A survey of digital preservation strategies

Feng Luan*, Mads Nygård*
Department of Computer and Information Science, Norwegian University of Science and Technology, N-7031, Trondheim, Norway
Thomas Mestl+
Research and Innovation, Det Norske Veritas, N-1322, Høvik, Norway

* luan@idi.ntnu.no; mads@idi.ntnu.no

+ thomas.mestl@dnv.com

Abstract

Digital information is created easily by a wide range of electronic devices. But it is surprisingly difficult to preserve digital information over a long period, because both technology as well as relevant organizational context change over time. Without careful protection, there is a real risk that these data can neither be retrieved from old storage media, nor that its content be rendered from old file formats or understood. In the future, our society may indeed be confronted with a digital 'black hole'. From the 1990s, scientists and curators started to design long-term digital preservation systems and strategies that shall enable the survival of digital information. This survey article gives an overview of the current main preservation strategies. We further discuss and compare those strategies. Finally, we design two procedures for making information accessible and manipulable over time. At the end of the article, some emerging challenges for preservation strategies are addressed briefly.

Introduction

Digital preservation is the exploration of systematic approaches to ingest, archive, and disseminate digitized information. The development of digital preservation is heavily influenced by the use of electronic devices such as personal computers, electronic game devices, mobiles, digital cameras, digital recorders, and digital TVs. In the last decade, increased dependency of our modern society on digital information has lead scientists beginning to worry about digital preservation issues.

Earlier research work on digital preservation was carried out by the Commission on Preservation and Access and the Research Libraries Group (RLG) in 1994. This task force summarized the basic requirements of preservation and issued a report (D. Waters and J. Garrett 1996). Between 1996 and 2002, many libraries, archives, and data centres started to create digital repositories for their digitized information. The Consultative Committee for Space Data System (CCSDS) proposed an infrastructure model—the Open Archive Information System (OAIS) (The Consultative Committee for Space Data Systems 2002). Later, in 2003, the OAIS became an ISO 14721:2003 standard. Besides defining the infrastructure, efforts were more focused on preservation policies that would define a trustworthy preservation system. For example, R. L. Dale and **B.** Ambacher (2007) defined the trusted digital repositories audit and certification (TRAC); while S. Dobratz, A. Hanger, K. Huth, et al. (2006) defined the nestor criteria.

Preservation metadata is another central element in digital preservation. In terms of the OAIS, B. Lavoie and R. Dale (2002) compiled a metadata dictionary, which is an aggregation of the CURL Exemplars in Digital Archive project (CEDARS), the National Library of Australia (NLA), the Networked European Deposit Library (NEDLIB), and the Online Computer Library Center, Inc. (OCLC). The National Library of New Zealand also proposed a metadata dictionary (the National Library of New Zealand 2003) for automatic metadata extraction. The PREMIS (PREservation Metadata: Implementations Strategies) working group from 2005 started to define a preservation metadata standard. The latest version is the PREMIS 2.0. In total, there are 13 preservation systems[1] that have deployed PREMIS 2.0.

[1] http://www.loc.gov/standards/premis/premis-registry.php

This article focuses on the very core of a preservation system, i.e., how to ensure that digital information can survive for a long period and the existing strategies. The readers will hopefully get a quick review about:

- popular preservation systems around the world
- various preservation strategies
- selecting the most appropriate preservation strategy
- possible challenges for the preservation strategies

The article starts with an overview of related work on digital preservation, including the OAIS standard, preservation systems, and research work groups, in Section 2. Threats to digital information are summarized in Section 3, and various preservation strategies are introduced in Section 4, followed by a further discussion and analysis of the strategies in Section 5. Finally, we summarize future research works in Section 6.

Related work

The OAIS standard

The CCSDS published a report, the Reference Model for an Open Archival Information System (OAIS) (The Consultative Committee for Space Data Systems 2002). The OAIS became an ISO Archiving Standard in 2003. Many preservation systems or repositories adapted the OAIS standard as their infrastructural model. Figure 1 gives an overview of the OAIS standard.

Feng
Delet

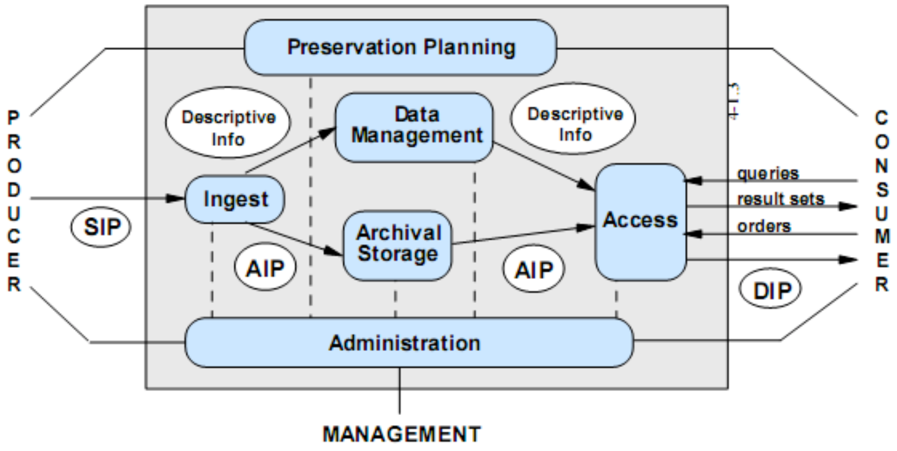

Figure 1 Overview of the OAIS standard, from (The Consultative Committee for Space Data Systems 2002)

The OAIS standard defines three roles for a person: producer, management, and consumer. A producer produces digital information. As long as the producer wants to submit digital information to a preservation system, the producer should bundle the digital information with metadata into a submission information package (SIP) based on the requirements of the preservation system. After receiving the SIP, the preservation system will divide the SIP to an AIP (Archival Information Package) and related descriptive information. The AIP will be maintained in the long-term archival storage. The description information will be stored in a database of data management so that any consumer can browse through them.

