Digital Archiving

David Nathan, SOAS

1 Introduction

This chapter is about digital archives and digital archiving of language materials. The term ‘digital archive’ is used here to refer to a *facility* that has been established with the primary goal of preserving digital data. In this sense, ‘digital archive’ does not refer to backup, original, compressed files or files that have been set aside and not subject to further change. From the archives’ point of view, activities include appraising and giving feedback on submitted materials, and then, for those materials which are accepted, their curation, preservation and dissemination, all of which involve processes and equipment unique to the digital domain. Some archives are involved in supporting activities such as training and software development.

From the language documenter’s point of view, digital archiving is a diverse set of activities including creating, selecting, preparing and documenting materials for deposit with the digital archive. Many of these activities should be understood as aspects of data management, rather than required only for archiving.

2 Digital data

Strictly speaking, digital data is something that “happens” rather than actually “exists”. Digital information is stored as physical (i.e. analogue) changes on carriers, such as tiny holes in plastic disks or changes in magnetic fields on metal disks. A computer can read the disks, interpret them as sequences of symbols, and then present that data in a form that is comprehensible to an agent that understands the symbols (for example, some software, or a human).

Many forms of data are not digital, such as an audio recording on a cassette. One could digitise its audio information by playing the cassette and turning its analogue audio signal into sequences of symbolic values. Equally, digital information does not have to be stored on computers. The digitised audio information could be printed on paper (or carved into stone) as zeros and ones, or as barcodes, which would be preferable from a purely preservation point of view, since magnetic and optical storage is quite fragile. Traditional written and printed content on paper can thus be regarded as in a sense ‘digital’. In the case of our cassette example, however, each minute of digitised audio would occupy over 10,000 paper pages!

Of course, ‘digital data’ usually refers to computer-readable files, not symbols written on paper (or stone), and the term will be used henceforth in that conventional way. Unlike physical information-bearing objects like books, digital data is inherently separated from its means of storage. Digital data cannot be directly experienced by human senses - it needs hardware, software and interfaces to render it accessible via screens, headphones, or touch. This separation - the abstraction of content from its physical form - is the fundamental property of digitisation, and it is what enables the copying, transmission, modification, linking, networking, searching and combining of data. But it also introduces severe obstacles to preservation. Objects such as paper - or even gradually degrading tapes and film reels - do not suddenly fail in the way that

every computer system will given sufficient time. It is hard to envisage preserving our digital data as long as the Rosetta Stone (over 2,000 years) or some Australian Aboriginal art (over 40,000 years) have endured, although that is the goal of digital archives.

3 The digital dividend

In the past, many people who made field recordings did not archive them, in the sense of depositing them in a digital archive facility. This was partly because they were focussed on other things, such as writing up grammars and other linguistic descriptions. It has been estimated[1] that 90% of the world's recorded cultural heritage materials, many of them unique and irreplaceable, lie stranded on researchers' shelves, unknown to their originating communities and to the wider world, and irretrievably decaying.

Recently, this unhappy situation is improving. Following general recognition of the consequences of language endangerment and loss, the discipline of documentary linguistics has emerged, together with specialist archives to support it. Both the discipline and its archives rely extensively on digital technologies, which are now central elements of every phase of language documentation, research, preservation and dissemination. Audio and video recording, data management, and many other activities including transcription and lexicography, are all performed using computers and other electronic devices. With the exception of hand written field notes, most researchers write nowadays using computers, thereby creating digital files. They do this because they welcome the ease with which computers allow the revision, searching, copying, sending and printing of those files compared to paper-based materials. More complex processes, such as restructuring or modification of information in files, or combining contents of different files, are now possible using databases, spreadsheets, or other computer programs. While it was always possible to manipulate data manually, computers enable processes such as reorganising or sending large amounts of data that would have previously been so time-consuming that they were rarely pursued. Software enables flexible integration of text with media, formerly only possible in highly specialised areas such as film production.[2] Thus, being digital makes data fluid and adaptable, so that resources such as audio, texts, lexica etc. can, often without too much work, be quickly repurposed for important and urgent tasks in language documentation. By using digital resources we can exploit rapidly growing Internet-based communications not only to make data available but also to collaborate with distant others in developing materials to support languages.

Despite all these capabilities representing a thorough transformation from the data management methodologies of 20 years ago, there is only one reason why it is *necessary* to use digital technology to archive endangered language documentation; long term preservation of audio and video is only possible if they are held in digital form. Until recently, audio and video were captured by recorders/cameras which turn their energy into electronic signals[3] and then use those signals to physically shape the properties of some carrier medium; for example, the magnetic patterns on a cassette tape. There is an unbroken causal and physical chain between the original energy

[1] By Dietrich Schüller of the Vienna Phonogrammarchiv.
[2] Although see Woodbury, this vol., on the (hard-copy) publishing of texts time-aligned to cassette timestamps in the Boasian tradition.
[3] Electronic signals as varying levels of energy, not as digital data.

source and the media carrier, which is therefore an ‘analogue’ of the original event. Inevitably, recording and playback processes are mediated by the nature, quality and performance of the actual objects involved (the recorder and the tapes), so that no rendition can ever be said to be “perfect”.

