STANDARDS FOR DIGITAL LIBRARIES AND ARCHIVES: DIGITAL LONGEVITY

H.H.J. Lorist and K. van der Meer

Section Information Systems, Delft University of Technology, PO Box 356,
2600 AJ Delft, The Netherlands
winfvdm@is.twi.tudelft.nl

Abstract. There is a lack of durable thinking on digital longevity. For this we require cross-domain communication and acceptation of standardized concepts, procedures, document formats and IT-tools. Four kinds of standards are required for preservation of digital documents over an indefinite period of time: standards for the architecture, document format standards for preservation, metadata standards for preservation of the access and standards for interoperability.
This article describes concisely the relevant standards and comments on their related longevity aspects. It concludes by explaining how the identified 'standard-domains' can be used to create software building blocks (IT-tools) of merged standards, gaining a higher level of recognition and enabling easy communication, acceptance, reuse and reduced development costs. Component Based Development (CBD) is indicated as a suitable system development methodology in order to create such building blocks for a digital Record Keeping System (RKS) based on these standards.

1 INTRODUCTION

Longevity of digital documents is finally a hot issue. At long last there is a critical mass on a solution-oriented discussion on digital preservation. The Dutch Structured problem analysis and research plan [1], and Avoiding technological quicksand [2] are examples of problem analysis; the progress on Encoded Archival Description (EAD), PADI and the Cedars project are only a few among at least 21 major international initiatives.
This variety of initiatives and solutions does not by itself establish a durable way of thinking about digital longevity. Moreover, all kinds of actors play a role in digital document preservation, ranging from librarians and archivists to information technology (IT) specialists, each having a different perception of different types of problems.

1.1 So far not too good

So far four main strategies have been suggested in order to try and tackle the problem of digital formlessness.
One solution is technology preservation, whereby obsolete hardware and software are kept in a museum set-up in order to serve as a backup when a certain media type cannot be read on current hardware anymore. Unfortunately even in a museum hardware will not have eternal life. The people that could work with this material have a limited lifespan as well.

Another solution that is being scrutinized is the emulation strategy, meaning that the bit stream of the software environment (platform and software) is preserved and interpreted with the help of an emulation specification.
A third strategy is migration, which uses the principal that bits and bytes can be transformed into a different format. Unfortunately, some aspects of document properties are lost on migration. This results in the loss of part of the evidential value of the original digital record.
And four: transfer all digital material to paper or similar more durable media types like microfilm, though such huge collections are difficult to manage and it is not clear how to represent hypertext links to dynamic Internet resources, moving images and sound.
It is likely that in practice a mixture of these strategies will have to be used and that there is not a monolithic black or white solution towards digital longevity. In this respect we agree with the manager of the Cedars project, Kelly Russell, who made the same observation at the conference "Preservation 2000" [3].

2 VARIOUS ASPECTS

The strategy choice is only one aspect in preserving access to digital documents for the future. Other aspects are legislative, procedural and organisational aspects. These different problem fields [1] should be reflected in the development of new ICT tools in order to support digital preservation over time.
Special problems with IT and information systems are the speed of developments and the ageing of the equipment, software and records. Lack of cross-domain communication and acceptance of standard concepts, standard procedures and IT-standards (both for documents and infrastructure) are at least as important as the strategy choice, and a major bottleneck of the digital preservation problem.
Standardization is in itself a promising way to slow down the speed of the ageing process, especially since documents are already being exchanged in a standardised way. Alas, standards are not explicitly designed for digital longevity.
So at first the relation of standards and digital document longevity is investigated.

2.1 Use of document standards

Standards for digital documents are used for a variety of goals. They ensure that some other person (or machine) can read the document; that some person can read the documents (s)he drew up in the past and that people can collaborate on the creation of documents.
Standards will enhance interoperability, interdisciplinary consensus on concepts, techniques and procedures and platform independence.
There are disadvantages. The use of standards will cost money. One needs to put a lot of cross-domain effort into them (as was shown by EDI experiences). The eLib Standards Guideline [4] contains the warning that several versions of a standard may be in use, suppliers may offer their own 'added value' to standards, standards may be implemented in different ways, and some standards have too many features, resulting in different subsets. And of course, standards are not the final answer to the digital longevity problem.
But the advantages of standards are evident.