When a consumer wants to access an AIP, the preservation system retrieves the AIP from the archival storage and transfers it to the access module. The access module then transforms the AIP to the DIP (Dissemination Information Package), and sends the DIP off to the consumer.

A manager with proper management rights is responsible for any necessary maintenance task in the preservation system. The manager should design the charter of the preservation system such as the access scope for producers and consumers, and the agreements with producers and consumers. In addition, the manager should design the preservation plan so that all AIPs can be accessed and manipulated in future.

Preservation systems

Most of the well-known preservation systems are open-source systems that can be used free of charge by libraries and digital repositories. Some systems also offer free discussion groups.

LOCKSS

LOCKSS[2] (**Lo**ts of **C**opies **K**eep **S**tuff **S**afe) is an open-source system under the auspices of the Stanford University. LOCKSS is a distributed, peer-to-peer preservation system that is able to manage multiple copies at remote data repositories. The system was released in 2004 and had been tested at more than 50 libraries worldwide.

Eprints

Eprints[3] is a set of open-source software applications for building open access services. Eprints was developed by the University of Southampton in UK. The current version is Eprint 3, offering improvements in the architecture, automatic extraction of metadata, access control, flexible workflows, format support, and thumbnails.

DSpace

DSpace[4] is an open-source software system developed by the Massachusetts Institute of Technology Libraries and Hewlett-Packard. The objective of DSpace is to provide a repository for research data sets and educational materials. DSpace can preserve not only digital journals and digitized documents, but also 3D digital objects, research data sets, and films.

e-Depot

e-Depot (E. Oltmans and H. V. Wijngaarden 2004) is a long-term preservation system developed by the National Library of Netherlands. The core part of e-Depot is the DIAS[5] that was developed by IBM. The objective of e-Depot is to maintain the integrity of stored digital objects so that these objects are accessible.

FEDORA

FEDORA[6] (**F**lexible **E**xtensible **D**igital **O**bject **R**epository **A**rchitecture) is constructed under an open-source digital preservation infrastructure. FEDORA began in 1997, headed by the Cornell University and University of Virginia. In the latest version, FEDORA provides not only the basic functions of preservation systems, but also a model using semantic techniques, i.e., the Resource Description Framework (RDF) maintaining relations between digital objects.

iRODS

iRODS[7] (**i**ntegrated **R**ule **O**riented **D**ata **S**ystem) is an open-source data grid software system. Developed by the Data Intensive Cyber Environments (DICE) research group and collaborators, it is the successor of the Storage Resource Broker (SRB). A peculiar feature of iRODS is its ability to represent the data preservation policies in a set of rules. Thereby, iRODS can interpret the rules and execute a sequence of pre-defined actions based on a given situation.

Recent research work groups

Current research groups that focus on the long-term preservation of digital information are:

PADI

The National Library of Australia's Preservation Access to Digital Information8 (PADI) is a gateway to digital preservation issues. PADI provides rich resources about every aspect of a preservation system such as

[2] http://lockss.stanford.edu/lockss/Home

[3] http://www.eprints.org/

[4] http://www.dspace.org/

[5] http://www-935.ibm.com/services/nl/dias/is/implementation_services.html

[6] http://www.fedora-commons.org/

[7] https://www.irods.org/

[8] http://www.nla.gov.au/padi/index.html

archiving, preservation strategies, data documentation and metadata, intellectual property rights management, format and media, management, digitization, approaches at national libraries, and digital preservation tools.

InterPARES

InterPARES[9] (**Inter**national Research on **P**ermanent **A**uthentic **R**ecords in **E**lectronic **S**ystems), starting in 1999, has currently entered the third phase. In total, 15 teams from different countries are included in InterPARES. The objective of InterPARES is to develop the necessary knowledge for survival of digital record over a long-term period. InterPARES mainly works on the theoretical foundation for a preservation system, including aspects such as access, creation, maintenance, and security.

DCC

DCC[10] (**D**igital **C**urator **C**entre) is a discussion centre for digital curators. It is funded by the Joint Information Systems Committee (JISC). DCC provides rich knowledge resources about digital preservation such as a curation reference manual, curation lifecycle model, policies and legal reports, case studies, tools and applications, standards, publications, and a curation journal.

CAMiLEON

CAMiLEON[11] (**C**reative **A**rchiving at **Mi**chigan & **L**eeds: **E**mulating the **O**ld on the **N**ew) was funded by the NSF/JISC. CAMiLEON explores various ways to keep the original functionality and 'look and feel' of digital objects. The emulation and migration on access, which will be introduced in Section 4, are key outputs of CAMiLEON.

PLANETS

PLANETS[12] (**P**reservation and **L**ong-Term **A**ccess through **Net**worked **S**ervices) is funded by the European Union under the Sixth Framework Programme of 2006. The PLANETS objective is to improve decision-making in long-term preservation, so that the valued digital objects can be accessed. It has investigated several digital preservation challenges such as a preservation plan, a set of characterizations of digital objects, preservation actions, an interoperability framework, and a test bed.

