The role of standards in digital longevity

K. van der Meer(1) and J.J.M. Uijlenbroek(2)
(1) Delft University of Technology
Zuidplantsoen 4
2628 BZ Delft
The Netherlands
(2) Het Expertise Centrum
Jan Willem Frisolaan 3
2517 JS 's-Gravenhage
The Netherlands

Introduction

An old joke says, "No one has contributed so much to history as historians have". The mission of historians is to create history by constructing and analysing of what has happened. Sometimes the reconstructions seem to deviate from the historical facts.
The advance of information and communication technology (ICT) has brought something new. Electronic documents and records will be unreadable within a few decades. A major task of computer scientists will be to preserve history by constructing and analysing an enormous amount of old electronic records and documents. Within a few years, no one will have contributed so much to history as computer scientists have…
ICT has made document handling easier. One can simply move, copy, cut and paste electronic document information, and construct a new document by using available document parts. It is difficult to believe that digital longevity could be a problem. It is an oxymoron that just under these circumstances document information seems to taint and wither under one's hands.
The digital longevity problem was noticed quite a few years ago. The Dutch Bureau of Digital Longevity was founded in 1991, the book 'Preserving the Present' was published in 1993 and Rothenberg's article appeared in the Scientific American in 1995. Recently, many organisations needed to convert electronic documents or had even to face their loss. But in spite of these, still today many people find it difficult to appraise the relevance of the questions. Proposals of answers are even more difficult to appraise.

Key factors

If we want to guarantee that an electronic document will be readable in 50 years, how can we attain that? On what data carrier can it be preserved? If it is migrated to an other carrier and to an other platform, can its provenance be proved? What will its legal status be? How to recognise electronic annotations to it? On a higher level, what key factors are pertinent to the development of a policy and strategies to preserve archival electronic records? Could an archive issue a Service Level Agreement (SLA) to promise to society that certain documents can still be accessed and serviced in 50 years' time?
An analysis of vital questions leads to the following types of problems.

- Can different levels of authenticity be identified, and if so, what will the use be of them? What roles could exist for electronic records with different levels of authenticity? What requirements must be set to provenance?
- What requirements to hardware and software should be set?

- What responsibilities will be put upon electronic archive management? How can appraisal of electronic records be realised in practice? How can quality requirements be developed that are applicable to control and management of electronic archives?
- What types of costs are made? Can benefits be realised? How do the costs and benefits of electronic record management relate to those of paper record management and how to deal with costs of the ICT infrastructure? What models of cost accounting are applicable for electronic archives?
- What standards are applicable to electronic documents and electronic records and what about longevity of standards?

Standards are vital
Document sharing proves to be a new and strong trend. Open organisations exchange lots of documents with their environments (and closed organisations may hardly be viable much longer, one more consequence of ICT). The sheer reason of many documents' existence is their serving as a means of communication between organisational units. Records witness transactions between organisations, stem from CSCW, or originate from outside the organisation that keeps the records. Document sharing depends literally on standardisation. The design of interorganisational document information systems depends on standardisation of electronic document functionality.
The importance of standards for preservation has, however, been a matter of dispute.
Firstly, it has been stated that standards cannot solve the problem of digital longevity. Still, in our opinion, standardisation is one of the key factors.
Secondly, it has been stated that standards are subject to change. There is a point. Technical developments and improvements continue to lead to new possibilities and hence to new standards. So it must be expected that standards will age.
Nevertheless: standards must answer longevity requirements, else they are void. This leads to the question, how standards and their substrates (documents) can be carried through time. For a first answer to that question, let us browse through some pertinent standards and comment on the longevity aspect. We limit ourselves to standards on (provision of) document information.

Families of standards
There are quite a few important groups of standards for electronic records. They are given here.

1. Document structuring.

- Standard Generalised Mark-up Language SGML (ISO 8879) for the logical document structure,
- Document Style Specification and Semantics Language, DSSSL (ISO 10179) for the document lay-out structure,
- HyperText Mark-up Language, HTML (ISO DIS 15445) for the mark-up and hyperlink facilities of Internet documents,
- Cascading Style Sheets, CSS (no ISO standard), viz. the DSSSL for HTML,
- eXtensible Mark-up Language, XML (not yet an ISO standard), a less complicated SGML,
- HyTime for Hypermedia and Time-based documents (ISO 10744),
- XHTML, standards for multimedia documents, XSL, Xpointer etc. for XML, …

In 1986 ISO accepted both SGML and one more standard for the document structure: Office Document Architecture, ODA (ISO 8613). Its purpose was comparable to SGML and DSSSL

together. Its implementation was expensive and ODA was hardly used at all. This standard came too early.
SGML needs a 100% DTD modelling. In practice, that is a lot of work. That is why SGML was not widely accepted. It is accepted now mainly in the fields of manuals of complex equipment and publishers. But the need for document structuring is persistent and now XML evolves. XML is a much simplified and slightly enriched (hyperlinking) SGML.