Since analogue media carriers cannot physically last forever (or even stay exactly the same from one usage to another), preservation of the content requires it to be copied from one carrier to another, making long term preservation in principle impossible:

> In the analogue domain, the primary information suffers an increase in degradation each time it is copied. Only the digital domain offers the possibility of lossless copying when refreshing or migrating recordings … For the long-term preservation of the primary information contained on an analogue carrier it is necessary, therefore, to first transfer it to the digital domain.
> [IASA 2005:5]

Without digital technologies, it would be possible neither to rescue the legacy materials already recorded, nor to preserve those that are presently being recorded. We are fortunate that the availability of digital technologies coincided with the growth of interest in language documentation and archiving. The long term preservation of audio and video is *impossible* without digital technologies; only through them will future generations be able to hear the sounds of endangered languages that are spoken today.

4 Encoding digital data

As we saw above, the essence of digital processing is the storage and processing of symbols. Computers thus provide a natural and efficient means of working with orthographic text. In fact, they could be said to be machines that “re-nativise” text as a medium of popular communication following a century of the dominance of sound and image through the analogue technologies of radio, cinema, and television (cf Levinson 1999, chapter 4). Text data is so compact that storage costs are negligible; it is easily transmissible, searchable and able to be copied and manipulated. Texts of almost any kind provide few challenges to the digital archive – provided that its symbols are properly encoded.

In digital form, a text is stored as a linear sequence of binary symbols (usually thought of as 0 and 1). There are several layers of encoding that take place between this stored sequence of drab 0s and 1s (‘bits’) and the varied orthographic and typographic information found on the typical screen or printed page. Looking from the screen display inwards, what we see are, firstly, glyphs, which are character images drawn from fonts; each glyph is specified by a number at the software level; those numbers are drawn from a “character set” that lists correspondences between a set of character concepts and a range of numbers, and in turn those latter numbers are packaged together into sequences of symbols that we call, appropriately enough, ‘files’. Figure xx shows examples: in the first row, the character concept ‘Latin capital A’ is allocated to number 65 in the ASCII character set; the next two rows show how the same underlying “data” is to be understood as different characters depending on how it is packaged by association with a character set; the final row shows that the concept ‘Latin small schwa’ is allocated to number 601 in Unicode. For more on character encoding, see Gippert 2006, Wood nd, and Korpela nd.

binary digits (in file)	*decimal equivalent*	*character set*	*character concept*		*glyph, in font Arial Unicode*
01000001	65	ASCII	Latin capital A	=	A
11111110	254	ISO 8859-1 or Latin 1	Lowercase “thorn”	=	Þ
11111110	254	ISO 8859-9 or Latin 5 (Turkish)	Lowercase “s” with cedilla	=	ş
0000001001011001	601	Unicode	Latin small schwa	=	ə

Figure xx. Binary symbols in files are mapped onto orthographic characters through standardised character sets or encodings.

But not all text files are created equal. So-called ‘plain text’ files work as just described, and a computer only has to know how the basic sequence is chunked into units (e.g. into groups of eight or sixteen bits[4]) and how to turn each of those units into numbers and consequently characters (which will, as the examples in Figure xx show, only be guaranteed to be as intended if the character encoding has been explicitly specified).

There are other kinds of files that are packaged differently, into formats typically called ‘proprietary’, usually because they are used by commercial, or proprietary, software. These formats allow for more complex types of information than just sequences of characters - for example formatting in varying sizes and colours, spacing, tables, other layout options, and even images – none of which can be represented as a sequence of characters that corresponds to the content. And in turn, specialised software is needed to create and view such files.

An example of a proprietary format is Microsoft Word. We tend to think little of the complexity it adds because the world of print is so familiar, but such ‘proprietary formats’ (and especially those used by commercial software) do present several challenges for digital archiving. Firstly, they encourage documenters to rely on typographic conventions instead of writing down knowledge explicitly in its own terms. When knowledge is transparently and explicitly provided independent of format, layout and need for specific software, we have the best chance of ensuring that the content is accessible long after today’s software (and its manufacturers) are forgotten.

Secondly, the archive, and any user of the materials (as well as the archive managing the materials), may be required to use the same software that created the files in order to view them, which limits accessibility of the content. Finally, the software manufacturer may change its formats over time (i.e. change the way it packages and renders the content), so that in order to retain access to the data content, an archive either needs to preserve and make available the relevant software versions (which may not be feasible due to expense or copyright), or to be aware of formats that are becoming defunct and migrate all content to another format while still possible. All of these complexities create a resource burden for an archive and jeopardise long term preservation, so archives strongly prefer to receive “plain text” in which any additional structural or formatting information is encoded in standard, explicit and open formats such as XML.