CASPAR

CASPAR[13] (**C**ultural, **A**rtistic and **S**cientific knowledge for **P**reservation, **A**ccess and **R**etrieval) is another project funded under the Sixth Framework Programme. CASPAR, implemented based on the OAIS guidelines, focuses on:

- what metadata are needed to describe the representation information and other relevant information
- how to integrate digital intellectual property rights for the preserved digital information
- how to integrate authentication and accreditation into the long-term preservation mechanism

SHAMAN

SHAMAN[14] (**S**ustaining **H**eritage **A**ccess through **M**ultivalent **A**rchivi**N**g) is financed by the European Union within the Seventh Framework Programme. The goal of SHAMAN is to develop new approaches for digital preservation. Those approaches shall not only guarantee preservation of digital content, but will also keep track of the digital content's integrity, authenticity, semantics, and usage context.

[9] http://www.interpares.org/

[10] http://www.dcc.ac.uk/

[11] http://www2.si.umich.edu/CAMILEON/

[12] http://www.planets-project.eu/

[13] http://www.casparpreserves.eu/

[14] http://shaman-ip.eu/shaman/

Threats to preservation of digital information

Traditional information carriers such as paper and stones can preserve information over long periods, for example, several decades or even hundreds of years. In contrast, our experience with digital information being generated in the last half century has shown us that digital information is confronted with considerably more threats than traditional information. The reasons are mainly due to its dependence on both storage media and interpretation software. Table 1 gives a short description of the main threats.

Feng
Delet

Table 1 Threats to digital information

Threat	Description
Disasters/accidents	A natural/man-made disaster or accident destroys the preservation system.
Storage media fault	A storage medium has faults that make it unreadable.
Hardware/storage media obsolescence	Hardware or storage medium becomes too old, making it difficult to find a replacement part or get technical support.
Software/format obsolescence	If software/format is too old, the preservation system will replace it.
Malicious attacks	Malicious attacks on a preservation system can result in information modification or even loss of information.
Lack of context	A preservation system stores too little context information for correct interpretation/understanding of the preserved information.
Lack of authenticity	Since very few evidences are stored, the preservation system cannot prove authenticity of the preserved digital information.
Financial problems	Too little funding could threaten the necessary operation of a preservation system.

Disasters/accidents

Disasters (e.g. flooding, earthquakes or terrorist attacks) and accidents (e.g. fire) could damage the necessary electronic or storage equipment leading to (partial) loss of digital information.

Storage media fault

Storage media are the core components of any preservation system as they hold the digital information in terms of physical characteristics, for example, magnetism and optics. The quality of those characteristics degrades as time passes. Thereby, faults would appear in a storage medium. Such faults are usually correctable unless the number of faults reaches a critical point, when (parts of) the information might be lost.

Hardware/storage media obsolescence

Hardware refers to any component of a computer system. The rapid evolution of hardware and storage media can make it difficult to access information on old storage media. For example, floppy disks were the standard storage media from mid-1970s to late 1990s, but 5¼-inch floppy disc readers are no longer available now. It will be only a matter of time before optical disks are replaced by memory sticks or online storage.

Software/format obsolescence

Software provides a way to organize and manipulate digital information according to pre-defined format structure(s), whereas formats must rely on software for correct interpretation of the sequence of 0s and 1s (bits). Software and file formats largely depend on each other. Any change in one, must be accompanied by an adaption of the other. The current rapid development in computing technology such as functionalities, increased performance, and new and better graphical interface, can result in a situation where old file formats can no longer be interpreted correctly by the newer software, and vice versa.

Malicious attacks

A preservation system that stores valuable information for an organization represents a potential target for malicious attacks such as a disgruntled employee and institutional espionage.

Lack of context

The context provides necessary background information for correct interpretation, understanding, and use of digital information. This usually includes a profile of digital information, information about the creation intention, and relations to other digital information. Lesser the context provided with the digital information, more difficult it will be for a future user to understand the data.

Lack of authenticity

Authenticity provides information on whether the preserved information is same as (or similar to) the original, and whether any modification on the information is done by authorized actions/personnel. Authenticity information includes annotated text, screenshots, annotation to screenshots, video clips, digital signatures, access control, and so on. If the evidence is incomplete or lost, trustworthiness of the information is in doubt.

Financial problems

Digital preservation requires a high volume of funding, for example, hardware/software update, operational costs, personnel, information acquisition, insurance, unavailability, and cost of losing document (A. Crespo and H. Garcia-Molina 2001). Therefore, any successful long-term preservation system depends on steady financial support.

Preservation strategies

As digital information is highly dependent on computer technology, its preservation will necessarily have to be different from that of traditional, or printed information. It is not enough to just preserve the digital storage media (equivalent to paper), but interpretation and manipulation software (equivalent to the language and understanding capability of the human reader), as well. Analysis of a generic read-and-write process for digital information may provide some clues for potential preservation strategies.

Consider the situation depicted in Figure 3 (below) where a text entry (e.g. 'World Digital Libraries: an International Journal') shall be modified. The text data are stored in terms of the ASCII code. In the first phase, the text data are read from the storage media, and the bit stream is sent off via the storage driver into a software application (left side of figure). In the second phase, the software translates the bit stream into a human readable and understandable rendering (right side of the figure). After having modified the journal entry, this process is reversed.

Feng
Delet

This simple generic example highlights two fundamental aspects in digital preservation: storage accessibility (phase 1) and bits manipulability (phase 2). The storage accessibility requirement shall ensure the correct reading and writing to and from the data storage, whereas the bits manipulability requirement shall guarantee that the data can be made operational.

With this perspective in mind, four strategies could be envisioned that would maintain storage accessibility, and six strategies for maintaining bits manipulability. In the next section, we will shortly describe these preservation strategies. The taxonomy of the 10 preservation strategies is shown in Table 2.

