

# PROFICIENT, PERMANENT, OR PERTINENT: AIMING FOR SUSTAINABILITY

**David Nathan**  
*ELAR, SOAS*

## Introduction

The concept of sustainability is a broad but timely way to refocus developments in language documentation as it is practised in the digital era. Sustainability can be considered as the sum of three factors: good data collection and management, robust preservation properties, and the relevance of materials. These three factors can vary independently and receive different weighting depending on the context. The inherently shareable nature of digital data, combined with the high cost of creating and storing it, means that audience and demand are significant criteria for pursuing sustainability in the wider sense. Within the area of endangered languages, the three major factors correspond to different activities (each crucial). Good data collection and management is essential for creating quality multipurpose records, for supporting analysis and cross-linguistic comparison, and the preparation of multimedia resources; data must be preservable in the long term due to its uniqueness and irreproducibility; and, finally, data must be mobilised to create usable resources that can directly help language support efforts. Notice that different techniques, tools, formats and methodologies might be applied in each case.

## Proficiency

In the many expositions of methodology for documentation, one of the commonly recurring themes is the selection of data standards, formats and tools. We have learnt from publications (Bird & Simons 2003; Gippert et al., 2006; Farrer & Langendoen 2003), workshops (DoBeS, HRELP), websites<sup>1</sup>, electronic lists<sup>2</sup> and blogs<sup>3</sup> about the value of data portability, and about the slowly growing number of standards, software, and techniques that help achieve it. Nevertheless it is nearly a decade since this theme arose in our discipline, and many would agree that, despite the amount of available advice, some of us find ourselves frustrated with low uptake among ordinary working linguists ('OWLs'), and the latter themselves are somewhat mystified by it.

Patchiness of uptake can be seen in various ways. Archives such as ELAR<sup>4</sup> get a view into linguists' understandings of data and metadata management through the shape of deposits. From ELAR's (admittedly limited) sample, some generalisations could be made: depositors do know that metadata is important; the interweaving of data and metadata within deposited resources remains confusing; and technologies such as XML markup have become poorly-applied talismans of 'good' practice in data and metadata management (see below). We see the lack of integration of relevant technologies into relevant linguistics courses, and relatively low depositor compliance with systems such as DoBeS' IMDI metadata domain (Klassmann, Offenga, Broeder & Skiba, 2006).

Of course there is also some progress, which could be attributed to several factors in addition to the sources of advice and expertise mentioned above: firstly, the preferences and skills of particular individuals; secondly, the Endangered Languages specialism, where compliance could be attributed to an environment currently enjoying relatively generous funding, with the resulting influence that funders can have on the practices of their grantees through training, regulations, and so forth. Thirdly, some progress has been fuelled by the priorities of areas such as typology, where researchers do wish to have large amounts of samples across languages structured and encoded in comparable form.

But it is the patchiness of uptake and skills that is of concern. One reason for it is that focus on standards, format and tools alone is not enough. Linguists need to know the rationale, capabilities, and strengths and weaknesses of them, so that they can make choices that are also informed by their own local goals and resources. Even more importantly, they need to know how to use them proficiently. Any of them, used poorly, will provide a worse result than some potentially inferior method used well.

Making data machine-readable is a fundamental goal of many recommendations for digital text data. It is important because machine readability provides the ability to exchange, process and restructure data. In turn, these processes underlie strategies for long-term preservation, and the use of tools to deliver usable products. Yet many linguists find it hard to come to grips with the idea of making data machine readable, because they are unaware that this involves scrupulous attention to the formal representation of data, and unclear on the benefits (which are often irrelevant to researchers, who are usually not directly involved in projects

to transform data). Many see machine readability as some kind of unavoidable by-product or constraint resulting from the use of particular software tools, rather than as an independent goal. Yet balance is needed: it is unfair to expect linguists to put in large amounts of work to make materials machine readable only for the benefit of those who have data-crunching agendas.

Mark-up—interspersing special labels or instructions in a stream of data—is one of the main ways of implementing machine readability.<sup>5</sup> Its purpose, however, is often misunderstood. We have found cases of use/abuse<sup>6</sup> such as:

- elements used, but the format of their content inconsistent or irrational (for example, dates expressed as `<date>21/9/04</date>` in one case, `<date>June 5</date>` elsewhere), or non-admissible characters included;
- XML provided that is generated by and relevant only to particular exporting software (for example, Filemaker Pro);
- generated XML that presents an inadequate representation of linguistic data structures (Shoebox, Toolbox);
- marked-up documents that include incompatible mixture of markup systems; for example, `\author <K Subranayam>` (which mixes SIL FOSF and XML);
- attempts to stuff enormous amounts of semantics into filenames, to the extent that system limitations are exceeded or information is lost;<sup>7</sup> spaces and non-Unicode characters used, missing extensions, and so forth.