[4] A group of 8 bits is called a byte.

Most archives will request metadata to accompany deposited materials. Chapter xx discusses metadata in detail, describing how researchers' contextual knowledge and the assumptions and conventions they use in writing up data should be included together with the data. Since the role of metadata is to facilitate the preservation, understanding, administration, and appropriate usages of data (Nathan 2006), it is even more crucial that metadata is provided in transparent formats that do not rely on specific software.

4.1 Non-text materials

The archival value of images is frequently underestimated. Photographs of fieldwork settings - consultants, objects, environment, events, and equipment setups - can all be very useful both for contextualising linguistic data and in their own right as documentations of the language community's life. Images are easy to store and use; for many purposes, a few photographs could be equally, if not more effective than video, while consuming far few resources. Other sources of images include field notes (especially useful if they contain drawings, diagrams, or examples of consultants' handwriting) or written materials found in the community. Today's digital cameras, used under good lighting, have sufficient resolution to make good quality images if scanning is not possible. All images should be accompanied by captions and descriptions, and linked to the relevant texts and recordings.

Turning to time-based media, we have seen that digitisation provides the only route to the future for audio and video. The format options for audio are now quite stable. Audio should be provided to archives in the form of WAV files (also known as linear PCM[5]) which involve no compression.[6] The trend in language documentation is towards capturing the full spatial 'image' of speakers' voices in their real-world acoustic contexts, so stereo is preferred. Currently, the most common parameters used in these files are a sampling rate of 44.1 KHz and a bit depth of 16 bits. Some archives are starting to recommend parameters of 48 KHz and 24 bits, and these are expected to become standards for audio over the next decade. Note that these figures apply to digitally originals - i.e. "born digital" recordings. When analogue materials such as tapes are digitised then higher resolutions (sampling rates and bit depth) should be used to capture the "undesirable artefacts" arising from the carrier due to its physical manufacture, storage or handling. Accurately capturing these "artefacts" increases the likelihood of being able to use software to successfully identify and remove them if restoration is attempted in the future (IASA 2005:6).[7] However, the current standard of 44.1 KHz/16 bits is sufficient to represent the full acoustic detail of human speech, and all computers and software support it, so it is likely to remain a practical and acceptable choice for some time.[8] For further details, background and recommendations regarding sampling rate and bit depth parameters, see IASA 2005:6.

There are, of course, other audio formats such as the ubiquitous MP3. MP3 files are compressed but listenable versions which are useful as dissemination copies, or for playing back in portable players, but should never be used for primary recording since

[5] A variant of WAV that contains preservation-oriented and other metadata embedded within the file is called BWF.
[6] Strictly speaking, digitisation involves initial sampling of an audio signal which could be regarded as a kind of compression; however, providing the sampling rate and accuracy are high enough, the full range of acoustic information that humans can hear is retained.
[7] For example, 96 KHz, 24 bit. See IASA TC-03 (2005), p 6.
[8] In any case, conversions between resolutions are relatively straightforward.

there is no reason to strip out various frequencies from the original acoustic data in order to make the file size smaller. The main archiving requirement, however, is that audio should be delivered to the archive in its original form, with appropriate metadata; it should not be covertly converted to a different format. If, for example, audio is recorded originally as MP3, but then converted to WAV (perhaps in an ill-fated attempt to keep an archivist happy), the actual audio information remains compressed; what was originally lost cannot be restored by the format conversion. The archive receives no record of the initial compression, which may cause problems for preservation and for future attempts to create compressed listening copies.

The situation for video is totally different from that of audio; the formats and parameters are far from stable. As of 2010, the technology is rampant with format variants from different manufacturers, and archives are forced to store highly compressed versions for purely practical reasons of size and the cost of storage. Discussion of video formats is beyond the scope of this chapter, and they are in any case undergoing rapid change at the time of writing - but see Section xx for further reflections on archiving digital video.

Digital audio and video must be accompanied by text-based metadata (and transcription, annotation or other associated text information) that can be listed, sorted, and searched so that users can identify media content. Without such text data, those searching for information are forced to play media files right through to get an indication of their content; the media resource is effectively hidden, unfindable and unusable, forever. The richness of the information that accompanies media files should be proportionate to their documentary value and the high costs of storing large media resources.

5 Archive strategies

Archives have traditionally made decisions about which materials they accept for deposit, based on their collection policy. An archive might be devoted to preserving materials for a particular community, group, or region, or its policies might be orientated to particular genres of materials. When an archive is established to hold digital materials, its procedures, equipment, and management of the deposits will be tailored to the specific needs of the digital domain, including appraisal in order to select those materials that have both sufficient value and are feasible to ingest and preserve. (see Conathan Chapter re Appraisal XX).