For preservation of digital documents over time, preferably in a Record Keeping System (RKS), the following types of standards are pertinent:

- Standards in order to serve as a reference model (architecture) describing functionality and behaviour of a digital library/archive, procedures and concepts;
- Standards in order to preserve the digital document object format plus presentation;
- Metadata standards in order to preserve access to the content describing the technical context, provenance and semantics, enabling future interpretation of the documents;
- Standards for interoperability.

2.2 Standards for mutual understanding, concepts and procedures

ISO DIS 15489. In a distributed digital environment the management of digital collections cannot be the responsibility of just one central organization. In such an environment it is important to agree on concepts, definitions and procedures. For this we now have a Draft International Standard on Records Management, ISO/DIS 15489 [5], which is soon expected to become a final draft international standard (DIS). This standard enables organizations to standardize the terms and definitions used in records management, regulatory environment, policies and responsibilities, records management requirements, design and implementation of a records system, records management processes and controls, monitoring and auditing and training.

AS 4390. The Australian AS 4390 [6] has been used as example for the ISO/DIS 15489, which has been adapted to North American and European archive cultures. AS 4390 was issued in 1996 and like its ISO cousin presents strategies and operational guidelines of best practices. AS 4390 is also applicable to all types of records (including a digital environment).

DoD 5015.2-STD. A similar standard is being developed by the Department of Defence of the United States of America. The DoD 5015.2-STD [7] provides implementing and procedural guidance on the management of document management systems. The Dutch government recommends its use in new legislation [8], which extends the current records management law [9].

OAIS. The Open Archival Information System [10], developed by the Consultative Committee for Space Data Systems (CCSDS) of the NASA, has gained a lot of attention [3]. The OAIS reference model describes both the information flow and archival requisites and is being reviewed as an ISO Draft International Standard (DIS). Various national libraries, among which the Dutch (NEDLIB project [11], the Australian (PANDORA project, [12]) and university libraries of the United Kingdom (Cedars project) study the implementation of this generic architecture.

Figure 1 shows the most generic level of the OAIS reference model. The model defines the main actors, functionalities and data flows. Details can be found in [10].

2.3 Preservation of content

Standards for the content in the RKS should delay the ageing process of the semantic and/or physical recoverability of the document that is being preserved. A standard can

only provide such longevity when the standard itself does not change or when backward compatibility is provided. Therefore only standards with a future for preservation should be selected. Portable Document Format (PDF) and the extensible Markup Language [13] are the most frequently used standards for the document format in electronic archives/libraries.

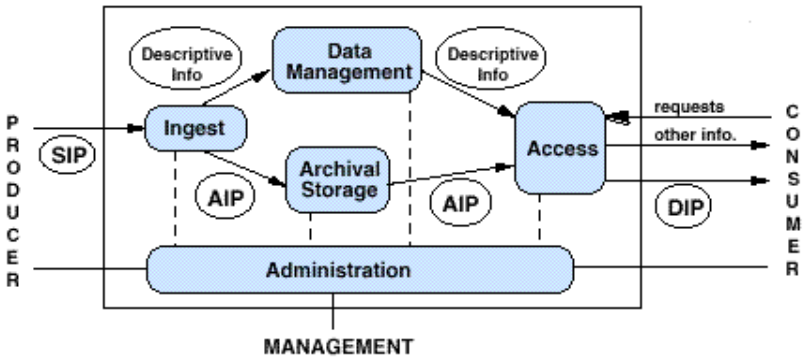

Figure 1. The OAIS reference model

Source: Reference Model for an Open Archival Information System (OAIS) '99

PDF. Portable Document Format is the proprietary standard format of Adobe and has become the 'de facto' document standard. It uses the image model of the Postscript language in order to depict text and images as exact copies of the original.
Unfortunately, users have no formal influence on improvements for newer versions and therefore backward compatibility is not guaranteed. The search facilities of PDF are currently not meeting the requirements of digital archives and libraries.

