

# Our digital heritage: how authentic should it be?

Titia van der Werf - Davelaar  
Research & Development  
National Library of the Netherlands  
Titia.vanderwerf@kb.nl

***Abstract:***

*Together with the National Library of the Netherlands, IBM has developed the concept of a generic preservation layer model (PLM) that can be implemented as a preservation management tool. It enables us to assess preservation strategies and how they affect the authenticity of digital objects. The National Library of the Netherlands has identified a number of deposit principles that bear on preservation and authenticity criteria for electronic publications. They take account of the specific nature of electronic publications and electronic publishing. This paper shows how both the generic approach and the specific application domain approach can lead to a decision-making framework for digital heritage institutions.*

## Introduction

Library collections rapidly grow to contain an ever increasing diversity of digital material: published and unpublished, commercially available, collected as institutional and personal donations, born-digital and digitised, distributed via offline media, hosted by third parties and captured from the web. The sheer variety of digital manifestations, formats and dissemination modes is so overwhelming that in discussing the preservation of it all, one easily doesn't see the wood for the trees. Safekeeping our growing digital heritage into the future is rightfully conceived to be a daunting task. As computer scientists, archivists and librarians join their expertise to address this challenge, digital preservation theories and experiments are steadily yielding improved insights, methods and techniques that can be put to use in real digital environments.

The merits and drawbacks of different preservation technologies, such as format conversion, canonicalisation (Lynch 1999), migration (Dollar 1999) and hardware emulation (Rothenberg 1995), have been discussed in professional journals. Test-beds have been set-up (Testbed Digitale Duurzaamheid) and experiments carried out (Cedars Project) (Rothenberg, 2000) (Lorie, 2001). They have led to a better understanding of the properties of digital objects and how they are affected by different preservation strategies. All these efforts point to the need to develop requirements for integrity and authenticity. Which part of digital objects do we need to keep into the future? Is it the "essence" of a document? Should we be able to "render" the "digital original" in its native digital environment?

## Digital heritage custodians' requirements

Research into user needs is now underway (Hedstrom, 2001) with the aim of gaining more understanding of end-user requirements and to evaluate how well different preservation methods meet those requirements. Another complementary and ongoing approach is the assessment of custodial requirements by libraries, archives and other memory organisations. These organisations have mission-critical requirements and they need to make those more explicit concerning the preservation of digital heritage. I am not referring to managerial requirements for simple, affordable, and easily implemented preservation methods. Mission-critical requirements of digital heritage custodians go beyond managerial requirements and beyond the individual, by nature temporary, end-user requirements. They express the set of criteria institutions need to meet in order to fulfil their preservation responsibilities adequately.

## Integrity and Authenticity

Assessing the integrity and authenticity of records is crucial to archival practice. In a digital environment this requirement has been exacerbated. "Digital information technology creates significant risks that electronic records may be altered, either inadvertently or intentionally. Therefore, in the case of records maintained in electronic systems, the presumption of authenticity must be supported by evidence that a record is what it purports to be and has not been modified or corrupted in essential respects." (InterPARES 2001)

The Copyright office in the United States is another example where stringent authenticity requirements apply, because of the use of deposited material as evidence in copyright lawsuits. Other organisations that do not have the task of keeping material for evidentiary reasons have less compelling requirements for the verification of authenticity.

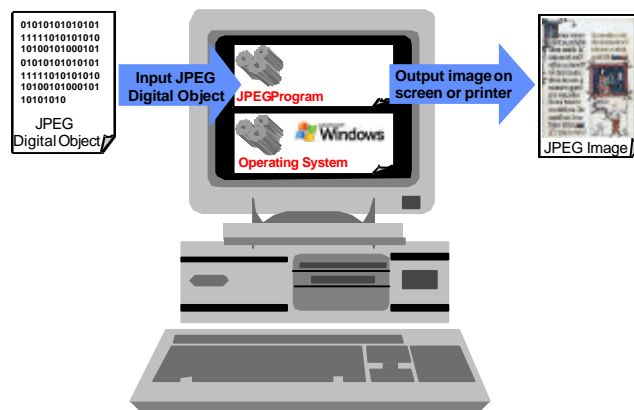
Still, all custodians need to address the issue of authenticity and integrity in the digital environment and they need to do so in an explicit way. They need to do so because digital objects cannot be accessed on the long term without undergoing some intentional transformation to make them accessible. In other words a digital original does not exist. A digital object is never “authentic”. This makes it all the more important to be able to evaluate the authenticity of a digital object: how authentic are the properties of a digital object? How much loss of authenticity has a transformation incurred upon a digital object?

