TECHNICAL REPORT

SURVIVABILITY OF DIGITAL RECORDS
(WP: READ)

DNV TECHNICAL REPORT 2007-1623

REPORT NO: 2007- 1623
REVISION NO 1

DET NORSKE VERITAS

TECHNICAL REPORT

Date of first issue: 04.07.2007	Project No: LongRec	DET NORSKE VERITAS AS 1322 Høvik Norway Tel: Fax: http://www.dnv.com NO 945 748 931 MVA
Approved by: Heidi Brovold	Organisational unit: Brino913	
Client: NFR	Client ref.: 176818 / I40	

Summary:
This report gives an overview of challenges with respect to reading digital files over a long time period. It also describes some of the more successful approaches that are in use.
A short description of the most commonly used storage media is given together with an assessment of their suitability for longterm storage. A similar description of frequently used file types are given together with some recommendations about format choices. Currently, it can be concluded that file format obsolescence appears to be of much greater concern than media obsolescence.
Longterm storage of digital data should always be accompanied with storing so-called preservation metadata giving information about technical format details, structure and use of the digital content; the history of all actions performed on the resource including changes and decisions; authenticity information such as technical features or custody history; and the responsibilities and rights information applicable to preservation actions.
File migration, i.e. moving a file from one storage media to another, and file conversion , i.e. moving the file content from one format to another, has been chosen as the main focus areas in the LongRec READ WP. There exist various strategies that have been chosen to tackle the challenges posed by migration and conversion, two of them are shortly sketched in one of the appendixes.

Report No: DNV Technical Report 2007-1623	Subject Group:	
Report title: Survivability of Digital Records (READ)		
Work carried out by: Thomas Mestl (DNV) Knut Nymoen (BBS) Jan-Ivar Bøyum (BBS) Lars Gaustad (NB)		
Work verified by: Jon ØlnesJon Ølnes		
Date of this revision: 28. mar. 2008	Revision No: 2	Number of pages: 53

Indexing terms

Keywords migration, conversion, file format, obsolence, preservation,
Service Area
Market Sector
☒ Yes = **Unrestricted distribution (internal and external)**

Table of Contents *Page*

1 INTRODUCTION 1

2 DIGITAL REPOSITORY 4

3 ORGANISATION 6

4 STORAGE MEDIA 7
4.1 Magnetic 8
4.2 Magneto-Optical (MO) disks 9
4.3 Optical 11
4.4 Solid State 13
4.5 Next generation storage media: holography 13
4.6 Conclusion & recommendations 14

5 MIGRATION – COPYING FILE TO NEW MEDIA 15
5.1 Some selected patents 15
5.2 Challenges of migration: 16

6 FILE FORMATS FOR LONG-TERM STORAGE 18
6.1 File format categorisation and identification 18
6.2 Basic File Formats 20
6.2.1 Wrappers and containers 20
6.2.2 Raster Graphics 21
6.2.3 Vector Graphics 21
6.2.4 Audio File 21
6.2.5 Video File 22
6.2.6 Text files 23
6.2.7 Databases 23
6.2.8 Biometric File Formats 24
6.2.9 Signature formats 25
6.3 File format obsolescence 25
6.4 Recommendation of file format choice for longterm storage 26

7 PRESERVATION METADATA AND METADATA FRAMEWORK 29
7.1 Preservation Metadata Extraction tools 31
7.2 Digital Signatures Metadata 31

8 CONVERSION – CHANGE OF FILE FORMAT 33

9 APPENDIX A: FILE TYPE CATEGORISATION 35

10 APPENDIX B: CEDARS PRESERVATION METADATA SET 36

11 APPENDIX C: PREMIS PRESERVATION METADATA SET 38

12 APPENDIX D: ARCHIVAL, MIGRATION AND CONVERSION STRATEGIES OF THE LONGREC PROJECT PARTNERS: BBS AND NATIONAL LIBRARY 40

13 APPENDIX E: STATE OF THE ART FOR DIGITAL SIGNING (BBS) 45

Preface:

This State-of-the-Art report is a part of the LongRec (Long-Term Records Management) project run by Det Norske Veritas (DNV) in collaboration with a number of case partners, commercialization partners and research partners. The primary objective of LongRec is the *persistent, reliable and trustworthy long-term archival of digital information records with emphasis on availability and use of the information*. The project's public web site is at http://research.dnv.com/longrec/

LongRec is a three year project (2007-2009) partly funded by the Norwegian Research Council. The project constitutes the Norwegian team of the InterPARES 3 project, http://www.interpares.org

LongRec addresses several research challenges[1], each of which is assigned a short name (in parentheses below): records transition survival (READ), long-term usage (FIND), preservation of semantic value (UNDERSTAND), preservation of evidential value (TRUST) and legal, social, and cultural framework (COMPLIANCE). Each research challenge is addressed by:

- General studies compiling state of the art and best practice of the area.
- Research on selected sub-topics, performed by the research partners and by one PhD student for each research challenge.
- One or more case studies with LongRec case partner(s).
- Studies on opportunities for products and services done together with the commercialization partners.

[1] We refer to the project's web site http://research.dnv.com/longrec for a description of the research challenges.

1 INTRODUCTION

Digital preservation has been defined as "the planning, resource allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable". In addition, the issue of authenticity may be added to this definition. Digital preservation addresses thereby hardware (e.g. storage media, reading and processing hardware), software (e.g. reading and processing), data models (e.g. file formats, preservation metadata) and involved processes (e.g. migration or conversion procedures, emulation strategies, obsolescence detection, quality assurance, and all kind of documentations).

A digital record is defined as a record created or received and/or maintained by means of digital computer technology[2,3]. A digital record is thereby not just the digital equivalent of a paper document but can virtually be anything that can be created and stored on a computer. In this respect digital records are not tangible objects but a combination of hardware, software and computer files. This combination is necessary to be able to use the records. The digital record “lives” (i.e. is only useful within a software environment) but is consigned to a physical carrier medium for storage. This dual nature of digital records tremendously complicates the long term survival problem. It both requires addressing the physical media and its durability or lack thereof, i.e. storage media and hardware equipment, and it requires maintaining a suitable living condition for the record, i.e. the program and/or operating system. No wonder the survival of digital records for decades or centuries is yet an unresolved problem.

Data Storage: physical properties, e.g. holes, magnetism, color etc. code data on hardware media (paper, magnetic tape, plastic etc.).

Data Reader: a suitable reader (hardware and usually software) is needed to transform the physical signal e.g. hole into a electrical signal (bit stream).

Data Processor: a hardware environment where the bit stream can be loaded into the desired software.

File Migration: the data in a file are either copied directly or via a computer from one media to a new media. The data storage format may have to be changed due to format obsolescence.

Figure 1: Data storage and processing always includes hardware. Either the storage medium, e.g. tape, disc, or the reader, e.g. floppy, disc drive, and the data processing equipment will degrade and become susceptible for failures or break downs and will have to be maintained or replaced.

Data Storage Media:

Digital data are either stored on volatile or non-volatile media. With volatile storage, e.g. RAM, all data is lost when the power is switched off, whereas with non-volatile storage, e.g. punch card, magnetic tape, hologram, the data will persist for a period of time without the need for a power supply. For longterm data storage only non-volatile media will be of interest. As electrical signals (= series of 0 and 1 = data stream) cannot be preserved directly they have to be transformed into a more permanent form. Punching holes in a paper tape or card, or utilising the magnetising properties of ferromagnetic tapes, floppy discs or hard drives was and is still a common practice. The natural decay rate of the storage medium, e.g. paper, tape plastic or the physical signal, e.g. dye, gives the time limitation of the storage medium.

[2] UBC Project, Glossary, March 1997.
[3] http://www.interpares.org

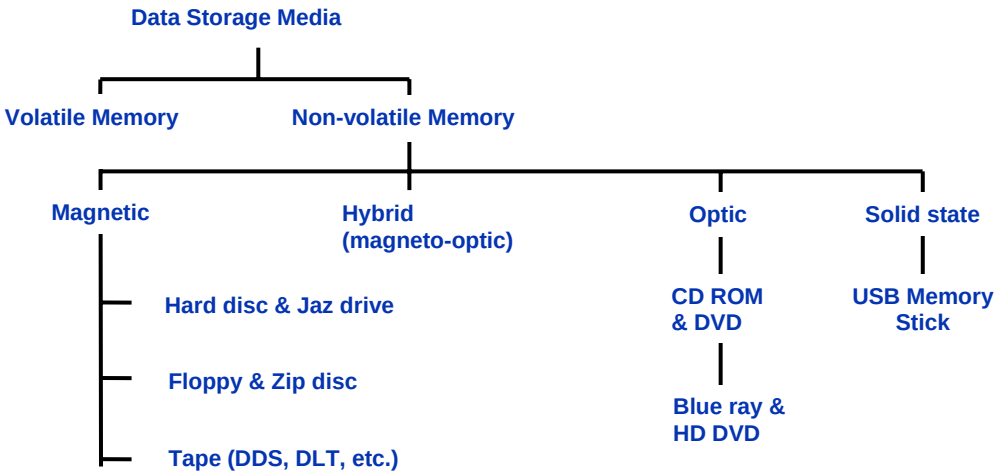


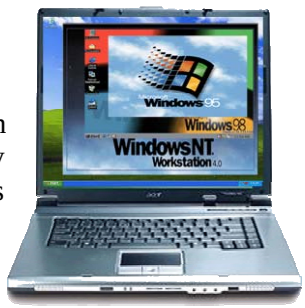
Figure 2: "There are two main technologies in use for storing digital data: volatile and non-volatile. Volatile storage loses its data when the power is switched off inside a computer (e.g. RAM), whereas non-volatile storage will retain the data for a period of time without the need for a power supply (for example, a CD-R)"[4].

Data Reading Hardware

Once data is stored on a medium it can only be read, i.e. transformed into electrical signals by a suitable reader specifically designed for this medium and design, e.g. punched tape reader. Without such a reader it can be extremely difficult to access the signals. The storage media must therefore always be considered together with the corresponding reading hardware, e.g. the old 5¼-inch floppy disc. Improvements in storage technology (capacity, size, speed) will sooner or later make a storage medium obsolete. Data might therefore be migrated to a newer storage media. As production, sales and support ceases the only hope is to frequently visit flea markets on the search for desired hardware or spare parts.

Data Processing Hardware

The majority of data files are in proprietary formats and can therefore only be interpreted correctly by the original or corresponding software. Such software often requires the right environment both with respect to processor hardware profile as well as operating system. For instance data games developed for Comodore 64 can usually not be run on modern computers. There are three options available for old software:



- **"Hardware museum":** a machine park of computers together with their operating system is kept and maintained in order to process old data files.
- **Emulation:** the old hardware as software is mimicked on more powerful modern computers. This requires, of course, that the old program and data are already digitally available in the modern system. When the current computer becomes obsolete a new emulator has to be written for it that runs on the next generation computer. In this way there will be an ever growing stack of emulators.
- **Universal Virtual Computer (UVC)**[5] concept which is based on elements from both migration and emualtion. This approach will be platform-independent regardless of future technological changes.

File Format

A file format defines the internal structure and encoding of a digital object. In a long-term preservation perspective a digital object is very often considered as a file or a bit stream that is packed together with preservation data into a single unit, often referred as object encapsulation.

[4] http://mercury.soas.ac.uk/it/docs/cop/cop-storage.htm#data
[5] http://www.nla.gov.au/padi/topics/492.html

A digital object may be a file, or a bit stream embedded within a file'[6]. All file formats fall into the eight MIME types high level file format categorisation provided by IANA[7]:

- Application
- Audio
- Image
- Message
- Model
- Multipart
- Text
- Video

No exact number can be given for the total number of available file formats as new formats come into existence every day, but there are over 1.000 file-formats registered in registries like PRONOM, filext or MyFileFormat[8]. Some file formats are preferable over others as they may have publicly available format documentation, a large user base, support service or even a guaranteed supported life time. Nevertheless, over time software applications will be further developed resulting usually in altered or even new file formats. Formats that were frequently used 10 years ago are turning unreadable, so will many of today's formats in 10 years. Signatures or certificates applied to files may become weak due to enhanced decryption power, improved algorithms or CAs are no longer present in the market. There are basically 2 strategies to render the file content accessible:

- **Conversion:** the content of a file is transferred into another file format. In this process there is a risk that some if not all information is lost or altered. Great care has to be taken to maintain the quality and integrity of the file.
- **Object Encapsulation:** is a technique of grouping together a digital object, e.g. a file, and anything else necessary to provide access to that object[9]. Encapsulation can be achieved by using physical or logical structures called "containers" or "wrappers" to provide a relationship between all information components, supporting information and software specifications.

Preservation metadata

Preservation metadata are intended to support and facilitate the long-term retention of digital information. In contrast to descriptive metadata schemas (e.g. MARC[10], Dublin Core[11]), which are used to describe the context of digital objects, preservation metadata are intended to store technical details on the format, structure, the history of all actions performed on the resource including changes and decisions, the authenticity information such as technical features or custody history, and the responsibilities and rights information applicable to preservation actions[12].

[6] Brown, A. (2006). Digital Preservation Technical Paper 2. The PRONOM Unique Identifier Scheme. *DPTP-02*, Issue2, p. 1-9. http://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf
[7] http://www.iana.org/assignments/media-types/
[8] Rauch et al. (2007) File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats http://elpub.scix.net/data/works/att/122_elpub2007.content.pdf
[9] http://www.nla.gov.au/padi/topics/20.html
[10] http://www.loc.gov/marc/
[11] http://purl.org/metadata/dublin_core_elements
[12] http://www.nla.gov.au/padi/topics/32.html

2 DIGITAL REPOSITORY

Repositories are collections of digital objects. (Trusted) digital repositories differ from other collections of digital objects such as directories, catalogues, databases by the following characteristics[13]:

- content is deposited in a repository, whether by the content creator, owner or third party,
- the repository architecture manages content as well as metadata,
- the repository offers a minimum set of basic services e.g. put, get, search, access control,
- the repository must be sustainable and trusted, well-supported and well-managed.

According to [14] all trusted digital repositories must:

- accept responsibility for the long-term maintenance of digital objects on behalf of its depositors and for the benefit of current and future users;
- have an organizational system that supports not only long-term viability of the repository, but also the digital information for which it has responsibility;
- demonstrate fiscal responsibility and sustainability;
- design its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited within it;
- establish methodologies for system evaluation that meet community expectations of trustworthiness;
- be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly;
- have policies, practices, and performance that can be audited and measured.

The following list shows available software solutions for digital preservation, for more info it is referred to [15].

Name	Description
DIAS (Digital Information Archiving System) http://www-5.ibm.com/nl/dias/	The DIAS (Digital Information Archiving System) solution provides a flexible and scalable open deposit library solution for storing and retrieving massive amounts of electronic documents and multimedia files. It conforms with the ISO Reference OAIS standard and supports physical and logical digital preservation.
DPS (Digital Preservation System) http://www.exlibrisgroup.com	The DPS is a preservation solution for digital objects. The system conforms to the OAIS standard recognised by ISO and supports many of the standards in the library environment (METS, PREMIS, MARC, DC, OAI-PMH etc.). It is designed to support the acquiring, validation, ingest, storage, management, preservation and dissemination of different types of digital objects.
CDS Invenio http://cdsware.cern.ch/invenio/index.html	Developed by CERN and CDS Invenio is designed to run an electronic preprint server, online library catalogue or a document system on the Web.
DSpace http://www.dspace.org	The DSpace digital repository system was designed to capture, store, index, preserve, and provide access to institutional digital research materials. It can accept all forms of digital materials, ranging from text, images and datasets to websites, multimedia, video and audio files. DSpace can be used in a variety of ways, including as an institutional repository, elearning objects or e-theses repository, an electronic records management

[13] Heery, R (2005) Digital Repositories Review. http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf
[14] Trusted Digital Repositories: Attributes and Responsibilities. *RLG-OCLC Report*, 2002 http://www.rlg.org/legacy/longterm/repositories.pdf
[15] (2007) D. 6.1 Market and Technology Trends Analysis *DigitalPreservationEurope (DPE)* http://www.digitalpreservationeurope.eu/publications/dpe-market-analysis.pdf

	system, a digital asset management system, and a digital preservation system.
EPrints http://www.eprints.org/software	The EPrints software was designed as repository software for e-prints, electronic versions of research articles, in either pre-print or post-print versions (or both).
Fedora http://www.fedora.info	The Fedora digital object repository management system is based on the Flexible Extensible Digital Object and Repository Architecture. The current version of the software provides a repository that can handle one million objects efficiently.
Greenstone Digital Library Software http://www.greenstone.org/cgi-bin/library	Greenstone is a suite of software for building and distributing digital library collections.
LOCKSS (Lots of Copies Keep Stuff Safe) http://www.lockss.org/lockss/Home	LOCKSS offers an easy and inexpensive way to collect, store, preserve, and provide access to their own, local copy of authorised content they purchase.

For instance, the Florida Digital Archives (FDA) accepts any file format, but only files in supported formats will receive full preservation services with the aim of ensuring the continued usability of the file. Files in unsupported formats will be preserved in their original (submitted) version only (bit-level preservation)[16].

Archival System Standards & Frameworks

It is referred to ref [17] for issues regarding criteria and checklists for audit and certification of trustworthy repositories. Work is currently ongoing towards a full ISO standardization of trustworthy repositories but this may still take several years.

