4. BACKUP PRACTICES FOR DIGITAL PRESERVATION

4.1. Digital Preservation:

Digital content is fragile and susceptible to decay. Digital obsolescence is a nasty trap in today’s life, making a lot of material no longer readable due to technological obsolescence and the deterioration of digital storage. In this situation digital preservation is an ever-pressing issue. The prospect of losing digital content is therefore highly alarming.

According to Stewart Brand “Digital storage is easy; digital preservation is hard. Preservation means keeping the stored information cataloged, accessible and useable on current media, which requires constant effort and expense.”

Digital preservation is biggest challenge facing by libraries.

According to Open Archival Information System and Giaretta, two key elements are central in digital preservation:

Usability and understandability over time of a digital object;

Authenticity over time of a digital object.

Digital Preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. (DPC Handbook, 2008)
Margaret Hedstrom (1998) defines “digital preservation 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”.

According to Trusted Digital Repositories (TDR) “ Digital preservation encompasses a broad range of activities designed to extend the usable life of machine readable computer files and protect them from media failure, physical loss and obsolescence”.

ALA annual conference Washington D.C.(2007) ALCTS ,Preservation and Reformatting Section defined digital preservation as “ the preservation which combines policies, strategies and actions that ensure access to digital content over time.”

Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term. The key goals of digital preservation include:

- **Usability** – the intellectual content of the item must remain usable via the delivery mechanism of current technology;

- **Authenticity** – the provenance of the content must be proven and the content an authentic replica of the original;
Discoverability – the content must have logical bibliographic metadata so that the content can be found by end-users through time; and

Accessibility – the content must be available for use to the appropriate community.

Long-term preservation - Continued access to digital materials, indefinitely.

Medium-term preservation - Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.

Short-term preservation - Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.

Susan Lazinger (2001) cites the following reasons why digital preservation is a pressing issue:

- Uncontrolled accumulation of data
- Accidental destruction of data
- Unauthorized tampering of data
- Lack of metadata and systems of documentation
- Lack of empowering mechanisms to institutions
- Technological Obsolescence & Physical Deterioration (Hedstorm & Montogomery, 1998)
### 4.1.1 Threats to digital information:

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

(Source: A survey of digital preservation strategies by Feng Luan & Mads Nygård 2010)
The preservation of digital content can be threatened in several different ways such as floods, earthquakes, and political upheaval. The report of “EU-project PARSE.Insight” lists seven threats.

- Users may be unable to understand or use the data, e.g. because of the semantics, format, processes, or algorithms involved.
- Non-maintainability of essential hardware, software, or support environment may make the information inaccessible.
- The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity.
- Access and use restrictions may not be respected in the future, jeopardizing proper reuse.
- Loss of ability to identify the location of data.
- The current custodian of the data, whether an organization or project, may cease to exist at some point in the future.
- The ones we trust to look after the digital holdings may let us down.

4.1.2 Strategies for digital preservation:

In 2006, the Online Computer Library Center developed a four-point strategy for the long-term preservation of digital objects that consisted of:

- Assessing the risks for loss of content posed by technology variables such as commonly used proprietary file formats and software applications.
➢ Evaluating the digital content objects to determine what type and degree of format conversion or other preservation actions should be applied.

➢ Determining the appropriate metadata needed for each object type and how it is associated with the objects.

➢ Providing access to the content.

There are several additional strategies that individuals and organizations may use to actively combat the loss of digital information. Such as Migration, Normalisation, Technological Hardware preservation, Emulation, Virtualisation, Encapsulation, Cloud Computing etc.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Strategy</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage accessibility</td>
<td>Auditing</td>
<td>The state of the storage media is periodically inspected by an auditing application.</td>
</tr>
<tr>
<td>maintenance</td>
<td>Refreshment</td>
<td>The hardware/storage media are periodically replaced by a new and healthy alternative.</td>
</tr>
<tr>
<td>Filling</td>
<td>Replication</td>
<td>Digital information should have multiple copies.</td>
</tr>
<tr>
<td>Federation</td>
<td></td>
<td>Every preservation system within the federation is connected in a distributed environment.</td>
</tr>
</tbody>
</table>
Bits manipulability maintenance

<table>
<thead>
<tr>
<th>Computer museum</th>
<th>Necessary hardware/software is preserved for accessing and manipulating digital information from the old storage medium.</th>
</tr>
</thead>
</table>

Emulation

An emulation application is used to simulate the function of old computer environment. Moreover, the original programme can be executed in the emulation application.

Encapsulation

The format specification is embedded with the digital information. The preservation system can develop new software in terms of the stored format specification.

Universal virtual computer

The digital information and the necessary software are converted to a set of particular instructions, which can be translated by a virtual computer.

Batch migration

The preservation system should periodically transform the old formats to new formats.

Migration on access

The format transformation is executed only when readers try to manipulate the digital information.

(Source: A survey of digital preservation strategies by Feng Luan & Mads Nygård, 2010)
Storage accessibility maintenance:

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

Auditing:

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

Refreshment is a procedure where an old storage medium is replaced by a new one by copying the bits from the old medium to the new. Refreshment includes replacement of a storage medium and updating the entire storage system.

Replication:

Replication refers to an approach where the preserved information has several copies using different formats that might be saved at different places.

Federation:

Federation refers to a distributed network environment, where each independent preservation system shares its digital information with other preservation systems. If the digital information is not readable, the preservation system can recover this information from other preservation systems. Federation is a promising approach because it not only overcomes the threats to accessibility, but also reduces the dangers related to financial problems.

Computer museum:

The computer museum strategy requires that the whole computer system, including hardware and software, should be preserved, so that the preserved digital information can be read and manipulated in the original environment.
Emulation:

Emulation uses a software application to imitate the function of an old hardware component. In this way, the preservation system is able to execute the old software without the real hardware component.

Encapsulation:

In encapsulation, a new software application is developed that is able to manipulate the preserved digital information. Encapsulation requires that the archive package must preserve the digital information and relevant format specifications.

Batch migration:

Batch migration aims to overcome the threats of software obsolescence and format obsolescence in that the preservation system periodically transforms digital information from one format to another.

Migration on access:

Transformation from the original format to a current one is only executed when the preserved digital information is accessed.

4.1.3 Preservation systems:

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

LOCKSS is an open-source system under the auspices of the Stanford University. LOCKSS is a distributed, peer-to-peer preservation system that is able to manage multiple copies at remote data repositories. The system was released in 2004 and had been tested at more than 50 libraries worldwide.

Eprints

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

DSpace

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

e-Depot

e-Depot is a long-term preservation system developed by the National Library of Netherlands. The objective of e-Depot is to maintain the integrity of stored digital objects so that these objects are accessible.
**FEDORA - Flexible Extensible Digital Object Repository Architecture**

FEDORA is constructed under an open-source digital preservation infrastructure. FEDORA began in 1997, headed by the Cornell University and University of Virginia. In the latest version, FEDORA provides not only the basic functions of preservation systems, but also a model using semantic techniques, i.e., the Resource Description Framework (RDF) maintaining relations between digital objects.

**iRODS integrated Rule Oriented Data System**

iRODS is an open-source data grid software system. Developed by the Data Intensive Cyber Environments (DICE) research group and collaborators, it is the successor of the Storage Resource Broker (SRB). A peculiar feature of iRODS is its ability to represent the data preservation policies in a set of rules. Thereby, iRODS can interpret the rules and execute a sequence of pre-defined actions based on a given situation.
Current research groups that focus on the long-term preservation of digital information are:

**PADI**  Preservation Access to Digital Information

The National Library of Australia’s Preservation Access to Digital Information (PADI) is a gateway to digital preservation issues. PADI provides rich resources about every aspect of a preservation system such as archiving, preservation strategies, data documentation and metadata, intellectual property rights management, format and media, management, digitization, approaches at national libraries, and digital preservation tools.

**InterPARES**  International Research on Permanent Authentic Records in Electronic Systems

InterPARES, starting in 1999, has currently entered the third phase. In total, 15 teams from different countries are included in InterPARES. The objective of InterPARES is to develop the necessary knowledge for survival of digital record over a long-term period. InterPARES mainly works on the theoretical foundation for a preservation system, including aspects such as access, creation, maintenance, and security.

**DCC**  Digital Curator Centre

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

CAMiLEON

Creative Archiving at Michigan & Leeds: Emulating the Old on the New

CAMiLEON was funded by the NSF/JISC. CAMiLEON explores various ways to keep the original functionality and ‘look and feel’ of digital objects.

PLANETS

Preservation and Long-Term Access through Networked Services

PLANETS is funded by the European Union under the Sixth Framework Programme of 2006. The PLANETS objective is to improve decision-making in long-term preservation, so that the valued digital objects can be accessed. It has investigated several digital preservation challenges such as a preservation plan, a set of characterizations of digital objects, preservation actions, an interoperability framework, and a test bed.

CASPAR

Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval

CASPAR is another project funded under the Sixth Framework Programme. CASPAR, implemented based on the OAIS guidelines, focuses on:

- what metadata are needed to describe the representation information and other relevant information
how to integrate digital intellectual property rights for the preserved digital information

how to integrate authentication and accreditation into the long-term preservation mechanism

**SHAMAN Sustaining Heritage Access through Multivalent Archiving**

SHAMAN is financed by the European Union within the Seventh Framework Programme. The goal of SHAMAN is to develop new approaches for digital preservation. Those approaches shall not only guarantee preservation of digital content, but will also keep track of the digital content’s integrity, authenticity, semantics, and usage context.