By looking at digital objects in a generic way we can assess how authenticity impacts on their generic properties. This can help digital heritage custodians to make the right choices in the light of their own specific mission critical requirements.

## Generic properties of digital objects

Research, led by IBM, into the long-term preservation of digital objects has resulted in the development of a generic preservation layer model (PLM) that can be implemented as a preservation management tool. The layered model in itself has been used earlier by the OAIS standardisation initiative and by NEDLIB.

From a technical point of view, preserving digital documents and collections really boils down to keeping the bits and bytes. This holds true whether we want to keep an MPEG sound file, a GIF picture, an HTML text, or more complex multimedia animations and interactive computer games. In spite of their extrinsic diversity, digital objects intrinsically consist of bit-streams. Bit-streams are represented as strings of digits: zeroes and ones. They can be processed by computer systems only. Without software interpreting the digital object and hardware to give it a physical representation, digital objects as such would be meaningless. Figure 1 shows the components involved in rendering digital objects.



**Rendering a Digital Object**

*Figure 1*

Four basic layers have been identified as playing a role in the rendering process (Diessen, 2001):

1. **Data format layer:** the format defines the structure of the bit-stream, i.e. the intangible digital object.
2. **Application layer:** software applications are used to create, modify and view information in its intended format.
3. **Operating system layer:** the operating system provides for the shared functionality, such as interfacing to peripherals and file management, needed by application software.

4. **Hardware layer:** the hardware is the computer platform on which the intangible digital object is rendered into real physical objects, such as a print or a screen representation.

This layered model shows the chain of software and hardware dependencies digital objects carry with them. It is called the preservation layer model because it is used as a tool to manage the preservation of the rendering process. Through time the chain breaks in different places, as software and hardware components become obsolete. One broken link is enough to make a digital object unusable.

The layered model applies to all digital objects, irrespective of their specific format and usage. This suggests that it can be used in a generic way. It can be used for many purposes:

- as a basic tool to map technical dependencies between computer hardware, operating systems, application software and file formats,
- to assess the longevity of a given file format or application software,
- to identify digital objects in danger of becoming inaccessible because of obsolescence in one of the layers,
- to generate possible view paths for rendering a digital object.

In particular, it can be used as a meaningful framework to assess preservation strategies and how they affect the authenticity of digital objects.

The table below gives an example for detailing such a framework.

Layer	Preservation risks	Preservation strategies	Impact on authenticity
Carrier	Medium decay	Medium refreshing	Carrier no longer original No impact on the bit stream
Bit stream	Bit stream corruption	Backup and restore	None (copying bit streams is such a ubiquitous computational act that it makes no sense to speak of the original bit stream)
Data Format	Obsolescence	<ul style="list-style-type: none"> <li>• Format conversion</li> <li>• Canonicalisation</li> </ul>	Format no longer original Possible impact on structure, presentation and content.
Application Software	Obsolescence	<ul style="list-style-type: none"> <li>• Migration (porting, upgrading)</li> <li>• Emulation</li> </ul>	Application no longer original Possible impact on functionality and presentation (migration) Possible impact on functionality (emulation)
Operating system	Obsolescence	<ul style="list-style-type: none"> <li>• Migration (porting, upgrading)</li> <li>• Emulation</li> </ul>	Operating system no longer original Possible impact on functionality and presentation (migration) Possible impact on functionality (emulation)
Hardware	Obsolescence	Emulation	Hardware no longer original Possible impact on i/o devices and other peripherals (emulation)

*Table 1: Framework for assessing the impact of preservation strategies on the authenticity of digital objects*

Any change at any layer entails an infringement on the authenticity of the original object. Which infringements and changes are acceptable depends on the preservation perspective taken. It is obvious that a computer museum will want to keep physical carriers and hardware platforms in their original state, but to be able to run old software they may take an interest in the emulation strategy. Data archives will focus on the data format layer. Because their mission is to facilitate re-use of old data, they may take a special interest in format conversion and canonicalisation.

## **Application domain specific requirements**

If the PLM helps us to position preservation strategies and to assess their impact on the authenticity of digital resources, a complementary framework is needed to be able to make choices about the importance of authentic properties within a given application domain. In other words it is necessary to look into the characteristics of application domains. To give an example from our physical world, take the Christmas tree. The decorated tree is completely different from its counterpart in a pine-forest. We look at the same tree from a very different perspective. The way we evaluate a tree in a given context, again, is different from the way we would evaluate a tree as such. The same holds true for digital objects. The HTML home page of a web site has different properties from the HTML title page of a scientific paper. Both share the same HTML file properties, the differences bear on the application domain. The differences between application domains account for digital diversity. From these observations it becomes obvious that we need to define preservation criteria in relation to the application area under consideration.