New partnerships between granting bodies and archives hold great promise for the growth of digital data management in the documentation field and for the strength of resultant archived collections. A small number of archives (e.g. DoBeS, ELAR) are now affiliated with organisations that fund documentation of endangered languages. These archives are tasked with preserving the outcomes of funded projects. Their respective granting bodies (DoBeS and ELDP[9]) want to ensure that the outcomes of their funded research are securely and visibly preserved. In addition, these funder/archive partnerships provide training and technical support of various kinds throughout the lifespan of documentation projects, so that there is potentially greater interaction and co-operation between researchers and archives than is generally found. While conventional archives typically receive materials “in the absence of creators

[9] ELDP disburses funds provided by Arcadia.

and collectors" (Conathan, Appraisal para 2 xx), these new partnerships allow documenters and archives to inform each other.

These partnerships also give new roles to archives. To the extent that archives inform their granting agency's policies and procedures, they can influence the nature and quality of their collections by, for example, specifying the skills, methods, processes and equipment that should be evident for an application to be successful. On the other hand, the archive's responsibility to grantees may result in having to deal with problems that an independent archive would not face. For example, ELDP applications in some years saw an escalation in the applicants' intended numbers of hours of audio and video recordings, presumably in order to make their applications look more attractive to the funder. But should these intentions come to fruition, the archive's planned capacity for curation and storage will be stretched or exceeded.

6 Standards and diversity

Standards are important for the effective operation of digital archives. Standards are promoted in pursuit of three goals: quality, interoperability, and the integrity of the archive's collections.

Some standards provide benchmarks for the quality of resources according to the expectations of a given field or for a particular task. These may be quantitative, such as the requirement for audio to have an adequate sampling rate (at least 44.1 KHz). Or they may be categorical, for example the mandated use of Unicode characters (but mainly because it increases interoperability; see below).

But most quality issues are qualitative and context dependent, such as the accuracy and listenability of an audio recording, the clarity and explicitness of the representation of data, or the accuracy of a transcription. It is these qualitative questions that have been patchily addressed in the theory and practice of language documentation. It remains unclear to what extent they are desiderata to be addressed through linguistic curricula or other training, or whether they should be addressed by archives. The limited attention paid to them in linguistic curricula has led to digital archives frequently being identified as the sources of standards, in turn leading to excessive focus on technical parameters, at the expense of qualitative evaluation.[10]

Metadata schemes such as OLAC (xx ref Good) use standard and conventional sets of categories, which archives use to populate their catalogues and to serve as 'finding aids' to make resources discoverable (Bird and Simons 2003 xx). In this way, metadata functions in the same way as catalogue records for books in a library, which have categories including author, title, date, ISBN, and publisher that all users expect to find. In addition to these standard categories, specialist libraries create additional metadata to serve the particular needs of their clients.

Despite the widespread use of compact library-like metadata schemas such as OLAC, the set of categories required for capturing the context and significance of documentation materials is not yet delineated. Because language documentation is an emerging field, in contrast to the maturity of libraries and publishing, it is peremptory to constrain documenters to particular schemes. ELAR encourages documenters to

[10] Elsewhere I have called this "archivism" (Dobrin et al 2008 xx).

design metadata to reflect their own research environments and needs. Following four years of operation, the metadata received by ELAR shows that categories vary according to the particularities of each project's goals, participants' skills and preferences, and the nature of communities, cultures and settings:

- each documentation project can have its own unique "recipe" for metadata, depending on factors such as the language's typology, consultant knowledge, and community values
- each language documenter has his/her own skills and priorities that determine what metadata categories they use and how they encode them
- ELAR's goal of maximising quality and quantity of metadata for each deposit requires the encouragement of diversity.[11]

It is thus necessary to distinguish between metadata schemes that are used across the board by an archive or group of archives, and the broader and varied sets of metadata that assiduous documenters provide.

Returning to the analogy with libraries, archivists and depositors function as 'joint librarians' in the digital archiving of endangered languages materials. In fact, depositors play the major role, because they are the ones who know the details of the fieldwork situation and the data. The depositor, not the archivist, has access to the language content, consultants and the language community in order to provide metadata such as speaker details, access conditions, ethnographic context, and captions for photographs.

7 Digital archives and their services

The policies and technologies of today's digital language archives have their origins in the earliest digital libraries (Arms 2001 xx). Later, an architecture developed by the Open Archives Initiative (OAIS 2002) highlighted the importance of identifying an archive's intended user groups (its "designated communities") in order to provide them with versions and formats appropriate to their needs. More recently, several archives have been established which are specifically dedicated to endangered languages materials, including AILLA, DoBeS, ELAR, LACITO, Paradisec and others (see Appendix xx). The associated initiatives OLAC and EMELD have vigorously promoted within the linguistic community the importance of creating digital data that is technically robust and flexibly re-usable, accompanied by metadata that, suitably catalogued, enables users to discover and access materials (Bird and Simons 2003:563).