### 4.1.4 Digital Preservation Initiatives:

According to Kirchhoff (2008) Best practices and standards in digital preservation are continuing to evolve, but a number of guidelines do exist, including:

- **OAIS** (Reference Model for an Open Archival Information System, ISO 14721:2003) - a high-level framework for designing a preservation organization

- **PREMIS** (Preservation Metadata: Implementation Strategies) - a data dictionary and documentation describing the metadata necessary for preserved digital content
➢ **TRAC** (Trustworthy Repositories Audit & Certification: Criteria and Checklist) - a set of digital preservation best practice criteria that can be used to evaluate repositories;

➢ **DRAMBORA** (Digital Repository Audit Method Based on Risk Assessment) – a risk management methodology that allows a repository to run an internal audit in order to assess its capabilities, weaknesses, and strengths;

➢ **nestor** (Network of Expertise in Long-Term Storage of Digital Resources) Catalogue of Criteria for Trusted Digital Repositories—a checklist similar to TRAC to assess the technical and organizational trustworthiness of a digital repository;

➢ **DPC Handbook** (Digital Preservation Coalition) – a detailed guide to the management of and long-term access to digital objects.

**Digital preservation and backup:**

Digital content is simple to create using digital devices but it is surprisingly difficult to preserve it over long time as quick advances in technology. In this scenario, digital libraries are worrying about digital preservation issue. It’s high time to work in direction to tackle with a digital “black hole or digital dark”.

Looking to the situation, Backup is one of the important aspects in digital preservation.
2. Backup:

According to Wikipedia (2013), a backup refers to making copies of data so that these additional copies may used to restore the original after a data loss event.

A backup refers to making a copy of important data for the purpose of data recovery. The word "data" refers to anything stored on a computer system: images, programs, documents, videos, etc. Should the important data get damaged or lost, a properly made backup will restore it all. Taking backups of important data can prevent loss of valuable work and the time needed to recreate it.

Dawid Michalczyk, 2013

Thrones (2008) defines backup as an extra copy of data from a computer system.

In data processing, to make a second copy of an important data file in case the original is lost, damaged, or destroyed. Also refers to computer files, equipment, and procedures created and maintained specifically for use in the event of loss or failure of normal systems. In a more general sense, any strategy designed to be implemented if a preferred method or system fails.


A copy of information held on a computer that is stored separately from the computer.

Cambridge Dictionary Online, 2013

Backup is the activity of copying files or databases so that they will be preserved in case of equipment failure or other catastrophe.

SearchStorage.com, 2013
The process of copying all data from one computer system onto another devices such as Network, Tape, CD, USB storage devices or other archive medium.

Swinburne University of Technology, 2012

4.2.1 Importance of Backup:

According to Smart (2012) modern libraries and information centres rely heavily or completely on computers for various library activities, programmes & services and storage of information resources. All kinds of devices have varying degrees of reliability; no device can guarantee failure-proof operations forever. The libraries and information centres should be able to continue functioning in the event of device failure. Thus backups are useful primarily for two purposes:

- To restore data after disaster called disaster recovery.
- To restore files after they have been accidentally deleted/ corrupted.
Fig. 4.1

A good database backup strategy prevents data loss and minimizes system downtime.

(Source: Importance of backup, Tiwari P. 2012)

Disaster Recovery

Physical errors (Such as hardware failure)

Only data saved on tape can be recovered

External factors (Such as fire or water damage)

Only tapes stored in a safe location can be recovered

Fig. 4.2 (Source: Disaster recovery Tiwari P., 2012)
Recovery from logical error:  

**Fig: 4.3**

**Point-in-time recovery**

- **Logical errors** (such as a deleted table)
- **Lost information**  
  12:50 - 4:00

- **Time when logical error occurred**  
  12:50

- **Time when database is stopped for recovery**  
  4:00

- **You must recover the complete database to a point in time before the error**

(Point in time recovery is very critical in a system landscape with data dependencies between systems)

(Source: Recovery from logical error Tiwari P., 2012)

**DataOne (2013) gives following reason for taking backup:**

1. Limit or negate loss of data, some of which may not be reproducible

2. Save time, money, productivity

3. Help prepare for disasters
   - a. Accidental deletions
   - b. Fires, natural disasters
   - c. Software bugs, hardware failures

4. Reproduce results of past procedures (if they were based on older files)

5. Respond to data requests

6. Limit liability
With reference to Disaster management, Gayatri Doctor (2012) says that

Only backup provides:

- Cost-effective portability of data
- Protection against complete data loss
- Protection against viruses and data corruption
- Managed multiple point-in-time copies of data
- Integrated archival management
- Integrated desktop/laptop protection

According to Miriam (2004) to prevent the loss of data and loss of clients, let alone work product, regular backups are imperative. It should be incorporated into daily routines. Besides this checkup about functioning of the backup system (hardware & software) should be done; the data and associated software are readable and usable. Otherwise, the organization has to create the data again.

The whole idea behind data and program backup is to prevent loss of data from the computer system when the server or network crashes. Careful premeditated steps must be taken to ensure the safeguarding of data and computer systems. Backup plan focus on the prevention of loss first and foremost.

When an organization loses its data, the financial implications can be enormous. One study examined the cost of lost data and concluded that “every cent of data backup is worth $2500 of data re-entry.” Another cost study reported that employees on average cost $36 per hour; imagine the amount of productivity lost should your
department, building, or institution lose its ability to access computer services for even one hour. Time studies show that for every hour of data lost, it takes one day to reenter it within the course of a normal workday; and for every day of lost data, it will take a week to reenter the data. Just think about the types of work you perform in the course of a normal day, and then multiply that by the stress of performing work under less than normal conditions in a strange environment. Now add to this the loss of your data files and the need to re-input data and re-create original ideas, projects, catalogs, and websites. The reconstruction of data is time-consuming and extremely expensive, and many times can be avoided. Thus there is urgent need to develop proper backup practices.

More important data—greater need of backing up this data.

Global backup survey 2008-09 reveals that data loss is very common, 66 % of internet users have suffered from serious data loss.

Data may loss accidentally or intentionally due to human errors, hardware or software malfunction, criminal activity, natural disasters (fire, flood, earthquake), manmade disasters (terrorism) etc. (Thrones, 2008)

In that situation, backup of relevant & important data is valuable for continuous functioning of library and information centres.
4.2.2 Types of Backup: (www.typesofbackup.com, 2013)

There are various types of backup of digital content. These are as following with a brief explanation of their meaning, common examples, advantages and disadvantages of each backup type.

1. Full Backup
2. Incremental Backup
3. Differential Backups
4. Mirror Backups
5. Full PC Backup or Full Computer Backup
6. Local Backups
7. Offsite Backup
8. Online Backup
9. Remote Backup
10. Cloud Backup

4.2.2.1 Full backup:

In full backup, all the files and folders selected for the backup will be backed up. It is an initial or first backup followed with subsequent incremental or differential backups. After several incremental or differential backups, fresh full back up is again taken.
Example of a Full Backup:

Library setup a full backup job or task to be done every evening from Monday to Friday. Assume library do its initial backup on Monday, this first backup will contain its entire list of files and folders in its backup job. On Tuesday, at the next backup run, the entire list of files and folders will be copied again. On Wednesday, the entire list of files and folders will be copied again and the cycle continues like this. At each backup run, all files designated in the backup job will be backed up again. This includes files and folders that have not changed.

Advantages:

1. Restores are fast and easy to manage as the entire list of files and folders are in one backup set.

2. Easy to maintain and restore different versions.

Disadvantages

1. Backups can take very long as each file is backed up again every time the full backup is run.

2. Consumes the most storage space compared to incremental and differential backups. The exact same files are being stored repeatedly resulting in inefficient use of storage.
4.2.2.2 Incremental backup:

Incremental backup is a backup of all changes made since the last backup. The last backup can be a full backup or simply the last incremental backup.

With incremental backups, one full backup is done first and subsequent backup runs are just the changed files and new files added since the last backup.

Example of an Incremental Backup:

Library setup an Incremental backup job or task to be done every evening from Monday to Friday. Assume library perform its first backup on Monday. This first backup will be a full backup since it haven’t done any backups prior to this. On Tuesday, the incremental backup will only backup the files that have changed since Monday and any new files added to the backup folders. On Wednesday only the changes and new files since Tuesdays backup will be copied. The cycle continues this way.

Advantages:

1. Much faster backups

2. Efficient uses of storage space as files are not duplicated. Much less storage space used compared to running full backups and even differential backups.
Disadvantages:

1. Restores are slower than with a full backup and differential backups.

2. Restores are a little more complicated. All backup sets (first full backup and all incremental backups) are needed to perform a restore.

4.2.3 Differential backups:

Differential backups fall in the middle between full backups and incremental backup. A differential backup is a backup of all changes made since the last full backup. With differential backups, one full backup is done first and subsequent backup runs are the changes made since the last full backup. The result is a much faster backup than a full backup for each backup run. Storage space used is less than a full backup but more than with Incremental backups. Restores are slower than with a full backup but usually faster than with Incremental backups.

Example of A Differential Backup:

Library setup a differential backup job or task to be done every evening from Monday to Friday. Assume it perform its first backup on Monday. This first backup will be a full backup since it haven’t done any backups prior to this. On Tuesday, the differential backup will only backup the files that have changed since Monday and any new files added to the backup folders. On Wednesday, the files changed and files added since Monday’s full backup will be copied again. While Wednesday’s backup does not include the files from the first full backup, it still contains the files backed up on Tuesday.
Chapter 4 Backup Practices for Digital Preservation

Advantages:

1. Much faster backups than full backups

2. More efficient use of storage space than full backups since only files changed since the last full backup will be copied on each differential backup run.

3. Faster restores than incremental backups

Disadvantages:

1. Backups are slower than incremental backups

2. Not as efficient use of storage space as compared to incremental backups. All files added or edited after the initial full backup will be duplicated again with each subsequent differential backup.