2. Printer language standards.
The printer language standard is Portable Document Format, PDF. It is based upon the older PostScript. PDF is not an official standard, but is proprietary to Adobe®.
PDF is an example of a de facto standard. The ISO printer language standard is SPDL, ISO 10180. SPDL does not seem to have many advantages over PDF. SPDL is hardly supported. Everyone seems to have an Acrobat reader for PDF, but more software is necessary to use all features of PDF. The longevity of this standard is possibly proportional to the longevity of the owner organisation.

3. Interoperability of documents.
- ODMA (Open Document Management API) and
- DMA (Document Management Alliance) are standards aiming at interoperability: access to documents in a distributed environment, open to different document management software and different applications. Of ODMA, a one-to-many solution, version 2.0 is getting accepted in the market. Of DMA, the many-to-many option, version 1.0 seems to lag behind. These standards are owned by AIIM.
- WebDAV, a distributed authoring protocol for Internet documents on which much work has been done in 1997-98, is ‘owned’ by the W3C.

At this moment there is not really a standard for document interoperability that is stable enough and has sufficient mass. The support may be limited if the supplier vanishes.

4. Meta data elements.
A lot of work has been done on definition of meta data since about 1970. Examples of sets are the following.
- MARC set,
- ISAD (G) set
- Pittsburgh set,
- Dublin core.

MARC is well known in the library. It is too limited for record management. UNESCO set up ISAD (G) in 1994; despite the high patronage there is only a very limited installed base of ISAD (G). The Pittsburgh set and Dublin core (for electronic library materials) have not yet crystallised out. This standardisation process is converging only very slowly.

5. Content access for information retrieval.
Content access will probably be defined by ISO (at the moment a Draft International Standard, DIS) 23950. This is a standard that supersedes ANSI Z39.50 and narrower terms (there are related ANSI standards) and some of the ISO standards in the range of ISO 10160 – 10166.

The ISO 10160 up to and including ISO 10166 standards and Z39.50 were widely known and although there were apparently differences in implementation, their content seems to have been respected. The ISO 23950 standard could be a basis for a long-time standard.

6. Information content structures.

The information content structures have been standardised by the thesaurus. There are two of these: the monolingual, ISO 2788, and the multilingual, ISO 5964. ISO 2788 is very widely known and many packages conform to that standard.

7. Character sets.

The list of standardised character sets contains ASCII (ISO 646), ISO 8859 (the Latin series), ISO 6937 (Teletext and Videotex) and finally UNICODE (ISO 10646). ISO 8859 and ISO 6937 include ISO 646; ISO 10646 includes the other standards. Apparently, there is backward compatibility to the older standards (that are in use). That standard could last the whole coming millennium.

8. Others.

A kind of catalogue for document information standards that are not very relevant from the viewpoint of ICT can be found at http://www.iso.ch/liste/TC46SC9.html, http://www.iso.ch/liste/TC46SC4.html and http://www2.echo.lu/oii/en/archives.html.
ICT standards that have less importance for document information are http, CORBA, COM / DCOM, OLE, DLL, ISO 9660 for CD-ROM and ISO 11560 for WORM, the Workflow Management Coalition standards etc. All those standards are out of scope for our purpose.

“The nice thing about standards is that there are so many different ones to choose from”. This common quote proves to be wholly unjustified. There are only few standards for preservation of electronic records that we would dare to advise now for digital longevity. A SLA for electronic records can, alas, not be issued at this moment.
For that kind of agreements and for the very goal of preservation of digital documents and records, existing and new standards in this area should explicitly meet approval, acceptance and implementation for digital longevity. They may have to be explicitly anchored to digital longevity purposes. Once standards are suited, there is a form of healthcare needed for standards, e.g. by insisting on backward compatibility of new developments to existing and useful standards, although other tools could be thought of, too.

Some notions concerning the evaluation of standards

In order to evaluate the operation of some families of standards from the digital longevity point of view, the following parts are distinguished concerning digital documents:

- a documents content which is the dominant aspect in standards that enable document structuring;
- a documents description on which meta data standards focus;
- the accessibility of documents, the central issues in standards concerning the interoperability (e.g. remote access) and standardised thesauruses.

It must be determined to what extent a stored document should correspond to its original. In order to ensure to future generations that documents are accessible in its original format (implying complete reuse) one has to store not only a document, but also its creating environment (not only

MS-Word). The cost of this demand is expected to be very high. Perhaps, in most cases sooner or later one has to accept some loss of information. The issue here is not to search for the 'super' standard on digital longevity, but to enable future computer-scientists-and-historians to reconstruct a document like historians today may reconstruct a document and the information therein from the Middle ages.

If one drops the demand regarding a document's original format but insists on reusability, one ends up with a standard from the XML-family. If one drops the reusability a PDF or TIFF-like formats satisfy.
The less complex a standards is, the more information is lost when it is used to store documents (but the lower the cost to convert a digital document from one version to a more recent version may be). Archives relate this to the appraisal question: which documents are so important (from an organisational, cultural or evidential point of view) that they must be preserved, and how will they have to be preserved, for what purpose will they be used.
But the basic trade off is between the current cost of preventing loss of information and the cost of reconstructing (e.g. converting) a document.