Similar problems arise for audio recording and processing, where certain fundamentalisms have taken centre stage, such as minimum format parameters and admonitions against using compressed formats. At ELAR, we have done several tests together with fieldworkers showing that the quality of recordings is influenced by the choice and handling of microphones, and the management of recording environments and processes, far more than any other factor including compressed vs. uncompressed formats. Of course, all things being equal, the value of uncompressed signals and the disadvantages of compressed ones must be acknowledged; however, for the OWLs, the danger is that simple rules replace broader understanding of the chain of processes involved in creating good audio. Deeper, flexible understandings get subsumed to

compliance with 'archive formats' while linguists remain in the dark about the wider range of choices and how they can optimise the end results given the resources and constraints that they have to work with.

Fieldworkers' lives have been made easier by the availability of ever-smaller digital devices (recorders, data storage, and so forth). The fact that several of these devices can provide 'approved' formats has given rise to a growing fetish about (small) size and weight, these having now become primary desiderata for choosing equipment. Quality, reliability, flexibility and compatibility have become secondary; and choices among inescapably analogue devices such as microphones are seen as less important.

We might also re-examine where audio fits into the epistemology of our field. Dietrich Schüller observed that the general approach taken by linguistics to recording as 'data gathering' is some of the least scientific of any field, due to the lack of explicit goals, such as to capture signals that are valid representations of audio environments.<sup>8</sup> If we extrapolate this further (using Schüller's arguments for recording in stereo or binaural/ORTF because that is how humans hear), and agree that the quality datum for linguistic audio recordings is evaluation by a human listener (as opposed to, say, a bird, or some computer signal processor), then we could potentially *increase* quality overall by encouraging linguists to use critical pairs of human ears as the ultimate measure of quality in all phases of audio work.

It is understandable that the ascent of the digital age—and the benefits it provides especially to those of us in information industries—should make us pay more attention to various quantified parameters. However, so far, much of the advice and discussion available to fieldworkers is overly concerned with choices amongst such parameters. We hear, for example, about acceptable audio resolutions but much less about techniques for making excellent recordings; nor do we see discussion about issues such as rates of usage of particular materials in relation to their storage costs. Such measures ought to be relevant in environments that make limited-resource social applications dependent upon digital preservation, for which curation and storage require significant and continuous funding. The preservation of physical objects has never seen the extremes in resource demand that we see today. Compare text and video: a one-hour video can occupy literally billions of times the space occupied by a typical text; for each single book, the video corresponds to covering the whole surface area of England with books. We need some measures to decide what is worth

preserving, whether in terms of proficiency, quality, uniqueness, demand, or other significance.

Ultimately, proficiency involves art as well as science, even for the simplest technologies. Consider the case of photography. Technically, we know that uncompressed digital images make better archive objects (precisely because uncompressed formats best support future usages in publication or production, that is, underlining their potential pertinence); nevertheless, most of our archives do accept heavily compressed JPEG images as a concession to the fact that these are the native or default output of most digital cameras.<sup>9</sup> Good images require correct shutter speed, aperture, focus, and so forth (even if automatically selected), but an image that is interesting, beautiful or useful will start from aesthetic appreciation of the subject, composition and the fall of light. We have recently seen several images that are neither technically nor aesthetically good, although improvement of the latter alone would have produced images of some value. And yet, photography is much less demanding than audio and video, and all of these are increasingly part of the documenter's toolkit.

While recent criticisms<sup>10</sup> of the use of computers in preparation of linguistic data are self-contradicting, it would be wise for us to consider our proficiencies at individual and collective level, not, of course, because there is any sanity in avoiding the use of computers, but because we can use them better. I suggest these maxims: (i) don't use new technologies to do badly what we already know how to do well; and (ii) information technology should be an *amplifier* for the efforts of those of us who work in the knowledge industries, not an end in itself, and certainly not used to amplify our shortcomings!