3. Restores are slower than with full backups.

4. Restores are a little more complicated than full backups but simpler than incremental backups. Only the full backup set and the last differential backup are needed to perform a restore.
Table: 4.3 Incremental vs Differential vs Full Backup

<table>
<thead>
<tr>
<th></th>
<th>Incremental backup</th>
<th>Differential backups</th>
<th>Full backup</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is it</strong></td>
<td>A backup of all changed and new files since the last backup</td>
<td>A backup of all changed and new files since the last full backup</td>
<td>A backup of all files in a specified backup set or job</td>
</tr>
<tr>
<td><strong>Backup Speed</strong></td>
<td>Fastest</td>
<td>Faster</td>
<td>Slowest</td>
</tr>
<tr>
<td><strong>Restore Speed</strong></td>
<td>Slowest</td>
<td>Faster</td>
<td>Fastest</td>
</tr>
<tr>
<td><strong>Storage Needed</strong></td>
<td>Least</td>
<td>More</td>
<td>Most</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>• Faster backups&lt;br&gt;• Less storage space used.&lt;br&gt;• No duplicate files</td>
<td>• Faster &amp; simpler restores then incremental backup&lt;br&gt;• Only needs the first full backup and last differential backup to restore</td>
<td>• Fastest restore&lt;br&gt;• Only needs the last full backup set to restore</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>• Slowest restores&lt;br&gt;• Needs all backup sets full + all increments to restores</td>
<td>• Slower backups&lt;br&gt;• Still stores a lot of duplicate files</td>
<td>• Needs the most storage space&lt;br&gt;• Inefficient storage with a lots of duplicates stored</td>
</tr>
</tbody>
</table>

(Source: http://typesofbackup.com, 2013)
4.2.2.4 Mirror backups:

Mirror backups are as the name suggests a mirror of the source being backed up. With mirror backups, when a file in the source is deleted, that file is eventually also deleted in the mirror backup. Because of this, mirror backups should be used with caution as a file that is deleted by accident, sabotage or through a virus may also cause that same file in mirror to be deleted as well. Some do not consider a mirror to be a backup.

Many online backup services offer a mirror backup with a 30 day delete. This means that when you delete a file on your source, that file is kept on the storage server for at least 30 days before it is eventually deleted. This helps strike a balance offering a level of safety while not allowing the backups to keep growing since online storage can be relatively expensive.

Many backup software utilities do provide support for mirror backups.

Advantages:

The backup is clean and does not contain old and obsolete files.

Disadvantages:

There is a chance that files in the source deleted accidentally, by sabotage or through a virus may also be deleted from the backup mirror.

4.2.2.5 Full PC Backup or Full Computer Backup:

Full PC backup or full computer backup typically involves backing up entire images of the computer’s hard drives rather than individual files and folders. The
drive image is like a snapshot of the drive. It may be stored compressed or uncompressed.

With other file backups, only the user’s document, pictures, videos and music files can be restored while the operating system, programs etc need to be reinstalled from is source download or disc media.

With the full PC backup however, library can restore the hard drives to its exact state when the backup was done. Hence, not only can the documents, pictures, videos and audio files be restored but the operating system, hardware drivers, system files, registry, programs, emails etc. In other words, a full PC backup can restore a crashed computer to its exact state at the time the backup was made.

Full PC backups are sometimes called “Drive Image Backups”

Examples Full PC Backup:

There are several vendors that offer full PC backup and restore software. These software’s have to be installed on the computer being backed up.

Typically one of the first things library need to do before it start the backup is to create a bootable rescue disk or DVD. When a computer has crashed but still has a workable hard disk, this rescue disk will be used to boot up the computer with minimum features to start the restore of the hard disk. Once the restore is complete, the computer can then boot up on its own primary hard disk again and would have all its files, programs, database settings etc just as it was at the time of the backup.
Chapter 4  
Backup Practices for Digital Preservation

Advantages:

1. A crashed computer can be restored in minutes with all programs databases emails etc intact. No need to install the operating system, programs and perform settings etc.

2. Ideal backup solution for a hard drive failure.

Disadvantages:

1. May not be able to restore on a completely new computer with a different motherboard, CPU, Display adapters, sound card etc.

2. Any problems that were present on the computer (like viruses, or mis-configured drivers, unused programs etc) at the time of the backup may still be present after a full restore.

4.2.2.6 Local Backups:

A local backup is any backup where the storage medium is kept close at hand. Typically, the storage medium is plugged in directly to the source computer being backed up or is connected through a local area network to the source being backed up.

Examples of Local Backups:

- Internal hard drive
- External hard drive (desktop or portable)
- Optical drives like CDs, DVDs and Blue Ray discs
- Solid state drives like USB Thumb drives or Flash Drives
Network Attached Storage (NAS)

Backup to a shared folder on another computer over the network

Magnetic tape drives

Advantages:

1. Offers good protection from hard drive failures, virus attacks, accidental deletes and deliberate employee sabotage on the source data.

2. Very fast backup and very fast restore.

3. Storage cost can be very cheap when the right storage medium is used like external hard drives

4. Data transfer cost to the storage medium can be negligible or very cheap

5. Since the backups are stored close by, they are very conveniently obtained whenever needed for backups and restore.

6. Full internal control over the backup storage media and the security of the data on it. There is no need to entrust the storage media to third parties.

Disadvantages:

Since the backup is stored close by to the source, it does not offer good protections against theft, fire, flood, earthquakes and other natural disasters. When the source is damaged by any of these circumstances, there’s a good chance the backup will be also damaged.
4.2.2.7 Offsite Backup:

Any backup where the backup storage medium is kept at a different geographic location from the source is known as an offsite backup. The backup may be done locally at first on the usual storage devices but once the storage medium is brought to another location, it becomes an offsite backup.

Examples of Offsite Backups:

1. When the backup drives of the office computers are taken home by the Administrator or Manager for safekeeping.

2. When the backup drive is kept at another branch office at another site or building.

3. When the backup drive is kept at a bank safe deposit box.

4. When the backup is sent across the Internet or network connection to a storage facility or data centre in another city. (Also known as online backup or remote backup)

Offsite backups offer all of the same protection as local backups with added protection against theft, fire, floods, earthquakes, hurricanes and other natural disasters. Putting the backup media in the next room as the source would not be considered an offsite backup as the backup does not offer this added protection.

Advantages:

Offers additional protection when compared to local backup such as protection from theft, fire, flood, earthquakes, hurricanes and more.
Disadvantages:

1. Except for online backups, it requires more due diligence to bring the storage media to the offsite location.

2. May cost more as people usually need to rotate between several storage devices. For example when keeping in a bank deposit box, people usually use 2 or 3 hard drives and rotate between them. So at least one drive will be in storage at any time while the other is removed to perform the backup.

3. Because of increased handling of the storage devices, the risk of damaging delicate hard disk is higher. (does not apply to online storage).

4.2.2.8 Online Backup:

An online backup is a backup done on an ongoing basis to a storage medium that is always connected to the source being backed up. The term “online” refers to the storage device or facility being always connected. Typically the storage medium or facility is located offsite and connected to the backup source by a network or Internet connection. It does not involve human intervention to plug in drives and storage media for backups to run.

Many commercial data centers now offer this as a subscription service to consumers. The storage data centers are located away from the source being backed up and the data is sent from the source to the storage center securely over the Internet.
Typically a client application is installed on the source computer being backed up. Users can define what folders and files they want to backup and at one time of the day they want the backups to run. The data may be compressed and encrypted before being sent over the Internet to the storage data center.

The storage facility is a commercial data center located away from the source computers being backed up. Typically they are built to certain fire and earthquake safety specifications. They have higher security standards with CCTV and round the clock monitoring. They typically have backup generators to deal with grid power outages and the facility is temperature controlled. Data is not just stored in one physical media but replicated across several devices. These facilities are usually serviced by multiple redundant internet connection so there is no single point of failure to bring the service down.

Examples of Online Backups:

There are many commercial service providers offering this as a subscription service. Some market it under the term Remote Backup or Cloud Backup

Advantages:

1. Offers the best protection against fires, theft and natural disasters.

2. Because data is replicated across several storage media, the risk of data loss from hardware failure is very low.

3. Because backups are frequent or continuous, data loss is very minimal compared to other backups that are run less frequently.
4. Because it is online, it requires little human or manual interaction after it is setup.

Disadvantages:

1. Is a more expensive option then local backups.

2. Initial or first backups can be a slow process spanning a few days or weeks depending on internet connection speed and the amount of data backed up.

3. Can be slow to restore.

4.2.2.9 Remote Backups:

Remote backups are a form of offsite backup with a difference being that you can access, restore or administer the backups while located at your source location or other physical location. The term “remote” refers to the ability to control or administer the backups from another location. You do not need to be physically present at the backup storage facility to access the backups.

Putting your backup hard drive at your bank safe deposit box would not be considered a remote backup. You cannot administer or access it without making a trip to the bank. The term “remote backup” is often used loosely and interchangeably with “online backup” and “cloud backup”.

Examples of Remote Backup:

1. Subscription backup services provided by commercial data centers.

2. Backup to an offsite FTP Server.
Advantages:

1. Much better protection from natural disasters than local backups.

2. Easier administration as it does not need a physical trip to the offsite backup location.

Disadvantages:

1. More expensive than local backups

2. Can take longer to backup and restore than local backups

4.2.2.10 Cloud Backup:

Cloud backup is a term often used loosely and interchangeably with Online Backup and Remote Backup. This is a type of backup where data is backed up to a storage server or facility connected to the source via the internet. With the proper login credentials, that backup can then be accessed securely from any other computer with an internet connection. The term “cloud” refers to the backup storage facility being accessible from the internet.

Examples of Cloud Backup:

1. Subscription backup services provided by commercial data centers.

2. Backup to an offsite FTP Server over the internet.

Advantages:

1. Since this is an offsite backup, it offers protection from fire, floods, earthquakes and other natural disasters.

2. Able to easily connect and access the backup with just an internet connection.
3. Data is replicated across several storage devices and usually serviced by multiple internet connections so the system is not at the mercy of a single point of failure.