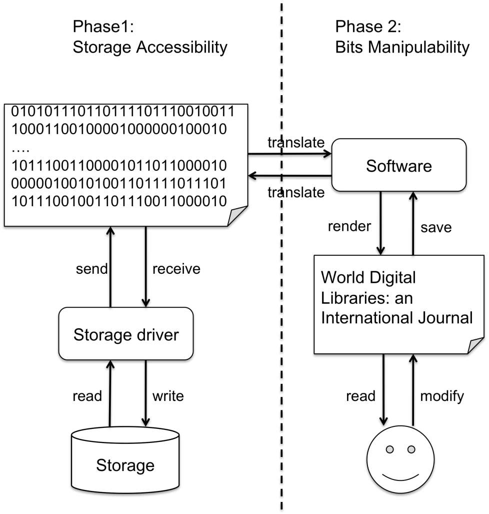

Figure 3 Typical phases when using digital information

Feng
Delet

Table 3 Taxonomy of preservation strategies

Feng
Delet

Classification	Strategy	Description
Storage accessibility maintenance	1. Auditing	The state of the storage media is periodically inspected by an auditing application.
	2. Refreshment	The hardware/storage media are periodically replaced by a new and healthy alternative.
	3. Replication	Digital information should have multiple copies.
	4. Federation	Every preservation system within the federation is connected in a distributed environment.
Bits manipulability maintenance	1. Computer museum	Necessary hardware/software is preserved for accessing and manipulating digital information from the old storage medium.
	2. Emulation	An emulation application is used to simulate the function of old computer environment. Moreover, the original programme can be executed in the emulation application.
	3. Encapsulation	The format specification is embedded with the digital information. The preservation system can develop new software in terms of the stored format specification.
	4. Universal virtual computer	The digital information and the necessary software are converted to a set of particular instructions, which can be translated by a virtual computer.
	5. Batch migration	The preservation system should periodically transform the old formats to new formats.
	6. Migration on access	The format transformation is executed only when readers try to manipulate the digital information.

Storage accessibility maintenance

Storage accessibility requires that digital information can be completely retrieved from — or written back to — a storage medium. Bit error or hardware faults can threaten storage accessibility. M. Baker, M. Shah, D. Rosenthal, et al. (2005) have analysed the reliability of storage media, and they argue that storage reliability can be increased by using better storage media, auditing data more frequently, automatic storage media repair, providing hot space drives, increasing the number of replicas, and increasing independence of the replicas. Accessibility could, therefore, be guaranteed by auditing, refreshment, replication, and federation.

Auditing

Auditing is an inspection procedure that looks for faults in a storage medium. Manual auditing is not feasible for large-scale preservation systems, where the volume of information may be too large. In addition, it may be difficult for administrators or readers to understand all the audit information. Once a fault is found, the preservation system should repair the storage as soon as possible. Otherwise, the information might be lost. For example, if all copies of digital information have various faults at the same time, the preservation system may not be able to recover them anymore. Thereby, auditing needs not only an application to find faults, but also an application to repair the faults as soon as possible.

Refreshment

Refreshment is a procedure where an old storage medium is replaced by a new one by copying the bits from the old medium to the new. This term differs from the OAIS' definition that simply refers to copying between storage media within the same type and does not permit any alteration to the storage mapping infrastructure. However, we deem that refreshment can be performed between any two storage media, and that the storage mapping infrastructure can also be modified. Thus, refreshment includes replacement of a storage medium and updating the entire storage system. Refreshment is considered relatively simple, but it is not easy to determine the time when the refreshment should be executed. Most preservation systems perform refreshment every 3–4 years to avoid expiry of warranty of the storage media. However, at that time, most of the storage media are still in a perfect working condition, and hence, a lot of money is wasted in the process.

Replication

Replication refers to an approach where the preserved information has several copies using different formats that might be saved at different places. This definition also differs from that of the OAIS. The difference lies in that the OAIS requires package information, content information, and preservation description information cannot be changed, whilst we deem that digital information should use different formats. In this way, it results in enhancement of reliability of digital information. Similar to refreshment, replication is a simple but expensive approach, as it requires more advanced technology and larger storage space for replicas. Most preservation systems use the RAID (Redundant Array Independent Disks) technique, which offers duplication of data and a better I/O performance. The preservation systems also allow usage of several formats. For instance, the National Library of Norway uses the PDF format for their publications, XML format for full-text search, TIFF format for preservation, and JPEG2000 format for dissemination on the Internet.

Federation

Federation refers to a distributed network environment, where each independent preservation system shares its digital information with other preservation systems. If the digital information is not readable, the preservation system can recover this information from other preservation systems. Federation is a promising approach because it not only overcomes the threats to accessibility, but also reduces the dangers related to financial problems. Recent preservation systems such as the LOCKSS and DSpace prefer building a federation environment. However, there are still some problems related to this approach such as synchronization issues among the systems, and access right and confidentiality issues.

Bits manipulability maintenance

Bits manipulability requires that users can successfully manipulate the bit stream in a manner understandable by humans. The key in this requirement is the technical environment, which supports software (e.g. format interpreter) that can be run. Software or format obsolescence is the typical threat to tackle which six practical strategies can be proposed.

Computer museum

The computer museum strategy requires that the whole computer system, including hardware and software, should be preserved, so that the preserved digital information can be read and manipulated in the original environment. For example, the National Library of Norway preserves old audio devices since they often get old audio media from other organizations or individuals. This approach is only suitable in the short-term, because:

- Every hardware component has a limited lifetime.
- Currently, most hardware and software are proprietary. It is impossible for a preservation system to produce the old hardware and software by itself.
- Hardware and software vendors might not support all of their previous products.
- In some situations, it is more expensive to maintain old hardware/software than to buy a new one.

