The Use of Metadata and Preservation Methods for Continuous Access to Digital Data

Groenewald, R
Digitisation Coordinator
Merensky Library, University of Pretoria

Breytenbach, A
Metadata Specialist
Veterinary Library, University of Pretoria

Abstract

Data loss prevention starts with the creation of a digital object. However, methods to minimize the loss of digital data are often ignored; the use of metadata structures embedded in digital objects from the outset thereof are recommended as a starting point towards good preservation principles.

The need to create awareness on the issue of digital preservation was promoted by the authors at various occasions during 2008 as the number of incidents of data loss and costs involved continue to be of concern to conservators. Whether the loss occurs by a malicious attempt, or an inadvertent mistake, it can be diminishing either personally or to the institute/company where it occurs.

Digital objects should be archived with metadata about the object and the creation thereof. Metadata need not necessarily be structured and controlled when used by individuals or small groups for preservation of self-owned data. The metadata content however, should describe the object, the method of creation and technologies used in the creation. All changes to the document should be captured in the preservation metadata.

Future access to digital content does not only depend on one preservation method but on a sequence of strategies and methods applied to the digital content. This paper discusses the use of metadata principles and the implementation of tools for the preservation of documents stored on personal computers.

Introduction

The Grainer Engineering Library indicated 1994 as the year that the Digital Libraries Initiative begun. According to the Library Timeline "the goal was to develop widely usable Web technology to effectively search technical documents on the Internet."

Libraries and museums and curators traditionally are the custodians of valuable artefacts and information. These valuables were acquired from individuals and other institutions, stored with well-managed conservation practices to ensure the long-term access thereof. Archives and libraries are now facing new access and preservation issues as personal papers that include digital media are being donated to them. (Steve Kolowich)

The term metadata, as an application method for digital document preservation, has established itself over the last few years, but is mostly applied in the form of descriptive metadata. Preservation metadata methods have not yet receive the same intensity of application to electronic documents, although these metadata sets are crucial for the preservation of the document format and to retain the significant properties (look and feel) of a document.

Information created in digital format and selected for archiving, need to be preserved in the format of creation without any restrictions embedded in the document. Material intended for web use is usually disseminated for fast, easy and clear access through the internet; this document does not serve the archival purpose. Although dedicated "web formats" also need to be able to migrate and have a sustainable "object life" it does not form part of the focus of this paper.
Social networks have an influence on the flow of digital data, personal documentation such as photos, life experiences (travel, events and personal opinions) are posted to blogs, wikis and other social websites such as Flickr and Facebook. The online method is used as replacements for former diaries, photo albums, letter boxes and filing systems of personal documentation. Metadata in the form of social tagging is usually very basic and descriptive in nature in these applications, thus neglecting preservation metadata. Web 2.0 tools orchestrate a clear shift to the individual as a publisher, and move the balance of power in relation to information away from the organisation.

When students and workshop participants were asked in 2008, to indicate whether they still have the first photo taken by them on a cell phone, the response was overwhelming negative. This is in contradiction to the paper print of many older “first photos” which are still in existence.

The PoWR Handbook, funded by JISC for web preservation, concentrates on strategies for the preservation of web material. In this handbook the MoSCoW approach to the selection of digital archival material is encouraged for digital preservation strategies.

M : Things you/institution must preserve
S : Things you should preserve, if at all possible
C : Things you could preserve, if it does not affect anything else
W : Things you won’t preserve

These principles can be applied in the selection of all digital material for long-term storage and archival purposes.

Methods of Study

1. Questionnaire

A questionnaire was circulated on two separate list-serves, of which the members mostly originated from South Africa. The purpose of the questionnaire was to determine the magnitude of awareness on digital preservation.

The outcome of the questionnaire indicates a lack of knowledge on preservation strategies and the management of digital objects on personal computers, as well as a need for training in basic digital preservation methods. It further indicates the usage of personal computers (office or at home) for storage of electronic documentation. Format types used, are indicated in a variety of different well known formats, with no indication towards unique formats for datasets etc.

2. Literature studies and visits

Literature studies were done on several strategies, policies and best practices. The information was mostly accessed online. The amount of articles recently published on the topic indicates a concern towards digital fragility and the need to create awareness amongst creators of digital content.

