

Future-proofing of the Royal Anthropological Institute's digital archives: Case for Support

Aims and objectives

The ability to digitize documents has a number of important benefits but it has one potential major pitfall which is the danger of obsolescence (Borghoff, Rödiger et al. (2006)). The objective of this proposal is to develop a suitably 'future-proof' format for the preservation of digitized archives of the Royal Anthropological Institution (RAI) – and other similar artefacts.

The dilemma is well summarized by Rothenberg (1999) 'Digital documents last forever—or five years, whichever comes first.' It is well described in Borghoff, Rödiger et al. (2006), particularly Chapter 1.

Advantages of digitization include high-resolution facsimiles which can be accessed by multiple users without access to the original. This helps in the preservation of the original, protecting it from any damage during handling. The digital facsimile takes up essentially no physical space. Copies can be instantaneously communicated anywhere in the world. As many copies as desired can be made (although this can also be a disadvantage because of copyright and intellectual property right problems). The digital copy may also have features additional to the original because of the ability to process the digital representation. For instance, it may be enhanced, perhaps improving the legibility of a faded original. Also, properties of the document can be processed, for example to automatically catalogue and index it, facilitating instantaneous searches of the whole collection.

Digitization essentially amounts to reducing a document to a series of *bits*, zeros and ones. A large part of the power of digitization comes from the fact that the format is so simple and flexible and there is a range of technology available to process it. Computers are the general technology to process any form of digital data, but mp3 players, phones, video recorders, televisions, in fact almost all electronic technology handled digital data. There is no reason to suppose that digital technology will not continue to dominate for a long time to come and thus digitized information ought to be very safe. However in practice there are dangers of potential obsolescence. There are two aspects to this: *hardware obsolescence* and *software obsolescence*.

All digital hardware can store and process bits. Long-term storage at this level amounts to being able to preserve the bits that represent a particular artefact. Most storage media are prone to degradation over time, but can be easily refreshed. That is, as long as the record (bits) are intact, they can be read and copied to a new, fresh medium. The real problem with hardware obsolescence derives from the development of the technology; as new (better) storage technologies are developed, the equipment to read the old technologies becomes unavailable.

This is easy to see in the context of the personal computer. Very early PCs can fitted with floppy disc drives, which used 5¼ -inch diameter discs (which were truly 'floppy') for storage. Soon, though, the 3½-inch floppy (with greater capacity – and a hard case) became the standard. There was a period when a user might buy a PC with one drive of each kind, and thereby be able to transfer their old data to the new format, and avoid hardware obsolescence. Suppose they did not, though. With a PC fitted only with a 3½-inch drive any data on old, 5¼-inch floppies had become inaccessible. The bits still existed, they just could not be accessed.

Probably the best-known example of hardware obsolescence is the Blue Peter Doomsday Book. This was an attempt by the BBC children's programme to create a year 2000 version of William I's catalogue of the country. The Blue Peter version was stored on the best available mass storage device of the time, a form of laser disc. Within a few years, though, no drives were available to read these discs, making their contents inaccessible. To all intents and purposes, the new Doomsday Book no longer existed.

The problem with hardware is a consequence of Moore's Law (Moore (1965)). This states that computing power doubles every 18 months. First observed in the 1960s, it has remained true ever since and has survived numerous predictions of its limits and consequent demise. Not usually seen as a 'problem', it drives the development of ever better hardware, and that in turn creates obsolescence: this year's hardware is better, bigger and faster than last year's and we want make full use of it.

There are a number of approaches to hardware future-proofing, but they amount more to a management requirement rather than a technical solution. They can be summarized as:

Migration

This amounts to periodically moving the data from last year's storage medium to this year's. Perhaps the most obvious approach, it is not always as straight-forward as might be assumed. Maintaining the integrity of the data during migration can be difficult. (TFADI (1996))

Preservation

This amounts to maintaining a 'computer museum'. Instead of throwing away old hardware, it is retained and used to access the old data. Maintaining the hardware and keeping it working is clearly difficult. (Borghoff, Rödiger et al. (2006))

Emulation

An alternative to keeping the old hardware working is to use software on new hardware which simulates the old hardware. This is an approach favoured by games enthusiasts. They wish to play classic games as released on computers such as the Sinclair Spectrum, so they have emulations which make their modern PC behave as if it is a Spectrum – and then they can run their original games software. (Rothenberg (1995); Rothenberg (1999); Rothenberg (2000); Rothenberg and Bikson (1999))

Had one of the above techniques been applied to the Doomsday Book, though, it would not necessarily have been a complete answer. While they would have retained the *bits* that represent the Doomsday Book, there is still a question as to whether anyone could have viewed the book. That would depend on the *format* of the information contained in it and whether the *software* on their new (CD-rom-equipped) computer could interpret it. This is the question of *software obsolescence*.