## Permanence

Formats and encoding schemes change over time. One important approach to dealing with this, based on 'preserving the byte stream',<sup>11</sup> assumes that future researchers will be sufficiently motivated to decode the underlying character sequences in digital data whose software environment has long disappeared. In its simplest form, this approach is not compatible with the maintenance of resources' pertinence, because the data is typically not easily accessible.<sup>12</sup>

Nevertheless, permanence (or preservation) has plenty to do with the physical substrate—the *carrier*—on which data is inscribed. In isolation,

magnetic digital disks are one of the most vulnerable of all means of storage. Their strengths for preservation come from their ability to provide rapid access to data from which unlimited verified-perfect copies can be made and then transmitted at low cost.<sup>13</sup> Their strengths, then, derive from properties of the here-and-now; and depend on the level of demand for them, and what resources (such as funding and political support) are available to support that demand. New forms of data storage may come along, but in the meantime, at least, attempts to preserve data for the long term should look at overall ecologies of preservation, rather than particular carriers such as nickel medallions<sup>14</sup> that may be almost physically permanent but have none of the here-and-now strengths of magnetic digital data.

## Pertinence

Providing materials that meet real demand, or materials that can generate interest or demand, is also an indispensable part of a strategy for sustainability. A bit of pertinence goes a long way; for example, consider the fact that several 'Shakespeare' texts have survived despite having lost some of their most crucial metadata (that is, the author's name, evidenced by ongoing controversies about authorship). In the area of endangered languages, there is surely an ethical principle that materials elicited from communities must be made quickly available and be capable of supporting language strengthening (Grinevald, 2003). Here, then, pertinence means 'relevant for the language community's aims and efforts for their language'; and it could be extended to apply not only to communities but also to other agencies such as educational authorities and language planners.

There are also pragmatic reasons for ensuring that our data is pertinent. Data collection and preparation is time consuming and electronic preservation is expensive, so funding needs to be attracted. Funders want to know who are the audiences for our materials, whether researchers, language communities, educationalists, or the general public. In addition, linguists may also gain from seeking new audiences.

Being relevant might include creating products such as websites and multimedia. Here, practical decisions and compromises often need to be made, for example about how to handle characters and fonts. Some concrete publishing or pedagogical projects simply cannot be fulfilled while observing 'best practices' such as use of Unicode (Csató and Nathan, forthcoming), but such products are part of the landscape of sustainability, because they will bring new uses and attract new audiences to our data.

Some language archives are attempting to promote greater levels of interaction with depositors and other users, by providing new facilities, such as enabling users 'to add layers of interpretation, annotations and commentary.'<sup>15</sup> A number of other things could be done to encourage linguists and others to be stakeholders in the management and preservation of digital language data. Currently, efforts are under way to convince institutions to give academic credential to data corpora and deposits. Once this has been achieved, data will enter the peer-evaluation cycle. Better recognition of such data should lead to more usage of it and in turn an increase in their willingness to prepare and deposit materials. The latter has to be a huge priority: Schüller has estimated that 80% of all ethno-linguistic recordings are sitting, degrading and undiscoverable, on researchers' private shelves (Schüller, 2004).

More linguistic data needs to reach language communities so that they too become stakeholders. Most linguistic data never reaches communities, although ethical frameworks and funding guidelines are changing practices. Few materials are made with community audiences in mind; for example, ELDP, currently the largest funder of endangered languages documentation, only allows a small component of a grant to be allocated to publishing community-oriented materials.<sup>16</sup>

Pertinence should not just be a hook or a selling point. A leading linguist advisor to the NSF stated on public radio that the purpose of NSF documentation was to create 'fun and motivating learning resources', despite the theoretical and typological emphasis that is presented in linguistics circles.<sup>17</sup> Some projects have started with a rationale about endangered languages but then did not specifically address language endangerment. Since we do not know how long the current funding situation for endangered languages will remain generally positive, and the task is so huge, we should discourage the 'pertinence factor' of language endangerment being used in ways that do not benefit languages and communities.

## Presentation genres and linguistic data

The traditions of our field do not provide the genres required to make all our materials pertinent to all their potential audiences; new genres of expression are needed. In documentary linguistics we are constantly reminded that data is not only written sentence examples, lexical information, and so forth, but also the source events, such as recordings. In addition, this 'real data' should be made accessible to create a truly

scientific discipline (Bird & Liberman, 1999; Thieberger, 2004). Newly developed software such as ELAN<sup>18</sup> and Audiamus<sup>19</sup> attest to the need for new genres, but their narrow targets (text and audio/video alignment) leave much territory uncovered, especially interfaces for presentation and exploration (Nathan, 2006).

Current representational technologies such as XML started out as systems for describing static texts, where content and presentation enfold each other. Although their goal is to separate content from presentation, and they are increasingly multimedia-capable, these technologies continue to be associated with software interfaces that are bound by a very small number of metaphors (Cooper, 1995). A preliminary survey, for example, of interface objects used to control language audio yields the following rather unadventurous list:

- button/icon
- timeline
- player controls (and advanced/alternative player controls such as jog)
- cartoons/speech balloons
- text/hyperlink/page

This is important for two reasons. Firstly, despite the emphasis of documentary linguistics on real language performances, we find no genuinely new genres for presenting such data—this surely limits the potential for attracting users to the data and mobilising it in support of languages. Secondly, if new interfaces (such as, for example, speech balloons that could be manipulated/stretched to hear various parts of utterances/dialogues), become available, they could influence how recordings are made and transcribed, and how data is structured.