ISO 14721:2002, the *Open Archival Information System Reference Model* provides a high-level reference model or framework identifying the participants in digital preservation, their roles and responsibilities, and the kinds of information to be exchanged during the course of deposit and ingest into and dissemination from a digital repository.

Model Requirements for the Management of Electronic Records (MoReq)[18] is based on ISO 15489-1 & 2: 2001 (Records Management) and ISO 23081-1: 2004 (Information and documentation - records management processes) and addresses at a general implementation level. An update (MoReq 2) will be available in 2008[19].

Noark standing for Norsk arkivsystem[20] (= Norwegian Arcival System[21]) states application specific requirements for electronic archive systems applicable for Norwegian public administration. The specification addresses

- information content (what information shall be registered and shall be retrievable)
- data structure (design of the individual data elements and their mutual relationship)
- functionality (what functions shall the system attend to).

An update, Noark-5 is currently under development and planned to be completed in 2008.

[16] FCLA Digital Archive (FDA) Policy Guide. *Florida Center for Library Automation,* 2004 www.fcla.edu/digitalArchive/pdfs/DigitalArchivePolicyGuide1_1.pdf.
[17] Trustworthy Repositories Audit & Certification: Criteria and Checklist (2007) CRL, The Center for Research Libraries or OCLC Online Computer Library Center, Inc.
[18] http://www.cornwell.co.uk/edrm/moreq.asp
[19] http://www.moreq2.eu/
[20] http://www.riksarkivet.no/arkivverket/lover/elarkiv/noark-4.html
[21] http://www.arkivverket.no/arkivverket/lover/elarkiv/noark-4/english.html

3 ORGANISATION

Not only hardware, software or the file format will become obsolete but also the organization(s) or organisational processes that are supposed to maintain and manage these records in a trusted repository. This is especially critical for digitally signed objects that are based on the existence and validity of certificates and their issuing authorities/companies.

Mergers, acquisitions, partly sell-offs or closures may significantly affect the trustworthiness or even the survival of digital data as this may have an impact on vital processes in the data preservation chain.

The risk originating from organisations may be reduced by a multi-site redundant storage approach[22] but clearly not entirely removed. In fact, other risks could be introduced such as copyright infringements, ownership issues or the gradual emergence of diverging copies due to use of different conversion software.

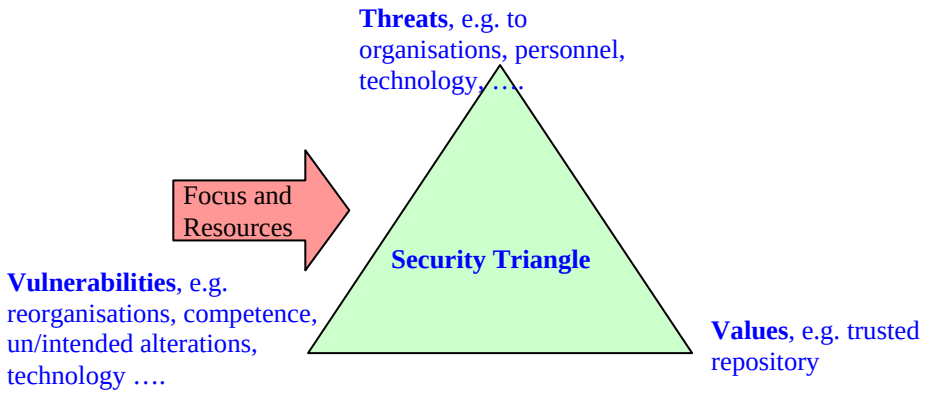


Figure 1: The security triangle visualizes the 3 areas that determine security in general. To mitigate risks it can either be focussed on the *Values* that shall be protected, *Threats* to the values and *Vulnerabilities* of the system.

A similar approach as in the security area may be used for identifying threats and vulnerabilities that may arise when trying to protect a given company value. These threats and vulnerabilities will vary considerably depending on the company's (or authority's) business value.

Various methodologies have been developed that build on a risk management approach to digital preservation. The Virtual Remote Control methodology (VRC) from Cornell University Library offers a compilation of tools for monitoring and identifying potential risks of loss of Web-based information[23].

In a similar risk based approach, Lawrence et al. showed that the levels of risk originating from different formats as well as from organizational, hardware, software, and metadata issues can actually be determined[24].

[22] Larry Masinter, Michael Welch; A System for Long-Term Document Preservation. *Proceedings of IS&T Archiving 2006 conference.*
[23] http://prism.library.cornell.edu/VRC/
[24] Lawrence,G. et al. Risk Management of Digital Information: A File Format Investigation. *Council on Library and Information Resources*, (2000) http://www.clir.org/pubs/abstract/pub93abst.html

4 STORAGE MEDIA

The choice of physical storage media will depend on a series of user defined preferences. If the data shall be archived, i.e. seldom used, magnetic tapes have traditionally been the favourite choice, if however the data will be frequently accessed disk drives have often been chosen. The following issues should be addressed when choosing a storage medium:

- Longevity
- Capacity
- Viability
- Obsolescence
- Cost
- Susceptibility

It is argued that longer longevity than 10 years of the carrier medium may not be necessary as technological obsolescence of the reading hardware may require copying data to a newer storage media. Similar considerations will have to be done dependent on the amount of data, security issues, company policy, automation, cost and of course expected life time of storage medium. For further reading it is referred to [25],[26], (see also the "Chamber of Horror" of obsolete or endangered storage media [27]).

Natural degrading processes such as dye degradation, aging of the carrier material will influence the lifetime of the storage media and its content. The life time of *digitally* stored data depends essentially on the following parameters[28]:

1. Recording procedure
2. Stability of the recording (magnetic, optical, magneto-optical, mechanical)
3. Storage conditions of the data carriers
4. Frequency of reading, wear caused through the reading of the data
5. Availability of suitable readers
6. Availability of decoding software

The table below gives a coarse overview over the expected lifetime of the most frequently used storage media. For an online implementation of a quick guidance with respect to different storage media its exposed risks and recommended temperature range it is referred to [29] and [30].

Table 1: Electronic records formats and life expectancy in years[31]

Media	Expected life time in yrs	Practical life time in yrs
Magnet tape- ½"-3480/3490 data recording	10-30	2 - 7
Digital linear tape	10-30	2 - 7
¼" cartridge	5-30	2 - 7
CD-ROM (yellow book standard)	5-100	10 - 15
CD-WORM (pits on bimetallic alloy thin film)	100	10 - 15

[25] Brown, A. (2003) Selecting Storage Media for Long-Term Preservation. *National Archives. Guidance Note*. http://www.nationalarchives.gov.uk/documents/selecting_storage_media.pdf
[26] http://palimpsest.stanford.edu/bytopic/electronic-records/electronic-storage-media/
[27] http://www.library.cornell.edu/iris/tutorial/dpm/oldmedia/
[28] Harken, H. (2007) Physikalisch-Technische Bundesanstalt. www.ptb.de/en/org/2/25/251/lifetime.pdf
[29] http://www.climatenotebook.org/MSQR/wheel_final.html
[30] Brown, A. (2003) Care, Handling and Storage of Removable Media *The National Archives, UK.* http://www.nationalarchives.gov.uk/documents/media_care.pdf
[31] http://www.geocities.com/jen_simm/elecrec.doc

Digital storage media can be grouped into three main categories, disk, tape, and solid state, which again can be split further into many levels of subcategories. Storage media cover both integrated storage, i.e. drive and media as a single unit, as well as removable media.

An important factor when comparing technologies is the applicability of each technology. The table shows that a hard disk RAID has a low cost per GB but a poor archive life and needs regular backups taken. WORM tape also offers a very low cost per GB but it's access time is very poor for organisations wishing to retrieve information quickly. At the current state of technology it seems that only optical storage can offer the desired 50 years of archival life and the relatively short retrieval times required by organisations. Any choice of storage technology will include some form of multiple copies or/and back-ups.

Table 2: Storage technology comparison

Technology	Capacity GB	Speed	Access Time	Archive Life	Power	Special requirements	TCO	Cost Gbyte
Blu-ray	50	Good	V Good	50	Low	No	Low	High
Plasmon UDO	30	Average	Good	50	Low	No	Low	Med.
DVD	4.7/9.4	Poor	Average	50	Low	No	Low	Low
TAPE	>800	V Good	V Poor	30	Medium	No	Medium	Low
RAID	>1TB	Excellent	Excellent	2	High	Needs backing up	V High	Low

	Maintenance	Reliability	Data Compliant	Media Sides	Optical	Ruggedised*	Data Format
Blu-ray	Low	High	Yes	Single	Yes	No	RW/WORM
Plasmon UDO	Low	High	Yes	Double	Yes	Yes	RW/WORM
DVD	Low	High	Yes	Single/Double	Yes	No	WORM
TAPE	Medium	Medium	Yes**	Single	Yes	No	RW/WORM**
RAID	High	Low	No	NA	No	No	RW

* hardened against rough handling

** Dependant on tape technology, RW = Read/Write, WORM = Write Once Read Many

4.1 Magnetic

WORM TAPE - For years tape has been used for backing up information and restoring.

Magnetic tape has historically been used for data storage – especially backup - for over 50 years. Tape drives can usually store up to 200-400 GB of data with a data transfer rate of about 80 MB/s. In 2007, the highest capacity tape cartridges can store 800 GB of data without using compression[32]. In this respect, tape has a high capacity and low cost per Gbyte; it is let down by it's access time which can be from 30 secs to many minutes to retrieve an individual file Although the storage of large amounts of data on tapes can be substantially less expensive than disk or other data storage options the usability of higher capacity tapes is limited by the relatively low data transfer rate.

Now many companies including Sony, Quantum and the LTO consortium have developed WORM tape. They offer a lower-cost way to store sensitive and regulated data and comply with rapidly growing regulations like HIPAA, Sarbanes-Oxley and SEC 17A-4. Data is written to the WORM tape by using a combination of hardware and software embedded in the drive and the tape cartridge. This communication prevents any intentional alterations or over-writes, even if the tape is extracted from its original cartridge and placed into a different non-WORM cartridge. These tapes are usually offered in three types of packaging: open reels, cassettes and cartridges. The omission of a recording-enabled hole from the media cartridge also safeguards against

[32] http://en.wikipedia.org/wiki/Digital_Linear_Tape

accidental overwrites. Further data integrity is ensured through the use of unique serial codes assigned to the cartridge during the manufacturing process.

WORM tape has at least a 30 years life. If it is stored in non ideal environmental conditions the stored signal can degrade due to the contact of the tape with the next layer in the tape coil. This was especially of concern for analogue recordings where this phenomenon lead to a decrease of the signal to noise ratio. With respect to the long-term storage of digital signals on tape this copying effect of adjacent tape sections is however of no concern anymore. Long-term effects like deterioration of the carrier material may still remain a concern[33]. For further info about protecting and handling magnetic media it is referred to[34].

The rapid improvement in disk storage density and price reduction together with relatively little new developments in tape storage technology have reduced the market share of tape storage products[35].

Magnetic Discs - These type of storage media consist of rotating discs where the data is coded as magnetic signals (fixed hard drives, removable hard disk packs, floppies, Zip, Jaz). Various kinds of floppy discs have already become obsolete and almost all new PCs are shipped without a floppy reader.

Hard drives benefit from an exponential increase in storage density but with similar falling prices. For instance, between 2002 and 2007 the capacity of a hard drive has increased from 61.4 GB costing $398 to 250 GB costing $129[36]. It is expected that storage density on hard drives will increase between 60% and 100% annually while storage pricing will fall between 35% and 40% at the same time. Currently, the price level is around $0.80 per GB and a further increase in storage density and price decline can be expected.
The main benefit of hard drives against optical storage media is the relative simplicity of data migration across networks[37].

RAID Arrays - A RAID (**R**edundant **A**rrays of **I**nexpensive **D**iscs) has long been used for storing and retrieving frequently used information very fast. A typical RAID has 5 or more disk drives and is configured with a RAID level, typically RAID 5, this can sustain two drive failures before data loss occurs. Newer RAID systems offer RAID 6, this can sustain three drive failures before data loss. A RAID system is constantly consuming power even when not being accessed, no matter how reliable a RAID system is, "it will break". A RAID array is a mechanical device with constantly moving parts, over time these parts wear out and when they do the RAID system fails. Many experts recommend that RAID should only be used for fast access to information and should not be considered as an archive medium for 5+ years.

4.2 Magneto-Optical (MO) disks

MO disks have been used for data backup for two decades. Many proprietary systems use them, but there is a growing market for MO discs complying with ISO standards that make them useful as storage media for audio and audiovisual files. Drives complying with ISO/IEC standards are capable of handling MO disks from a wide range of manufacturers.

As MO disks have a much smaller presence in the market than CD-R or DVD-R it must be surmised that they are at greater risk of format obsolescence. Such possibilities must be considered as a part of any archival appraisal. Although MO disks share similarities to data tape storage, and many of the issues associated with data tapes, such as development roadmaps and

[33] http://www.nationalarchives.gov.uk/documents/selecting_storage_media.pdf
[34] http://www.aa.gov.au/recordkeeping/rkpubs/advices/advice5.html
[35] http://en.wikipedia.org/wiki/Magnetic_tape_data_storage#Viability
[36] http://www.littletechshoppe.com/ns1625/winchest.html
[37] FP6-IST-507336 PrestoSpace Deliverable D12.6 Survey of Digital Formats for Storage

version compatibility need to be considered, the MO discs have been backward compatible for two decades, and ISO standardisation may well be seen as a further guarantee for future compatibility.

MO disks come in two physical sizes; 3.5” which are single sided and aimed at the consumer market and double-sided 5.25” which are aimed at the professional archiving market. The recording disk is held in a rugged dustproof shell at all times, the whole cartridge is inserted into the drive. The drives may be standalones or mounted in a computer cabinet in the same manner as CD or DVD drives. The drives communicate with the computer through protocols such as IDE and SCSI for internal drives and Firewire or USB for stand-alones.

For quite a number of years MO disks had only small data capacity, however recent development has produced disks that may hold up to 10 GB of information. This is the equivalent of 15 CD-Rs or 2 DVD-Rs, or approximately 10 hours of 48 kHz 24 bit linear PCM files. Mainstream products with a capacity of up to 5.2 GB and 9.1 GB are currently available from major market players, and developments are planned. Most disks are rewritable, but there are also disks that are write-once read-many (WORM) and those are the appropriate disks for archival purposes. They are marketed with the abbreviation CCW (Continuous Composite Write). CCW WORM disks are specifically designed for use as write-once media in ISO Standard 5.25" multifunction drives, data cannot be altered without detection due to specific qualities of the recording layer. This is due to a feature where the disk signals the optical drive not to rewrite media sectors.

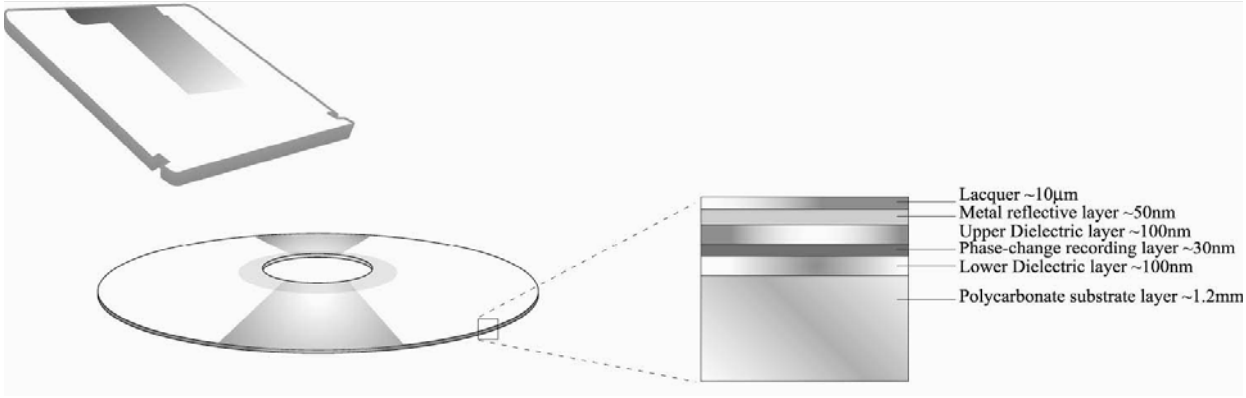


Figure 2: A schematic view of a MO disk

MO technology operates on a combination of magnetic and optical principles. While a hard disk or a tape can be magnetised at any temperature, the magnetic coating used on MO media is designed to be magnetically stable at room temperature, making the data unchangeable unless the disk is heated to above a temperature level called the Curie point, usually around 200 degrees centigrade. The MO drives use a laser to target and heat specific small regions of magnetic particles. This accurate technique allows MO media a higher packing density than other magnetic devices. Once heated, the magnetic particles can easily have their direction changed. When reading, a weaker laser beam directed at the layer will then alter its rotation due to a phenomenon called the Kerr effect, and read 1s and 0s as recorded. Figure to the left shows the schematics of the recording device.

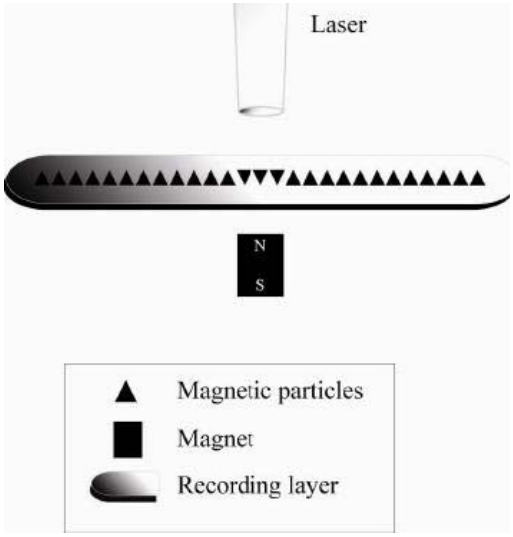


Figure 3: of the MO recording device.