4. When the service is provided by a good commercial data center, service is managed and protection is un-paralleled.

Disadvantages:

1. More expensive than local backups

2. Can take longer to backup and restore

4.2.3 Network Systems:

Before explaining the different type of storage media for backup, it is useful to know how the storage can be connected to local and external networks, making it possible to access the common data set from various computers. A network system setup is independent of the type of storage; whether it is Hard Disk arrays, tape libraries, optical disk jukeboxes or even Solid State Drive arrays. In order to access files that are stored remotely, i.e., not directly on the computer you are working from (host computer), there are different network setups to access storage devices. These networks connect the storage devices with the host computer, making it possible to retrieve the files needed. Different types of network setups are described in the following:
4.2.3.1 NAS  (Network Attached Storage), a NAS server is a preconfigured file server with internal hard disks, which makes its storage capacity available via LAN (Troppens 2009). This type of server system was developed for file sharing. It is relatively easy to install and expand a NAS server system. Provided it does not become a giant system, with many individual departments connected, it is also easy to maintain making it a cheap solution. However, NAS servers are not suitable for providing storage space for databases that need frequent access, because each file must pass through the internal system of the file servers twice before the files arrive at the computer where they are required (Troppens 2009). This makes the system slow and not suitable for quick access to databases, backup procedures or multimedia applications. It could be a good network solution for digital preservation specifically.
Fig. 4.4 Network Attached Storage

(Source: Google Images, 2013)
4.2.3.2 SAN (Storage Area Network) is an advanced network connecting storage devices to servers, making the devices appear like locally attached disks to the operating system (Tate 2006). The network basically “cheats” the file system into believing the disk is locally connected, even though it’s not. A SAN consists of a communication infrastructure, providing physical connections and a management layer, which organizes the connections, storage elements, and computer systems to make data transfer more stable (Tate 2006). Usually SAN is the type of network used for a cloud server setup. Both NAS and SAN networks can be achieved by using for example Fibre Channel or iSCSI (Internet Small Computer System Interface) technology. Fibre Channel is at present the predominant storage network technology and allows serial transmissions for high speed and long distances, low rate of transmission errors and little delay of transmitted data (Troppens 2009). The transmission through Fibre Channel is only possible if the storage devices have some kind of connection point, called a port. The connection between different ports can be set up in different ways, depending on storage size and the complexity of the network. iSCSI is a form of technology used for serial input and output just like Fibre Channel. Fibre Channel and iSCSI are often referred to as I/O buses (Input/output), because they transfer data between components in a computer/server.
4. 2.4 Backup Devices (Storage Device for Backup):

The first computer backups were made onto big reels of magnetic tape, and even paper: punch cards and paper tape. In the next era, backups were mostly stored on floppy disks of various sizes. But today’s PCs don’t even have floppy drives, not to mention punch card processing devices. Backups are now written on CDs, hard drives, flash drives or via network. But some technologies, such as tape backup, still remain very popular and develop.

Fig.:4.5 Timeline: Data Backup Storage

(Source: www.backuphistory.com, 2013)
Backup devices: Fig. 4.6
4.2.4.1 Preferred properties of backup media:

According to Ravn-Grove 2010 backup device should have following properties:

1. **Easy to use**: The more information each medium can contain, the fewer units of media are needed. A small number of media is easier to handle and maintain.

2. **Easy to copy**: How fast and easy is it to copy between two media of the same type? Can it be done without the risk of losing data? How easy is it to extract data from the media to new storage technology? The easier it is, the more time and money one saves.

3. **Widespread usage**: The more widespread the technology is, the longer it will typically survive on the market. It will be easier to obtain new equipment that can read or utilise the media as well as having updates or repairs done.

4. **Low price**: What is the price of obtaining new versions of the same media? What does it cost, and does the media need a reader in addition? How much storage space can you get for your money and how much do you need in the future? Think of the long term, it might be a cheap start up solution, but over time it might become expensive. Another solution might be better.

5. **Durability**: What is the life expectancy of the media, before any risk of errors occur? Make sure it is a stable format that has a well-known life cycle.
4.2.4.2 The following topics should be considered when investing in backup media:

1. **Storage facilities**: Size of room needed, temperature and ventilation possibilities, risk of flood etc.

2. **Maintenance**: Personnel to keep track of the hardware, test and check files integrity.

3. **Need for expansion**: If new storage is needed is there room for it? Can it be integrated with the system? When developing new equipment, its size tends to decrease, so one might be able to use the same storage space with more data capacity.

4. **Environmentally**: Does it make a lot of noise, heat up, or use a lot of electricity?

5. **Budget**: What does it cost to set the system up, compared to its maintenance, including the electricity bill and so forth?

These criteria can be held up against the different available solutions in order to find the most suitable one for one's digital content. At the same time it is recommended that one should have more than one storage type, to avoid difficulties retrieving data if any errors occur on specific hardware. Which makes it possible to “**mix and match**” to best suit the four criteria of any institution:

1. Size of data collection

2. Budget

3. Internal staff/expertise

4. Use of data (for access or preservation)
4.2.4.3 Backup Media:

4.2.4.3.1 Data Tapes:

The technology of data tapes (magnetic tapes with the capacity of storing digital data) has been around for almost thirty years (Mearian 2012). Since its invention, a large range of different types have been developed, some of them already obsolete (e.g., DAT tapes) while others are still going strong. Today magnetic data tapes are one of the most widely used hardware storage devices for digital files in larger archives, together with hard disk drives. There are two types of tape housing design: the single-hub and the dual-hub cassette (Bradley 2009). Single-hub tapes ensure minimum friction and accurate tape guidance, but the dual-hub is faster, can contain more data on the same amount of tape and is easier to spool through. The most common dual-hub cassettes are the 8mm helical scan tapes by Mammoth and Sony AIT. However, a quick search on the Internet shows that the dual-hub tape technology is no longer being produced; with Sony declaring in 2009 they would completely stop the development of AIT in 2010 (Murai 2009). If an institution has data stored on helical scan tapes it is therefore recommended that it be migrated onto other storage media before equipment becomes too difficult to procure. Single-hub LTO type is the most well known option when choosing tapes.

LTO (Linear Tape Open)

The most widely used tape in the field of digital preservation is LTO Ultrium (Linear Tape Open). The word “linear” refers to the way the tape drive reads or
writes the tape, i.e., it is read with a fixed head and not with a helically / rotary head (Bradley 2009). The LTO Ultrium was first developed in 1990 as collaboration between HP, Quantum and IBM. It is an open format, in order to provide a wider range of products that might be compatible with the tapes, and also to make it easier for different vendors to make tapes and reader/writers with the same properties. When creating a digital archive it is always good to pick open standards, because it there is a lower risk of obsolescence, a wider choice of brand and therefore lower prices.

Fig.4.7

![LTO-4 tape and a LTO-5 reader/writer](image)

Regardless of the brand, the LTO tapes should work in a similar way; so an institution can have more than one brand of LTO tapes. What is important is that all drives (reader/writers) are compatible with the tapes and the indexing systems one
chooses. It can be difficult for the drives to read some types of tapes and it will be costly to invest in many different drives. Since 2000, new generations of LTO tapes have been made every two-three years and the last generation was released in 2010. This was generation five, and was called LTO-5. According to the manufacturers it has a data capacity of 1.5 TB (terabyte) for uncompressed material and a transfer rate of 140 MB/s (Pease et al 2010, Mearian 2012). LTO-6 was released in December 2012 and has a capacity up to 8 TB and data rates of 525 MB/sec.

**Fig:4.8**

Development of LTO tape storage capacity from www.ultrium.com, 2013

Before generation five, tapes didn’t have any file indexing system, making it difficult to search and find files on the specific tapes. However, with the release of LTO-5, a system called Linear Tape File System (LTFS), it is now possible to create self-describing tapes (Pease et al 2010). This means that there is a built-in index on each tape with a file system making it as easy to find files as on a USB stick. One
should be aware, however, that the LTFS software is hardware specific, i.e., not all readers/writers can understand the software. There is LTFS specifically for HP and specifically for IBM, and they are not compatible. There are some open source LTFS programs, but they are not quite on the same usability level. Since generation three there has been the so-called WORM (Write Once Read Many), which protects against overwriting data that has already been added, and which made the LTO tape usable for restricted documents and preservation masters that need to be kept unedited.

If a specific brand of LTO is not being produced anymore or there are several new generations of tapes, the existing tapes should be updated. One should also make sure that newer drives are compatible with older tapes so they can still be read. It’s important to be aware of the backward compatibility characteristics of the tapes. The LTO Ultrium compatibility is defined with two concepts demonstrating investment protection.

1) An Ultrium drive can read data from a cartridge in its own generation and two prior generations.

2) An Ultrium drive can write data to a cartridge in its own generation and to a cartridge from the immediately prior generation in the prior generation format.

As long as the drivers and readers are available, then in principle one doesn’t have to make any changes but there will be a higher risk of unreadable material. Therefore it is recommended converting to a newer generation of tapes every four-five years.
4.2.4.3.2 Autoloaders / tape libraries

Fig.4.9 Autoloaders / tape libraries


For a small collection of data a manual backup procedure will be fine, but an institute with a larger collection of tapes could benefit from so-called automated robotics (Bradley 2009). There are different sizes and complexities to robotic libraries. They are basically closed Shelving systems from which each tape can be taken out automatically and put into a drive. The drive can be set to backup any possible changes or additions from a server, as well as check for errors on the tapes, which allows for quicker access and also backup of any changes. A small autoloader can only hold a limited amount of tapes and it will be necessary to manually swap tapes if the tape collection is bigger than the capacity. Data should also be copied on tape or other storage devices stored offline, to ensure that any errors occurring during automated backup isn’t fatal for the files concerned, although this does require some
sort of regular plan and manual work. For a small institution with an increasing amount of tapes it can become an issue to manually handle all the tapes. However, at the same time autoloaders might not be an option because they can be expensive to invest in and use more space than the tapes alone. Also setting the system up requires people with expertise. For institutions with larger storage room capacity and technical staff it is a valuable help to keep track of tapes and check for errors automatically.

Solutions for optimum preservation of data on data tapes are as follows:

1. Have data redundancy (several copies of the same data).

2. Check files on tapes and the actual tapes every six-twelve months or on an automated basis to make sure the files and tapes are still working correctly.