Emulation

Emulation uses a software application to imitate the function of an old hardware component. In this way, the preservation system is able to execute the old software without the real hardware component. When using emulation, the relevant software and file formats can be handled at the same time. This makes emulation a useful strategy for complex digital information such as databases and computer games. However, S. Granger (2000) has argued that emulation is not suitable for long-term preservation because:

- The preservation task becomes more complex, since the preservation target switches to the whole computer system rather than just digital information.
- Information management becomes harder because the preserved information is dispersed over various emulation applications.
- It is difficult to develop an emulation application. Thus, the preservation system has to rely on an external entity.

Encapsulation

In encapsulation, a new software application is developed that is able to manipulate the preserved digital information. Encapsulation requires that the archive package must preserve the digital information and relevant format specifications. The main preservation task of encapsulation is to develop a new software application. Encapsulation is not feasible in the long term according to (A. Waugh, R. Wilkinson, B. Hills, et al. 2000). They wrote about three major challenges:

- There may be no application that can automatically generate the archive package for the encapsulation. The preservation system should automatically perform this function.
- The storage overhead of encapsulating the format specification with digital information might be a problem.
- Some formats are proprietary and their specifications may be unpublished. A part of the encapsulation approach might include a transformation from unpublished, proprietary formats to published, open formats.

In addition, we may point out another challenge, i.e. the formats in the preservation system are too many to create the manipulation applications as time passes by. The cost for developing such applications might surpass the preservation system's ability. This situation might be mitigated by converting the original format to some standard format.

Universal virtual computer

The universal virtual computer (UVC) proposed by R. A. Lorie (2001; 2002) is an innovative approach. UVC is similar to the JAVA Virtual Machine. Digital information and the relevant and original software applications are compiled with a set of special instructions of UVC. For each computer generation, the digital information can be manipulated by the original application in UVC. The National Library of Netherland has used such an UVC for JPEG and PDF files manipulation, but more pilot projects are needed to test this approach for other formats. Moreover, Lorie mentioned that UVC may be hard to optimize.

Batch migration

Batch migration aims to overcome the threats of software obsolescence and format obsolescence in that the preservation system periodically transforms digital information from one format to another. The process seems simple, but requires a careful experiment and validation phase, where alternative solutions should be tested. S. Strodl, C. Becker, R. Neumayer, et al. (2007) used utility analysis to determine the preferred format. Several case studies at many European libraries show that utility analysis is a feasible approach for batch migration. However, the disadvantages of the process are:

- Digital information is not preserved in its original format, and consequently a degradation of the digital information might occur.
- The time to execute the batch migration will become a challenge. In the latest system update, it took the National Library of Norway more than three months to do refreshment. When doing batch migration, more time is needed. A dangerous scenario could emerge where the old batch migration is still not finished before a new batch migration has to be started.

Migration on access

Migration on access was proposed by P. Mellor, P. Wheatley, and D. M. Sergeant (2002) where the transformation from the original format to a current one is only executed when the preserved digital information is accessed. D. S. H. Rosenthal, T. Lipkis, T. S. Robertson, et al. (2005) believe that a characteristic feature of migration on access is to have the original file that maintains its authenticity and integrity. Moreover, migration cost can be considerably reduced. However,

- Migration on access will cause additional delay while accessing.
- Migration on access will increase the workload for the preservation system.
- Migration on access should be closely integrated with the dissemination process.
- There must exist a file format converter from the original format to the current format.

Further discussion

The 10 preservation strategies presented above are further analysed in this section. Their intended use will be discussed later on.

How to use the strategies to maintain storage accessibility?

Table 4 (below) illustrates the advantages and disadvantages of the aforementioned strategies for maintaining accessibility. The strategies cannot be ranked and should have the same priority in a preservation system. For instance, the current preservation policy of the National Library of Norway is to use replication and refreshment, where XML files, PDF files, TIFF files, and JPEG2000 files are created for a digital information object. All these files are packaged and stored in a long-term storage system that consists of two tapes and one RAID storage system at two different places. For refreshment, they replace some tapes and hard disks every 3–4 years, when the warranty expires.

Feng
Delet

Table 4 Advantages and disadvantages of strategies for maintaining accessibility

Feng
Delet

	Advantages	**Disadvantages**
Auditing	• Cheap • Find faults in time • Postpones refreshment	• Increased workload
Refreshment	• Simple • High reliability	• Cost of buying new storage • Time consuming • Unknown refreshment time point
Replication	• Simple • High reliability	• Cost of obtaining large storage space • Cost of obtaining advanced technology
Federation	• Independence • High reliability	• Complex • Published digital information

Replication and refreshment are widely used by other preservation systems. However, since no audit application is used, threats like latent, irrecoverable faults might exist. The best preservation policy might be to use all four strategies together. Figure 5 illustrates such an approach where four strategies are deployed simultaneously for a preservation system.

Feng
Delet

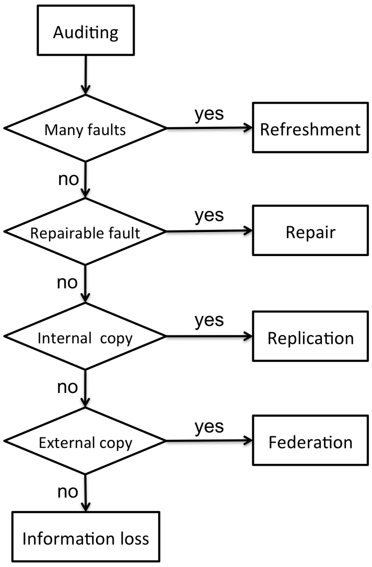

Figure 5 Preservation approach when using all four strategies

Feng
Delet

How to use the strategies to maintain bits manipulability?