4.3 Optical

CD and DVD can store more than 650MB and at least 4.7GB of data respectively[38] because of different recording densities. The standard CD track pitch is 1.6±0.1µm whereas for a DVD it is 0.74 ±0.01 µm. The higher densities of DVDs make them very sensitive to jittering. For a good introduction to various DVD formats and a guideline for Care and Handling of CDs and DVDs as physical objects see [39], [40].

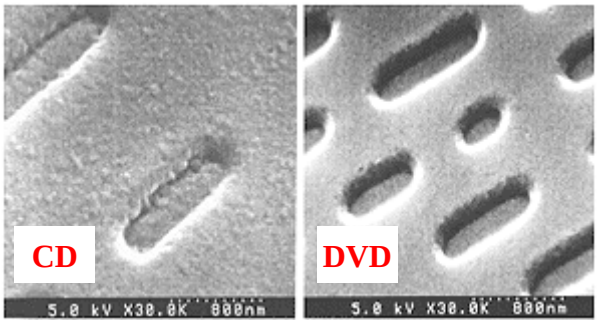


Figure 3: DVDs have a much denser pits pattern then CDs.

Optical discs can roughly be divided into two groups:

- Pressed CD/DVD ROM (read only) have a life expectancy of 100 to 200 years or more. Whether this becomes true remains to be seen.
- Recordable CD/DVD-R (write once), CD/DVD RW (rewritable): life expectancy of 25 years or more

The lifetime of optical discs depends strongly on assuring and maintaining correct storage conditions as exemplified by the exponentially decreasing lifetime of Kodak writable CDs with rising temperatures[41].

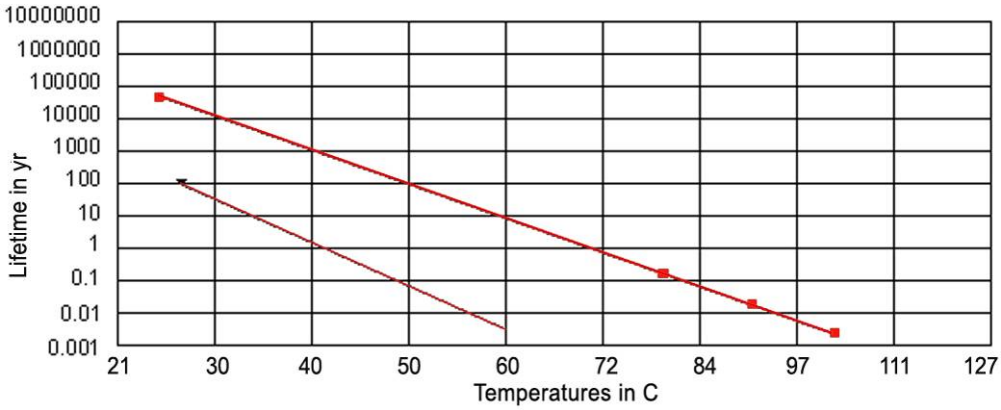


Figure 4: Predicted theoretical life time of KODAK Writable CD and Photo CD Media under varying temperatures. Under controlled conditions, the best estimate of expected lifetime of these products is 217 years[41].

Much of the attention is focused on longlivety of the physical media but it might however be of more importance to address the longlivety of the signal stored in the media – which may have quite different lifetimes.

CD - The first CD (**C**ompact **D**isc) was released 17. August 1982 in the stores and has gained great popularity as a long-term storage media due to its low price, high storage capacity and high access speed. In 2007 CDs are still the standard playback medium for commercial audio recordings, whereas DVDs have become more popular than VHS tapes. After 25 years of market presence it is now speculated when these optical discs will be phased out. The obsolescence could start from the fact that most of the music today is stored in disc drives or even on the Internet and played from portable mp3 players rendering the need for CDs and CDs player unnecessary.

[38] http://h71036.www7.hp.com/hho/cache/733-0-0-225-121.html
[39] http://pioneer.jp/crdl/tech/dvd/2-e.html
[40] Byers, f. (2003) Care and Handling of CDs and DVDs —A Guide for Librarians and Archivists. *NIST*. http://www.itl.nist.gov/iad/894.05/docs/CDandDVDCareandHandlingGuide.pdf
[41] Stinson et al. (1995) Lifetime of KODAK Writable CD and Photo CD Media. *Eastman Kodak Company* http://www.cd-info.com/CDIC/Technology/CD-R/Media/Kodak.html

There exist series of international standards and manufacturers' specifications with respect to optical discs, e.g. ISO 18925, ISO/AWI 18938 or ISO 9660, for a more detailed list of standards together with a detailed overview about risks associated with the use of recordable CDs and DVDs, and recommendations with respect to disc choice, recording, reproduction and maintenance, see reference[42].

DVD – A DVD (**D**igital **V**ersatile **D**isc) holds 4.7GB of data on a single side and 9.4GB using double sided media. The media is un-protected and requires a rotation mechanism to flip the media over. A DVD jukebox is ideal for storing approximately 1TB of data, above this the cost becomes expensive compared to other optical technologies. Although the technology is primarily WORM (DVD-R), the creation of the pits causes a chemical change in the media. Due to the way the disks are written more care must be taken when archiving DVD media. The other DVD technology that is available and is a better technology than DVD-R is DVD-RAM. It was developed by Panasonic. The DVD-RAM disc structure is laid out the same as a hard disk with sectors/tracks and error correction, is supplied in a ruggedised caddy or as bare media. This is more suited for archiving data than DVD-R.

UDO – (**U**ltra **D**ensity **O**ptical) Developed by Plasmon to replace MO (Magneto Optical) technology. It holds 15GB per side making 30GB per disk; as the media is double sided it requires a rotation mechanism to flip the media over. It can read/write data at 8MB/sec, uses Reed Solomon error correction and has 8K logical sector size. The Plasmon UDO (Ultra Density Optical) uses phase-change technology to write data, where a powerful laser heats a substrate to one of two heat points: at one heat level, the substrate turns into a crystalline structure; at another heat level, the crystalline breaks down to a less reflective amorphous state. A less-powerful laser is used to read the data without altering it. The phase-change method should provide for faster write times, higher storage densities, and a higher read/write life cycle (the number of writes that a spot can withstand before it can no longer change its state reliably). The Plasmon UDO media is housed in a ruggedised shock/dust proof caddy.

Blu-ray - The next generation optical discs are the high density DVDs (HD DVD) and Blue ray discs (BD). This technology uses blue-violet laser for reading/writing and with the laser's considerably shorter wave length much higher storage densities of 15 GB (HD DVD) and 25 GB (BD) can be achieved. Doubling or tripling disc capacity can be obtained by adding additional disc layers. At the current time HD DVD and BD are rival formats and only time will show which establishes itself as the preferred technology in the market[43]. According to the analysis agency Forrester, this format war will last well into 2009 — maybe even longer[44].

BD was developed to enable recording, rewriting and playback of high-definition video (HD), as well as storing large amounts of data. The format offers more than five times the storage capacity of traditional DVDs and can hold up to 25GB on a single-layer disc and 50GB on a dual-layer disc. For more general information about Blu-ray, please see[45].

According to the Blu-ray Disc specification, 1x speed is defined as 36Mbps and to satisfy movie quality requirements a data transfer rate of 2x (72Mbps) is expected to be seen. Blu-ray also has the potential for much higher speeds, as a result of the larger numerical aperture (NA) adopted by Blu-ray Discs. The large NA value effectively means that Blu-ray will require less recording power and lower disc rotation speed than DVD and HD-DVD to achieve the same data transfer rate. While the media itself limited the recording speed in the past, the only limiting factor for Blu-ray is the capacity of the hardware. Assuming a maximum disc rotation speed of 10,000

[42] Bradley, K. (2006) Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections - Strategies and Alternatives. *UNESCO.* http://unesdoc.unesco.org/images/0014/001477/147782E.pdf
[43] http://en.wikipedia.org/wiki/Blue_ray
[44] http://arstechnica.com/news.ars/post/20070925-hd-dvd-and-blu-ray-deadlock-to-continue-into-2009-at-least.html
[45] http://www.peripheralstorage.com/html/blu-ray_disc_format.html

RPM, then 12x at the outer diameter should offer about 400Mbps. This is why the Blu-ray Disc Association (BDA) already has plans to raise the speed to 8x (288Mbps) or more in the future.

For Blu-Ray and HD DVD there is a chance that they will neither be cross compatible nor backwards compatible towards DVD and CD players. Although optical discs have a long media life time and can store large amounts of data there is the issue of conflicting formats and compatibility in a large high quality archive.

4.4 Solid State

The third category of storage media are so-called solid state media such as CompactFlash, Memory Stick, Smart Media (digital camera memory), USB memory key or stick, pen drives, keychain drives. Flash drives (IDE and SCSI) using standard hard disk form factors are often used for industrial or military purpose (capacity currently up to 80GB and even 640GB). The currently achievable read and write speeds are 800MB/s and 600MB/s respectively[46]. For a detailed description of these types of media it is referred to the white-paper from Memorex[47].

Solid state memories can sustain only a limited number of write and erase cycles before failure. The lifetime is estimated to last several 100,000 r/w cycles. Obsolescence of solid state memory devices may therefore mainly be due to the emergence of technologically better products (smaller, higher, new or non-standardised ways of access). So far, no information has been found about the expected lifetime of data on such devices).

4.5 Next generation storage media: holography

Tapestry from InPhase[48]**:** The discs holds currently 300 GB and it is expected that a commercial version of 1.6 terabytes will be on the marked in 2010[49]. The Tapestry HDS-300R (write only) 300GB holographic disc costs about $180/disc and $18,000 for the drive.

Holographic Versatile Disc (HVD): HVD discs is supposed to hold 3.9 terabytes (TB) corresponding to 830 times the capacity of a DVD with a high data transfer rate, about 1 gigabit/s[50],[51]. No product information available.

Within the next 2 years holographic storage systems will be storing the equivalent of 63 DVD's on a single holographic disc of 300 Gigabytes capacity or 35 hours of broadcast quality video. With a 50 year archive life and a road map to 1.6TB per disc holographic storage promises to be the content distribution and archive format of the future.

Holographic storage may be the one furthest down the developmental pipe in contrast to other storage techniques such as storing digital signals on a single atomic[52] or molecular level[53], e.g. proteins.

[46] http://www.fusionio.com/index.html
[47] http://www.memorex.com/downloads/whitepapers/WhitePaper_Flash_Cards_Drives_Dec06.pdf
[48] http://www.inphase-technologies.com/
[49] http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9011144&source=rss_news10
[50] http://www.manifest-tech.com/media_dvd/dvd_holo.htm#Holographic%20Techology
[51] http://en.wikipedia.org/wiki/Holographic_Versatile_Disc
[52] http://www-03.ibm.com/press/us/en/pressrelease/22254.wss
[53] http://pubs.acs.org/subscribe/journals/tcaw/10/i06/html/06comp.html

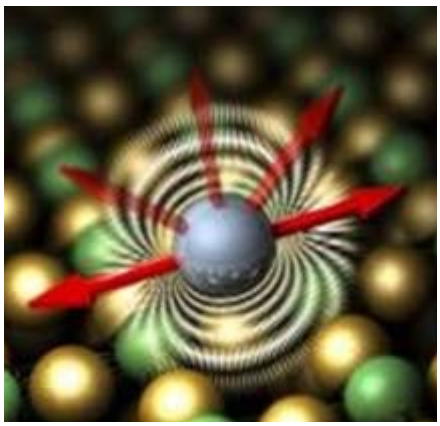


Figure 5: Atomic storage gives the ultimate storage density with 1,000 trillion bits of information in an iPod, according to IBM estimates.

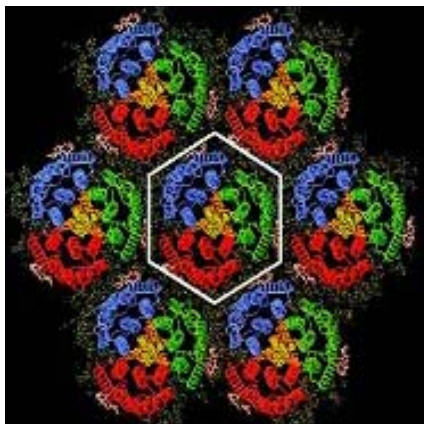


Figure 6: Membrane proteins are being used to generate the first protein-based information storage system to store terabytes of information *(Image: V Renugopalakrishnan)*[54]

4.6 Conclusion & recommendations

Based on the experiences so far it may be concluded that file format obsolescence appears currently to be of much greater concern than media obsolescence.

- Today most data are stored on hard discs. It is generally recommended that data stored on magnetic media are refreshed regularly depending on the frequency of use.
- A standard recommendation is to create duplicates in an offsite location using more than one kind of backup software and different kind of storage technology to write the copies so as to safeguard against software bugs and hardware failure [55].
- Some recommend consolidation and limiting the number of different media types in the collection. Inventory existing digital holdings and quantify their significant properties; maintain that inventory as the collection grows [55].
- Develop a timetable for evaluating holdings including integrity checks of the bitlevel data, media refreshing and retention evaluation.
- Implement a technology watch protocol to ensure that no media type, file format or standard becomes obsolete before objects associated with any of the above have been addressed sufficiently.

[54] http://www.abc.net.au/science/news/stories/s1680304.htm
[55] http://www.chin.gc.ca/English/Pdf/Digital_Content/Digital_Preservation/digital_preservation.pdf

5 MIGRATION – COPYING FILES TO NEW MEDIA

Data migration is often referred to as the process of transferring files between (same or different) storage media types without altering the file formats. File migration has many parallels to file conversion, i.e. copying data from one file format into another, as by all these processes should be performed quality assurance procedures. Migration may be divided into two categories:

- **Refreshment:** A very basic migration strategy of copying the information to a fresh physical carrier of the same type.
- **Media change:** Copying the information from a less stable medium, such as a floppy disc to a more stable one, such as optical disc or hard drive.

For both migration strategies a *migration path* should be clearly defined and be described. The migration path shall document how an organization will safely and completely transfer long-term and archival records from one generation of hardware and software to another generation. The strategy should be written and maintained with the system documentation. Current strategies for data migration include[56]:

- upgrading equipment and software as technology evolves and periodically recopying optical or magnetic storage media as required;
- recopying optical or magnetic storage media based upon projected longevity and/or periodic verification of the records;
- or, transferring the data from an obsolete generation of optical or magnetic storage media to a newly-emerging technology, in some cases bypassing the intermediate generations that are mature but at risk of becoming obsolete.

According to the PrestoSpace[57] EU project optical storage media have managed to change dramatically between the onset of CDs (1982) to Blue-Ray or HD DVD (2004) leading to less compatibility across formats and hardware. They argue that using optical media storage will result in a multitude of conflicting formats and compatibility issues, i.e. neither cross compatible nor backwards compatible towards DVD and CD players. Hard drives, in comparison, have changed little in terms of technology which means that data migration from one hard drive to another is much less problematic than with optical media storage.

5.1 Some selected patents

A quick search with search key words: data + migration + (verification v media), at the European Patent Office revealed a series of patents within the field:

- Data migration system and method for migrating data from a first storage media to a second storage media (patent (2005): WO2005066845 [58]).
- A method and system for effectively and rapidly migrating recorded content from one storage-media to a second storage-media (patent (2006): WO2006012328)[59].
- A method for concurrent data migration includes classifying files to be migrated into plural jobs, selecting media to which to migrate each job, and using plural drives concurrently to write the jobs to the media. (patent (2005): US2005033932[60])

[56] http://www.cslib.org/publicrecords/optical.htm
[57] FP6-IST-507336 PrestoSpace Deliverable D12.6 Survey of Digital Formats for Storage (2006) www.prestospace.org/project/deliverables/D12-6.pdf
[58] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=WO2005066845&F=0&QPN=WO2005066845
[59] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=WO2006012328&F=0

- Virtual control unit and method for controlling data migration: To perform data migration processing among a plurality of storage media without changing identification information of a volume which a host computer recognizes as an object to be accessed. (patent (2004): JP2004227558[61]).

- To provide a simple data migration method and device for simplifying verification by migrating data without stopping this system between new and old systems capable of using common data. (patent (2005): JP2005266973[62])

- A hierarchical data storage system including a policy based migration engine to select a migration policy and migrate data from a first set of removable storage media, such as tape cartridges, to a second set of removable storage media in accordance with the migration policy. (patent (2006): US2006101084[63]).

5.2 Challenges of migration:

Technical Challenges - Any copying of files introduces the risk of alteration which, in the worst case, can render the file unreadable. The minimum quality assurance should always be a quality check of the migrated files, sometimes a system check may even be required.

Refreshment and migration may be very time consuming. The migration of data from a tape to an optical storage media is mainly dependent on the write speed, e.g. 100 TB on tape and a typically write speed of 4 MB/s would result in 290 days of migration[64].

Organisational Challenges - Companies are seeking well-managed storage systems to comply with regulatory requirements valid for the type of data stored and the applicable law in the country or countries the data applies to. Whatever technology selected, an upgrade or migration to more commoditized platforms also presents a strategic opportunity to either dispose of data that are no longer needed or to convert them into newer format types.