3. Due to the previously explained new generation of tapes one should base the replacement of tapes on the backwards compatibility and availability of drives for that generation of tapes.

Pros and cons

Pros:

1. Low prices.

2. Scalability.


4. Transfer rates relatively high.

5. Low error rates compared to hard disk drives.
6. Life span of thirty years (not taking into account the reader/writer equipment).

7. After generation LTO-5 the possibility of internal file index for faster access to the right data

8. Supports compression and encryption.

9. Has WORM.

Cons:

1. Writes in a linear way, meaning that excessive accessing is not recommended, because the tape will be worn faster.

2. Access times are long.


4. Can become an expensive solution if robotic autoloaders are needed.

5. The reader/writer equipment is not compatible with future generations and needs to be updated.

(Pease et al 2010, the linear file system)

4.2.4.3.3 Hard disk drives

Most people know that a computer is usually supplied with a hard disk drive (HDD) to store data and programs (although one can also get computers with solid state drive (SDD)). For laptops these are usually 2.5 inch HDD because they are small enough to fit into a computer and made for low energy consumption. A hard disk can also be in a desktop computer or as an external device, as with a 3.5 inch HDD, which
is bigger and craves a more sufficient power supply. The HDDs come with different storage capacities and speeds and for archival purposes there are several hard disk drives to choose from. A basic explanation of a hard disk drive is that it contains a number of hard platters/disks with a magnetic surface that can rotate rapidly with help from a spindle. On these disks the data is recorded or read through the magnetization of ferromagnetic material done by read/write heads that control the level of magnetising (Hoel 1999, p.153).

**Fig:4.10** Hard Disk Drive

![Hard Disk Drive Diagram](Source: Google images, 2013)

The rotation of the disks is measured in RPM (rotation per minute) and has been continuously developed to go faster and faster. According to a study in 2007 HDDs can spin with 15,000 RPM (Anderson 2007). This calls for a stronger motor,
and also for features which make sure the data can be read as fast as possible, keeping the temperature down and the mechanical parts in place.

**Connecting an external HDD**

There are different methods of connecting an external HDD to a computer:

- **USB** (Universal Serial Bus)
- **Firewire**
- **SCSI** (Small Computer System Interface)
- **SATA** (Serial Advanced Technology Attachment)
- **SAS** (Serial Attached SCSI)

Usually USB and Firewire connected HDDs are for plugging into personal computers. They are easy to plug in to any computer with that type of connection. SATA is typically the type of HDD built into a personal computer, although today a SATA HDD can also be made for external enterprise/professional use together with SCSI (Bradley, 2009). This would typically be larger storage and server systems.

**SCSI and SATA technology recommended**

SCSI and SATA facilitating HDDs are usually said to be the most reliable type of disk drives (Schroeder 2007). It is not recommended to use USB and Firewire connected HDDs for archiving and long-term storage purposes (Hunter 2011). As they are not made to be connected to larger systems they are difficult to monitor and can be unreliable. The functionality and features of the two recommended types of HDD technologies are slightly differentiated, resulting in a price difference. Some
claim that the SCSI is more reliable and it is also more often seen in large storage system contexts, with the use of SAS. But HDD with SATA is becoming more and more common, maybe because it costs less, but also because it has been shown that SCSI and SATA HDDs have equal replacement rates, indicating that the failure rates are approximately similar, provided the replacement occurs when the disk actually fails (Schroeder 2007). It has been difficult to retrieve publications after 2007 about the development in these areas, so it is not known at this point whether there has been a significant change in the technology. In practice SCSI is becoming less popular because of the pricing and it often has less capacity than the SATA solution.

**Server systems**

Depending on the size of the archive needed, the storage capacity and numbers of HDDs can be regulated. They can all be connected to each other as a unified storage system and linked to servers and individual computers of a whole institution. This allows for quick access and backup, also for HDD stored on other locations. It’s possible, and for some institutions enough, to have a 'bunch' of individual disks that needs to be maintained separately, but they all have to be individually connected to a server. To make this easier they can be put together in a so-called JBoD (Just a Bunch of Disks), which means that the individual disks are put together at one connection point. This can lead to slow access if several disks are activated at the same time, or if one disk fails, and there is no automatic backup of the data on that specific disk. So for larger collections and for safety reasons a RAID
(Redundant Array of Independent Disks) is recommended. This is a system that brings together a number of hard disks to form one virtual hard disk (Troppens 2009). Data is distributed across the drives in one of several ways called "RAID levels", but the server sees it as one big disk. RAID systems are faster and more fault-tolerant than individual physical hard disks because if a HDD fails it can be exchanged without data being lost. The data can be reconstructed with the help of the other disks, because it is distributed and connected between the disks. How this is built up depends on the level of RAID.

There also exists a system called Massive Array of Idle Disk (MAID), a storage technology like RAID but with energy management. MAID turns the HDDs off to save energy as soon as the content on the disks is not in use. It has also been said that a MAID could increase the life expectancy of the HDDs, because they would not spin constantly. However, Google published an article in 2007 saying that the usage and higher temperature of a HDD does not seem to correlate with the disk failure rate, making this argument less reasonable (Pinheiro 2007). MAID seems to be an advanced system and is not very well known, so it still isn’t clear if it increases the life expectancy of HDDs, and as such is not a very widespread system. Most institutes go for a RAID setup, although there could be beneficial energy savings using a MAID system.

In terms of which RAID level is suitable for long-term storage, it is important to focus on the level that provides the highest fault tolerance and lowest risk of data
loss, and then on its overall performance and speed. There are many different setups, unfortunately none of them with the perfect, risk free solution.

**HDD failures according to Wikipedia are:**

1. **Head crash:** The reader/writer head comes in contact with the rotating platters, due to mechanical shock or other reason
2. **Bad sectors:** The magnetic sectors on the platters might become faulty because of magnetic fields.
3. **Circuit failure:** Parts of the electronic circuit might fail
4. **Motor failure:** Electric motors might fail or burn out
5. **Miscellaneous mechanical failures:** Mechanically moving parts do have a risk of break or fail.

Solutions to avoid the loss of data during a failure of any kind of storage media, are data backup, data redundancy (several copies of the same data) and SMART. It is recommended that one change a hard disk after three-five years, since manufacturers have estimated the life span to be 40,000 hours (Bradley 2009, p.105), which is approximately four-six years. It has been remarked that after five years failure rates seem to increase (Schroeder 2007).

**Pros and cons**

**Pros:**

1. Random and quick access.
2. Contains file index managed by file index systems.
3. Multiple files can be opened and used at once by multiple users.
4. Portable across platforms and operating systems.

5. Scalable.

Cons:

1. For scaling content it can be an expensive solution.

2. Energy costly.

3. SATA HDD systems have been proved to have higher error rates than magnetic tapes.

4. Life span of only five years and an initial infant failure risk.

4.2.4.3.4 Solid State Drives:

SSD (Solid state drive), also called Flash Drives, are differently built than hard disk drives and optical media. They don’t have any moving parts; instead they are based on semiconductor chips (Chen et al 2009). The overall performance and stability is good, because the chips don’t heat up and aren’t prone to mechanical failures. A SSD has fast reading times, low energy usage and can be extremely small in relation to the amount of storage. Therefore they have become increasingly popular in consumer products, such as MP3 players, mobile phones, laptops and even external storage. The main reason not to implement SSDs for larger storage networks has usually been the high prices, but in 2008 Google announced plans to exchange some of their HDD systems with SSD, mainly due to energy savings (Claburn 2008). Recently the cloud host company CloudSigma has started slowly moving all its data onto SSDs, claiming that they give fewer inconsistencies in performance (Clark 2013). Although SSDs could be a fair competitor for HDDs, not many tests on their
durability have yet been conducted. Some say the hardware parts used in SSDs are the same as in HDDs and therefore are sceptical as to the overall hardware stability of SSDs (Marks, 2012). A couple of tests showed that although a SSD at some levels is more effective than a HDD, there are still some elements that need improvement.

The SSD handles random writing and reading differently than the HDD. During reading SSDs are very fast, but the writing speed is seriously reduced, increasing the power consumption considerably. Also a SSD can find random files quickly, but it is not particularly fast when the files are searched in a more sequential manner, which is important when searching large databases (Dumitru 2007, Chen et al. 2009). Some solutions are being proposed to deal with such issues, which might improve the function of the SSD, but at the same time increase price.

**Single-Level Cell Flash recommended:**

There are different ways a SSD can be built up, either through a MLC flash (Multi-Level Cell) or a SLC flash (Single-Level Cell). MLC basically means that each memory element of the SSD (called a cell) is capable of storing more than one bit of information. A SLC on the other hand can only store one bit per cell (Spanjer 2008). If SSDs are considered for long term preservation of digital content, it is recommended that only SLC flash is used, because it has lower error rates, higher performance speed, higher number of program erase cycles and uses less energy than the MLC flash (Spanjer 2008). Program-erase (PE) cycles are a sequence of events in which data is written, erased and then rewritten to a SSD.
A SSD has a limited amount of PE cycles because each cycle causes physical damage to the medium (Kwanghee, 2009), eventually causing the SSD to malfunction. Apparently an SLC SSD can handle a higher number of PE cycles, but the exact life span depends on how many times the data is overwritten. The more you overwrite the quicker the SSD gets worn out. Manufactures claim the life cycle of an SSD is five-seven years (Hodges, 2011).