XML. The logics of an XML, Extensible markup language documents are set down in the Document Type Definition (DTD) or the XML Schema language [14]. Despite the developments, XML is not a too brilliant choice for preservation. Backward compatibility of XML is limited; a parser for XML's predecessor SGML cannot easily read an XML file and an XML parser is not always without effort able to handle a SGML file.
Moreover, for information presentation on an output device a layout structure is needed. The extensible Stylesheet Language [15] has been developed, but W3C has not yet accepted XSL as a Recommendation. So, XML cannot yet guarantee preservation of the original appearance, as is necessary for the evidential value of digital documents.
Nor PDF nor XML have been developed to provide digital longevity. As a consequence, storage of authentic records in a RKS is a problem.

Data carriers. There is a long list of standards for data carriers. Standards that can help to prolong the physical storage of bit-streams range from CD Rom, magnetic tape, optical/magnetic and DVD standards. An extensive overview can be found on the ISO homepage [16]. Unfortunately market forces sometimes benefit from creating 'new standards'. Betamax, for example, lost the videotape war of VHS. In the end the

users were the ones that had to pay for this. This shows one of the dangers of lack of standardization.

2.4 Standards for preservation of access

Preservation of the bit-stream of the document in some format is not sufficient to preserve a digital document over a indefinite period of time. One also needs a description of the digital object or the various parts that belong to one object (in a digital environment often various elements of a file are being stored in different 'physical' places in the information system and are linked by cross-reference).
This so-called metadata (data about data) should have included the contextual information that is necessary to manage, retrieve and interpret the electronic information over time.
In order to make this research redundant for the future, guarding this metadata is vital. There are several existing metadata schemes that have been standardised, each with a specific goal and application domain.

Dublin core. The Dublin Core [17] has especially been developed for cross-disciplinary networked resource discovery. Its current status is 'Draft American National Standard' and has been given the much revealing name ANSI/NISO Z39.85-200x. It exists of 15 elements selected carefully for resource description and discovery. Since it has been developed to suit multiple domains the elements are rather generic contrary to the higher granularity of MARC. For some stakeholders this rather abstract set may not be sufficient for creating access to their collections.

MARC. Secondly, the Machine-Readable Cataloguing (MARC) standards are under supervision of the Machine-Readable Bibliographic Information Committee in conjunction with Network Development and MARC Standards Office Library of Congress [18]. MARC defines the representation and communication of bibliographic and related information in machine-readable form. "The MARC record contains a guide to its data, or little "signposts," before each piece of bibliographic information." [19].
The advantage of using MARC metadata is that one does not have to develop a domain specific method of organizing bibliographic information, which saves work and enables cooperation and exchange of catalogue data with other libraries. MARC is an industry-wide standard, thereby facilitating permanent access to the records through time.

ISAD(G). A third standard suitable for defining and managing access to the actual content is the General International Standard Archival Description [20]. It contains general rules for archival description (i.e. the entities) in order to ensure creation of consistent archival description, retrieval facilities, sharing and cooperation and integration of descriptions into a unified information system.
ISAD(G) is based on the long history of the archival theory. Its 26 elements are multi-levelled starting with the general level (fonds), to the less specific 'series-level', which is subdivided in to 'files' consisting of 'items'.
The advantage of this standard is the fact that archivists understand it and that it represents the structure of the archive in relation to the context (e.g. business processes) of the descriptions.

In an archival utopia we would like to see one metadata standard, but these four are the most promising ones to use for guarding access.

Z39.50. Z39.50 is a related standard under supervision of the Z39.50 Maintenance Agency (Library of Congress) [21]. Z39.50 and its possible successor ISO 23950 enable searching in heterogeneous information systems on different platforms in a distributed environment.
The client (called Origin) and server (called Target) architecture enables the communication between different computer systems by defining 11 facilities, which exist of several services (Initialisation, Search and Present). After a query has been performed the results can be presented in various record syntaxes like SUTRS (Simple Unstructured Text Record Syntax based on ASCII), XML or MARC. Unfortunately, in spite of its success several extensions have been defined creating occasional problems when the various information systems do not exactly 'speak' the same language [22].

2.5 Interoperability standards

Also a standard is needed to concatenate RKS's, a type of standards pertaining to interoperability of different RKS's although not pertaining to properties of the digital documents themselves.

ODMA. The Open Document Management API [23] enables integration of proprietary desktop applications (e.g. MS Word) into the ODMA compliant document management system. Many document management software packages (DMS's) are already ODMA compliant.