The prospect of data migration can be overwhelming. Some of the common conditions that would be found analysing situations where data migration will be needed are familiar to many IT managers:

- ✓ **Lack of clear definition of requirements for all data.** Data rules should focus on security, availability and recoverability. It's easy to imagine that documents with mandatory data (based on the law) and optional data could be intermingled, making it difficult to determine which data is important and which isn't.
- ✓ **Distributed islands of data.** Often, a business unit will implement a new application and request that the infrastructure for it remains close at hand. Unfortunately, organizational politics can worsen this phenomenon in the IT department, too.
- ✓ **Funding constraints.** Tight budgets may limit technology decisions and options. A company may invest in technology for projected bottom-line benefits, only to find that other factors will interfere with hoped-for business impact.
- ✓ **Lack of expertise in heterogeneous storage environments.** With each storage system vendor's support limited to its own products, incompatibility between storage technologies becomes the problem of the IT manager.

The selected technology for data storage systems can, and will affect the data migration strategy. The way the storage system organizes the data will have a great impact on the strategy for data migration. An example will be a storage system based on HSM (Hierarchical Storage Management) versus file system based storage. The HSM system will theoretically be the same

[60] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=US2005033932&F=0
[61] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=JP2004227558&F=0
[62] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=JP2005266973&F=0
[63] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=US2006101084&F=0
[64] http://www.gwdg.de/forschung/veranstaltungen/workshops/langzeitarchivierung/2006/slides/Migration_Scheller.pdf

as the file system, but since the HSM uses pointers into slower media as for instance tape, the data migration process in this case will be much more time consuming.

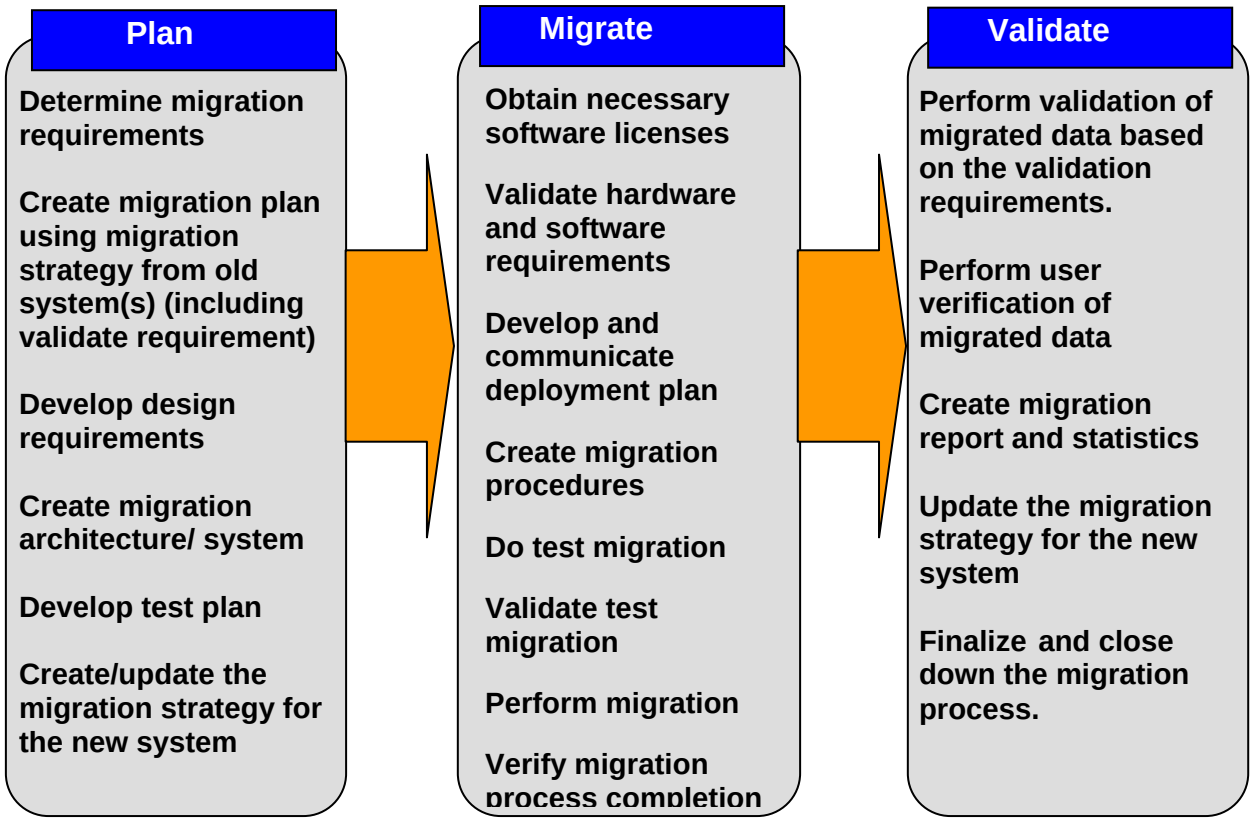
To avoid most problems with future data migration a data migration strategy document should be a part of the design / development process for a new system. This strategy should address the following issues:

- ✓ **Data classification**. The metadata rules for security and availability. The rules have to include specification of which data is required by the law (mandatory data), and which data that are optional. The data classification is to be described based on the existing rules at the time of implementing the storage system. If the system metadata are changed at a later stage, the strategy document has to be updated to reflect the current system.
- ✓ **Hardware / software description**. A description of the hardware and software used to build the system, including storage size requirement. This should also include a recommended export procedure, and if possible a basic source code or a recommendation to tools for exporting data from the system.
- ✓ ***System environment***. This part describes the environment needed to run the system such as power requirements, network requirements and other similar environment variables.

The strategy document should be input to a future migration plan and migration process for data from the system. When the strategy turns into a migration plan some time in the future it should be remembered that the following three key issues have to be addressed:

- ✓ The actual migration plan
- ✓ The migration itself.
- ✓ The validation of the migrated data

The figure below shows an overview of a complete migration process.



For additional information about the chosen migration strategies in some LongRec case partners it is referred to Appendix D.

6 FILE FORMATS FOR THE LONG-TERM STORAGE

A file format specifies a bit stream that is read/written from/to a non-volatile storage medium, see ref [65] for recommendations on file format design. Without the exact knowledge of the stream of 0s and 1s, files cannot be interpreted correctly and the data content in file can in the worst case be completely useless. The file format must therefore be seen as a way of decrypting the content of a file. Formats can be distinguished between two different groups: Proprietary and open formats.

- **Proprietary formats** i.e. some company or organisation owns the file specifications and do not want to make them public. The format code is usually not available to the end user and restrictions for using and modifying any proprietary file format may apply.

- **Open formats** are always fully documented, no license or license fees required, and the user can freely modify the format structure.

Anyway, it is usually difficult to acquire complete and reliable file format specifications from a single source and general resources such as software developers, public FTP archives, monographs and Internet discussion lists will have to be approached[66]. A good starting point is the ongoing activities that collect, document and store all kind of digital formats such as Global Digital Format Registry[67] or PRONOM[68] or others see [69,70,,71,72,73,74,75].

6.1 File format categorisation and identification

There are thousands of different file formats and e.g. IANA[69] provides a categorization into eight main groups which, of course can be broken down further into sub-categories, see Appendix A. As there are many file formats with different versioning, a challenge is to recognise the encountered format correctly. There exist various ways of (not always uniquely) identifying a file format such as Filename extension[76], Magic number[77], Explicit metadata, OS type-codes, Uniform Type Identifiers (UTIs), OS/2 and POSIX Extended Attributes, PRONOM Unique Identifiers[78], MIME types or File format identifiers. For a nice overview see [79].

Table 2: Methods of file format identification

Identifier	Characteristics
File extension	The characters following the period after the file name give info about the type of file e.g. *.html, *.GIF, *doc. For a list see[80]
Magic number	A number or an ASCII string within or at the beginning of a file identifies it uniquely.
Explicit metadata	Information about the format is explicitly stored in the file system.

[65] http://www.magicdb.org/filedesign.html
[66] Greg Lawrence, William Kehoe, Oya Y. Rieger, William Walters, and Anne R. Kenney, Risk Management of Digital Information: A File Format Investigation. Washington, DC: Council on Library and Information Resources, 2000.
[67] Global Digital Format Registry http://hul.harvard.edu/gdfr
[68] PRONOM, UK National Archives http://www.nationalarchives.gov.uk/pronom
[69] Internet Assigned Names Authority (IANA) MIME type registry.www.iana.org/assignments/media-types
[70] http://www.wotsit.org
[71] http://en.wikipedia.org/wiki/List_of_file_formats
[72] http://dmoz.org/Computers/Data_Formats/
[73] http://www.matisse.net/files/formats.html
[74] http://www.digitalpreservation.gov/formats/index.shtml
[75] http://www.magicdb.org/stdfiles.html
[76] http://filext.com, http://www.file-extensions.org/
[77] http://www.garykessler.net/library/file_sigs.html
[78] http://www.nationalarchives.gov.uk/pronom/#
[79] http://en.wikipedia.org/wiki/File_format
[80] http://www.file-extensions.org/

Mac OS type-codes	Store information for *creator* and *type* as part of the directory entry for each file.
Mac OS X Uniform Type Identifiers (UTIs)	Use strings as identifier of file types and are based on registered domain names. UTIs as replacement for Mac OS type-codes.
OS/2 Extended Attributes	Use a set of triplets with an unique name, a coded type for the value and a value.
POSIX extended attributes	Allow for an arbitrary list of unique "name=value" strings.
PRONOM Unique Identifiers (PUIDs)	PUIDs can be expressed as Uniform Resource Identifiers using the info:pronom/ namespace (developed by The National Archives of the UK)
MIME types	consist of a standardized system of identifiers (managed by IANA) consisting of a *type* and a *sub-type*, separated by a slash — for instance, text/html or image/gif. Unfortunately there are several MIME versions.
File format identifiers (FFIDs)	A string of digits in form: NNNNNNNNN-XX-YYYYYYY, where “N…” indicates the organisation origin/maintainer, “XX” categorize the type of file in hexadecimal, and ”Y...” is the usual file extension.

JSTOR[81] and the Harvard University Library[82] have developed JHOVE (JSTOR/Harvard Object Validation Environment[83]), a tool to automate the format-specific identification, validation and characterization of file formats. It provides answers to questions like:

1. Identification
    a. "I have an object; what format is it?"
2. Validation
    a. "I have an object that purports to be of format *F*; is it?"
    b. "I have an object of format *F*; does it meet profile *P* of *F*?"
    c. "I have an object of format *F* and external metadata about *F* in schema *S*; are they consistent?"
3. Characterization
    a. "I have an object of format *F*; what are its salient properties (given in schema *S*)?"

A initial implementation of the JHOVE framework includes the modules for correct recognition, validation and characterisation of arbitrary byte streams, ASCII and UTF-8 encoded text, GIF, JPEG2000, and JPEG, and TIFF images, AIFF and WAVE audio, PDF, HTML, and XML; and text and XML output handlers.

The DROID (Digital Record Object Identification) software tool developed by The National Archives allows automated batch identification of file formats, DROID can be downloaded from [84].

The National Library of New Zealand developed an open-source Metadata Extraction Tool, for more info it is referred to section 7.1.

A less comprehensive but awarded tool for file format identification based on their binary signatures is TrID, downloadable as freeware from[85]. It is extensible and can be trained to recognize new formats in a fast and automatic way.

[81] http://www.jstor.org/
[82] http://hul.harvard.edu/
[83] http://hul.harvard.edu/jhove/index.html
[84] http://droid.sourceforge.net
[85] http://mark0.net/soft-trid-e.html

6.2 Basic File Formats

File formats can roughly be categorised into two groups based on compression methods applied:

- Lossless data compression applies data compression algorithms that allow exact reconstruction of the original data from the compressed data, for instance ZIP, PNG, GIF or JPEG 2000.
- Lossy data compression, some data are lost during compression. Lossy compression is quite common to reduce file size of multimedia data (audio, video, still images), and especially in streaming media and internet telephony, for instance JPG, MPEG, WMA.

If a lossless file format is chosen then the probability that it can be read also after 10 years is quite high. However, when using lossy file formats the readability of the file depends directly on the existence and availability of codecs (the term *codec* is a combination of "Compressor" and "Decompressor") and driver software for the "controllers" and operating systems of the future – which has shown to be of great difficulty[86]. Also the fact that this compression is often done with patented algorithms may decrease the lifetime of the file format.

6.2.1 Wrappers and containers

A container format, also called wrapper, is a computer file format that can hold various types of data. These wrappers can contain different types of audio or video codecs, video streams, subtitles, chapter-information, and metadata.

Some containers are exclusively for audio, others only for video and some are both. The table below lists some of the more popular containers.

Abbr.	Description	proprietary
WAV[87]	Waveform audio format is an audio container usually holding uncompressed audio data. It is readable by virtually all audio software programs and has become a de facto standard. Recommended for long-term file storage.	y
TIFF[88].	Tagged Image File Format is an image container only. It incorporates perhaps the most comprehensive metadata support of any raster format, allowing the addition of a wide variety of technical and resource discovery information to be included. TIFF is designed to be an extensible format, and new tags can be registered with Adobe.	y
AVI[89]	The Audio Video Interleaved is a special case of the RIFF (Resource Interchange File Format). AVI is a container that can hold both audio and video data. It is the most commonly used audio/video container but its actual appearance is very dependent on the used coding. According to [90] AVI files should be converted to more stable formats since Microsoft declared that they will soon no longer support this format.	y
MXF[91]	The Material eXchange Format (MXF) is an open file format targeted at the interchange of audio-visual material with associated data and metadata.[92]	n
ZIP[93]	popular data compression and archival format.	n
AIFF	See section 6.2.4	n

See appendix A for an overview of file format categories. In the following only a selection of the most common file formats are given together with references on further information.

[86] Harken, H. (2007) Physikalisch-Technische Bundesanstalt. www.ptb.de/en/org/2/25/251/lifetime.pdf
[87] http://www.sonicspot.com/guide/wavefiles.html
[88] http://partners.adobe.com/public/developer/tiff/index.html
[89] http://www.digitalpreservation.gov/formats/fdd/fdd000059.shtml
[90] Peters McLellan, E. 2007. Selecting Digital File Formats for Long-Term Preservation. *InterPARES 2 General Study 11 Final Report*
[91] http://mxf.info/
[92] Devlin B. The Material eXchange
[93]: http://www.pkware.com/documents/casestudies/APPNOTE.TXT

6.2.2 Raster Graphics

Abbr.	Description	proprietary
GIF	Graphics Interchange Format, proprietary format, full technical specifications for GIF version 89a are available from CompuServe Incorporated[94].	y
JPEG	JPEG itself is not a file format, but rather an image compression algorithm. The official file format specification is called SPIFF (Still Picture Interchange File Format, ISO 10918-1). Full technical specifications available at[95].	n
JPEG 2000	JPEG 2000 is a replacement for the JPEG algorithm. Full technical specifications are an international standard (ISO/IEC 15444 Part 1)	n
PNG	Portable Network Graphics. The PNG specification is entirely public domain and free to use[96].	n
BMP	Windows Bitmap. A formal technical specification for BMP has not been released by Microsoft.	y
Tiff	It is an image container only, ref 6.2.1	y

The Research Libraries Group and Digital Library Federation recommend the TIFF format ref 6.2.1. Although proprietary it may currently be considered as the one most suited for archival. For additional information see [97, 98, 99].

6.2.3 Vector Graphics

Abbr.	Description	proprietary
DWG	AutoCAD Drawing Format has become a de facto standard for vector graphics. A new The DWG specification is revised with each release of AutoCAD[100]	y
DXF	AutoCAD Drawing Exchange Format. The frequent specification changes can cause compatibility problems. Full technical specifications available at [101].	y
SVG	Scalable Vector Graphics is a format for describing two-dimensional graphics using XML. I is an open, non-proprietary format which is rapidly becoming a major standard for vector imagery[102].	n

6.2.4 Audio File

Audio signals are converted into digital signals by sampling the audio signals of the individual channels with a certain samplings rate. These digital signals can then either be stored uncompressed or compressed to reduce the file size. Refer to the 'Digital Images Archiving Study' and 'Moving Pictures and Sound Archiving Study'[103] for preservation and archiving methodologies.

Abbr.	Description	proprietary
AIFF[104]	The AIFF file type contains uncompressed data, it is actually a wrapper. It was developed by Apple Computer. Recommended for long-term file storage.	n
MP3[105]	MP3 files are MPEG files with audio layer 3 which corresponds to a coding scheme for the	n

[94] http://www.w3.org/Graphics/GIF/spec-gif89a.txt
[95] www.jpeg.org
[96] http://www.w3.org/TR/PNG
[97] Peters McLellan, E. 2007. Selecting Digital File Formats for Long-Term Preservation. *InterPARES 2 General Study 11 Final Report*
[98] http://www.nationalarchives.gov.uk/documents/graphic_file_formats.pdf
[99] http://www.library.cornell.edu/preservation/tutorial/presentation/table7-1.html
[100] http://www.opendesign.com/downloads/guest.htm
[101] www.autodesk.com/dxf
[102] http://www.w3.org/TR/SVG/
[103] http://ahds.ac.uk/about/projects/archiving-studies/index.htm
[104] http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/AIFF/AIFF.html

	compression of audio signals.	
WAV[106]	Waveform audio format is a Microsoft and IBM audio file format standard for storing audio on PCs. WAV files can be encoded with a variety of codecs to reduce the file size.	y

6.2.5 Video File

There exists a multitude of various moving pictures or video file formats. The library of congress lists 69 different moving image formats[107]. Many newer versions are not even encompassed by this list.