Avoiding software obsolescence is the real focus of the planned research. The RAI documents are currently being stored in *tiff* (tagged image file format) format. This is a well-established format, first developed in 1992 and currently in version 6.0 of its specification. There is a lot of software for handling and processing files in this format, but the very fact that it is in its sixth incarnation may be illustrative of the problem. It is quite unlikely that a modern program, written to handle tiff 6.0 documents would be able to read a document in the original tiff format. In other words, a document which had been digitized in 1992 might now be unreadable. More

to the point, in 2026 the documents that the RAI is currently digitizing will be inaccessible – unless they are translated to a safer format.

An analogy can be drawn with language. Suppose an archaeologist uncovers a text inscribed in stone, written in some ancient, unknown language. In this case the 'hardware' is very robust – stone – and the characters written on it are entirely legible. However, if the language is no longer understood, the message in the writing is meaningless. What is required is a Rosetta Stone, which essentially tells us how to translate from the unknown language to one which we do know. Avoiding software obsolescence amounts to creating a representation which includes its own Rosetta Stone.

Background

The RAI has an important historical collection of several hundred manuscripts, most of which are unavailable anywhere else, and are of immense value to researchers into the history of anthropology and related subjects. While maintaining its copyright of these documents, the RAI is committed to making them available to whomever expresses an interest, and is especially keen to aid those researching their own cultural histories. Information gathered by the anthropologists of the past on the local culture and language is proving to be of great interest to such communities, who may be located far from London and have limited resources for travel. Thus, there is a great value in making digital copies available.

A small grant has been received recently, and work has begun on digitizing a collection of papers and drawings relating to the culture of Malekula made by A. B. Deacon in the 1920s. This collection is of particular interest to the Vanuatu Cultural Centre, and the RAI plans to share the result of the digitization with the Centre. Note that it is impossible to date many of the documents in the collection. Thus, the Classification section of the application mentions the 20th Century – when the documents were created – but they record history much further back in a way which is not covered by that form.

The RAI is a small organization which does not have a large budget to spend on digitization projects. Outsourcing this kind of work is out of the question. Also it is important that it gets value for the money that it spends. A digital archive which survives only a few years would be unacceptable. Thus it must find a way of working which is inexpensive but robust.

The PI is a Senior Lecturer in Computer Science at the University of York. His main research interest is in multimodal representations of information. He has been Principle Investigator on a number of projects funded by the Engineering and Physical Sciences Research Council (EPSRC), most recently *Sonification Of Cervical Smear Data To Improve Screening Accuracy*, 2005-2009. He has also been a partner in projects funded by the European Commission, *Lambda*, and *Maths*. For further details, please refer to his CV and List of Publications.

Research methods

Future-proofing is an active area of technological research. The objective of this, one-year, project will *not* be to develop new techniques. Rather it will be to investigate which formats and methods are most appropriate to ensure the preservation of the RAI's collections. This could serve as a case-study which could be applied to other future-proofing exercises for the RAI and other similar organizations. There are other similarly small organizations with the same need for economic solutions.

The following workplan is proposed, broken down into work packages.

The work will be carried out by one Research Associate (RA). The RA will be based at the University of York and supervised by the Principal Investigator (PI). However, he or she will, of course, work closely with the RAI, in London. All of the work will be undertaken with reference to the OAIS Reference Model (ISO (2003)).

WP0: Management

For a project involving one RA, the management requirements are minimal. It will be possible to maintain close regular communication between all the parties involved. In particular, though, the RA and PI will be co-located and able to maintain regular contact. It will be important, though, to also maintain communication with management at the RAI, to ensure that progress remains in line with their requirements. To that end, there will be regular meetings.

In addition a wiki will be established as a medium of communication between all the parties involved.

WP1: Establishment

There will be a meeting of all involved (RA, PI and staff involved from the RAI) to establish the objectives of the project and the particular constraints and requirements for the RAI's digitized collections. Samples of the existing digitized documents will be collected and examined.