The strategies for maintaining bits manipulability lead to other priorities for a preservation system. The administrator of the preservation system usually chooses one strategy as the main one. Other strategies are only used for some special digital information objects in a short period. The required material or documents for each strategy are illustrated in Table 6 (below). In total, all the strategies need six types of materials and documents:

Feng
Delet

- **Change in bits** means for a given strategy how many times the bits of the digital object have been changed over its lifetime. Possible answers are '*no change*', '*changed once*' or '*changed many times*'.
- **Hardware specification** means whether a strategy should have hardware specifications. Possible answers are '*no*' or '*yes*'.
- **Hardware component** means whether a strategy should have previous hardware components. The possible answers are '*no*' or '*yes*'.
- **Format specification** means which type of format specification should the strategy have. Possible answers are '*the original format specifications*', '*the latest format specifications*', '*the mediatory format specifications*' or '*N/A*'.
- **Format converter** means which type of format converter should the strategy have. Possible answers are '*from the original format to a mediatory format*', '*from the format currently used to a new format*', '*from the original format to a new format*' or '*N/A*'.
- **Format interpreter** means which type of format interpreter application should the strategy have. Possible answers are '*the original interpreter*', '*the current interpreter*' or '*the new interpreter*'.

Table 6 Requirements related to the strategies for maintaining manipulability

Feng
Delet

	Computer museum	**Emulation**	**Encapsulation**	**UVC**	**Batch migration**	**Migration on access**
Change in bits	No change	No change	No change	Changes once	Changes many times	Changes once
Hardware specification	No	Yes	No	No	No	No
Hardware component	Yes	No	No	No	No	No
Format specification	N/A	N/A	The original format's specifications	The mediatory format's specifications	The latest format's specifications	The original format's specifications
Format converter	N/A	N/A	N/A	From the original format to a mediatory format	From the format currently used to a new format	From the original format to a new format
Format interpreter	The original interpreter	The original interpreter	The new interpreter	The new interpreter	The current interpreter	The current interpreter

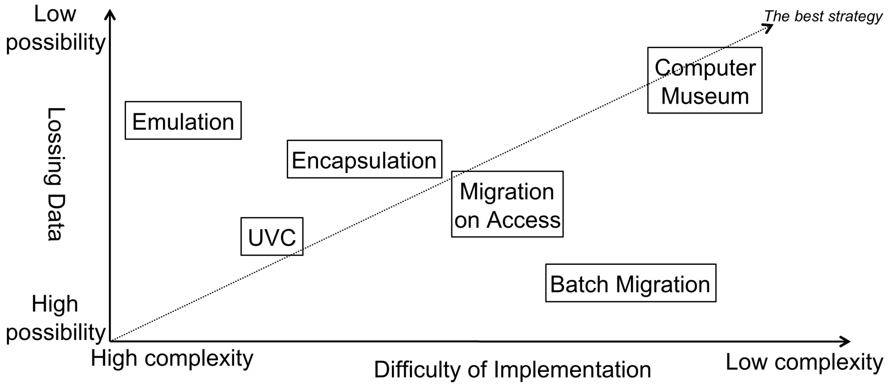

Figure 7 Ranking the strategies for maintaining manipulability

Feng
Delet

Based on these criteria, two dimensions, i.e., the possibility of losing data and the difficulty of implementation, are used to rank the strategies (see Figure 7). The possibility of losing data depends on change in bits, format converter, and format interpreter. Change in bits and format converter include the format transformation procedure. If a format is not compatible with another format, some parts of the digital information object may be lost during format transformation. Moreover, old formats have a higher possibility of incomprehension. This is because the technique may be so old that little is known about it. Even the specification might be difficult to understand. Format interpreter includes developing interpretation procedures

Feng
Delet

for a preserved format specification. The interpreter procedure might not work well if the format's specification is not well documented or it is too complex to understand. Thereby, some content of the digital information may not be rendered.

The difficulty of implementation will depend on hardware specification, hardware component, format specification, format converter, and format interpreter. To implement any strategy, the first difficulty is to preserve relevant material and documents. The less material and documents a strategy needs, easier is its implementation. The second difficulty may lie in understanding the relevant technique specifications, including hardware specification and format specification. In general, the older a specification is, the more difficult it might be to understand. The third and final difficulty is the implementation target. For instance, emulation needs emulation applications; batch migration and migration on access need format transformation applications; while encapsulation and UVC need format interpretation applications. From practical experience, it can be concluded that emulation application is the hardest to implement. Format transformation applications might also be difficult since developers need to know the relation between the two formats. It seems that the easiest implementation is the format interpretation application.

Figure 8 (below) further illustrates the above-mentioned preservation strategies with the help of a selection diagram. The selection depends on the preservation period, i.e. one, several, and many computer generations. If the digital information is to be preserved for just one computer generation, then the computer museum might be the ideal choice as the necessary hardware and software may still be easily obtainable from the market. However, this strategy does not scale up for a longer preservation period, as compatible hardware and software alternatives become increasingly rare with time.