Abbr.	Description	proprietary
AVI	Audio Video Interleaved is a special case of the RIFF (Resource Interchange File Format). AVI is said to be an audio/video format but in fact it is a container for such data[108, 109]. It is the most commonly used audio/video container but its actual appearance is very dependent on the used coding According to [110] AVI files should be converted to more stable formats since Microsoft declared that they will soon no longer support this format.	y
MPEG1	Coding of moving pictures and associated audio for digital storage media with bitrates from 500 Kbp/s to 2Mbp/s. It is commonly used on CDs and MP3 files. Open standard developed by ISO technical program JTC 1/SC 29 (WG11)[111]	n
MPEG2	The format was initially developed to serve the transmission of compressed television programs via broadcast, cablecast, and satellite, and subsequently adopted for DVD production and for some online delivery systems[112] with bitrates from 4 to 8 Mbp/s. MPEG-2 has already become obsolescent.	n
MPEG4:	The standard for multimedia for the fixed and mobile web, ISO/IEC 14496-2:2004, with bitrates from 200Kbp/s to 2Mbp/s.	n
MPEG7	Also named "Multimedia Content Description Interface" developed for description and search of audio and video multimedia content [113]	n
MOV	In the Apple Quicktime Format the description of the media is stored separately from the media data.[114] MOV is a proprietary format developed by Apple Computer, Inc. but fully documented[115].	y
WMV	Windows Media Video File Format based in ASF (Advanced Systems Format) which is fully documented[116].	y
MJPEG-2000	Motion JPEG 2000 compresses (around 1:3) each frame separately as a JPEG image and is completely lossless [117]. ISO/IEC Intl. Std. 15444, Information technology – JPEG 2000 image coding system, particularly Part 3: Motion JPEG 2000 (Sept. 2002, with subsequent amendments).	n

For guidelines on preservation of audio objects see also e.g. ref [118]

[105] http://www.wotsit.org/getfile.asp?file=mpeg3&sc=242289287
[106] http://standards.jisc.ac.uk/catalogue/WAV.phtml
[107] http://www.digitalpreservation.gov/formats/fdd/video_fdd.shtml
[108] http://www.digitalpreservation.gov/formats/fdd/fdd000059.shtml
[109] www.n-visual.com/
[110] Peters McLellan, E. 2007. Selecting Digital File Formats for Long-Term Preservation. *InterPARES 2 General Study 11 Final Report*
[111] http://www.iso.org/iso/en/CatalogueListPage.CatalogueList?COMMID=148&scopelist=PROGRAMME
[112] http://www.digitalpreservation.gov/formats/fdd/fdd000028.shtml
[113] http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
[114] http://developer.apple.com/technicalqas/QuickTime/
[115] http://developer.apple.com/documentation/QuickTime/QTFF/qtff.pdf
[116] http://www.microsoft.com/windows/windowsmedia/forpros/format/asfspec.aspx
[117] Pearson, Glenn and Michael Gill, "An Evaluation of Motion JPEG 2000 for Video Archiving", Proc. Archiving 2005, IS & T (www.imaging.org), pp. 237-243.
[118] Bradley, K (2004) Ed., Guidelines on the Production and Preservation of Digital Audio Objects (=IASA-TC 04, IASA Technical Committee, Standards, Recommended Practices and Strategies. Available from http://www.iasa-web.org

6.2.6 Text files

Abbr.	Description	proprietary
ASCII [119]	ASCII (American Standard Code for Information Interchange) text format tends to be the most portable format because it is supported by almost all applications on most platforms. ASCII, or plain text files contain data made up of ASCII characters (*.txt). Very limited in terms of formatting and multimedia support. Problems can occur in rendering the text when transferring files between computers which use different coded character sets.	n
PDF/A [120]	Portable Document Format (PDF) from Adobe. Its specification has been openly published. The PDF/A-1 (ISO 19005-1:2005 *Document management – Electronic document file format for long-term preservation*) standard is based on Adobe's PDF Reference 1.4. PDF/A does not support encryption, LZW compression, embedded files, external content references, transparency, multimedia and JavaScript. Other pdf standards are PDF/X for pre-press data exchange (ISO 15930 parts 1, 3, 4, 5 and 6); PDF/E for engineering, architectural and GIS documents; PDF/UA for handicapped accessibility. The problem is that the feature rich nature of PDF can create difficulties in preserving information over the long term. For example, PDF documents are not necessarily self-contained; some files depend on system fonts and other content drawn from outside the file. As technology changes, these external dependencies can cause information to be lost. Because of these reasons the PDF/A (A for archival) was developed.	n
Doc[121]	Native Microsoft Word format is not stable and changes in various ways. Wordart also changed drastically in a recent version causing problems with documents that used it when moving in either direction. The DOC format's specifications are not available for public download but may be received by contacting Microsoft.	y
OXML [122]	Open XML-based file format from Microsoft for electronic documents such as spreadsheets, charts, presentations and word processing documents. This standard has not yet been approved by ISO (stand September 2007).	n
ODF [123]	Open Document Format for Office Applications (*ISO/IEC 26300*) is a file format for electronic office documents, such as spreadsheets, charts, presentations and word processing documents.	n
HTML [124]	Hypertext Markup Language, is the predominant markup language for web pages. It provides a means to describe the structure of text-based information in a document — by denoting certain text as headings, paragraphs, lists, and so on — and to supplement that text with *interactive forms*, embedded *images*, and other objects.	n
RTF[125]	Rich Text Format is a proprietary document file for cross-platform document interchange. Most word processors are able to read and write RTF documents.	y

6.2.7 Databases

Basically, a database system consists of three components[126]:

- the database itself (the actual content);
- DataBase Management System, DBMS (for example, Oracle 9i);
- the database applications. This incorporates both the graphical user interface and the functionality the user needs to search through and process the content of the database, as well as programs that function automatically to support the system in processing inputs and outputs.

[119] http://dio.cdlr.strath.ac.uk/standards/fileformats/textformats.html
[120] http://www.aiim.org/standards.asp?ID=25013
[121] http://www.microsoft.com/downloads/details.aspx?FamilyId=941B3470-3AE9-4AEE-8F43-C6BB74CD1466&displaylang=en
[122] http://www.ecma-international.org/publications/standards/Ecma-376.htm
[123] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485
[124] http://www.w3.org/html/
[125] https://www.microsoft.com/downloads/details.aspx?familyid=AC57DE32-17F0-4B46-9E4E-467EF9BC5540&displaylang=en
[126] From Digital Volatility to Digital Permanence : Preserving Databases (version 1.0), 2003. Digital Preservation Testbed. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-databases-en.pdf

A variety of DataBase Management Systems are in use depending on the underlying data model:[127]

(a) Relational database - Oracle Databases, Microsoft Access

(b) Hierarchical database – a database organized in the form of a tree structure

(c) Native XML database – Tamino

(d) Object database - the Testbed System uses Oracle iFS to map objects to an underlying relational database. So whilst this is not an example of a native object database it is an example of how object related commands could be issued to a database.

(e) Network database

According to Christensen[140] 'One major disadvantage is that database systems aren't suited for long-term storage. Database systems are under continuous development and improvement and normally only the most recent versions are supported. The archive becomes crucially dependent on the selected database system and the dependencies of this system. The consequences of reliance on one database system are comparable to the consequences of reliance on one operating system.' A more detailed discussion of the challenges and potential solutions pathways are given in [103].

According to the Dutch Digital Preservation Testbed initiative, there are virtually no practical studies that have been carried out, either at a national or an international level, into technical approaches to the durable preservation of databases. They recommend that an approach must be found that addresses organisational, legal and technical issues.

6.2.8 Biometric File Formats

The Common Biometric Exchange File Format (CBEFF) describes a set of data elements necessary to support biometric technologies[128]. The current version of CBEFF is defined in NISTIR 6529A[129] as the standard data structure/format for communicating biometric data. The international version of CBEFF has become an ISO standard[130] (ISO/IEC 19785 and ISO 19794-1, …, 6) at the end of 2005, for a detailed draft version (2003) see[131].

The CBEFF can accommodate any biometric technology and standards are or have been developed for a series of biometric identifiers such as Finger Pattern-Based Interchange Format (ANSI/INCITS 377-2004) or Signature / Sign Data Interchange Format (ANSI/INCITS 395-2005), see [132].

CBEFF is forward compatible and includes the definition of format and content for data elements such as:

- A biometric data header that contains such information as version number, length of data, whether the data is encrypted or not, etc., for each biometric type available to the application or system;
- Biometric data (content not specified);
- Any other required biometric data or data structures.

[127] Verdegem, R. (2003) Databases Preservation Issues. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/longterm_preservation_of_databases.pdf

[128] (2001) http://www.oasis-open.org/committees/xcbf/docs/NISTR6529-CBEFF.pdf

[129] http://www.itl.nist.gov/div893/biometrics/documents/NISTIR6529A.pdf

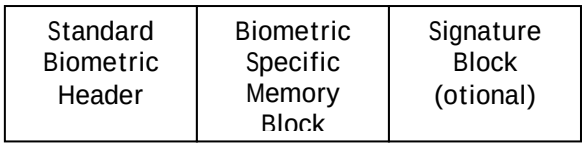
[130] http://iso.w3j.com/ISO-Sheet124_list_76.html

[131] http://isotc.iso.org/livelink/livelink/3923839/JTC001-SC37-N-352.pdf?func=doc.Fetch&nodeid=3923839

[132] http://www.bioapi.org/history.asp

The OASIS XML Common Biometric Format (XCBF) Technical Committee is defining a common set of secure XML encodings for the patron formats as specified in CBEFF (i.e. NISTIR 6529)[133]

The CBEFF file standard consists of 3 sections:



6.2.9 Signature formats

Digitally signed objects will become more common in the future. In a long-term perspective their main challenge will be to keep their validity even if the signer or verifying party later attempts to deny it (i.e., repudiates the validity of the signature). Four different long-term signature formats based on XML signatures were developed:

- Basic Electronic Signatures (XAdES-BES),
- Explicit Policy based Electronic Signatures (XAdES-EPES).
- Electronic Signature with Time (XAdES-T),
- Electronic Signature with Complete Validation Data References (XAdES-C)

These XML signatures implement ETSI TS 101 903 V1.3.1 (2005-05), "XML Advanced Electronic Signatures (XAdES)" standard, for more detail it is referred to [134]. ETSI TS 101 733 v1.6.3, September 2005. “CMS Advanced Electronic Signatures (CAdES). RFC3126 is technically equivalent to an earlier version (1.2.2) of ETSI TS 101 733.

6.3 File format obsolescence

No matter what file format is chosen there will always be the danger that it will become obsolete as time goes by. The reasons can be as diverse as:

1. The format itself is superseded by another one or evolves in complexity.
2. The format "take up" is low or industry fails to create compatible software.
3. The format fails, stagnates, or is no longer compatible with the current environment.
4. Software supporting the format fails in the marketplace, is no longer supported by vendors or is bought by a competitor and withdrawn.

Stanescu at the Online Computer Library Center describes the INFORM (INvestigation of FOrmats based on Risk Management) and has developed a methodology measuring the durability of various file formats. Such a risk assessment will have to address (excerpt from [135]):

> 1. **Digital object format** - risks introduced by the specification itself, but also including compression algorithms, proprietary (closed) vs. open formats, DRM (copy protection), encryption, digital signatures.
> 2. **Software** - risks introduced by necessary software components such as operating systems, applications, library dependencies, archive implementations, migration programs, implementations of compression algorithms, encryption and digital signatures.

[133] http://www.cesg.gov.uk/site/ast/index.cfm?menuSelected=4&subMenu=4&displayPage=413
[134] XAdES Long-Term Signature Format Profile Version 1.0 (2006) *Next Generation Electronic Commerce Promotion Council of Japan (ECOM)* http://www.ecom.jp/LongTermStrage/en/XAdES01_e.pdf
[135] Stanescu, Andreas, “Assessing the durability of formats in a digital preservation environment,” *D-Lib Magazine* 10:11 (November 2004). http://www.dlib.org/dlib/november04/stanescu/11stanescu.html

3. **Hardware** - risks introduced by necessary hardware components including type of media (CD, DVD, magnetic disk, tape, WORM), CPU, I/O cards, peripherals.
4. **Associated organizations** - risks related to the organizations supporting in some fashion the classes identified above, including the archive, beneficiary community, content owners, vendors, open source community.
5. **Digital archive** - risks introduced by the digital archive itself (i.e., architecture, processes, organizational structures).
6. **Migration and derivative-based preservation plans** - risks introduced by the migration process itself, not covered in any other category.

It may not be necessary to continuously monitor format obsolescence, a more cost efficient strategy would be to follow leading organisations and libraries and convert obsolete files into newer formats – called conversion – whenever they do.

The National Archives of Australia have just released version 4.0 of the digital preservation software Xena[136] which is designed to:

- Determine file formats
- Convert files into standards based, open formats for preservation

Cornell's Virtual Remote Control (VRC), the Global Digital Format Registry (GDFR), and VersionTracker. A project called PANIC tries to integrate those services using web service techniques. (PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services). The results of this approach are available as a software tool called AONS (Automatic Obsolescence Notification System) from [137].

6.4 Recommendation of file format choice for the long-term storage

It goes without saying that open file formats should be preferred in a long-term storage setting but many programs do not offer any choices with respect to storing a file in different formats whereas others may allow it. Different file formats do not only store the data in different bit streams but may actually not allow storing all the desired information. The most obvious example is MS Word where a lot of document information is lost when storing the document in e.g. plain text. In general, conversion between (or storing in) different formats usually leads to some loss of information. This loss could be encoded functionality, data or meta data.

The user will also have to consider how the stored file shall be used in the future. If it is really necessary to have all the functionality encoded in the file then a format shall be chosen that encodes for that whereas if only viewing or machine reading is required than some presentation or ASCII encoding would be suitable enough. Also storage space requirements, length of intended storage time ect. may give some constraints for format choice. According to the US Library of Congress there are 7 factors that should be considered when evaluating a digital format as they influence the feasibility and cost of preserving the information content in the face of future transitions to other formats (excerpt from [138]).

1. Disclosure. Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.
2. Adoption. Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.
3. Transparency. Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.

[136] http://xena.sourceforge.net/
[137] http://www.apsr.edu.au/aons or http://www.apsr.edu.au/aons2
[138] http://www.digitalpreservation.gov/formats

4. Self-documentation. Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.
5. External Dependencies. Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.
6. Impact of Patents. Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.
7. Technical Protection Mechanisms. Implementation of mechanisms such as encryption that prevent the preservation of content by a trusted repository.

Other organisations and projects have derived similar recommendations with respect to selection of long-term file formats [139, 140]. According to them the format should

- be an open format
- be simple to describe, understand and implement
- not depend on specific hardware
- not depend on specific operating systems
- not depend on proprietary software
- be robust against single points of failure
- suited for long-term storage
- be OAIS compatible
- support all important Internet protocols
- support meta-data
- be easy to verify and maintain
- be simple to backup
- support recording of access limitations
- support authenticity information
- be possible to retrieve the original bit-stream
- be possible to delete material from the archive
- be easy to locate archived data
- support format transformations
- support data compression
- support duplicate reduction
- the format should be efficient

The UK National Archive developed an on-line information system about data file formats and their supporting software products called PRONOM[141]. It gives detailed information about what file formats individual software products can read and write. It goes without saying that their list is incomplete. It may also be referred to the InterPARES study 'Selecting Digital File Formats for Long-Term Preservation' [142] for issues related to selection criteria wrt. long-term storage.

In general, widespread adoption, non-proprietary origin, published specifications, interoperability and lack of compression (or lossless compression) appear to be the most important aspects when selecting digital file formats for long-term storage. However, it is often not possible to identify formats having all these attributes and compromises will have to be made that will differ between institutions depending on their long-term preservation policy. InterPARES2 gives the following recommendations for forming one's own preservation policy:

[139] Christensen, S. S. (2004) Archival data format requirements. *Stats Bibliotek Report of the Royal Library, Denmark.* http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf
[140] Christensen, S. (2004) Archival Data Format Requirements. *The Royal Library, Denmark.* http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf
[141] http://www.nationalarchives.gov.uk/pronom/#
[142] http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf

1. Define/clarify the terminology such as *open*, *standard*, *stable* and *well-documented*, in the policy documents.

2. Distinguish between file formats, wrapper (or container) formats, and tagged formats such as XML-tagged files, and ensure that version, encoding and other characteristics are understood and fully specified.

3. For XML files, require that the files be well-formed and valid and accompanied by the relevant DTDs or schemas.

4. Choose widely-used, non-proprietary, platform-independent formats with freely available specifications where possible.

5. Specify whether compressed files are acceptable, and if so, state the type of compression permitted. Where possible, choose lossless compression techniques that conform to accepted international standards.

6. If it is not feasible to choose formats with the characteristics listed in recommendation 4, choose formats that are being preserved at other digital repositories and collaborate with these other repositories to develop preservation plans for them.

7. Where possible, work with records creators to ensure that they use software programs that create records in formats that meet the criteria listed in recommendation 4.

8. Prioritize the relative importance of each format type and the resources allocated to supporting that format. Identify formats that the institution will not support and ensure creators/depositors are informed of this.