There are alternative providers of digital preservation. There may be national, sector-based, or institutional facilities in your country that can offer preservation. In the UK, the trajectory of these has been uncertain, as funding is influenced by economic circumstances and the perceived value of competing disciplinary areas. Then there is the possibility of managed outsourcing. Companies such as Amazon provide mass data storage (through its "Simple Storage Service", which allows for the customisation of access), but current sentiments would rule against trusting commercial companies with collections of irreplaceable and culturally sensitive data.

[11] Of course this also imposes costs. Additional work is required to integrate eclectic sets of metadata into a catalogue.

Note, though, that we are dependent on commercial businesses for the supply of storage appliances, networks, and communication services, and that we have gained from competition between them. In the future, it might be feasible to outsource data storage to companies with domain specialisations. These companies could provide appropriate levels of service, commitment and trust, in the same way that, for example, we are generally satisfied to store our shared documents with Google and our money with banks.

Given the scale of language endangerment, within a few years digital language archives are likely to become the repositories of much of the world's linguistic and cultural heritage, and the major sources for research on and the revival of moribund or extinct languages. It is therefore important for archives to disseminate materials, functioning as specialist electronic libraries that are equipped to deal with the new genres of documentation that are marked by an emphasis on media, sensitivities and restrictions on access, and few alternatives channels for publication.

8 Access

The large investment in the creation, management and preservation of digital resources demands appropriate resulting benefits. Access and distribution flow naturally from the existing digital infrastructure, since data can be copied cheaply and perfectly, and quickly transmitted to most parts of the world. However, access has to be managed, and digital archives have to steer a narrow path between reasons for data to be freely accessible, on the one hand, and on the other hand, to be protected or closed. There are several constituencies on the "open" side, starting with the source language communities and those who wish to assist them in language maintenance and revitalisation activities – such groups should not be prevented from accessing data that they morally "own" or which can facilitate their efforts. The second is the scientific community which champions openness and the neutrality of data. Thirdly, it is frequently argued that the public should have access to the outcomes of publicly-funded research.

Despite all these compelling arguments, factors at the core of language endangerment argue against across-the-board free access to data. Firstly, there is the nature of the data itself. Since language documentation consists ideally of recordings of spontaneous communication in everyday social contexts, such recordings can be expected to contain instances of private, embarrassing, secret, sacred, or other restricted content that may cause harm to the speakers or others.

At the Endangered Languages Archive (ELAR), we use the term 'protocol' as shorthand for the concepts and processes that apply to the formulation and implementation of language speakers' rights and sensitivities. Corpus linguistics has long taken note of protocol; for example, recorded subjects are asked whether their identity can be revealed and measures such as anonymisation are taken where necessary. Protocol issues are heightened in endangered languages situation, which typically involve small communities under socioeconomic, political, or military pressures. In such communities it is almost impossible to be anonymous; even the slightest bit of apparently harmless information can reveal someone's identity, whether to another community member or to some hostile external agency.

See section xx for a description of an innovative method for implementing flexible access control.

Language documentation's often private or sensitive content means that some protection of intellectual property and/or copyright is required. Many archives have statements which those accessing materials must agree to, typically prohibiting commercial use or republishing without explicit written permission. The level of protection required depends on the nature of the resource and the goals of the information providers. Some language communities welcome the opportunity to showcase their language and culture to the wider world. One way of specifying permitted usage and distribution is by means of Creative Commons licences. The Creative Commons initiative, with its catch cry "*some* rights reserved", is more oriented to facilitating the sharing of resources for personal, non-profit and creative usages, and has various formulations that require acknowledgement only, or that restrict usage and distribution to non-profit purposes. The licences also formulate varying controls on the creation and onward distribution of materials that incorporate some or all of a given resource but add additional content - "derivative works" - a category that could apply to analyses of linguistic materials, lexical material that is reorganised or combined with other data, and various types of multimedia.

What about technical solutions for controlling access and distribution? While there exist some technologies that can protect files from unauthorised copying and distribution, such as mechanisms for Digital Rights Management and audio "watermarks", these are in general tailored for use by large companies to protect commercial music and similar products, and language archives running on limited budgets are unlikely to be able to implement them. In any case, such technologies are in themselves a threat to the robustness of short and long term archiving, since they involve encrypting the contents of files, often by secret and proprietary methods. A digital archivist's perspective is that long term preservation is best facilitated by keeping resources in standard and transparent formats, and designing protection and distribution systems based around generally accepted behaviour. The risks of inappropriate access should be imposed at the point of managing access to resources, rather than through solutions that modify the resources themselves. An archive with sufficient technical and financial resources could preserve originals as well as creating protected versions for dissemination, but the costs of implementing such a system and keeping pace with changing technologies are unrealistic for most archives. Nevertheless, the evidence so far suggests that the actual level of unauthorised copying in the domain of indigenous cultural/intellectual property is actually very low.