**Fig.4.11** Solid state drive

![SSD Image](Source: Google images, 2013)

**Pros and Cons**

**Pros:**

1. Low power consumption.
2. Compact size.
3. Shock resistant.
4. High performance for random data access.

*(Based on Chen et al 2009)*
Cons:

1. High prices.

2. Low performance in data writing (this can be improved with certain built-in software).

3. Uncertain lifespan

4.2.4.3.5 Optical Disks:

Commonly known storage media are optical disks such as CDs (Compact Disc) and DVDs (Digital Versatile Disc). A CD can contain 700 MB of information and a DVD maximum 17.08 GB or gigabyte (if it is double sided and has two data layers) (Byers 2003 p. 28). If they aren’t pre-recorded at the manufacturer (CD-ROM), they can either be recordable (CD-R, DVD-R) or rewritable (CD-RW, DVD-RW). A typical CD consists of at least three layers: a polycarbonate base, a reflective layer, usually of aluminum or gold, and a protective coating (Adelstein 2009). A DVD has even more complex layers, since it can be both double-sided with one or two layers of data. If the CD is recordable it will also have a laser-sensitive dye coated on the polycarbonate base. There are three commonly used dyes: cyanine, phthalocyanine and azo (Bradley 2009). When a CD-R is recorded a laser burns the dye and leaves transparent pits and non-transparent areas, which cannot be erased (Bradley 2009, Adelstein 2009). The areas will be reflected in the metallic layer when read by a laser in the CD/DVD drive. A Rewritable CD has germanium, antimony or tellurium as the recordable layer instead of organic dyes. When recording on a rewritable disc, the layer is heated
up and controllably cooled. This can be done again, hence the name rewritable, but only a limited amount of times. A DVD is usually more stable in terms of chemical and physical decay because the metal and dye layer are embedded at the core of the disk instead of at the surface (Iraci 2012). There are other types of optical disks such as Blu-ray; although they do have a larger data storage capacity they are susceptible to the same errors and problems as CDs and DVDs.

**Fig.4.12** Blu-ray Disc

![Blu-ray Disc](Source: Google images, 2013)

**Blu-ray Disc:** In 2002 the Blu-ray Disc Association was formed between Sony, Samsung and Pioneer, amongst others, and the Blu-ray format became available shortly after that. This was to compete with the DVD and the now already lost format HD-DVD (High Definition). Blu-ray ended up outcompeting the HD-DVD format. A two-layered Blu-ray disk can contain up to 50GB of data. Today you can even get a triple layer Blu-ray with a capacity up to 128GB (Hodges 2011). The name comes from the laser, which reads the Blu-ray disk, and is blue whereas the laser reads or writes CDs or DVDs is red and the rills of information are thinner and more compact. As for
using Blu-ray for long-term storage of data, WORM technology and playback equipment is backwards compatible with, for example, CDs and DVDs and early Blu-ray formats. The main difference between the disk types is the laser used to produce them: Blu-ray has a larger capacity of data storage and a higher data transfer rate. A rewritable Blu-ray disc can be rewritten up to 10,000 times (Bennet 2011). It is said to be more robust than DVDs, but it’s still a disk with layers of materials sensitive to the climate and wrong handling. Whether it stays on the market, only time can show... maybe the optical disk format will end up being replaced by another form of technology all together.

4.2.4.3.6 Cloud storage:

Cloud computing is an emerging technology for easy access and storage of files, software and platforms. To put it simply, using cloud computing means having a third party provider of digital storage, processing capacity, infrastructures, platforms and software programmes (Vaquero 2009). Different levels of using the cloud have specific terms, according to V.D. Hoeven 2012:

1. Infrastructure as a Service (IaaS): using storage/computing elsewhere.
2. Platform as a Service (PaaS): using predefined remote computer platforms running specific system software such as databases or software development environments (IDE).
3. Software as a Service (SaaS): using software tools or even complete suites remotely.

Infrastructure services in the cloud as storage service or cloud storage. Because the cloud is an external service, it is easier to scale to fit one's needs. A company
doesn’t have to invest in their own equipment and the service can be made to suit any size of data collection. The data content is still stored on hardware components (usually HDD or Tape systems), the difference is that the access goes through the Internet and the cloud host is responsible for maintaining the hardware. This way the IT department can focus on other issues rather than just maintaining old servers and installing new. If an institution doesn’t even have an ICT infrastructure the cloud is an easily scalable and relatively cheap solution for any IT related ne

**Table 4.4 Properties of cloud computing based on Vaquero 2009:**

<table>
<thead>
<tr>
<th>Properties of cloud computing</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Friendliness</td>
<td>Anyone should be able to understand the interface, and access his or her files.</td>
</tr>
<tr>
<td>Virtualization</td>
<td>You can access the content of the hardware remotely and see the data anywhere.</td>
</tr>
<tr>
<td>Internet Centric</td>
<td>Everything in the cloud is accessed through the internet. Without internet, no cloud content.</td>
</tr>
<tr>
<td>Variety of Resources</td>
<td>Choice between different levels of services such as infrastructures, platforms, software and applications. More or less complex solutions.</td>
</tr>
<tr>
<td>Automatic Adaptation</td>
<td>The system can be setup to automatically keep all files, calendars and emails up to date, no matter what device is in use.</td>
</tr>
</tbody>
</table>
Scalability | Without having to buy new equipment to integrate with the existing network you can get more storage space in a few minutes.
---|---
Pay per use and resource optimization | You don’t have to pay for maintenance or hire people for the IT department, you only pay for what you exactly need.

Often “the cloud” is mentioned as a service where everything is out of your hands and given to the cloud hosting company, but in fact there are several ways to use cloud computing:

**Public cloud:**

1. Access over the web by third party vendor.
2. Requires less start-up money, because you don’t have to invest in equipment yourself.
3. No operational or maintenance staff is required either.
4. The amount of storage and services is scalable without the cost of buying new equipment.

**Private cloud:**

1. The cloud is set up solely for one organisation, with same accessibility as the public version, but only for people connected to the organisation.
2. Gives control over all assets.
3. Costs more because you need staff and equipment to make the cloud work, although a third party vendor can also do this.

4. Not as scalable, because one can’t expand without buying new equipment.

**Hybrid:**

1. For maximum flexibility and same scalability as with the public cloud.

2. Combination of some data content in a public cloud, more confidential parts in closed, private cloud. Also this can be for backup reasons.

3. At the same time one can keep control over the content

(HP video, white paper)
Table 4.5 Risk of cloud computing in a preservation context, based on v.d. Hoeven 2012:

<table>
<thead>
<tr>
<th>Risk</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less control over data.</td>
<td>As managing digital assets is a core task of a cultural organisation the cloud represents a risk in this regard.</td>
</tr>
<tr>
<td>Copyright and legal issues.</td>
<td>Extra security measures are needed if collections contain personal information or allow only for restricted access due to copyright constraints. Data in the cloud has a higher risk of being hacked than offline data. Or there could be data not complying with local regulations. If a shared (available for multiple users) cloud is used, it must be defined explicitly at which geographical locations the data is stored to avoid legal issues.</td>
</tr>
<tr>
<td>Dependency and stability on a specific vendor.</td>
<td>The extended preservation period increases the likelihood of having to deal with multiple vendors during the lifetime of the data set. Migration of data and adaptation of underlying software is then required.</td>
</tr>
<tr>
<td>Loss of data, due to payment issues.</td>
<td>In case of financial problems, not paying the cloud provider will result in direct loss of data hosted elsewhere. This is not the case if you have your own storage infrastructure— which can be shut down while keeping the data on the media carriers.</td>
</tr>
<tr>
<td>Increase in access and availability.</td>
<td>Aspects that are contradictory to long-term preservation, where an organisation must be sustainable with a focus on preservation rather than providing short-term access.</td>
</tr>
<tr>
<td>Upload and migration issues.</td>
<td>To be able to upload or download large files such as video a fast and reliable internet connection is needed. It will be slower and more time-consuming than a regular in house network.</td>
</tr>
</tbody>
</table>
Fig.: 4.13 Cloud Storage Service Providers

(Source: Google images, 2013)

Recommendations when choosing cloud storage:

Questions to ask any cloud host company before selection (Findlay 2010):

1. How easy / difficult / costly will it be for them to meet any record-keeping requirements specified by your organisation, for example additional metadata fields?

2. Whether any additional charges would be levied by the service provider in the event of your organisation seeking to remove information from ‘the cloud’?

3. If they will commit to storing and processing your information in specific terms that are acceptable to your organisation?

4. Whether they will make a contractual commitment to obey privacy requirements on behalf of their customers – both local to the organisation and in the location or locations(s) where the information is stored?

5. Ask for an assurance that the service provider retains no copy of the records or information after the termination of the contract.
6. Whether they are regularly subjected to external security audit or certification processes?

7. How many administrators will have access to your files and details?

8. If they have measures such as multiple geographically separated backup sites in place, so that they can do a complete restoration of your records if needed, and how long this would take?

9. As well as complete restoration of data, how will they go about finding and restoring particular specified files or sets of files and what timeframes will they guarantee for this? For example, if someone accidentally deleted some files or if some data becomes corrupted.

10. When restoring files, can they ensure that the structure of these files (not just the content) and associated metadata is maintained?

11. What are the guaranteed service provision parameters offered by the provider (given the greater likelihood of service disruption due to provision over the internet) – and what action will they take in the event of service disruption? Do they offer any recompense?

12. Whether service providers subcontract part of their service offering to third parties and, if so, what contractual agreements they operate under?

13. Whether there are any relevant standards they are certified as meeting?
Pros and cons:

Cost/benefits

1. Can be cost saving
2. Reduced pressure on IT department for maintenance
3. Easy access, also outside office premises
4. Better and easier collaboration - geographically

Issues/Risk

1. Security
2. No preservation plan
3. No control and checking of the integrity of the files
4. Not yet ideal as a long term preservation solution
5. Less control over own content
4.2.5 Backup Software:

4.2.5.1 Free Backup Software(s):

AMANDA, ArecaBackup, BackupPC, Bacula, BoxBackup, cpio, CreateSynchronicity, CobianBackup(v11), CwrsyncRsynforWindows, DirSyncPro, DAR, dcfldd, dd, dump, duplicity, Duplicati, FlyBack, luckyBackup, Mondo, obnam, QtdSync, RedoBackup and Recovery, rdiff backup, rsync, tar, Zmanda Recovery Manager etc.