The overall recommendation is not to allow lossy compression but that is only a clear recommendation for professional quality material[143]. If compression formats have to be chosen then MPEG-4 (MPEG-2 has already become obsolete) and MJPEG 2000 should be chosen as multimedia formats.

Norway has specified its own requirements for public archived digital objects[144]

[143] D12.6 Survey of Digital Formats for Storage, *FP6-IST-507336 PrestoSpace Deliverable*, 2006
[144] http://www.lovdata.no/for/sf/kk/kk-19991201-1566.html

7 PRESERVATION METADATA AND METADATA FRAMEWORK

Preservation metadata is usually not seen as another category of metadata (see also Dublin Core Metadata Best Practices [145]), but as a combination of existing metadata sets that provide the information needed for long-term preservation of and permanent access to digital objects. Preservation metadata have to contain technical details on the format, structure and use of the digital content; the history of all actions performed on the resource including changes and decisions; authenticity information such as technical features or custody history; and the responsibilities and rights information applicable to preservation actions. Administrative metadata and technical metadata are generally considered to be the most important for preservation[146].

A preservation metadata framework gives an overview or description of the types of metadata that should be associated with an archived digital object. A more systematic approach to developing a preservation metadata framework might involve the specification of a formal information model such as 'The Open Archival Information System (OAIS) reference model (ISO14721). The OAIS reference model is a conceptual framework for a digital archive. The model establishes terminology and concepts relevant to digital archiving, identifies the key components and processes endemic to most digital archiving activity, and proposes an information model for digital objects and their associated metadata.

It makes however no presuppositions about the type of digital object managed by the archive, nor about the specifics of the technology employed by the archive to achieve its goal of preserving and maintaining access to the digital object over the long term[147]. For practical purposes an implementation of the high level information model has to be developed that gives a detailed specification of the required set of preservation metadata elements. The list below gives an indication of currently a large number of preservation metadata frameworks that have been developed so far[148].

- METS, the Metadata Encoding and Transmission Standard[149];
- the Australian Recordkeeping Metadata Schema; (PADI)[150]
- the New South Wales Recordkeeping Metadata Standard[151];
- the Recordkeeping Metadata Standard for Commonwealth Agencies[152];
- the South Australian Recordkeeping Metadata Standard[153];
- the VERS (Victorian Electronic Records Strategy) Metadata Scheme[154];
- National Library of New Zealand –Metadata Standards Framework – Preservation Metadata[155]
- the Record Keeping Metadata Requirements for the Government of Canada[156];
- the Arizona Electronic Recordkeeping Systems (ERS) Guidelines-IV Functional Requirements for Recordkeeping Systems;
- the Minnesota Recordkeeping Metadata Standard;
- the PERM Preservation Attributes;
- GILS (Government Information Locator Service)
- IS0 82045-2 Document Management Metadata;
- ISO 23081 Records Management Metadata Standard[157]

[145] Dublin Core Metadata Best Practices, *Ver. 2.1.1 CDP Metadata Working Group* (2006) http://www.cdpheritage.org/cdp/documents/cdpdcmbp.pdf

[146] Verheul, I. (2006) Networking for Digital Preservation. *Saur, ISBN 10: 3-598-21847-8* http://www.ifla.org/VI/7/pub/IFLAPublication-No119.pdf

[147] Preservation Metadata for Digital Objects: A Review of the State of the Art. A White Paper by the OCLC/RLG Working Group on Preservation Metadata (2001). http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf

[148] Gilliland-Swetland, A; McKemmish, S. A Metadata Schema registry for the Registration and Analysis of Recordkeeping and Preservation of Metadata. *Final Program and Proceedings. Society for Imaging Science and Technology. 2005, p 109-112.*

[149] http://www.loc.gov/standards/mets/

[150] http://www.nla.gov.au/padi/topics/30.html

[151] http://www.records.nsw.gov.au/publicsector/erk/metadata/rkmetadata.htm

[152] http://www.naa.gov.au/recordkeeping/control/rkms/contents.html

[153] http://www.archives.sa.gov.au/management/standards.html

[154] http://www.prov.vic.gov.au/vers/vers/default.htm

[155] http://www.natlib.govt.nz/catalogues/library-documents/preservation-metadata-revised

[156] http://www.imforumgi.gc.ca/meetings/2006/06-08/gcrmm-mgdgc/page01_e.asp

- TBITS 39 & 39.1 TBS-Std -TBS-GOL Metadata Standard
- the CEDARS metadata specification for preservation[158];
- MARC; ISO 23081-1: 2004[159]
- XrML[160], eXtensible rights Markup Language
- Open Digital Rights Language (ODRL);
- Digital Rights Expression Languages (DREL),
- Online information Exchange (ONE);
- Preservation Metadata - Networked European Deposit Library (NEDLIB) Metadata for Long Term Preservation;
- NLA Pandora Metadata Element set;
- RLG[161]
- NISO U9.87-2002 ***ALM*** 20-2002 Data Dictionary - Technical Metadata for Still Images,
- Metadata for Images in **XML** (MIX);
- a range of geospatial metadata standards;
- PREMIS metadata set[162]

According to an analysis performed by Gilliland-Swetland[148] in 2004 the schemas contain approximately 120 different fields which can be organised at different organisational hierarchic levels. At the highest level there are 11 elements: Registration, Identification, Accessibility, Rights, Provenance, Description, Analysis, Documentation, Relationships, Administration, and a general Note element. On a sub-level, metadata elements may address areas specific for the digital object e.g. an image: date, watermark, transcriber, resolution, producer, compression, capture device, source, capture details, colour change history, colour management, validation key, colour bar/greyscale bar, encryption, control targets.

To the author's knowledge there does not exist an all-encompassing preservation metadata standard. For an example of a metadata set it is referred to Appendix A: CEDARS preservation metadata element set[163]. A comprehensive review of the state-of-the-art on 'Preservation Metadata for Digital Objects' was made by the *OCLC/RLG Working Group on Preservation Metadata* and can be found at [164] and [165].
In an effort to identify relevant metadata sets *The Metadata and Archival Description Registry and Analysis System (*MADRAS [166, 167]) was initiated which was further developed into an analytical assessment tool that could be used to evaluate the current capabilities of registered metadata schemas.

[157] https://committees.standards.org.au/COMMITTEES/IT-021/N0001/ISO_23081-1_2006.pdf
[158] http://www.leeds.ac.uk/cedars/guideto/metadata/
[159] https://committees.standards.org.au/COMMITTEES/IT-021/N0001/ISO_23081-1_2006.pdf
[160] http://www.xrml.org/
[161] http://www.rlg.org/preserv/presmeta.html
[162] Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group (2005). http://www.loc.gov/standards/premis/
[163] Preservation Metadata for Digital Objects: A Review of the State of the Art. A White Paper by the OCLC/RLG Working Group on Preservation Metadata, 2001 http://cendicites.infointl.com/item300.html
[164] (2001) http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf
[165] (2002) http://www.oclc.org/research/projects/pmwg/pm_framework.pdf
[166] InterPARES 2: Description Cross-domain Report: Investigating the Roles and Requirements, Manifestations and Management of Metadata in the Creation of Reliable and Preservation of Authentic Digital Entities. *Ver. 2 Draft 2007*
[167] For further details on the development of MADRAS, see Anne J. Gilliland, Nadav Rouche, Joanne Evans and Lori Lindberg (2005). "Towards a Twenty-First Century Metadata Infrastructure Supporting the Creation, Preservation and Use of Trustworthy Records: Developing the InterPARES 2 Metadata Schema Registry," *Archival Science* 4(1): 43-78;

7.1 Preservation Metadata Extraction tools

The list below from 2004 contains available metadata extraction software, for more details see [168].

JHOVE[169]	JHOVE provides functions to identify, validate, and characterize digital objects, it has three main operational modes: • Format identification • Format validation • Format *characterization* is the process of retrieving the significant properties of an object of format X.
National Library of New Zealand Metadata Extract Tool[170]	The National Library developed an open-source Metadata Extraction Tool in 2003 (redeveloped in 2007)[171]. It programmatically extracts preservation metadata from the headers of a range of file formats, including PDF documents, image files, sound files and Microsoft Word documents. The output is a standard format (XML) for uploading into a preservation metadata repository.
DCM (Digital Collection Manager)[172]	The DCM is a database application that supports digitization workflows including upload and download of files to and from the Library's Digital Object Storage system (DOSS). The system records management and technical metadata about digital collection items, including relationships between parts of a work and between various copies of those parts (e.g. originals, masters, view copies etc.), records process information about creation of copies, and, for images, extracts relevant technical metadata from file headers.
OPUS[173]	OPUS is a commercial product which is designed to work with flatbed or planetary scanners to manage imaging workflow, including scanning, image post-processing, derivative creation and metadata creation. OPUS supports multisource metadata input, including technical metadata from image headers and descriptive and structural metadata via OCR and intelligent interpretation of scanned images. Metadata can be output to custom and standard formats including METS XML.
AONS[174]	AONS (Automatic Obsolescence Notification System) is a preservation metadata capture tool using Semantic Web services. It was developed by the PANIC[175] project.

7.2 Digital Signatures Metadata

Preservation repositories use digital signatures in three main ways[176]:

1. For submission to the repository, an agent (author or submitter) might sign an object to assert that it truly is the author or submitter.
2. For dissemination from the repository, the repository may sign an object to assert that it truly is the source of the dissemination.
3. For archival storage, a repository may sign an object so that it will be possible to confirm the origin and integrity of the data.

Usage of digital signatures upon submission and dissemination to and from the repository is already common today. The validation occurs relatively shortly after the signing and there is

[168] RLG Digi News, (2004) Volume 8, Number 5 http://www.rlg.org/en/page.php?Page_ID=20462#article5
[169] http://hul.harvard.edu/jhove/jhove.html
[170] http://www.natlib.govt.nz/about-us/news/all-news-items/metadata-extraction-tool-announced/
[171] http://meta-extractor.sourceforge.net/
[172] http://www.nla.gov.au/dsp
[173] http://imageaccess.com/dlsg/products_dline.htm
[174] http://www.metadata.net/panic
[175] http://www.metadata.net/panic
[176] Final Report of the PREMIS Working Group (2005) Data Dictionary for Preservation Metadata http://www.oclc.org/research/projects/pmwg/premis-final.pdf

little need to preserve the signature itself over time. According to PREMIS[177] in the first situation the storage repository may record the validation as an *Event*, and save related information needed to demonstrate provenance. In the dissemination situation the repository could record the signing as an *Event* but the use of the signature is the responsibility of the receiver. Only in the third case, where digital signatures are used by the repository as a tool to confirm the authenticity of its stored digital objects over time, must the signature itself and the information needed to validate the signature be preserved. PREMIS suggests the following metadata on signature information in order to be able to later validate a digital signature in a preservation repository:

- The digital signature itself.
- The name of the hash algorithm and encryption algorithm used to produce the digital signature.
- The parameters associated with these algorithms.
- The chain of certificates needed to validate the signature (if a certificate model is used to relate the signer and the signer's public key).

It is recommended that a repository also stores the definitions of the algorithms and relevant standards (e.g., for encoding the keys) so that these methods could be reimplemented if necessary. PREMIS signature metadata are:

signatureInformation

- signatureInformationEncoding
- signer
- signatureMethod
- signatureValue
- signatureValidationRules
- signatureProperties
- keyInformation
- keyType
- keyValue
- keyVerificationInformation

It is referred to the TRUST report in the LongRecp project for more detailed coverage of this topic.

[177] Final Report of the PREMIS Working Group 2005 http://www.loc.gov/standards/premis/

8 CONVERSION – CHANGE OF FILE FORMAT

Data conversion is the process where data coded in one format is changed to a different one. Data conversion is usually done when the old and the new system still co-exist. It is then easier to copy the file content over into the new file format. This process always contains the risk of data and information loss when for instance the new format does not support features of the old version. Conversion errors that appear in important or frequently used files are more easily identified and often trigger modification of the system or the software application. However, conversion errors that only appear in rarely accessed files do usually not lead to a system modification and hence file conversion as a means for long-term preservation is error prone. It is referred to "Risk Management of Digital Information: A File Format Investigation[178], for a detailed study of the impact of conversion.

In general, as there is no single, accepted and correct strategy that guarantees long-term preservation and access to all different types of digital objects various strategies should consequently be pursued. However, a common basis for all strategies is that long-term storage requires proper procedures for media refreshment and a good backup regime, i.e. several copies of the information.

The three main preservation strategies are:

- **Bit-level preservation:** independent of format obsolescence, requires secure storage, with proper procedures for backup and refreshment, e.g. PREMIS[190]
- **Normalisation:** conversion of formats into a set of acceptable formats. Some archives go as far as to convert all document files into TIFF images, e.g.[179].This strategy would imply that they are only readable by humans (or OCR) and that many media types cannot be handled.
- **Conversion**

Converting data from one format to another contains a measurable risk which of course will vary, sometimes significantly, depending on the source and target format and context. Lawrence et al showed that in fact it is possible to identify the levels of risk originating from different formats as well as from organizational, hardware, software, and metadata issues[180]. Interestingly, they were not able to recommend a cost-effective, off-the-shelf commercial software program where this risk based approach could be implemented into. From an automation point of view conversion and migration software should have the following functionality:

- Read the source file and analyze the differences between it and the target format.
- Identify and report the degree of risk if a mismatch occurs.
- Accurately convert the source file(s) to target specifications.
- Work on single files and large collections.
- Provide a record of its conversions for inclusion in the conversion project documentation.

Conversion test files

All file format conversion will have to be accompanied with quality assurance processes to make sure the conversion process has not altered the file content unexpectedly. As conversions usually encompass thousands of files this task must be automated. Usually, test files are designed that contain important features of a file, i.e. properties that shall be preserved. It may however be too time-consuming and costly to develop test files for all properties and some

[178] http://www.clir.org/pubs/reports/pub93/pub93.pdf
[179] Goethals, A. Action Plan: PDF 1.2. *FCLA*. (2003). http://www.fcla.edu/digitalArchive/pdfs/action_plans/pdf_1_2.pdf
[180] Lawrence,G. et al. Risk Management of Digital Information: A File Format Investigation. *Council on Library and Information Resources*, (2000) http://www.clir.org/pubs/abstract/pub93abst.html

prioritisation will have to be done depending on legislation or company policy. For a test file for Lotus 123 see ref[181].

Data Bases:

The *Digital Preservation Testbed*[182] has specified two sets of minimum authenticity requirements: one for the database itself, and one for the user application (optional).

- Data base:
  - The actual content of the tables must always be preserved.
  - The physical and logical structure of the data base must be preserved.
- User Application:
  - The onscreen representation must be preserved.
  - The content of the database displayed onscreen must be preserved.
  - The structural composition of the data as presented onscreen must be preserved.

Some selected patents (search keys: *file* + *content* + *conversion* + *format*)

- Document content and structure conversion: A system that can convert content and structure of a document from an original format into a target format irrespective of the functional specifics of the original format. (patent (2007): US2007192687[183])
- Online method and system for converting any file in any format into a pdf file for various uses: A method and system for enabling the conversion of the printable content of any file document in any file format type (1) from a client computer to a Portable Document Format (PDF) file (2) created on a central server. (patent (2004): WO2004070617[184])
- Method and apparatus for converting different format content into one or more common formats: A method and apparatus for converting different format content into one or more first common formats. This conversion method allows content that is received in multiple, different formats to be converted into one standard format (patent (2004): US2004170374[185])
- System for multimedia document and file processing and format conversion: An adaptive transformation and User Interface system enables transformation of a file or document (e.g. an SGML, XML, HTML or other multimedia file or document) from one format to another format. (patent (2002): US2002194227 [186])
- Automatic file format converter: Determining, prior to operation of an application program module, that a foreign file format is fully convertible to a native file format. Full conversion of the foreign file format means that a significant majority of the style and presentation of the content of the foreign file are preserved after conversion to the native file format. (patent (2001): US6260043[187])

For additional information about the chosen conversion strategies in some LongRec case partners it is referred to Appendix D.

[181] http://www.clir.org/pubs/reports/pub93/AppendixB.html
[182] From Digital Volatility to Digital Permanence: Preserving Databases (version 1.0), 2003. Digital Preservation Testbed. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-databases-en.pdf
[183] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=US2007192687&F=0
[184] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=WO2004070617&F=0
[185] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=US2004170374&F=0
[186] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=US2002194227&F=0
[187] http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=US6260043&F=0

9 APPENDIX A: FILE TYPE CATEGORISATION

Aa a good starting point for file type categorisation it is referred to Wikipedia[188].