9 Preservation issues

Digital archives have to take account of many factors to unsure long term preservation, from the broader political, organisational and financial issues that guarantee their sustainable operations, to budget and equipment planning, to technical details of scheduling automated tape backups. Full discussion of all of these is outside the scope of this chapter, and many tend to be generic to digital preservation of all kinds. Below, some topics that intersect with documentation are briefly described.

Prospects for hardware and storage

The history of digital technologies is a giddy progression of steady trends in hardware, sudden changes in architectures (such as operating systems), and unforeseen revolutions in the ways that the technology is used (e.g. the arrival of the World Wide Web, and the current transition to mass participation in it via Web 2.0). The predictable trends tend to be in hardware capabilities, such as the rapid but predictable increases in processor speeds[12] and data density (holding capacity per drive unit). These are of immense relevance and benefit to archives, especially language archives that need to survive on low budgets, because they allow the transmission and storage of more data, more quickly, and at less cost.

The price of conventional hard disk space continues to tumble, with the cost per megabyte halving about every two years. This has introduced new possibilities for mass data storage, for example (i) expanded use of redundancy techniques, which ensure that data can survive hardware failures; (ii) the use of disk rather than tape for backup; (iii) greater storage within a single appliance (currently allowing up to about 100TB in a single unit), which greatly reduces costs by simplifying systems and avoiding costly enterprise-level solutions that were until recently necessary for storing large volumes of data; and the feasibility of setting up project-local archives (note, however, that local data "archiving" should not be confused with the services of committed institutional archives that guarantee behind-the-scenes backup, data migration, data dissemination, and other services.).[13]

The years 2009-2010 also saw rapid reductions in price of solid state drives (SSDs). Although they currently still cost 10 times the price per megabyte as their corresponding conventional (magnetic) hard disk drives, their adoption in the laptop computer market is likely to drive further reductions, so that at some point in the future, archiving storage will transition to SSD technology. This in turn will have many positive implications for language archives' costs, robustness, and flexibility, due to SSD's inherent reliability, increased read/write speeds, reduced size, and a large reduction in energy costs (for both running and cooling).

Data migration

Earlier discussion in Section xx on digital encoding showed that the retrieval and meaning of digital data is dependent on character and file encoding. While some encodings (e.g. plain text as Unicode) are widely supported by a range of software, including open-source software, and openly accessible as ISO standards, it is inevitably impossible to guarantee that all files involved in language documentation will be stable or usable in the medium or long term. Vulnerable examples include media files, most particularly video (see further discussion below), proprietary formats (e.g. MS Word, Excel, Filemaker Pro and others). Other files needing special care include specialist linguistic materials such as Toolbox and ELAN files. While their underlying file formats may be enduring (e.g. ELAN uses Unicode, XML plain text), they may not be usable in the expected way when the software itself no longer

[12] This is more precisely known as "Moore's Law", which predicts that the number of transistor (basic processor) units that can be physically fitted together into a computer's CPU doubles every two years.
[13] See also the discussion about video.

runs on new versions of operating systems.[14] A central function of digital archives, therefore (and complementary with their role in educating the documentation community to use the most stable formats possible) is to catalogue the file preservation characteristics of all files in their collections, and, at the appropriate time, to migrate vulnerable files to new and safer formats.

Video

The benefits of digitisation are only fully realised when data and file formats have become stable. At the time of writing, digital video formats are volatile, varying with carrier type (e.g. hard disk vs. flash card vs. DV tape), camera manufacturer, and processing software. Video provides an interesting test case for the capabilities and limits of digital data management, storage, and delivery.

Video has many merits for language documentation, offering a record integrating audio together visual representation of language speakers, their gestures, body movements, locations and contexts. The breadth of this potential, however, invites many problems. Many documentation projects are not concerned with gestural or spatial information, so to shoot (and archive) video may not be a good use of resources. Even if projects do have aims that make video relevant, the filming methodology (or lack thereof) may not effectively capture the phenomena concerned. But most importantly, the costs and inconveniences of using video - whether from equipment purchase, electricity consumption, weight (including necessary accessories such as tripod), need for training, intrusion and distraction to both researchers and researched, capture, processing and storage - all provide bottlenecks and constraints on good outcomes.