## **Deposit library preservation requirements**

The deposit library represents such a specific application area. After several years of digital deposit practice, preservation research and experiments at the National Library of the Netherlands, insight into the nature of electronic publications and electronic publishing has grown.

## **Nature of electronic publications**

We can distinguish between two types of electronic publications: document-like publications and executable publications.

Document-like publications consist of entities that are very similar to those of printed publications. They embody data content, structure and presentation. This classifies them as belonging to the data format layer. They require reader functionality for viewing, scrolling and word searching. The readers belong to the application software layer. PDF, HTML, XML and SGML are typical examples of document-like publication formats. Image and sound formats also fall into this category.

It can be argued that reader, viewer and sound renderer functionality is universal and not specific to any data format. In other words document-like publications are not dependent on their original reader software, because any future reader can accommodate for the required functionality. An interesting aspect of this particular type of digital object is the way in which carrier, content, structure, presentation and functionality can be distinguished as separate entities. By contrast, a printed book contains all entities in one: you cannot discard the paper without the content, the pages and the layout. If a digital object can easily be decomposed into its consisting parts and some parts are more difficult to preserve than others, the question

as to which parts should be preserved in an original state and which ones can be replaced by newer technology arises.

Executable publications are stand-alone digital objects that only need an operating system to be executed. Content, structure, presentation and functionality are blended into one program that, when activated, manifests itself as one integrated entity. The consisting parts of such publications, usually coming in proprietary file formats, are not re-usable within other contexts. Educational software, computer games and web animations are typical examples of executable publications. They belong to the application software layer. It is generally recognised that these publications pose a greater preservation challenge than the document-like publications. Emulation is considered as one of the most promising technologies for the preservation of executable publications.

## **Deposit principles**

The National Library of the Netherlands has identified a number of deposit principles that bear on preservation and authenticity criteria for document-like publications.

### **For reading and viewing only**

Deposit copies of electronic publications need only be viewed in a document reader environment as opposed to edited and re-used in a document-processing environment. This principle conforms to the way in which publishers make their publications available. If publishers do provide processing functionality in exceptional cases, the deposit library should also try to make future re-use of data possible. In most cases however, publishers tend to disseminate their products making use of consumer market standards, such as PDF and HTML. Only in rare instances do they provide dedicated reader software or do they impose the use of a specific reader version. Often publishers can cater to several output formats, which leaves the deposit library with a choice.

The library should provide for the reader environment with the appropriate functionality (view, scroll, page down, print, download, word searching). In this sense it seems perfectly legitimate for the deposit library to aim to provide for generic viewing functionality through time and to opt for canonicalisation as a preservation strategy.

It should be stressed that this principle is very specific for the deposit regime of electronic publications, and does not necessarily apply to other memory institutions such as data archives – where reusability of content is the top requirement. Moreover it should be noted that publishers increasingly tend to publish research data together with the research report analysing the data. In this case, reusability of research data will be of interest to the deposit library as well.

### **Outside the dissemination environment**

Another principle, related to the previous one, is that the publisher's dissemination environment, with features such as graphic design, branding and advanced searching capabilities, is not considered to form an intrinsic part of the deposited publication. I am referring here to the publisher's own online information delivery services such as Elsevier's Science Direct or Academic Press's IDEAL service. In other words the deposit copy is considered an autonomous published entity that should be definable and identifiable outside its dissemination context as well. This allows deposited publications to be archived in a separate deposit environment with its own searching functionality supporting deposit collection uses.

The separation of archive versus dissemination environment ensures:

- agreement on well defined and identifiable published entity to be deposited;
- that the deposit environment does not compete with the added value of the publisher's dissemination environment (indexing, searching, usage rights enforcement mechanisms, etc.)
- that the deposit system develops its own searching environment and preservation functionality and caters to user needs over time.

The usefulness of the distinction between content and functionality, enabling the separate development of value-added functionality for dissemination purposes and for archival purposes, was also highlighted by Yale University Library and Elsevier Science, in their proposed approach for a collaborative project funded by the Andrew W. Mellon Foundation (Proposal 2000).