## Conclusion

This article has discussed a number of ways in which components of sustainability—proficiency, permanence, and pertinence—can complement the important principles of data portability (Bird & Simons, 2003) for the area of endangered languages. It has also highlighted some current problems and frustrations facing linguists, fieldworkers, and technical people who work with them. Addressing these problems means first recognising them. They might be dealt with in various ways including training and dissemination of advice about good recording, data modelling and representation, and pedagogical materials development; case studies in

project development, especially in regard to multidisciplinary teams and their workflows; increasing linguists' and communities' incentives for participation in digital data preparation and use; and, perhaps above all, integration of relevant skills into linguistics and fieldwork courses.

Resource sustainability is a result of proficiency of its preparation, its permanence, or its pertinence to a range of users. It does not require all of them. A stronger showing of one may compensate for weakness in another; for example, high demand may provide the impetus to add value to a poorly structured document. However, generally one would look for strength in all three areas: a very poor-quality recording is unlikely to ever attract high demand, no matter how compliant its format. In some cases, factors are opposed, such as when practical choices made in developing a publication militate against their long-term preservation. Despite compromises of standards or best practice, resources can only benefit from being better known. Some compromises may not be necessary once technologies such as Unicode or SMIL have fully matured.

There may be parallels between resource sustainability and language survival. Sustainability favours software that is 'open source' because it is open to participation and evolution; the same is true of human languages—the first and largest open source project ever undertaken. And perhaps on the other hand there is a parallel between a too-narrow focus on data standards, and exhortations to language purity, where both limit the scope for full participation. Sustainability provides a framework for clarifying and disentangling such issues in managing language resources. It takes us one step further than 'data' because it admits consideration of practical issues, and it connects meaningfully with the states of languages.

## Endnotes

<sup>1</sup> For example, E-MELD: <http://linguistlist.org/emeld> (E-MELD, n.d.).

<sup>2</sup> For example, Resource Network for Linguistic Diversity: <http://www.linguistics.unimelb.edu.au/thieberger/RNLD.html> (RNLD, n.d.).

<sup>3</sup> For example, Transient Languages and Cultures blog: <http://blogs.usyd.edu.au/elac> (University of Sydney, 2006).

<sup>4</sup> Hans Rausing Endangered Languages Project Archive (ELAR): <http://www.hrelp.org/archive> (HRELP, 2006).

<sup>5</sup> The other main method is relational databases.

<sup>6</sup> Some of these are examples of 'markup voodoo', or irrational expectations about what markup can achieve (the term was coined by Manfred Thaller).

<sup>7</sup> For example, information lost when names are truncated. Typing mistakes are also much more likely in long names. The practice of overloading filenames has been associated with the view that data is more robustly preservable if the filename identifies its content 'when all else fails,' but, this belief, like markup, has achieved talisman status for some researchers.

<sup>8</sup> ELAR Workshop: 'Audio Recording, Digitisation and Archiving,' conducted by Dietrich Schüller, Phonogrammarchiv, Austrian Academy of Sciences. Held at SOAS, February 13, 2006.

<sup>9</sup> Actually, the situation is getting worse with more use of video in fieldwork: we are receiving increasing numbers of images that are stills taken by researchers using video cameras as a "one unit does all" device.

<sup>10</sup> Message posted by Alexandra Aikhenvald on behalf of Robert Dixon (March 30, 2006), on Discussion List for ALT <LINGTYP@listserv.linguistlist.org>. Elsewhere, Aikhenvald even states, 'the current focus on computer-based "documentation" is akin to racism' (<http://www.latrobe.edu.au/rclt/StaffPages/aikhenvald%20downloads/Contents%20of%20issue.doc>).

<sup>11</sup> For example, the Cedars Project; see <http://www.leeds.ac.uk/cedars> (*Cedars*, n.d.).

<sup>12</sup> The Cedars project addresses this by emphasising the role of metadata in two ways: to outline the nature of the resource content, and to describe, or even reproduce, the relevant software environment.

<sup>13</sup> With the decreasing cost of hard disks, it has recently become feasible to implement backup strategies using disks rather than tapes, exploiting their abilities to quickly make continually verifiable backups.