4.2.5.2 Proprietary Backup Software(s):

@MAXSyncUp, ArgentumBackup, AcronisTrueImage, AsigraCloudBackup, Attix5OnlineBackup, ARCserveBackup, ARCserveD2D, Avamar, BackupExpress (BEX), Backup4all, BackupAssist, BackupExec, Bitser, Continuous Data Protection, Comodo Backup, Crashplan, Backup, Dolly, Double, DruvaInSync, DruvaPhoenix, EMCNetWorker (formerly Legato), GenieBackupManager, HandyBackup, HPDataProtector, IAOBackup, IBM TivoliStorageManager, IBM TivoliStorageManagerFastBack, IBM AggregateBackup and RecoverySystem, InMageDRScout, ImageforWindows, IperisBackup, LangmeierBackup, LazySave, Livedrive, MacBackupakaMobileMe, MacriumReflect, Memopal, Mozy, NetVaultBackup, NTBackup, NetBackup, Norton360, NortonGhost, Replay4, EMC RecoverPoint, Retrospect, ShadowProtect, System Center Data Protection Manager, SpiderOak, SyncBack, SyncToy, TimeMachine, TonidoBackup, TotalRecovery Pro, UltraBac, Unitrends, UraniumBackup, Ventis, BackupSuite2008,
Windows Home Server Computer Backup, Windows Backup and Restore, Yosemite Server Backup, Magic Mirror Sync etc.

**Cloud Storage Provider:**

Dropbox, Google Drive, SkyDrive, SugarSync, Bitcasa Infinite Drive, Justcloud.com, Livedrive, Crashplan, SOS online backup, Mozy, Carbonite, Hightail, Zip Cloud, iCloud, box etc.

4.2.6. Maintenance:

Maintenance of backed up data is important aspect of backup procedure. Digital preservation is about continuously checking data to keep it accessible and understandable. This means that physical parts as well as software parts have to be maintained, as is nicely summed up in the following quote.

"It is also critical to note that any current technology, data tape or otherwise, will require migration to a newer storage media in the next 5-10 years, otherwise the data tapes themselves will become inaccessible through either tape failure or format obsolescence."

*Bradley 2006 (UNESCO CD-DVD)*

There is always a risk of data being lost or inaccessible. This can happen in a number of ways, but one of the most common is hardware failure (Figure). The stability of hardware has probably been improved since this chart was made, but hardware are still mechanical devices that will degrade and at some point fail to work.

It is interesting to note that the second most common reason for data loss is human error. Either way it is vital to have a backup plan and a maintenance procedure to stay ahead of any possible issues. In short it entails:

1. Making sure there is someone who is responsible for taking care of the data, and that it is defined what this actually means.

2. Planning renewal and maintenance.

3. Making a decision on which strategy to use and when.

4. Having several copies of the same data.

5. Having the copies on at least two different types of hardware storage.

6. Storing backups at different locations.

7. At ingest, making sure there are a few well-used formats, well-described metadata and persistent identifiers, so the data can be found and identified.
8. That the plan and all equipment are evaluated every five years.

9. When evaluating the data and hardware there are some things to consider, in order safeguarding the proper content and also save money and time.

4.2.7 Backup Procedures:

When it comes to backing up procedures it is important to make a distinction between collections of data that are stored for long-term preservation and data that needs quick and easy access on a daily basis. The latter needs to be backed up, but not necessarily stored long term. It is recommended to have a copy for easy access available on a different server system than the actual preservation masters. It is also necessary to have several copies of the preservation masters (two or three) to reduce the risk of losing data. A more automated backup function that keeps an institution’s data accessible and up and running even with minor operating problems would be suitable for larger collections. RAID is a way of insuring that if one disk crashes the data on that disk can be reconstructed from the other disks. This backup system is made for quick recovery (which is the basis of backup) but benefits a long-term preservation system too. Robotic automation systems for tapes can also be used for easy and automated backup procedures and can be connected to a server, ensuring any changes are copied to a new tape. Regardless of whether the system is automated or not, the important thing is that there is a distinction between backup copies and preservation masters, because the later should not need to be backed up very often. Each preservation master and also distribution copies should be copied and stored so
there are at least two. These two versions should be placed at two different locations.

In any preservation plan a procedure for dealing with errors or malfunction should be written down.

### 4.2.8 Error Checking Procedures:

To avoid any errors in the data, the storage devices in use should regularly be checked. This can be done with different codes and algorithms that the storage devices can utilize automatically. A common error is when a bit in a sequence of data changes; a so-called “bitflip”. This can occur when a single bit changes, or when multiple bits in a row change their value. The reason for this can’t always be explained, but it can be any number of reasons such as temperature changes, voltage, humidity, pressure, vibrations, power supply fluctuations, electronic interference, ground loops, cosmic rays, loose connections, aging components, resistive or capacitive variations and noise in the system (Alchesky).

**Error detection**

In short, this is a way to detect any discrepancies in the bits (a 1 has changed to 0 and so on). Just detecting the error is not enough, since a correction of the fault is necessary to maintain the data.

**Data scrubbing** - error correction reduces the risk of accumulating single correctable errors. Data scrubbing is the automatic reading of disks and checking for defective blocks.
**ECC memory** - Stands for Error Correcting Code memory, and is a protocol that can detect not only single but also multiple bit errors, and correct the errors on the fly. To benefit fully it is a good idea to set up the operating system to receive notifications when errors have been fixed. The errors could indicate that hardware needs to be changed or that there is something else wrong.

### 4.2.9 Evaluation of Content:

**Table: 4.6 Evaluation of Content: Based on Oswalt et. al. 2010**

<table>
<thead>
<tr>
<th>Evaluation of content</th>
<th>Considerations/modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is being stored? Is the data that is already stored still needed? (Is there new material, is some information outdated?)</td>
<td>If the data format is becoming out-dated and it is still worth preserving, it is necessary to consider migration or other methods for keeping the data accessible.</td>
</tr>
<tr>
<td>What is the data’s growth rate? (How much and how often is new data added to the collection)</td>
<td>Depending on how quickly the collection size is growing, more storage space is needed. It could be helpful using cloud hosting services to minimize the workload of maintaining larger and larger storage facilities or investing in an in-house ICT infrastructure.</td>
</tr>
<tr>
<td>Evaluation of content</td>
<td>Considerations/modifications</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>What are the trends in its growth? (Will more or less data be added in the future?)</td>
<td>If it is only a temporary growth, it can be enough to invest in a few extra drives or tapes, but if you know it will increase a lot over time, compare your present solution with other options to make daily maintenance easier.</td>
</tr>
<tr>
<td>How long do we retain this data?</td>
<td>Is there any of the content that is not meant to be stored for long term that can be handled differently? Or are some data only meant to be kept for legal purposes, in which case make sure they are not kept longer than needed?</td>
</tr>
<tr>
<td>How often is the data accessed?</td>
<td>If the data is on tapes, but needs to be accessed often, it should be considered if HDDs are a better option. If you have preservation master data on HDDs, but don’t use them often, consider making access copies for the HDD and invest in tapes for storing the master data.</td>
</tr>
<tr>
<td>Can some of the data be taken offline from the network to lower the cost of archival storage?</td>
<td>If a lot of the data is not accessed often, but needs to be stored for long term, it can be more efficient to keep it offline on tapes at a remote place.</td>
</tr>
<tr>
<td>Evaluation of content</td>
<td>Considerations/modifications</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Is the data accessible easily and quickly?</td>
<td>If there have been incidents where data needed couldn't be easily reached, maybe it is better to use a HDD server system or a cloud hosting service.</td>
</tr>
<tr>
<td>How often is the data backed up?</td>
<td>To make sure that the backups are up to date, they should be created every time any changes are made. Also onto remotely stored data.</td>
</tr>
<tr>
<td>Where are the backups stored?</td>
<td>It minimizes the risk of lost data if the data is stored on several media, at different locations. If you manually handle for example tapes, whenever there is an update, the tapes have to be moved back to their location, for safety precautions.</td>
</tr>
</tbody>
</table>

**4.2.10 Storage of backup:**

When storing data, it is worth thinking about the environment that the hardware storage devices will be exposed to. By storing them properly one can increase data life expectancy. An ideal room for storing any archival records (both analogue and digital) is in the middle of a building, without any windows. This provides the most stable climate, without too great a in temperature and UV light. A top floor location often means problems with hot summer days and cold winter days,
and there is also a risk of roofs leaking. In a basement or bottom floor location the risk of water or insect damage and humid air are often high. It is of course possible to choose such locations, but keep in mind it will require more equipment and time spent maintaining the temperature and humidity level stable. The room must be ventilated, especially with high spinning HDDs, as they emit a lot of heat. There should be no main water pipes near the selected room, because if the pipes are broken it can cause damage to the storage room.

The ideal storage environment: (Recommendations from Hunter, Guideline 3 on Digital Storage, Appendix 5, 2011)

1. Relative humidity maintained between 35%-40%.
2. Temperature maintained between 15-21°C (depending on hardware type).
3. Temperature and humidity monitoring.
4. Fire alarm.
5. Internal, non-water, zero-residue fire suppression system.
6. Restricted access to protect media from theft.
7. HEPA air filtration.
8. Magnetic shielding (specifically for tapes).
10. Backup power to facility.
11. UV shielding on lights and any windows.
12. No water pipes near or above storage area.
13. Storage area should not be made of concrete (increases humidity to near 100% during fire).

14. Storage area should be insulated with material that can maintain temperature within storage area for several hours in event of fire outside storage area.