### **Web publications**

Distinguishing between dissemination and preservation environments raises the issue what to do about web publications. In how far can and should web pages be preserved in their dissemination context? How can we define and delimit web publications for preservation purposes? Are web sites to be considered publications in their own right or are they publisher dissemination environments? Should a deposit of web publications support hyperlinks across web publications, should it provide web search engine functionality? Should it reflect the functionality of the web as it develops over time?

A starting point taken by the Koninklijke Bibliotheek on these issues, again as part of the long-term preservation study carried out by IBM-Netherlands, is that web archiving has a different aim than the deposit of electronic publications. Web archiving has grown to mean harvesting and preserving web pages from the Internet, with the aim to safeguard the web and its history for future generations. While web publications can and should be added to the deposit collection, web archiving is out of scope because it preserves much more than just publications: it preserves snap-shots of shopping malls and e-bazaars, glances into the invisible college of scientific communities and traces of online civic participation. It is an interesting strategy within the broader framework of safekeeping our cultural heritage - but strictly speaking web archiving does not fall under the mission of a deposit library.

In general, web publications are document-like publications. Some typical aspects of web publications such as the hyperlinks, the embedded advertisements and the interactive buttons with help and feedback functions, need special attention.

Hyperlinks are functional in the web. You have to maintain their functionality or else a web page reads like a table of contents. Internal links all belong intrinsically to the publication and should not pose a problem to preserve. The external links are more problematic. They tend to be less informational than in the print world, as URLs are increasingly used in place of full bibliographic citations. This aspect is narrowly related to the whole URI-issue on the web. It has been recognised as an organisational issue, where organisations need to take up their responsibility for producing well-behaved and persistent identifiers that ultimately resolve into locators. Deposit libraries can play an important role in this as providers of last-resort locators (Werf, 1999).

Advertisements, interactive buttons and other extrinsic features present in a web publication can be considered to be part of it. After all, print publications also carry advertisements. Deposit libraries have never wanted publishers to submit journal copies without advertisements.

## **Conclusion**

Research, led by IBM, into the long-term preservation of digital objects has resulted in the development of a generic preservation layer model (PLM) that can be implemented as a preservation management tool. It enables us to assess preservation strategies and how they affect the authenticity of digital objects in general, irrespective of differences in origin, usage, manifestation and context. This is a first step towards the development and deployment of ubiquitous preservation functions in digital environments. Within the application domain of deposit libraries, the National Library of the Netherlands has started to formulate mission critical requirements for preserving electronic publications. These include authenticity criteria for document-like objects and take account of the specific nature of the electronic publishing environment. It is the blend of the generic IT-based approach with the specific application domain approach that seems so promising to enable us to arrive at making the right choices.

## References

Cedars Project. United Kingdom.

<<http://www.leeds.ac.uk/cedars/testsites.htm>>

Diessen, Raymond van, 2001. Requirements specifications of DNEP subsystem preservation. IBM Report, draft version December 2001.

Dollar, Charles M., 1999. Authentic electronic records: strategies for long-term access. Cohasset Associates.

<<http://www.cohasset.com/main/library/dollarBook/>>

Hedstrom, Margaret and Lampe, C., 2001. Emulation vs. Migration: Do Users Care? *RLG DigiNews*, 5(6).

<<http://www.rlg.org/preserv/diginews/diginews5-6.html#feature1>>

Lynch, Clifford, 1999. Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information. *D-Lib Magazine*, 5 (9).

<<http://www.dlib.org/dlib/september99/09lynch.html>>

Lorie, Raymond, 2001. Proof-of-concept: a UVC-based approach for preserving digital documents. Raymond Lorie. IBM Report, draft version December 2001.

Proposal for a digital preservation collaboration between the Yale University Library and Elsevier Science. Version 4. Date: 30 September 2000

Requirements for assessing the authenticity of electronic records. Authenticity Task Force. InterPARES Project. Draft for public comment. July 2001.

Rothenberg, Jeff, 1995. Ensuring the Longevity of Digital Documents. *Scientific American*, January 1995.

Rothenberg, Jeff, 2000. An experiment in using emulation to preserve digital publications. NEDLIB Report.

<<http://www.kb.nl/nedlib/results/emulationpreservationreport.pdf>>

Testbed Digitale Bewaring. Netherlands. This is a joint programme of the Dutch Ministries of Home Affairs, Education, Culture and Sciences. It aims to look into the issues of safekeeping digital government information.

<<http://www.digitaleduurzaamheid.nl/>>

Werf, Titia van der, 1999. Identification, location and versioning of web resources. URI discussion paper. DONOR Report, March 1999.