<sup>14</sup> See the Rosetta Project's 'technology' page at <http://www.rosettaproject.org/about-us/disk/technology> (Rosetta Project, n.d.).

<sup>15</sup> See the pamphlet of the DAM-LR partners, *Live Archives: a checklist of principles and tasks* (DAM-LR, 2006).

<sup>16</sup> <http://www.hrelp.org/grants> (HRELP, 2006).

<sup>17</sup> <http://www.npr.org/templates/story/story.php?storyId=5557885>. NPR Talk of the Nation, 'Preserving Endangered Languages,' July 14 2006.

<sup>18</sup> See ELAN website: <http://www.mpi.nl/tools/elan.html> (Max Planck Institute for Psycholinguistics, 2006).

<sup>19</sup> See website: <http://www.linguistics.unimelb.edu.au/thieberger/audiamus.htm> (Thieberger, 2006).

## References

- Bird, S. & Liberman, M. (1999). *A formal framework for linguistic annotation. Technical Report MS-CIS-99-01*. Department of Computer and Information Science, University of Pennsylvania.
- Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language* 79, 557-582.
- Cooper, A. (1995). *About face: The essentials of user interface design*. Foster City, California: IDG.

- Csató, É. Á. & Nathan, D. (forthcoming). Multiliteracy, past and present, in the Karaim communities. In P. Austin (Ed.), *Language Documentation and Description, Vol 4*. London: Hans Rausing Endangered Languages Project, SOAS.
- Curl Exemplars in Digital Archives (Cedars)*. (n.d.). Retrieved October 27, 2006, from <http://www.leeds.ac.uk/cedars>
- Distributed Access Management for Language Resources (DAM-LR). (2006). *Live Archives: a checklist of principles and tasks*. Retrieved October 27, 2006, from [http://www.mpi.nl/DAM-LR/flyers/DLRA\\_Flyer\\_2006-04-23.pdf](http://www.mpi.nl/DAM-LR/flyers/DLRA_Flyer_2006-04-23.pdf)
- Electronic Metastructures for Endangered Languages Data (E-MELD)*. (n.d.). Retrieved October 25, 2006, from <http://emeld.org>
- Farrer, S., & Langendoen, T. (2003). A linguistic ontology for the semantic web. *Glott International* 7 (3), 1-4.
- Gippert, J., Himmelmann, N. & Mosel, U. (Eds.). (2006). *Essentials of language documentation*. Berlin: Mouton de Gruyter.
- Grinevald, C. (2003). Speakers and documentation of endangered languages. In: Peter Austin (Ed.) *Language Documentation and Description, Vol 1* (52-71). London: Hans Rausing Endangered Languages Project, SOAS.
- Hans Rausing Endangered Languages Project (HRELP)*. (2006). Retrieved October 25, 2006, from <http://www.hrelp.org>
- Klassmann, A., Offenga, F., Broeder, D., & Skiba, R. (2006). IMDI Metadata field usage at MPI. *Language Archives News* 8, 6. Retrieved October 30, 2006, from [http://www.mpi.nl/LAN/issues/lan\\_08.pdf](http://www.mpi.nl/LAN/issues/lan_08.pdf)
- Langendoen, T. (2006). *Preserving endangered languages* [podcast]. NPR Talk of the Nation, July 14 2006. Retrieved October 27, 2006, from <http://www.npr.org/templates/story/story.php?storyId=5557885>
- Max Planck Institute for Psycholinguistics. (2006). *Tools: ELAN*. Retrieved October 27, 2006, from <http://www.mpi.nl/tools/elan.html>
- Nathan, D. 2006. Thick interfaces: mobilising language documentation. In J. Gippert, N. Himmelmann & U. Mosel (Eds.), *Essentials of language documentation* (363-79). Berlin: Mouton de Gruyter.
- Resource Network for Linguistic Diversity (RNLD)*. (n.d.). Retrieved October 27, 2006, from <http://www.linguistics.unimelb.edu.au/thieberger/RNLD.html>
- Rosetta Project. (n.d.). *The Rosetta project: Building an archive of all documented human languages*. Retrieved October 27, 2006, from <http://www.rosettaproject.org>
- Schüller, D. 2004. *Audiovisual Archiving: Visions, Challenges, Strategies*. Presentation at DELAMAN workshop, MPI Nijmegen, November 2004.
- Thieberger, Nicholas. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In Peter Austin (ed). *Language documentation and description, Vol 2* (169-178). London: Hans Rausing Endangered Languages Project, SOAS.

University of Sydney. (2006). *Transient Languages and Cultures blog*. Retrieved October 27, 2006, from <http://blogs.usyd.edu.au/elac>