DMA. The Document Management Alliance [24] tries to solve the problem of the 'islands of information' that different proprietary DMS's create within organisations. DMA creates bridges between these islands on protocol level, for example by facilitating clients to access several different DMA compliant DMS's though one DMA server in a networked environment.

WebDAV. The Web Distributed Authoring and Versioning standard [25] is a different type of interoperability standard. Where as a browser using the Hyper Text Transfer Protocol (HTTP) can read networked documents, WebDAV enables authors to actually write (upload) document elements by extending the HTTP protocol. The advantages are big since it enables distributed authoring and version management, being exactly what one wants in a distributed environment.

3 IMPLEMENTATION

A RKS with the goal of preserving the content and its access for an indefinite period of time has to comply with a lot of requirements, a. o. concerning content, law (copyrights etc.), organisation and technique. The various preservation strategies and the different standards have to ensure the slowing down of the natural ageing process of the digital environment.

"The nice thing about standards is that there are so many different ones to choose from". This remark certainly sketches the current messy situation when it comes to standards for digital preservation.
In order to really stimulate durable thinking about digital preservation it is necessary to merge groups of standards into specific standard 'domains'. This will make the situation more comprehensible for the different stakeholders and will facilitate cross-domain communication, acceptance and use of such groups of standards.

A point in case. Do we have a choice between PDF and XML? A closer look shows that PDF and XML are fundamentally different. PDF offers an electronic description of a page, where as XML describes the elements and structure of a page. In fact, XML and PDF complete each other [26]. Both preservation of the structure (logics) and a copy of the exact representation are needed.
The various stakeholders within the digital preservation domain should put pressure on the responsible party in the market place to make PDF an open standard. And XML needs among others standardised and accepted presentation structures.
So, the ideal document format for a RKS does not yet exist. It might be X-PDF (extensible Portable Document Format, a merger of XML and PDF, to be invented). The actors in the automation process should not wait calmly and silently until eventually some and finally all RKS's have become unfit for the job. It is in everyone's best interest that it is realised that durable preservation of digital documents requires durability of these standards.

One more example: could X-PDF and Dublin Core or MARC form a coherent couple? Could XML Names RDF (Resource Description Framework) act as a kind of mortar between those two different types of building blocks, thus forming a new building block? Would they be stable and accepted? What is the granularity of a building block we should aim at?

An analysis shows that the idea of grouping IT related standards is related to the current practice for gaining a more abstract level in IT system development of Component Based Development (CBD) [27]. This requires components to comply with criteria like "high cohesion, low coupling, a well defined interface, independence of implementation and being an encapsulated abstraction of a understood thing" [28]. Reuse of these components reduces developments costs and speeds up the development process.
The way of thinking, the way of modelling and the way of working of CBD are valuable for the construction of groups of standards into logical building blocks. These building blocks then are to be used for a digital RKS for digital longevity.

The individual building blocks (think of the generic elements of the OAIS model) need to:
- Be part of a system that and have a clearly defined interface (sort of functional contract) with the other parts of the system;
- Be capable of pro-active behaviour;
- Cooperate with other parts of the system in order to fulfil its functions;
- Be independent of its implementation and be described by its conceptual model (OAIS).

The (generic) building block should be built to be recognisable, affordable, replaceable and extendible.

The current research is focused on the durable way of thinking as explained in this paper, whereby the modelling of the system takes place, using OAIS, the Unified Modelling Language (UML) [28] and the mentioned document and metadata standards.
For the actual implementation of the system a platform independent language like JAVA is preferred since it already makes use of the component based philosophy.
The resulting system based on the building blocks for digital longevity will be tested for various preservation strategies in real-life case studies.

4. CONCLUSION

Standards are essential for digital preservation. Though they were not developed for digital longevity, standards do facilitate communication and acceptance throughout the various involved domains and are useful to delay the ageing process. They should be merged together according to the identified categories.
Four groups of standards have been categorized into the following specific requirement domains for digital preservation:
- Standards for concepts, procedures and architecture;
- Standards for preservation of the digital document itself;
- Standards for preservation of access;
- Standards for interoperability.