High resolution video (such as captured directly from miniDV tape) is very large in volume; at least three or four times its typical distribution size and ranging from 10 to 100 times the size of audio of comparable length. In addition, the high resolution versions captured directly from cameras are often in proprietary formats specific to the particular brand of camera and/or software used for transfer and viewing. Thus, due to practical and theoretical limitations on language archives' data systems, the high resolution video that comes directly from cameras - the most informative version that would normally be preferred for archiving - cannot in general be preserved. Only compressed files such as those in MPEG format are sufficiently tractable in size and standard in format for both archiving and practical usage, so most archives have, to date, accepted video only in the highly compressed MPEG2 format. But this proves to be merely a short-term or misleading strategy, perhaps with a blind ending. Almost any subsequent processing of the video, including editing, subtitling, or re-rendering to other formats for migration or delivery, should be derived not from MPEG2 (archived) versions but from original high resolution versions, because editing is normally followed by re-rendering involving another compression, causing great loss of video quality. This leads to two conundrums, if not contradictions, for digital archiving of video.

[14] In the medium term, we cannot anticipate the fortunes of the organisations or companies that produce, maintain and/or sell software; in the longer term - hundreds of years and beyond – the likelihood of today's software remaining usable is close to zero.

First, the fundamental reason for adopting digital archiving - its long-term support for preservation of media data through the ability to make perfect copies (see Section xx) - is negated, due to the repeated re-encoding that will be needed as video formats continue to change. The loss over each generation of re-encoding simply recapitulates the original problem with analogue carriers.

Second, consider the scenario where researchers want to create some products from their video, for example to support language revitalisation. Editing should proceed from original high-resolution formats, which are unlikely to have been archived, and, if they have been retained at all, it is more likely that they have not been transmitted elsewhere but have been stored locally by the original researcher. So what we see is a reversal of general archiving strategies; in this case only the researcher, not the archive, is in a position to preserve the "best version".[15]

Archive assessment

There are currently a small number of dedicated digital archives for endangered languages documentation (see Appendix XX). This chapter has discussed only a few of the strategic and operational complexities that these archives must face.

Documenters wanting to archive their data need to choose a suitable archive facility. In part, they will do this by matching their type of materials with the collection policy of a relevant archive. More generally, depositors (and others; see below) might want to evaluate the qualities of archives before they trust their precious data to the curation, care, and custodianship of a particular archive facility.

Several initiatives have been set up to help such depositors, as well as to assist archives to assess their own digital preservation policies and practices. These include Drambora (www.repositoryaudit.eu), NINCH (www.ninch.org/programs/practice), Data Seal of Approval (www.datasealofapproval.org), and the Digital Curation Centre's toolkit (see www.dcc.ac.uk). They provide participating archives with document requirements or templates (e.g. policy and planning documents for access control, backup, security, disaster recovery, staffing, and funding) and various sets of operational criteria. While most digital endangered languages archives have not yet defined which initiative is the most suitable, nor uniformly subscribed to any of them, such assessment schemes are expected to play a greater role in the future, for example as funders require their grantees to archive with an approved archive, or archives form federations (Broeder 2008) with those that share similar goals and strategies.

Redefining language documentation archives

Section 8 "Access" described how protocol issues are highlighted by the nature of language documentation data. Many materials need to be subject to controlled access, and conditions of access can change over time and depend on who is seeking access.

Following the explosive growth of social networking, or "Web 2.0", between 2005-2010, people worldwide have proved keen to conduct interactions, negotiations and

[15] There are two positives, however: video editing and production is more likely to be appropriate in the context of the original project, or language community; and this situation provides a good incentive for the development of small-scale local or personal digital archives.

relationships via the World Wide Web. The use of social networking sites such as Facebook and MySpace are now embedded in lifestyles in both wealthy and poorer nations, and there are mature technologies that are open-source and freely available for adoption.

The Endangered Languages Archive at SOAS is currently pioneering the application of these social networking models to providing controlled access to endangered languages documentation (Nathan 2010). Via the archive's web-based catalogue system, depositors can manage access conditions, respond to access requests from individuals, and monitor the usage of their materials. By devolving access management to depositors, the system neatly addresses the sensitive nature of many archived materials, whilst also solving the problem of managing complex access conditions for an ever-growing collection with a fixed and small staffing level.

This new approach realises the transactional functions that are foregrounded for the depositors and users of a modern digital archive. Preservation functions are slowly receding into the background as essential but generic services that businesses, government and educational institutions all have to provide to carry out their work. The question for the future is not whether such an approach is likely to be widely adopted but how wide-reaching its effects will be; will, for example, blogs, wikis and media sharing websites also take a place in the language preservation and dissemination landscape? Whatever the precise outcome, the digital archive is no longer essentially defined by its data preservation function, but is reconceived as a forum for conducting relationships between information providers (usually the depositors) and information users (language speakers, linguists and others).

10 Conclusion

In this chapter we have seen how the abstract nature of digital data enables long term preservation of media resources as well as flexible usage and sharing of data of all kinds. On the other hand, storage and retrieval of digital data inevitably require complex processing and computing hardware, so that the feasibility of long term preservation depends on reducing the complexity of the layers that stand between the underlying data carriers and the users of the data. The future usefulness of resources depends on careful documentation of data at all levels, from the methods by which characters and files are stored, to rich descriptions of the resources and their contexts that enable their content to be identified, retrieved, and understood.