1 Archive and compressed
    1.1 Physical recordable media archiving
2 Computer-aided
    2.1 Computer-aided design (CAD)
    2.2 Electronic design automation (EDA)
    2.3 Test technology
3 Database
4 Document
5 Font file
6 Geographic information system
7 Graphical information organizers
8 Graphics
    8.1 Raster graphics
    8.2 Vector graphics
    8.3 3D graphics
9 Object code, executable files, shared and dynamically-linked libraries
10 Page description language
11 Presentation
12 Scientific data formats (data exchange)
    12.1 Chemical/biological file formats
13 Script
14 Signal data formats (non-audio)
15 Sound and music
    15.1 Lossless audio
    15.2 Lossy audio
    15.3 Other music formats
    15.4 Playlist formats
    15.5 Audio Editing & Music Production formats
16 Source code for computer programs
17 Spreadsheet
18 Tabulated data
19 Video
    19.1 Video Editing & Production formats
20 Video game data
21 Video game storage media
22 Virtual Machines
    22.1 Microsoft Virtual PC/Virtual Server
    22.2 EMC VMware ESX/GSX/Workstation/Player
    22.3 Parallels Workstation
23 Webpage
24 XML, markup language and other web standards-based file formats
25 Financial Records
26 Other

[188] http://en.wikipedia.org/wiki/List_of_file_formats

10 APPENDIX B: CEDARS PRESERVATION METADATA SET

This preservation metadata example set was taken from Preservation Metadata for Digital Objects: A Review of the State of the Art. A White Paper by the OCLC/RLG Working Group on Preservation Metadata, 2001[189]

Preservation Description Information
- **Reference Information**
  - Resource description
  - Existing metadata
    - Existing records
- **Context Information**
- Related information objects

Provenance Information
- History of origin
  - Reason for creation
  - Custody history
  - Change history before archiving
  - Original technical environments
    - Prerequisites
    - Procedures
    - Documentation
  - Reason for preservation
- Management history
  - Ingest process history
  - Administration history
    - Action history
    - Policy history
- Rights management
  - Negotiation history
  - Rights information
    - Copyright statement
      - Name of publisher
      - Date of publication
      - Place of publication
      - Rights warning
      - Contacts or rights holders
    - Actors
    - Actions
      - Permitted by statute
        - Legislation text pointer
      - Permitted by license
        - License text pointer
- **Fixity Information**
  - Authentication indicator

Content Information
- **Representation Information**
  - **Structure Information**
    - Underlying abstract form description

[189] http://cendicites.infointl.com/item300.html

- Transformer objects
  - Platform
  - Parameters
  - Render/analyze engines
  - Output format
  - Input format
- Render/analyze/convert objects
  - Platform
  - Parameters
  - Render/analyze engines
  - Output format
  - Input format

Semantic Information

- Render/analyze objects
  - Platform
  - Parameters
  - Render/analyze engines
  - Output format
  - Input format

Data Object

11 APPENDIX C: PREMIS PRESERVATION METADATA SET

This preservation metadata set was taken from PREMIS[190], 2005.

- objectIdentifier
    - objectIdentifierType
    - objectIdentifierValue
- preservationLevel
- objectCategory
- objectCharacteristics
    - compositionLevel
    - fixity
        - messageDigestAlgorithm
        - messageDigest
        - messageDigestOriginator
    - size
    - format
        - formatDesignation
            - formatName
            - formatVersion
            - formatRegistry
        - formatRegistryName
            - formatRegistryKey
        - formatRegistryRole
    - significantProperties
    - inhibitors
        - inhibitorType
        - inhibitorTarget
        - inhibitorKey
- creatingApplication
    - creatingApplicationName
    - creatingApplicationVersion
    - dateCreatedByApplication
- originalName
- storage
    - contentLocation
        - contentLocationType
        - contentLocationValue
    - storageMedium
- environment
    - environmentCharacteristic
    - environmentPurpose
    - environmentNote
    - dependency
        - dependencyName
        - dependencyIdentifier
            - dependencyIdentifierType
            - dependencyIdentifierValue
    - software

[190] http://www.loc.gov/standards/premis/

- swName
- swVersion
- swType
- swOtherInformation
- swDependency
- hardware
  - hwName
  - hwType
  - hwOtherInformation
- signatureInformation
  - signatureInformationEncoding
  - signer
  - signatureMethod
  - signatureValue
  - signatureValidationRules
  - signatureProperties
  - keyInformation
    - keyType
    - keyValue
    - keyVerificationInformation
- relationship
  - relationshipType
  - relationshipSubType
  - relatedObjectIdentification
    - relatedObjectIdentifierType
    - relatedObjectIdentifierValue
    - relatedObjectSequence
  - relatedEventIdentification
    - relatedEventIdentifierType
    - relatedEventIdentifierValue
    - relatedEventSequence
- linkingEventIdentifier
  - linkingEventIdentifierType
  - linkingEventIdentifierValue
- linkingIntellectualEntityIdentifier
  - linkingIntellectualEntityIdentifierType
  - linkingIntellectualEntityIdentifierValue
- linkingPermissionStatementIdentifier
  - linkingPermissionStatementIdentifierType
  - linkingPermissionStatementIdentifierValue

12 APPENDIX D: ARCHIVAL, MIGRATION AND CONVERSION STRATEGIES OF THE LONGREC PROJECT PARTNERS: BBS AND NATIONAL LIBRARY

This document gives a short summary on a higher level of the implemented migration and conversion strategies in the National Library (NB) and Banking and Business Systems (BBS) respective archival solutions.

12.1 Archive Architecture:

The National Library (NB) and Banking and Business Systems (BBS) have differing archival, migration and conversion strategies and it may therefore be worthwhile to present their solutions.

The National Library of Norway:
The digital repository treats files like digital objects, meaning the content to be preserved is married together with the appropriate preservation metadata. The digital repository of the National Library is in continuous development due to the ever growing amount of digital objects which puts strain on the scaling issue, and also to the fact that tools, practices and standards are improving all the time. The repository is implemented according to the Open Archive Information System model, ISO-standard ISO 14721:2003, a reference model for long-term preservation of digital objects. The OAIS model defines six areas of concern.

Ingest
Data Management
Archival Storage
Administration
Preservation Planning
Access

The relations are shown in this diagram:

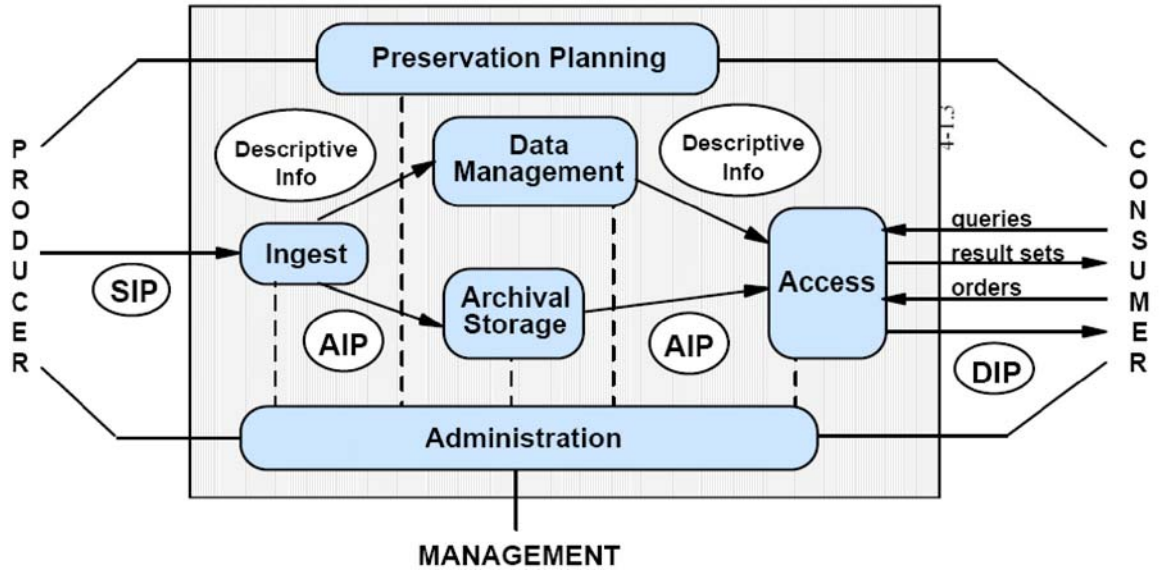


Figure 7: The repository solution chosen by the National Library is based on the Open Archive Information System model, ISO-standard ISO 14721:2003

SIP - Submission Information Package
AIP – Archival Information Package
DIP – Dissemination Information Package

The data ingest can in short be described as follows, for a more detailed description it is referred to the LongRec Case Study: Repository Records Management description. Building on this model the National Library receives and creates digital objects creating the necessary preservation metadata through the use of extraction tools like JHOVE and DROID, and static XML-files to populate METS conforming schema, built on the APSR/NLA METS schema which is built to implement PREMIS schemas and other defined METS schemas like MODS, MIX, LoCs AMD.xsd and VIDEOMD.xsd and other METS conforming schemas.

Data integrity is monitored on file movement through fixity checks, all processing to the file is registered as events in the PREMIS schema and all relevant preservation metadata are also imported into the bibliographic catalogues.

Similar approaches based on these standards and tools are chosen when planning or designing repositories by every major cultural heritage institution, e.g. Library of Congress, National Library of Australia, Libraries and Archives of Canada, British Library, National Library of New Zeeland, National Library of France, National Library of Germany.

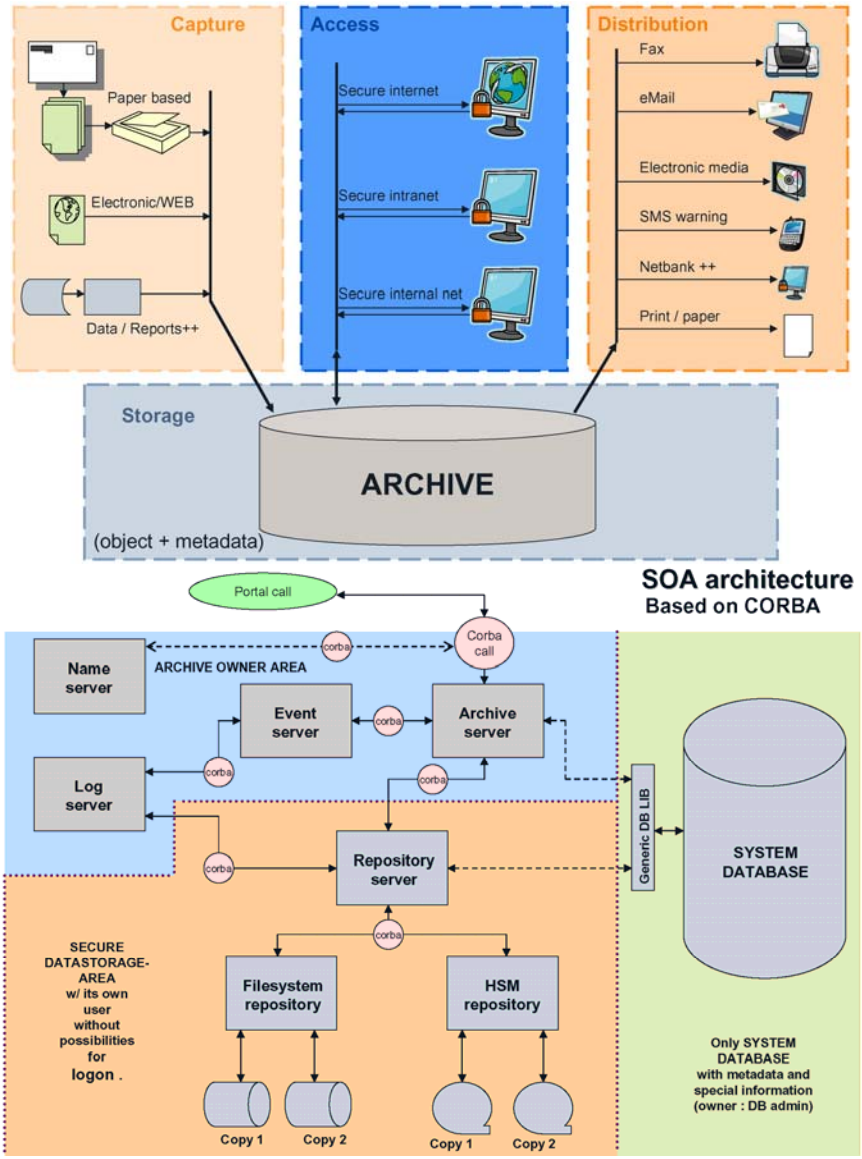
The major advantages are trustworthiness and security. The widespread usage of this approach guarantees further development of the methodology.

References:

PLANETS http://www.planets-project.eu/
DELOS http://delos.info/
DPE http://www.digitalpreservationeurope.eu/
DCC http://www.dcc.ac.uk/
PADI http://www.nla.gov.au/padi/
METS http://www.loc.gov/standards/mets/
OAIS http://nost.gsfc.nasa.gov/isoas/
http://kopal.langzeitarchivierung.de/downloads/kopal_DIAS_SIP_Interface_Specification.pdf
http://kopal.langzeitarchivierung.de/downloads/kopal_DIAS_DIP_Interface_Specification.pdf
PREMIS METS SCHEMA v1.1 http://www.loc.gov/standards/premis/v1
DROID http://droid.sourceforge.net/wiki/index.php/Introduction
JHOVE http://hul.harvard.edu/jhove/
Metadata extractor (NZ) http://www.natlib.govt.nz/about-us/current-initiatives/metadata-extraction-tool/?searchterm=extraction
KoLibRI http://kopal.langzeitarchivierung.de/index_koLibRI.php.en
SIP Manager http://wiki.epc.ub.uu.se/display/FV/SIP+Manager

Banking and Business Systems (BBS):

BBS handles all online money transactions in Norway. Their archive was designed for online access and handling of very large amounts of data. The archive contains about 7.4 Tera Bytes of living data and runs on 4 UNIX servers in 2 clusters. The archive consists of 8 physical archives containing 3 to 10 logical archives. Each physical archive has one file system (see figure below). Online access is absolutely core to BBS with 24/7 availability requirement. Online access is globaly from Singapore, Shanghai, Dallas, New York and northern countries with 2.314.881 archive accesses (during November 2007) and a maximum of 541.714 accesses in a 24-hour period (November 2007). The average retrieval time per object was 0.380 seconds. This average access time includes also format conversion from archive format to presentation format (e.g. TIFF to PNG or TIFF to PDF).



Advantages of chosen archive solution

SOA architecture for the archive environment

- The flexibility to change the environment based on new demands
- The APIs give us the possibilities to enrich the system with new functions
- The possibilities to tune the system for high performance demands
- The flexibilities to use other SOA methods such as Web Services.

File system software

- The built in functionality for data security (RAID etc)
- The admin tools that make it easy to add and remove disk on the fly.
- The built in functionality that gives automatic media conversion

The archival solution uses portals for batch input, batch output and online access. Emphasis was put on the development of an archive independent XML import / export format (w/Base64 objects) functionality.

12.2. Migration Strategies:

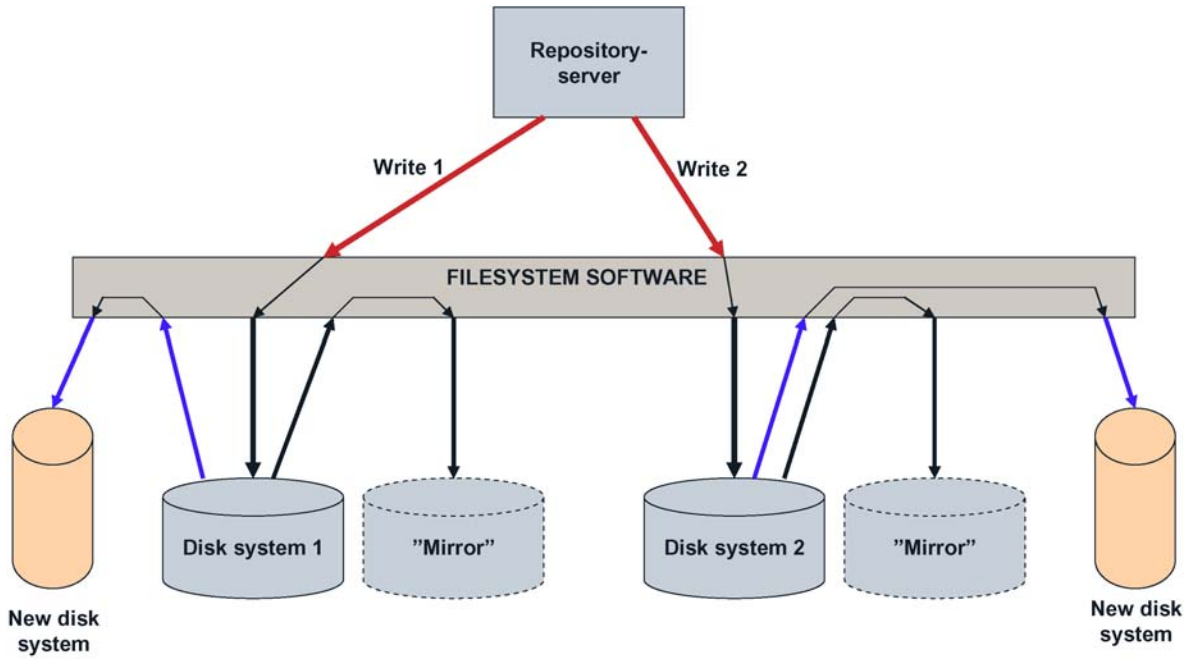
National Library:

The National Library utilizes the move functionality in the UNIX system for file migration including fixity checks. Migration of files to new storage systems have be done 'manually' by UNIX commands. The storage capacity can be fully utilised.

Banking and Business Systems (BBS):

BBS' migration strategy is based on mirror functionality in the file system and the RAID solution used. This is functionality that originally is used for resilience in that several copies of an information object can be stored on different disks. In ordinary operation, at least two copies are kept of each object. When migration is necessary, one or two new disks (can be of a new technology) are plugged in with mirroring towards the disks in use. All data are then automatically mirrored to the new disk(s) through the RAID systems. This mirroring/migration is performed in the background and has no effect on the response time or availability of the system.