The DCCs (Data Curation Centre) Preservation Life Cycle has been studied and an implementation of the practical usage thereof will be discussed during the presentation. A digital object need to be actively managed at each stage of its life, and preservation strategies implemented from the creation thereof. The model describes sequential activities to ensure that all necessary stages are preserved.

Although the primary uses of PREMIS (a data dictionary for preservation metadata) are for repository design, this report has been used as a starting point for the use of preservation metadata.

Personal visits to Libraries in the UK and Egypt also create an opportunity to learn more about preservation methods of digital content. A visit to the Wellcome Library gave important insight in the starting process of initiating preservation methods.

3. Software

During the study the authors investigate practical implementations to preserve and retrieve digital content by individuals, with the emphasis on preservation. Although our focus had been on
preservation we realised that preservation practices cannot be fully implemented without a good content management system and search facility. Different types of software have been identified for preservation usage.

Content Management
- Joomla
- Alfresco (archival software)

Format Preservation
- Xena

Web 2.0
- Web Curator Tool

Search Facilities
- Windows Explorer
- Copernic

4. Metadata

A digital object does not have any meaning to a human being unless the content is described with descriptive, structural and technical (or administrative) metadata. Preservation applications must be accompanied by metadata to be successful.

Descriptive and preservation metadata assigned to digital research objects contains valuable information that can electronically be tracked by using the abovementioned software tools Technical (or administrative) metadata that consists of two categories namely preservation and rights management metadata, are created to archive and sustain continuous access to data, from the origination of the digital asset to the storage of the final format of the object. This metadata aids in the long-term management of digital material and needs to be embedded in the planning processes.

4.1 The following technical (or administrative) metadata categories have been included in the research -

(a) **Preservation**

Preservation metadata contains archival information, which is needed for the long-term preservation of the object and the migration to other digital formats as software and hardware changes continuously.

(b) **Rights management**

Technological mechanisms (Technical Protection Measures (TPM)), which restrict the usage of a digital object, can be embedded in the rights management metadata. Rights management metadata capture the permission of usage of an object and include the ownership, license information, restrictions on access, special permissions and methods of payment (if applicable).

4.2 The **OAIS** [Open Archival Information System (ISO 14721:2002)] model introduces four new categories to the conventional standard metadata structure. These categories are grouped under the term Preservation Description Information (PDI).

(a) **Reference Information**

The reference information includes, and enumerate on specific identifiers which were assigned to the data, i.e. referencing such as ISBN number or Uniform Resource Name (URN)

(b) **Provenance Information**

The history of the content information (e.g., its origins, chain of custody, preservation actions and effects) is captured in the provenance information. This form of metadata helps to support a digital object’s authenticity and integrity that is important for record-keeping and publication.
(c) **Context Information**
The context information index the relationship of the content to its environment (reason for creation, relationship to other data objects)

(d) **Fixity Information**
The fixity fields will document the authentication mechanisms, which in turn will ensure that the data is unaltered or show the extent of manipulation (e.g., checksum, digital signature). The checksum information can be used to implicate change in a stored file.

Completion of the above metadata has been applied to a variety of digital objects, during the course of the study. The purpose was to test the compliance of the metadata sets against the anticipated outcome for preservation of digital objects. Although the results of the testing of the metadata were according to our expectations, a need for effective software for the management and retrieval of the objects through the metadata become clear.

The increasing amount of stored information impacts on the accessibility and preservation issues. The speed of retrieval and the ability to reproduce or retrieve information (within 24 hours in most cases) is a key factor for future deliverance of data. The use of metadata to index a document's content and history and thereby making it searchable, reduces the amount of time it will take to retrieve a particular document. According to Rick Lawhorn the total worldwide digital archive capacity in the commercial and government sectors will grow to more than 27,000 petabytes or 27 exabytes by 2010.

5. **Practical preservation applications**

The term digital preservation refers to the preservation of materials that are digitally born and documents created with the use of imaging and recording technologies. Various views on the definition of preservation, and what is meant by preservation, exist. For the purpose of this study digital preservation is the preservation of digital materials for a period long enough in order for the object to survive the next generation of technology and software change, without damage to the original content of the source, which in turn should be able to be preserved further in the newer format for future access.

Repositories normally contain re-formatted copies of digital content for web display, i.e. MSWord converted to PDF-format. In most instances, the original digital object contains the archival value which should be stored as the master, ensuring that all changes can be tracked and interactions as well as relationships to other documents retained. Back-up copies of archival documents can be stored on trusted external hard drives and/or DVDs stored in controlled temperature conditions, inserted in acid-free pockets. However, the life span of this storage ware is dependent to the supportive technology.