In an emerging field such as documentation of endangered languages, archives can draw on digital technologies and standards developed over the last 40 years, but they still have to provide discipline-specific facilities to meet the needs of their users. Maturing web technologies and new understandings of the role of digital archives in preservation and dissemination are recasting archives as amplifiers of the value of language documentation by linking documenters, their documentation materials, and the diverse users of these materials, now and into the future.

11 References

Arms, William. 2001. *Digital Libraries*. MIT Press: Massachusetts MA.

Bird, Steven and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. In *Language* 79, pp 557-582.

Broeder, Daan, David Nathan, Sven Strömqvist & Remco van Veenendaal. 2008. Building a Federation of Language Resource Repositories: the DAM-LR Project and its Continuation within CLARIN. In Proceedings of the Sixth International Language Resources and Evaluation (LREC 08, Marrakech, Morocco, 28-30 May 2008) Online at http://www.lrec-conf.org/proceedings/lrec2008/summaries/370.html

Dobrin, Lise M., Peter K. Austin & David Nathan. 2008. Dying to be counted: the commodification of endangered languages in documentary linguistics, *Language Documentation and Description* Vol 6. London: SOAS. [online at http://www.hrelp.org/publications/ldlt/papers/ldlt_08.pdf]

Gippert, Jost. 2006. 'Linguistic documentation and the encoding of text materials'. In Gippert et al (eds). pp 337-362.

Gippert, Jost, Nikolaus Himmelmann & Ulrike Mosel (eds). 2006. *Essentials of Language Documentation.* (Trends in Linguistics. Studies and Monographs, 178). Berlin: Mouton de Gruyter.

IASA (International Association of Sound and Audiovisual Archives) 2005. *Guidelines on the Production and Preservation of Digital Audio Objects*. Technical Committee, Standards, Recommended Practices and Strategies (IASA-TC 04, 2nd edition).

IASA (International Association of Sound and Audiovisual Archives) 2005. *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*. Technical Committee, Standards, Recommended Practices and Strategies (IASA-TC 03). Version 3, December 2005. Online at http://www.iasa-web.org/downloads/publications/tc03_english.pdf

Korpela, Jukka. nd. *A tutorial on character code issues.* http://www.cs.tut.fi/~jkorpela/chars.html. (accessed 15 June 2009).

Levinson, S. 1999. *digital mcluhan*. London: Routledge.

Nathan, David. 2010. Archives 2.0 for Endangered Languages: from Disk Space to MySpace. In *International Journal of Humanities and Arts Computing*, Volume 4 (Special issue).

Nathan, David. 2006. Proficient, Permanent, or Pertinent: Aiming for Sustainability. In Linda Barwick and Tom Honeyman (eds) *Sustainable data from Digital Sources: from creation to archive and back.* Sydney: Sydney University Press, pp 57-68.

OAIS 2002. Consultative Committee for Space Data Systems (CCSDS). CCSDS 650.0-B-1. *Reference Model for an Open Archival Information System (OAIS).* Blue Book. Issue 1. January 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed 19 April 2008).

Wood nd. Alan Wood. http://www.alanwood.net/unicode/. Accessed 30 June 09.

12 Appendix

A select list of archives for endangered languages that host digital materials.

Aboriginal Studies Electronic Data Archive, Australian Institute of Aboriginal and Torres Strait Islander Studies. http://www1.aiatsis.gov.au/ASEDA/

Alaskan Native Language Center Archives (ANLC) University of Alaska. http://www.alaska.edu/uaf/anlc/

Archive of the Indigenous Languages of Latin America (AILLA), University of Texas. http://www.ailla.utexas.org/site/welcome.html

Digital Endangered Languages and Musics Archives Network (DELAMAN). http://www.delaman.org/

Dokumentation Bedrohter Sprachen Archive (DoBeS), Max Planck Institute Nijmegen. http://www.mpi.nl/DOBES

Endangered Languages Archive (ELAR), School of Oriental and African Studies. http://www.hrelp.org

Langues et Civilisation et Traditions Orale (LACITO), Centre National de la Recherche Scientifique. http://lacito.vjf.cnrs.fr/archivage/index.htm

Leipzig Endangered Languages Archive (LELA), Max Planck Institute Leipzig. http://www.eva.mpg.de/lingua/resources/lela.php

Northeastern North American Indigenous Languages Archive, University of Buffalo. http://nnaila.org/

Pacific and Regional Archive for Digital Sources in Endangered Cultures (Paradisec), University of Melbourne/University of Sydney. http://paradisec.org.au/

Rosetta Project, Long Now Foundation. http://www.rosettaproject.org/