Current research aims at a Component Based Development-like implementation of groupings of related standards, based on the generic architecture (OAIS model).
These are then logical building blocks ensuring improved longevity of digital documents in a Records Keeping System that aims to carry the content, structure and context of digital document through time.

REFERENCES

[1] Structurering probleemveld Digitale duurzaamheid en onderzoeksopzet. Het Expertise Centrum, The Hague (1998).
[2] Rothenberg, J.: Avoiding technological quicksand. CLIR, Washington D.C. (1999).
[3] The Cedars Project: CURL exemplars in digital archives, Leeds, Oxford, Cambridge (2000). http://www.leeds.ac.uk/cedars/.
[4] Dempsey, L., et al: eLib Standards Guidelines, version 2.0, October 1998. http://www.ukoln.ac.uk/services/elib/papers/other/standards/version2.
[5] ISO/DIS 15489 on Records Management. International Organization for Standardization, Geneva (2001).
[6] Australian Standard for Records Management AS 4390. National Archives of Australia, Sydney (1996). http://www.naa.gov.au/recordkeeping/rkpubs/advices/advice25.html.
[7] Design criteria standard for electronic records management software applications: revision 1. Department of Defence, Washington D.C. (2000).

[8] Van der Ploeg, R.: Regeling geordende en toegankelijke staat archiefbescheiden. SDU, Den Haag (2000).
[9] Archiefwet (Archive Law). SDU, Den Haag (1995).
[10] Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems (NASA), Washington D.C. (2000). http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html
[11] Networked European Deposit Library; European National Libraries (NEDLIB). Koninklijke Bibliotheek, The Hague (2001). http://www.kb.nl/coop/nedlib/home.html.
[12] Preserving and Accessing Networked Documentary Resources of Australia (PANDORA). National Library of Australia, Canberra (2000). http://pandora.nla.gov.au/pandora/.
[13] Extensible Markup Language (XML) 1.0 (Second Edition) / World Wide Web Consortium (2000), www.w3.org/TR/REC-xml.html.
[14] XML Schema Part 0: Primer (W3C Candidate Recommendation). World Wide Web Consortium (2001). http://www.w3.org/TR/xmlschema-0/.
[15] Extensible Stylesheet language (XSL): version 1.0 (Working Draft). World Wide Web Consortium (2000). http://www.w3.org/TR/2000/WD-xsl-20000327/xslspec.html.
[16] International Organization for Standardization (ISO). http://www.iso.ch/.
[17] Dublin Core metadata initiative. OCLC Office of Research and Special Projects (1999). http://purl.org/DC/index.htm.
[18] Machine-Readable Bibliographic Information Committee and Network Development and MARC Standards Office. Library of Congress, Washington D.C. http://lcweb.loc.gov/marc/.
[19] Furrie, B.: Understanding MARC Bibliographic: Machine-Readable Cataloging. Library of Congress, Washington D.C. (1998). http://lcweb.loc.gov/marc/umb/.
[20] General international standard archival description, adopted by the Committee on Descriptive Standards (ISAD(G)). International Council on Archives (ICA), Stockholm, (1999).
[21] Z39.50 Maintenance Agency page. Library of Congress, Washington D.C. (2001). http://www.loc.gov/z3950/agency/.
[22] Lynch, C.: The Z39.50 Standard (1997). http://www.dlib.org/dlib/april97/04lynch.html.
[23] Open Document Management API (ODMA). Association for Image and Information Management (AIIM), (1994). http://www.aiim.org/odma/odma.htm.
[24] Document Management Alliance 1.0 Specification (DMA). Association for Image and Information Management (AIIM), (1997). http://www.aiim.org/dma/index.html.
[25] World Wide Web Distributed Authoring and Versioning (WebDAV). Internet Engineering Task Force (IETF, W3C) WebDAV Working Group, (2001). http://www.ics.uci.edu/pub/ietf/webdav/.
[26] Kasdorf, B.: SGML and PDF: Why we need both. Journal of Electronic Publication 3 (4) (1998). http://www.press.umich.edu/jep/03-04/kasdorf.html.
[27] Jonkers, H. et al: Component-Based Rapid Service Development: state-of-the-art and definitions. Telematica Instituut, Enschede (2000).
[28] Pooley, H.: Using UML: Software engineering with objects and components. Edinburgh (1998).