Feng
Delet

For preserving over many computer generations, batch migration to the latest computer technology seems to be the most viable solution. Currently, batch migration is the most widely used approach in long-term preservation. Researchers believe that migration could avoid loss of data. P. Wheatley (2001) summarized the batch migration activities for digitized materials of BBC. The Digital Preservation TestBed (2001) published a report on the practices of migration. P. Caplan (2007) introduced the migration solution in the Florida Digital Archive. C. Becker, H. Kulovits, M. Guttenbrunner, et al. (2009) described a systematic evaluation procedure to select the best format for batch migration. Distributed technology is also used for batch migration. For example, PANIC (J. Hunter and S. Choudhury 2006) and CRiB (M. Ferreira, A. Baptista, and J. Ramalho 2007) are two web migration projects, which use web service techniques to find format transformation services and automatically choose the most appropriate service to do the batch migration. In addition, some supplementary resources for batch migration are published online. For example, format information can be dynamically extracted from format registries such as GDFR[15] (Global Digital Format Registry) and PRONOM[16].

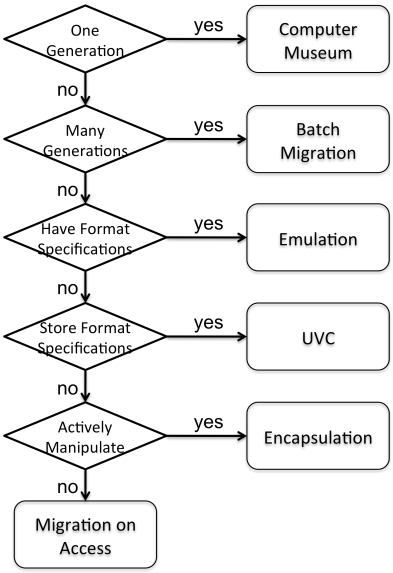

Figure 8 Selection of preservation strategies for bits manipulability maintenance

Feng
Delet

[15] www.gdfr.info/

[16] www.nationalarchives.gov.uk/PRONOM/Default.aspx

For preserving over several computer generations, emulation, UVC, encapsulation, and migration on access should be taken into account. If the preservation system does not have any format specification, then emulation may be chosen, as it does not require any format specification. If the preservation system has a format specification but does not choose to preserve it, UVC might be chosen since the format will be converted to UVC instructions. If the digital information can be actively manipulated, the preservation system should use encapsulation.

Concluding remarks

Preservation systems are a set of complex information systems, containing an ingest function, an archive function, a dissemination function, a search function, a user interface function, and a management function. From the management perspective, we have identified three main challenges:

Very large volume of data

J. Gantz, C. Chute, A. Manfrediz, et al. (2008) released a survey report on the state of the digital world. They estimated that in 2007, the total size of the digital world was 281 Exabyte[17], and that the annual growth in the digital world is approximately 60%. In 2011, the amount of digital information should be roughly equal to 1800 Exabyte. Moreover, they believe that the volume of digital objects might exceed the storage capacity. They have found that in 2008, the volume of information created exceeded the capability of storage media. If this is correct, then it can be expected that twice as much information will be generated in 2011.

Besides possible shortage in storage capacity, storage performance may be another challenge. In 2007, the National Library of Norway spent nearly three months to transfer all of its digital information objects to the new storage system. We tested the read/write performance of normal storage media, for example, a RAID storage system, a hard disk of desktops, and a memory stick. The RAID storage system had the highest performance because of its read-and-write speed of about 110MB/second. The hard disk had read-and-write speeds of about 50 MB/s. The memory stick, with read-and-write speeds of about 10MB/second and 1.5MB/second, respectively, was the slowest. Thus, if all the 957 Terabyte of data in the National Library of Norway (in 2007) were to be transferred, 116 days, 232 days, and 7743 days would be respectively required for the RAID storage system, the hard disk, and the memory stick. Therefore, as the amount of digital information rises exponentially, there is real concern that the time required for the next system replacement will far exceed the support time, i.e. the previous copying process of all data is still not finished before the next has to be initiated.

Lack of a comprehensive preservation strategy

Storage access and bits manipulability are two basic requirements. In order to ensure that the digital information objects are useful, two more requirements, i.e. content understandability and object trustworthiness are needed. The content understandability requirement means that readers should obtain relevant context information to understand the meaning of the content of the digital information object. The object trustworthiness requirement means that readers should be able to consider the preserved digital information objects as accurate and authentic in terms of the preserved evidences.

Among all the above-mentioned preservation strategies, emulation seems to be the most suitable and easiest strategy that could satisfy these four requirements. However, this strategy is difficult to implement.

Regarding the other preservation strategies, especially batch migration and migration on access, most efforts are employed at the levels of accessibility and manipulability. Little research is directed towards understandability and trustworthiness issues, thereby indicating that a more comprehensive strategy maybe needed.

What metadata should be preserved?

Metadata are data about data. They provide supplementary information about a digital information object. Currently, there are numerous metadata standards for information management such as the Dublin Core (The Dublin Core Metadata Initiative 2008), the Moreq 2 (Moreq2 2008), the ISO 23081 (ISO 23081 2009), the PREMIS (the PREMIS Editorial Committee 2008), and the OAIS metadata model (B. Lavoie and R. Dale 2002). Even though researchers agree that metadata are essential for a successful preservation strategy, however, till date, there are only a few metadata standards proposed. We believe that metadata research for a long-term preservation system should also cover the following aspects:

[17] 1 Exabyte = 1024^3 Gigabyte = 1024^6 Byte

- What metadata are necessary for preservation strategies?
- What metadata can improve authenticity of the preserved digital objects?
- What metadata should be recorded after a preservation strategy is carried out?
- How can the history of a preserved digital object be visualized?

Conclusion

Ingest, archive, dissemination, search, interface, and management are the important functions for any preservation system. This article focused on the management issues. We described and assessed 10 different preservation strategies. Auditing, refreshment, replication, and federation are often used together to keep bit integrity, i.e. storage accessibility maintenance. Computer museum, emulation, encapsulation, UVC, batch migration, and migration on access are often used to offer bits manipulability. The objective of the preservation system is to determine which of those six strategies should be selected.