Until so far the evaluation focussed on the documents content and possible loss of information. This line of reasoning is also assumed to be applicable to a document description, e.g. the meta data. Although meta data describes the context of a document, that is needed to understand and interpret a document, (partial) loss of meta data does not by definition makes a document obsolete. Compare it to the loss of information of paper documents from medieval times, where lots of meta data were lost and have been reconstructed by historians. Consequently, one has to determine which kinds of documents belong to the 'no loss of information' category and which kinds of documents can be stored with very little meta data (e.g. system generated meta data). Related questions pop up: what are the cost of adding relevant meta data compared to future recovering of lost meta data and which kinds of documents are so important that loss of meta data is cannot be allowed.

Regarding the accessibility of documents a new phenomenon comes up: a certain knowledge area structured in a thesaurus is by definition cultural and time based. Consequently, today's thesaurus may be of importance to future generations concerning the way a certain knowledge area was structured, but will not probably not satisfy their demands concerning accessibility. Future historians play an important role: they will know their field of study in order to gain accessibility to the document of today. Does that mean that preserving thesauruses to access today's documents for future generations is in fact a waste of resources? The same applies to interoperability: is it effective to preserve today's interoperability in order to ensure future interoperability?

Conclusion

Digital longevity of documents and records is a major problem. Standardisation is necessary to ensure that electronic records can be preserved in time. An overview is given of types of standards. Their behaviour in time is commented upon. It is concluded that they are not explicitly meant for longevity purposes. A conclusion is drawn with respect to their use for digital longevity: only few standards can be recommended now. However, the choice for a standard does not solve the problem of digital longevity permanently. It is expected that standards evolve during time and sooner or later one has to convert ones documents to a contemporary standard, probably leading to some lost of information. Regaining a document in its original state, appeals on the reconstruction capabilities of future historians. Consequently a trade off has to be made between the cost of keeping a document in its original format versus the cost of reconstructing is. The basic question is what to do now, to enable future historians to easily reconstruct our times. Choosing the right standard is only one thing to do. Work is in progress aiming at the common area of longevity and standards.

References

T.K. Bikson and E.J. Frinking: Preserving the Present / Het Heden Onthouden. SDU Publishers, 's-Gravenhage, 1993.

D.C. Blair and M.E. Maron: An evaluation of retrieval effectiveness for a full-text document-retrieval system. Commun. of the ACM 28, (1985), 289-299.

P. le Cerf, L. de Bremme and R. Schockaert: Standards for electronic document management. In: Proceedings of the DLM-forum, Brussels, December 1996. Luxembourg: Office for Official Publications of the EC, 1997. Pp. 217-222.

Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System. CCSDS 650.0-W-5.0 White book. April 21, 1999.

http://www.aiim.org/dma

M. Hedstrom: Research issues in migration and long-term preservation. Archives and Museum Informatics 11 (3-4), (1997), 287-291.

P. Horsman: Digital Longevity: policies on electronic records in the Netherlands. Archives and Museum Informatics 11, (3-4), (1997), 235-240.

B. Kasdorf: SGML and PDF – Why we need both. J. Electronic Publ. 3(4), (1998).

KPMG IT-Trends institute: De IT-antenne: Volgen van ontwikkelingen in een dynamische markt (The IT antenna: Developments in a dynamic market). KPMG, Washington DC, March 1999 (in Dutch).

Y. Marcoux and M. Svigny: Why SGML? Why now? J. Am. Soc. Info. Sci. 48(7), (1997), 584-592.

K. van der Meer: Documentaire informatiesystemen. 3e gewijzigde druk. NBLC, 's-Gravenhage. 1998. (in Dutch).

L. Moscato: Australian approaches to policy development and resulting research issues. Archives and Museum Informatics 11(3-4), (1997), 241-250.

http://www.activedoc.com/odma

P. Over, W.E. Moen, R. Denenberg and L. Stovel: Z39.50 Implementation experiences. NIST Special Publication 500-229, US Department of Commerce, NIST, Gaithersburg, 1995.

J. Rothenberg: Ensuring the longevity of digital documents. Scientific American 272 (1), (1995), 24-29.

J. Rothenberg: Avoiding technological quicksand: finding a viable technical foundation for digital preservation. A report to the Council of Library & Information Resources (CLIR), January 1999. (http://www.clir.org/pubs/reports/rothenberg/pub77.pdf).

G. O'Shea: Research issues in Australian approaches to policy development. Archives and Museum Informatics 11 (3-4), (1997), 251-257.

K. Tombs: Governmental, industry and user perspectives of achieving standard storage mechanisms for long-term archival activities. In: Proceedings of the DLM-forum, Brussels, December 1996. Luxembourg: Office for Official Publications of the EC, 1997. Pp. 210-216.

E.J. Whitehead: Collaborative authoring on the Web: Introducing WebDAV. Bulletin of the Am. Soc. Info. Sci. Oct/Nov 1998, 25-29.