Main advantage: mirroring is done automatically with integrity as a built-in function; main disadvantage: when migrating to larger disks, each new disk is a mirror of a smaller one, meaning that the disks cannot be filled. Thus storage capacity may not necessarily be fully utilised, and a storage structure where some disks are filled with old objects with new objects being written sequentially until disks are full is not possible.



12.3. Conversion Strategies

National Library:

The National Library may as a part of the ingest process convert from a submission format to the preservation format to be used. At present, no conversion is later done on the preservation format, which is kept unchanged across migrations. In addition to the preservation format, presentation format representations may be stored, or conversion from preservation to presentation format can be done when presentation of an object is requested. This however does not affect the preservation format. This strategy assures that the file content remains unchanged and no conversion errors will later be introduced if the ingest is successful. However, this approach requires availability of either the original software or other software being able to read the content correctly. In the long-term (decades, centuries) it may be a challenge to find and identify such proper reading software.

Banking and Business Systems (BBS):

BBS' strategy is to restrict the number of different file formats stored to a minimum. In addition these formats must be accepted and in widespread use in the industry (de-facto standards). BBS distinguishes between storage and display format. Currently BBS allows only storing of some carefully selected file formats, these are XML, PDF and TIFF. File conversion happens only between stored and displayed format, i.e. XML, PDF,Tiff -> TIFF, PNG, PDF, JPEG, XML and HTML, which however does not affect the storage format (see conversion table below).
New formats are carefully investigated to avoid costly traps like GIF, LZW etc.
Since the start of the electronic archive in 1992 there has not been any need to convert the formats since the software in use offers backwards capability wrt. format change.

BBS conversion table:

Storage format \ Presentation format	TIFF G4 (multi)	TIFF JPEG7	PNG	JPEG	PDF	XML	HTML	Comments
TIFF G3 1D	Yes	Yes(1)	Yes	Yes(1)	Yes	No	No	Convert to JPEG7 not recomended
TIFF G3 2D	Yes	Yes(1)	Yes	Yes(1)	Yes	No	No	Convert to JPEG7 not recomended
TIFF G4	Yes	Yes(1)	Yes	Yes(1)	Yes	No	No	Convert to JPEG7 not recomended
TIFF JPEG7	Yes(2)	Yes	Yes	Yes	Yes	No	No	Convert to 1 bit picture not recomended
PDF	No	No	No	No	Yes(3)	No	No	
XML	No	No	No	No	Yes(4)	Yes	Yes	
TEKST	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Text not used in BBS

(1) Scale to gray function used
(2) Removing grayscale/color information (poor quality)
(3) Only possibilities to have parts of the document returned (e.g. page 5 to 12)
(4) Special functions for some predefined XML data formats and archives

13 APPENDIX E: STATE OF THE ART FOR DIGITAL SIGNING (BBS)

(As seen) by Knut Nymoen, BBS at 16/10 2007.

1. Introduction

The following report gives an overview of the state of the art for digital signing as perceived from Tillitstjenester BBS Norway at 16/10 2007. It is not complete, and contains some open issues.

Note:

This document does not try to define terms in detail or use very specific definitions. E.g. the terms “digital signing” and “electronic signing” are considered practically equivalent.

2. Structuring the field

One of the major problems with electronic IDs and digital signing is the lack of standardisation of levels or models. The telecom/internet business has been using the 7-layered OSI model for 30 years, and everyone has to understand this model.

For electronic identity and digital signing, there is a lack of models, and correspondingly, people can find themselves discussing PKI details when they should have been discussing business processes. We have defined a rudimentary model for these purposes:

Business process level
    Workflows, applications
Value added level
    Portal components, hosted services, software
ID scheme
    The ID itself, e.g. a certificate
    Basic services and capabilities needed to operate the ID

This document will mainly analyse the levels “ID scheme” and “Value added level” related to digital signing

2.1. ID scheme

- Basic services/capabilities needed to operate the ID

An electronic ID has the same capabilities as a physical ID: It shall guarantee up to a certain level that the person is who he is, and that he can perform a number of additional actions.

For instance a passport: It identifies you and gives an added capability: All countries/areas require (at least) a passport to let you into the country/area.

For an electronic ID with signing capabilities, the added capability is that using the ID itself, you can produce an electronic signature that guarantees who signed the document.

Electronic IDs can exist in many shapes. Two of the most known schemes are username/passwords and PKI based IDs. The additional services are typically registration services, clients/infrastructure for validation of the ID and clients for generation of signatures.

Generally an electronic ID scheme must have a number of given capabilities:

- The format of the ID:
    - The ID itself: Smart card, centrally stored (username/password, BankID Norway), stored on PC (“soft certificate”), USB stick. (For instance BankID Sweden exists both as soft certificate and on smart card)

  - Inside the ID: personal identification number, birth date, name etc
- Intended capabilities: Signing, encryption, identification. For instance, a user name/password based scheme cannot be used for signing.
- Intended usage:
  - corporate use, personal use, banks etc
  - e-mail, signing, identification etc (must match the capabilities)
- Who can issue the ID, how is it issued, trust chain (i.e. CA function)
- What is the security level
  - at the ID issuer (networks, firewalls, personell etc)
  - when distributing the ID: E-mail, passport identification, driving licence etc. ,
  - when managing the ID (e.g. replacing a smart card when it is lost, revoking a certificate etc) (i.e. RA function)
  - when using the ID (PIN, One time password etc)
- Responsibilities for all parts, e.g. how persons shall keep a PIN code.
- Validation: How can the ID be validated, and issues like how long it will take from an ID is revoked to all validation services will have acknowledged this
- Legal matters: Liability etc.

For a PKI, most of these capabilities are described in the certificate policy (CP) or in the specific implementations stated in the Certification Practice Statement (CPS).

2.2. Value added level:

For the value added level, there are two main objectives
- Providing easier integration between business processes, i.e. providing “plug-ins” to a business process. Some examples are electronic invoicing, payment solutions etc.
- Adding value: Functionality like distribution service, automatic archiving, single sign-on etc. are adding value to a business process (and making it easier to integrate towards it)

The value added level can be delivered on different levels:
- Software components
- Hosted services, providing e.g. web services
- More complete solutions like portals providing e.g. an invoicing service

Typical relevant examples are
- Software components: http://www.nexussafe.com/pages/SV/MultiID.aspx, and Adobe/ PDF related products where you can sign a document.
- Hosted services: BBS TrustSign, http://www.bbs.no/bbs/forretningsomrader/tillitstjenester/trust_sign.htm
- Portals: www.chambersign.se

Solutions may typically be PC based where you sign a document on your PC and send via e-mail, or server-based where the systems can manage the process and archives the documents

2.3. Business processes

The business process level is the level where businesses handle their processes they make their living from. For a bank these are creating accounts, withdrawals, creating loans etc. An example of managing a new car insurance contract:

- Consumer dials the call center
- Call center checks the customer
- Call center provides offer to the customer

- Customer signs the offer (and requests e-invoicing)
- An e-mail is sent to the customer with receipt
- The back end systems receives the signed contract and creates the actual insurance with state "pending"
- E-invoice offer is sent to the customer
- The customer accepts e-invoice
- Back end systems receives the receipt, and set the insurance state to "accepted"

In this case, the plug-in from the value added level is "sign document" (and "create e-invoice"). An added value is to send an e-mail to the end user with a copy of the contract.

There are different types of players here

- Large corporations making their own solutions (including extensively customizing an ERP solution, and using consulting partners): Banks, telecoms etc.
- Niche players covering their own niche/field: Real estate, insurance etc.
- ERP: SAP etc.
- SME market who basically needs to purchase either a system adapted to their market (real estate, …) or using a low end ERP system which does not need a lot of customization.

3. General technical issues

This section gives a rudimentary technical overview of electronic signing.

The following general process is used.

All persons and companies to sign need a valid digital certificate that has the "cryptographic strength" required. In the following, it is assumed that the signers are presons.

- A document is created which shall be signed
- The person to sign the document reads the document
- The person signs the document electronically. This is a cryptographic process using the document + certificate as input, creating a hash of the document. A digital signature is then generated – which contains the hash, and, in some cases validity information about the certificate (proving that the certificate was valid at the time the document was signed).
- In some cases, the document + signature(s) are packed together into a structure, like the Estonian/Finnish OpenXades structure or the Norwegian SEID structure. The reason for doing this is to ensure that the document + signatures always are stored together.

3.1. WYSIWYG documents – over time

When signing digitally, the document is displayed to the user who signs the document. The user will then perceive e.g. that he reads some terms and conditions, or that he sees one price / one interest rate for a loan.

However, many document types may contain macros or other constructions that alter the displayed content. For instance a MS Word macro could be "interest rate = current year – 2000", which would hike the interest rate by 1% each year. Even for XML/XSL, you could write a XSL altering the displayed content like this.

It is crucial that the document types supported can be trusted in this respect. It can either be the document type itself (for instance, image files are easier to trust than MS Word), or technical/administrative procedures like checking the received documents for macros.

Maybe LONGREC will find that for trusting signed documents, image files must be used.

One additional aspect is that document formats evolve over time, adding functionality. The PDF format from 1998 probably contained fewer possibilities for altering the displayed content than the 2007 version. And that is probably why the PDF/A format was created – the original PDF format had become bloated and large.

3.2. Signing the document

When signing a document, the following process is performed (assuming that the signing is done utilizing an ordinary existing certificate)

- Checking the validity of the certificate. The certificate should not be expired, and should not be revoked from the certificate authority.
- Computing the hash for the document
- Optional: Obtaining the validity information for the certificate, and include this in the signature. This may be the OCSP response for the certificate (i.e. an "OK" that the certificate was valid at the time of signing)

3.3. Storing / sealing / packing documents and signatures

In many ways literature is mixing the document signing itself, with the sealing/storage of the documents.

With written contracts, you sign the document, each participant keeps his copy, and in case of legal issues, the two contracts are compared. That is, there are two or more original contracts and these are compared. As we see in the case with footballers John Mikel Obi and Morgan Andersen, we have no real way of evaluating if and how a written contract was tampered/ changed, and how to get back to the original. If they had signed the contract with a trusted third party attending, signed 3 contracts, and archived the 3rd contract with the trusted third party, they would have avoided the problem.

A digital signature gives a hash of the document so that you can evaluate whether the document is correct or not, but you cannot get back to the original document if something is wrong.

So: A number of mechanisms have been developed to cope with this problem.

OpenXades (Estonia):

Document + signatures are packed together, including the certificate validation
If the physical storage facility and accessibility is sufficiently hardened, this is might be an acceptable solution. A more detailed analysis may however be done.
www.openxades.org

SEID-SDO (Norway):

Document + signatures including certificate validation are packed together and sealed with a certificate. The advantage of the seal is that you can immediately see if someone has tampered with the whole structure.
www.npt.no/iKnowBase/Content/44963/**SEID**_Leveranse_3_v1.0.pdf
(NPT = Norwegian Post and Telecommunications Authority)

Denmark: Trusted 3rd party:

Actually Danish legal authorities have stated that "*the signature is useless as evidence"* (*"*Signaturen er derfor reelt værdiløs som bevis"), and is using a trusted 3rd party instead. See https://www.signatursekretariatet.dk/pdf/vejledninger/juridisk.pdf

So basically they have an archiving authority which stores the proofs from the digital signing, and trust this authority.

However, by reading the referred document juridisk.pdf, it seems that they have not seen the possibilities utilised by SEID-SDO and OpenXades by storing the certificate validation-result inside the packed structure.

In all these cases, the main goal is to be able to prove that the signing entity has signed
- a defined document
- at the given time
- with a defined ID

and that these proofs fulfil a set of legal requirements (which may be a bit different from country to country)

4. Players

ID level, Nordic region, major players:

Norway: BankID (bankid.no), Buypass, MinID (minside.no)
Sweden: TeliaSonera E-leg, Nordea E-leg, BankID (bankid.com)
Denmark: NetiD, OCES/Digital Signatur
Finland: TUPAS, Fineid

For the public sector, read the Modinis report “Modinis - Study on Identity - Management in eGovernment”, electronic version dated 28 february 2007.

Additionally read the Fraunhofer report “Study on PKI 2006 in Europe final.pdf”.

Some information is also found on: http://countryprofiles.wikispaces.com/EU+Inclusive+e-Government

However, these reports must be read with care. For instance, BankID Norway is not mentioned in the Fraunhofer report.

Value added level:
The market on the value added level is extremely fragmented, with lots of small companies, initiatives, and different certificates. The field is at an in early stage, and none of the consolidation done in banking, card transactions, telecom, consulting etc. has happened in this field. Additionally, there is a large gap between the public and private sector. See below for examples. (It is not viable to give a full overview.)

4.1 Types of solutions

There seems to be 3 types of solutions
- Public sector, services for the public and businesses
- Private sector, services for businesses
- Private sector, services for the public

Being desktop oriented, many companies sign a PDF and need a certificate installed in their browsers. Nordic companies have a broader scope, and aim to cover the whole population based on existing public IDs, cover server-based signing and other document types than PDF.

4.1.1. Public sector, services for the public and businesses

This is typically government portals providing a set of services like
- For businesses: paying VAT/registering applications (Norway: altinn)
- For the public: Change your address, file your tax return (Norway: MinSide)

For these services, you typically need your government issued e-ID. In some cases, a privately issued ID can also be used (Buypass Norway, BankID Sweden).

These services seem to be quite widespread and quite equal around the world, but the availability of the ID itself is often an issue. Estonia has been able to issue electronic IDs to the whole population, while an advanced country like Finland only has issued the public ID to 4-5% of the adult population.

4.1.2. Private sector, services for businesses

This is typically B2B services, and may either be "ad-hoc" services like signing with Adobe and agreeing that this is sufficient, or more managed services like signing with www.chambersign.se Typically the usage is scattered, and not very widespread.

4.1.3. Private sector, services for the public

These services are to a large extent concentrated around the Nordic countries (more analysis is required!), and are banking / payment related. Most services are authentication, but some are signing too.

For signing, the usage is concentrated around specific application like performing a payment in your netbank (Sparebank 1), signing applications for new accounts, signing applications for credit (Komplett.no), signing applications for credit cards (SEB) etc.

4.2. Standards/ standardizing bodies

This section lists the main standard/standardizing bodies relevant for electronic signatures.

ID level

ETSI:

XADES, ETSI TS 101 903
CADES, ETSI TS 101 733
These define formats for the actual signature
ETSI TS 101 456, Requirements for CAs' issuing Qualified certificates

Stork project

The STORK project is a harmonization project for the EU for interoperability of electronic IDs. They seem to not have a website, but presentations may be found: http://www.enisa.europa.eu/doc/pdf/Workshop/June2007/Presentations/auth_LeymanFrank_%20STORK%2013062007.pdf

Value added level

Adobe - ISO

The Adobe PDF standard is beeing transferred from Adobe to ISO, ISO 32000.
Additionally, the PDF/A standard is already an ISO standard, 19005-1.
These standards contains capabilities for digital signing. Details to be analysed.

Norwegian SEID standard

Defines format for packing a signed object, i.e. the document + signatures + a seal for the package. Mainly used by BankID

Openxades:

Estonian/Finnish variant of XADES which defines a way of packing a signed object, i.e. the document + signatures. Seems to be lacking the actual seal, this has to be investigated by a technical expert.

Oasis
DSS http://www.oasis-open.org/specs/index.php#dssv1.0
Provides a simple version of BBS TrustSign:
Sign a document via web service,
validate a signature via web service

4.3. Examples of companies, private and public, value added level

Some companies are listed on:
http://www.a-cert.at/php/cms_monitor.php?q=PUB-TEXT-A-CERT&s=16946ppq
http://www.a-cert.at/php/cms_monitor.php?q=PUB-TEXT-A-CERT&s=49655aci

International
Thales E-security
Adobe
Microsoft
Identrust
DNV Validation Authority http://va.dnv.com

NO
Signicat (former Kantega) + Ergogroup. Signicat
BBS
Public services: http://www.norway.no/temaside/ (click around)

SE
Chambersign Sweden. Uses the name from chambersign.com
(Formpipe http://www.formpipe.se/default_en.htm, they have now switched more to content management and higher level solutions)
A list of merchants/portals is given on:
http://www.bankid.com/BankidCom/Templates/LinkCollectionPage.aspx?id=91&epslanguage=SV

DK
TDC provides the public "Digital Signatur" certificate. http://privat.tdc.dk/digital/
PBS provides the private "Netid": http://www.pbs.dk/net-id
A list of merchants/portals is given on:
http://www.digitalsignatur.dk/visVirksomheder.asp?artikelID=637

DE
http://132.199.120.220/sigdb/scripts/AbfragenAuswahl.asp?ref=http://pc50461.uni-regensburg.de/ibi/de/brancheninfo/sigdb/
(or go to http://www.sigdb.ibi.de/ and use "Durchsuchen" for searching.
http://www.stepover.de/eSignatureOffice.92+M52de9b5fcb6.0.html
http://www.deutschepost.de/dpag?tab=1&skin=hi&check=yes&lang=de_DE&xmlFile=link1015461_63931
Mentana GMBH www.mentana.de
signotec GMBH http://www.signotec.com

NL
http://www.diginotar.com/

FR

Chambersign itself + A list of companies:
http://www.chambersign.fr/chambersign/partenaires.jsp
UKAscertia, www.ascertia.com

AUS

http://www.a-cert.at/

SLO

Crea, www.crea.si

US

Adobe
E-lock, www.elock.com
Topaz systems, http://www.topazsystems.com

- o0o -