6. **Metadata added to document**

Researchers, serious collectors of information and even users of information should know what guidelines to use for capture, management, storage and/or preservation of digital objects. The future of digitised and born digital material, require significant thought and action.

Adopting good practice at the outset of a document will increase the longevity of the digital content. Additional to automated generated metadata, a table containing metadata of the specific document can be included in the document as well as stored separately as a “side-car”.

The following is an example of additional information that can be added to the body of a digital document to explain the format, and workflow/history of a digital document for preservation.

<table>
<thead>
<tr>
<th>Document Title</th>
<th>The African Elephant: a digital collection of anatomical sketches as part of the University of Pretoria’s Institutional Repository - a case study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Breytenbach, Amelia and Groenewald, Ria</td>
</tr>
<tr>
<td>Description</td>
<td>Although several collections have been digitised and made available in the University of Pretoria’s Institutional Repository, a pilot study has not been done to measure the project management and workflow. The collections available in the repository at the time of this project were all long-term projects. There was a need to identify a project small enough to conform to normal project management requirements to use as an example to establish the planning and workflow of future projects. This paper offers practical help to libraries starting with digitisation, it supplies valuable information for project management, planning</td>
</tr>
</tbody>
</table>
of workflow and estimate time frames for completing a specific task in the digitization process.

An example of embedded metadata in a document such as PDF-format, this application in the software can be used for preservation metadata as it is searchable and will be migrated with the document format.

7. Management of digital content

The workflow of personal documentation can be managed through the use of a spreadsheet with metadata fields and a Content Management System (e.g. Joomla) with capturing functions, however, these functions need programming skills to be fully operational.

Apply consistent file naming conventions (abbreviations of file content without any spaces between letters), standardisation on formats used, and metadata descriptions. Valuable information on the stored information can be embedded in the object in the form of a “preservation indicator”. This will provide future users with the creators' assessment of the long-term intellectual and informational value of the object.

Compiling a spreadsheet for a collection i.e. images, will include all relevant preservation data, needed to recover a document. Description of the software used during the process, the workflow, editing dates, and
Specific manipulations done to the original images should be captured in such a document. This document needs to be stored separately from the original objects and preferably in more than one copy.

**Example of spreadsheet:**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Date</th>
<th>Time</th>
<th>Event</th>
<th>Location</th>
<th>Duration</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1.2022</td>
<td>9:00</td>
<td>30min</td>
<td>80.1</td>
<td>50.0</td>
<td>9:30-10:00</td>
<td>1990-1991 Dinosaur Excavation Site</td>
</tr>
<tr>
<td>6.2.2022</td>
<td>10:00</td>
<td>60min</td>
<td>80.2</td>
<td>51.0</td>
<td>10:30-11:30</td>
<td>1992-1993 Fossil Collection Site</td>
</tr>
<tr>
<td>6.3.2022</td>
<td>11:00</td>
<td>45min</td>
<td>80.3</td>
<td>52.0</td>
<td>11:45-12:30</td>
<td>1994-1995 Plant Specimen Site</td>
</tr>
</tbody>
</table>

**8. Search Functions**

The search function plays an important role in the retrieval of electronic documents. On a personal computer, the Microsoft Explorer search tool is commonly used for finding and retrieving information. However, this search engine is not effective enough in specific hits. The Copernic tool was found to be valuable for desktop searching. This software searches a variety of categories, i.e., email, contact lists, files, music, and the history of recently web visits with a structured display of the hits.

**Conclusions and Recommendations**

The study helped us to benchmark the current situation regarding preservation in South Africa against the global awareness and actions on this topic.

Negligence on format specifications and standardisation can cause huge data losses in the future and need further study towards a more simplified implementation of preservation strategies.

Storage and preservation of digital data need more attention in South Africa, and awareness towards preservation methods should be created amongst creators of digital content.

Training in the preservation of digital content and the actual delivery of plans and policies need to receive more attention in the corporate environment, especially towards digital content stored on personal computers.

Metadata, as a consistent, logical manner to keep documents accessible and usable over the long haul, need to be widely accepted and implemented, and the use thereof promoted to creators of digital objects.

**References**


