <language>Atinle</language>
	    <contributor>R. Gasket</contributor>
	      …*

6. Extended metadata standard.	As 5, with the metadata standard extended to capture everything that might be of interest to researchers or the speakers. For example, OLAC or IMDI compliant data, with accompanying structures modeling the relationship between the language and phenomena such as kinship terminologies, ethnobiological ontologies or geographical information.

* assuming a complete example would be at least minimally compliant with OLAC or IMDI.

7. References

Aristar, Anthony and Helen Dry (2001). The EMELD Project. *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia.

Austin, Peter K. (2002). Developing Interactive Knowledgebases for Australian Aboriginal Languages - Malyangapa. *Paper presented at the Workshop on Australian Aboriginal Languages*, University of Melbourne

Austin, Peter K. (2003). Introduction. In Peter K Austin (ed) *Language Documentation and Description Volume* 1: 6-14. London: SOAS

Bender, Emily M., Dan Flickinger, Jeff Good and Ivan A. Sag (2004). Montage: Leveraging Advances in Grammar Engineering, Linguistic Ontologies, and Mark-up for the Documentation of Underdescribed Languages. *Proceedings of the Workshop on First Steps for the Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation,* LREC 2004, Lisbon

Bradley, K. (ed.) (2004). *Guidelines on the production and preservation of digital audio objects.* IASA Technical Committee.

Bird, Steven and Gary Simons (2003). Seven Dimensions of Portability for Language Documentation and Description, *Language* 79/3: 557-582.

Checkland, P. and J. Scholes (1990). *Soft Systems Methodology in Action*, Chichester, UK: John Wiley and Sons Ltd

Csato, Eva A. and David Nathan (2003). Multimedia and documentation of endangered languages, in Peter K Austin (ed.) *Language Documentation and Description, Vol 1*:73-84. London: SOAS.

Duncan, W. R. (1996). *A Guide to the Project Management Body of Knowledge*. Upper Darby, PA: Project Management Institute

Grinevald, Colette (2003). Speakers and documentation of endangered languages. In Peter K Austin (ed). *Language Documentation and Description Volume 1*:52-72. London: SOAS.

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

Holton, Gary (2003). Approaches to digitization and annotation: A survey of language documentation materials in the Alaska Native Language Center Archive. *Proceedings of EMELD 2003*

Johnson, Heidi (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.

Munro, Robert (2004). Digital skills and obligations: is language documentation a new ICT discipline? *Proceedings of the 2nd Endangered Language Academic Program (ELAP) Workshop: Multidisciplinary approaches to language documentation*. London: SOAS.

Nathan, David (2004). Sound Recording: Microphones. In *Language Archives Newsletter*, 1/3, 6-9 [http://www.mpi.nl/LAN/vol_01/lan_v01_n03.pdf] last checked 03/08/05.

Nathan, David and Peter K. Austin (2004). Reconceiving metadata: language documentation through thick and thin. In Peter K. Austin (ed.). *Language Documentation and Description Volume 2*. 140-153. London: SOAS.

Wittenburg, Peter and Dan Broeder (2002). Metadata Overview and the Semantic Web. *Proceedings of the LREC "Resources and Tools in Field Linguistics" Workshop*. Las Palmas.

Wittenburg, Peter (2003). The DOBES model of language documentation. In Peter K. Austin (ed.). *Language Documentation and Description Vol 1*:122-139. London: SOAS.

Wolcott, Henry F. (1995). *The Art of Fieldwork*. Oxford: AltaMira Press.

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

Woodbury, Anthony C. and Nora C. England (2004). Training speakers of indigenous languages of Latin America at a US university. In Peter K Austin (ed.). *Language Documentation and Description Volume 2*:122-139. London: SOAS.