Acknowledgement

The research work for this article was funded by the Norwegian Research Council and our industry partners as part of the LongRec[18] project.

References

A. Crespo and H. Garcia-Molina. 2001. **Cost-driven design for archival repositories**. Proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries, Roanoke, Virginia, United States, ACM.

A. Waugh, R. Wilkinson, B. Hills, and J. Dell'oro. 2000. **Preserving digital information forever**. DL '00: Proceedings of the fifth ACM conference on digital libraries: 175–184.

B. Lavoie and R. Dale. 2002. **Preservation of Metadata and the OAIS Information Model. A Metadata Framework to Support the Preservation of Digital Objects.** Published by the OCLC/RLG Working Group on Preservation Metadata.

C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. 2009. **Systematic planning for digital preservation: evaluating potential strategies and building preservation plans**. *International Journal on Digital Libraries* **10**(4):157.

D. S. H. Rosenthal, T. Lipkis, T. S. Robertson, and S. Morabito. 2005. **Transparent Format Migration of Preserved Web Content**. *D-Lib* **11**(1). [From http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html, last accessed on 25 November 2010].

D. Waters and J. Garrett. 1996. **Preserving Digital Information: Report of the Task Force on Archiving of Digital Information**. Commissioned by the Commission on Preservation and Access and the Research Libraries Group, 71p.

E. Oltmans and H. V. Wijngaarden. 2004. **Digital preservation in practice: the e-Depot at the Koninklijke Bibliotheek**. *Vine* **34**: 21.

ISO 23081-2. 2009. **Information and documentation -- Managing metadata for records.** Published by the International Organization for Standardization, 33p.

J. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva. 2008. **The Diverse and Exploding Digital Universe: An updated forecast of worldwide information growth through 2011**. A report sponsored by EMC [From http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf, last accessed on 25 November 2010].

J. Hunter and S. Choudhury. 2006. **PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services**. *International Journal on Digital Libraries* **6**(2): 174–183.

M. Baker, M. Shah, D. Rosenthal, M. Roussopoulos, P. Maniatis, T. J. Giuli, and P. Bungale. 2005. **A Fresh Look at the Reliability of Long-term Digital Storage**. Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006, EuroSys '06, pp. 221–234, ACM, 2006.

[18] www.longrec.com

M. Ferreira, A. Baptista, and J. Ramalho. 2007. **An intelligent decision support system for digital preservation**. *International Journal on Digital Libraries* **6**(4): 295–304.

Moreq2. 2008. **The MoReq2 Specification and Metadata Model**. [From www.moreq2.eu/index.htm, last accessed on 6 August 2009].

P. Caplan. 2007. **The Florida Digital Archive and DAITSS: a working preservation repository based on format migration**. *International Journal on Digital Libraries* **6**(4): 305.

P. Mellor, P. Wheatley, and D. M. Sergeant. 2002 **Migration on Request, a Practical Technique for Preservation**. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, Springer-Verlag, pp. 516–526.

P. Wheatley. 2001. **Migration–a CAMiLEON discussion paper**. *Ariadne* **29**(2). [From http://www.ariadne.ac.uk/issue29/camileon/, last accessed on 25 November 2010].

R. A. Lorie. 2001. **Long term preservation of digital information**. Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries. Roanoke, Virginia, United States, ACM**:** 346–352.

R. A. Lorie. 2002. **A methodology and system for preserving digital data**. Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, Portland, Oregon, USA, ACM: 312–319.

R. L. Dale and B**.** Ambacher. 2007. **Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)**. Published by the Research Libraries Group and National Archives and Records Administration (NARA).

S. Dobratz, A. Hanger, K. Huth, M. Kaiser, C. Keitel, J. Klump, P. Rodig, S. Hohde-Enslin, A. Schoger, K. Schroder, S. Strathmann, and H. Wiesenmuller. 2006. **Catalogue of Criteria for Trusted Digital Repositories**. Published by the nestor Working Group -Trusted Repositories Certification.

S. Granger. 2000. **Emulation as a Digital Preservation Strategy**. *D-Lib* **6**(10). [From http://www.dlib.org/dlib/october00/granger/10granger.html, last accessed on 25 November 2010].

S. Strodl, C. Becker, R. Neumayer, and A. Rauber. 2007 **How to choose a digital preservation strategy: evaluating a preservation planning procedure**. Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, Vancouver, BC, Canada, ACM: 29–38.

The Consultative Committee for Space Data Systems. 2002. **The Reference Model for an Open Archival Information System (OAIS).** Published by the CCSDS Secretariat Program Integration Division (Code M-3) National Aeronautics and Space Administration Washington, DC 20546, USA, 148p.

The Dublin Core Metadata Initiative (DCMI). 2008. **Dublin core metadata element set**. [From http://dublincore.org/documents/dces/, last accessed on 25 November 2010].

The Digital Preservation TestBed. 2001. **Migration: Context and Current Status**. A report sponsored by the National Archives and the Ministry of the Interior and Kingdom Relations, 21 p.

The National Library of New Zealand. 2003. **Metadata Standards Framework - Preservation Metadata.** [From http://www.natlib.govt.nz/catalogues/library-documents/preservation-metadata-revised, last accessed on 25 November 2010], 50p.

The PREMIS Editorial Committee. 2008. **PREMIS data dictionary for preservation metadata**. 217p. [From http://www.loc.gov/standards/premis/v2/premis-2-0.pdf, last accessed on 25 November 2010].