WP2: Survey

Methods and techniques for software future-proofing will be investigated and evaluated. The objective will be to identify a small number of candidate techniques for further testing, in WP3.

At the same time, existing hardware-upgrade management policies will be investigated. These will feed into WP5.

Deliverable: Internal report describing the alternatives surveyed and selected.

WP3: Software evaluation

Candidate techniques identified in WP2 will be subjected to a thorough evaluation. This will assess properties such as:

- Strength of the future-proofing.
- Ease of translation, particularly from existing formats (e.g. tiff).
- Cost – including possible long-term licensing.
- Suitability for RAI's purposes.
- Flexibility of the format with regard to additional processing, e.g. enhancement, automatic cataloguing and indexing etc.

Deliverable: Internal report identifying and evaluating the chosen technique.

WP4: Proof of concept

Having identified the best technique, selected components of the existing digitized collection will be translated into the chosen format. The use of these artefacts will then be assessed with the collaboration of RAI members to evaluate the appropriateness of the format. Evaluation techniques will be chosen depending on the outcomes and the requirements of the RAI.

Deliverable: Internal report to the RAI.

WP5: Hardware management

A management policy will be devised for the maintenance of hardware compatibility into the future. This will be presented to the RAI and recommended for adoption.

Deliverable: Report to RAI.

WP6: Dissemination

A paper will be written, on the basis of WP4 and WP5. This will be published in a journal or presented at a conference accessible to other keepers of collections, similar to the RAI. Another paper will be written from more of a technical perspective for publication in the information systems literature. A website will also be established for dissemination.

In addition, the project will run a workshop to disseminate their results.

Particular effort will be made to share the results of the project with the Vanuatu Cultural Centre since most of the digitized collection consists of artefacts relating to Vanuatu.

Note that all results will be freely disseminated. While the contents of the RAI collections will have to remain protected as before, there will be no restrictions on any intellectual property generated within the project. The aim is for other institutions with similar needs to the RAI to benefit also.

Deliverables: Two or more publications. One workshop.

Work chart

Month	1	2	3	4	5	6	7	8	9	10	11	12
WP0												
WP1												
WP2												
WP3												
WP4												
WP5												
WP6												

Contribution to career development

The principal contribution to career development will be for the benefit of the RA employed. The individual has not been identified and so it is impossible to be definitive about the effect on their career, but it can be assumed that anyone suitable to be employed on this project will come out of it with a greater understanding of the technology and requirements of archive digitization. Since they will have to be from a technical background, working with archivists is likely to be a new experience from which they should obtain novel and rare insight. From a one-year project they will be in a good position to undertake further work in the digital preservation of artefacts.

From another point-of-view, the principal contact at the RAI does not have a lot of prior experience of technological research. On completion of this project she will be in a good position to undertake further projects of this kind.

This project also represents a new direction for the PI. While he has extensive experience of information technology research, this particular application is quite novel. The project should put him in a good position both to undertake similar

research in the future, but also to feed the results into some of his other areas of research, particularly multimodal representations.

References

- Borghoff, U. M., P. Rödiger, et al. (2006). *Long-Term Preservation of Digital Documents: Principles and Practices*. Berlin, Heidelberg, Springer-Verlag.
- ISO (2003) "Open archival information system - Reference model." International Standards Organization ISO/IEC 14721:2003,
- Moore, G. E. (1965). *Cramming more components onto integrated circuits*. *Electronics* **38**(8)
- Rothenberg, J. (1995). *Ensuring the longevity of digital documents*. *Scientific American* **272**(1): pp. 24-29
- Rothenberg, J. (1999) "Avoiding technological quicksand: Finding a viable technical foundation for digital preservation." Council on Library and Information Resources, (<http://www.clir.org/pubs/reports/rothenberg/introduction.html>)
- Rothenberg, J., Ed. (2000). *Using Emulation to Preserve Digital Documents*. The Hague, RAND-Europe and Koninklijke Bibliotheek.
- Rothenberg, J. and T. Bikson (1999) "Carrying authentic, understandable and usable digital records through time." Dutch National Archives and Ministry of the Interior, http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf
- TFADI (1996) "TFADI: Preserving digital information - final report." Task Force on Archiving of Digital Information, <http://www.oclc.org/research/activities/past/rlg/digpresstudy/final-report.pdf>

2,647 Words