Table 4.7 Preservation issues and tips for different storage media

<table>
<thead>
<tr>
<th>Digital media storage type</th>
<th>Magnetic Tape</th>
<th>Hard Disk Drives</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preservation issues</td>
<td>Avoid high intensity magnetic fields.</td>
<td>Regular maintenance checks are essential to detect potential wear-out and parts failure.</td>
<td>USB memory devices should never be used for anything other than the transfer of records from one system to another.</td>
</tr>
<tr>
<td></td>
<td>High humidity can cause mould growth.</td>
<td>Machines should be used in an air-conditioned environment.</td>
<td>MLC (Multi-Level Cells) flashes should also never be used to store digital content</td>
</tr>
<tr>
<td></td>
<td>Temperature and RH fluctuations can cause binder degradation, sticky tape, magnetic shedding and layer separation.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
for a longer period of time. The SLC type of solid-state drives is recommended.

<table>
<thead>
<tr>
<th>Ideal storage temperature and humidity</th>
<th>11°C - 50% RH. Arrange between 11-23°C with corresponding 50% or 20% RH can be accepted as long as there are few fluctuations.</th>
<th>Avoid temperatures below 20°C.</th>
<th>Ideal temperature is still being determined. A SSD has been seen working between -40 to 55°C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended time of renewal</td>
<td>Four-eight years (after one or two tape generations it is considered best to renew the tapes, although a single tape can last for about thirty years). All new tapes should be of the newest generation.</td>
<td>Five years.</td>
<td>Five years. Should retain information for a decade powered down, but recommended to refresh more frequently.</td>
</tr>
</tbody>
</table>
Displaying Handling and disaster Management:

| Handling and disaster Management | Never stack tapes high to reduce the chance of damage from tapes falling. Before loading check sticky labels on cartridges and never load a tape with a peeling label. Never touch the tape (either the recording or non recording surface). If using an automated tape library, ensure that the library supports partitioning26 so that older and newer generation media can both be supported. | Use RAID set up to cope with simultaneous failures on two disks. If in servers, they should be mounted in racks in machine rooms to make it easy to find and replace any disks. | Periodically verify data integrity (ideally write whole memory, then verify and then store). |

(Source: www.dca-project.eu Sofie Laier Henriksen & Other 2013)
4.2.11 Manipulation of Data and Dataset Optimization:

It is frequently useful or required to manipulate the data being backed up to optimize the backup process. These manipulations can provide many benefits including improved backup speed restore speed, data security, media usage and/or reduced bandwidth requirements.

1. Compression

Various schemes can be employed to shrink the size of the source data to be stored so that it uses less storage space. Compression is frequently a built-in feature of tape drive hardware.

2. Deduplication

When multiple similar systems are backed up to the same destination storage device, there exists the potential for much redundancy within the backed up data. For example, if 20 Windows workstations were backed up to the same data repository, they might share a common set of system files. The data repository only needs to store one copy of those files to be able to restore any one of those workstations. This technique can be applied at the file level or even on raw blocks of data, potentially resulting in a massive reduction in required storage space. Deduplication can occur on a server before any data moves to backup media, sometimes referred to as source/client side deduplication. This approach also reduces bandwidth required to send backup
data to its target media. The process can also occur at the target storage device, sometimes referred to as inline or back-end deduplication.

3. Duplication

Sometimes backup jobs are duplicated to a second set of storage media. This can be done to rearrange the backup images to optimize restore speed or to have a second copy at a different location or on a different storage medium.

4. Encryption

High capacity removable storage media such as backup tapes present a data security risk if they are lost or stolen. Encrypting the data on these media can mitigate this problem, but presents new problems. Encryption is a CPU intensive process that can slow down backup speeds, and the security of the encrypted backups is only as effective as the security of the key management policy.

5. Multiplexing

When there are many more computers to be backed up than there are destination storage devices, the ability to use a single storage device with several simultaneous backups can be useful.

6. Refactoring

The process of rearranging the backup sets in a data repository is known as refactoring. For example, if a backup system uses a single tape each day to store the incremental backups for all the protected computers, restoring one of
the computers could potentially require many tapes. Refactoring could be used to consolidate all the backups for a single computer onto a single tape. This is especially useful for backup systems that do incrementals forever style backups.

7. Staging

Sometimes backup jobs are copied to a staging disk before being copied to tape. This process is sometimes referred to as D2D2T, an acronym for Disk to Disk to Tape. This can be useful if there is a problem matching the speed of the final destination device with the source device as is frequently faced in network-based backup systems. It can also serve as a centralized location for applying other data manipulation techniques.

4. 2.12. Managing the Backup Process:

As long as new data is being created and changes are being made, backups will need to be performed at frequent intervals. Individuals and organizations with anything from one computer to thousands (or even millions?) of computer systems all require protection of data. The scales may be very different, but the objectives and limitations are essentially the same. Those who perform backups need to know how successful the backups are, regardless of scale.
Objectives

1. Recovery point objective (RPO)

   The point in time that the restarted infrastructure will reflect. Essentially, this is the roll-back that will be experienced as a result of the recovery. The most desirable RPO would be the point just prior to the data loss event. Making a more recent recovery point achievable requires increasing the frequency of synchronization between the source data and the backup repository.

2. Recovery time objective (RTO)

   The amount of time elapsed between disaster and restoration of business functions.

3. Data security

   In addition to preserving access to data for its owners, data must be restricted from unauthorized access. Backups must be performed in a manner that does not compromise the original owner's undertaking. This can be achieved with data encryption and proper media handling policies.

4.2.13 Limitations:

   An effective backup scheme will take into consideration the limitations of the situation.

1. Backup window

   The period of time when backups are permitted to run on a system is called the backup window. This is typically the time when the system sees the least
usage and the backup process will have the least amount of interference with normal operations. The backup window is usually planned with users' convenience in mind. If a backup extends past the defined backup window, a decision is made whether it is more beneficial to abort the backup or to lengthen the backup window.

2. Performance impact

All backup schemes have some performance impact on the system being backed up. For example, for the period of time that a computer system is being backed up, the hard drive is busy reading files for the purpose of backing up, and its full bandwidth is no longer available for other tasks. Such impacts should be analyzed.

3. Costs of hardware, software, labor

All types of storage media have a finite capacity with a real cost. Matching the correct amount of storage capacity (over time) with the backup needs is an important part of the design of a backup scheme. Any backup scheme has some labor requirement, but complicated schemes have considerably higher labor requirements. The cost of commercial backup software can also be considerable.

4. Network bandwidth

Distributed backup systems can be affected by limited network bandwidth.
4.2.14 Implementation:

Meeting the defined objectives in the face of the above limitations can be a difficult task. The tools and concepts below can make that task more achievable.

1. Scheduling

Using a job scheduler can greatly improve the reliability and consistency of backups by removing part of the human element. Many backup software packages include this functionality.

2. Authentication

Over the course of regular operations, the user accounts and/or system agents that perform the backups need to be authenticated at some level. The power to copy all data off of or onto a system requires unrestricted access. Using an authentication mechanism is a good way to prevent the backup scheme from being used for unauthorized activity.

3. Chain of trust

Removable storage media are physical items and must only be handled by trusted individuals. Establishing a chain of trusted individuals (and vendors) is critical to defining the security of the data.

4.2.15 Measuring The Process:

To ensure that the backup scheme is working as expected, key factors should be monitored and historical data maintained.

1. Backup validation
Provides information about the backup, and proves compliance to regulatory bodies outside the organization. Disaster, data complexity, data value and increasing dependence upon ever-growing volumes of data all contribute to the anxiety around and dependence upon successful backups to ensure business continuity. Thus many organizations rely on third-party or "independent" solutions to test, validate, and optimize their backup operations (backup reporting).

2. Reporting

In larger configurations, reports are useful for monitoring media usage, device status, errors, vault coordination and other information about the backup process.

3. Logging

In addition to the history of computer generated reports, activity and change logs are useful for monitoring backup system events.

4. Validation

Many backup programs use checksums or hashes to validate that the data was accurately copied. These offer several advantages. First, they allow data integrity to be verified without reference to the original file: if the file as stored on the backup medium has the same checksum as the saved value, then it is very probably correct. Second, some backup programs can use checksums to
avoid making redundant copies of files, and thus improve backup speed. This is particularly useful for the de-duplication process.

5. Monitored backup

Processes are monitored by a third party monitoring center, which alerts users to any errors that occur during automated backups. Monitored backup requires software capable of pinging the monitoring center's servers in the case of errors. Some monitoring services also allow collection of historical metadata, that can be used for Storage Resource Management purposes like projection of data growth, locating redundant primary storage capacity and reclaimable backup capacity.

4.2.16. Backup Policy:

An organisation's procedures and rules for ensuring that adequate amounts and types of backups are made, including suitably frequent testing of the process for restoring the original production system from the backup copies.

(Wikipedia, 2013)

A backup policy is typically described as a principle or rule to guide decisions and achieve a rational outcome from some sort of interruption in the availability and integrity of protected data and systems.

A backup policy is defined as a set of rules to archive the required backup goals with the commonly associated terms such as: what, where, how, how much, how often and others.
At the highest level, the backup policy matches the requirements of common terms such as recovery point objective (RPO) and recovery time objective (RTO) with constraints like the backup window and budgetary parameters.

*(Backup Academy, 2013)*

The purpose of backup policy is to ensure that there is a consistent and reliable method for recovering data.

Create a backup policy that clearly identifies: *(DataOne, 2013)*

1. Roles,
2. Responsibilities,
3. Where the data is backed up,
4. How often the files are backed up,
5. How to access the files,
6. Recommended file formats to be used, and
7. Policies for migrating data to assure data are not lost due to media degradation or changing formats or programs
8. Review your backup policy and plan periodically to ensure it is still valid and applicable

*Ali Navid Akhtar & Others*(2012) discuss Points for Backup Plan:

Decide what needs to be backup

Determine the appropriate backup type to use for your data.

Establish an appropriate backup schedule and window.
Decide where to store backups.

Develop a backup retention policy.

With reference to this, Symantic (2011) says:

- Don’t wait until it’s too late; identify critical resources
- Protect information completely
- Get employees involved
- Test frequently
- Review your plan

**Conclusion:**

Preservation of collections is fundamental to the success of the library: if the library cannot preserve them, it will fail in its core mission to enable access for future generations. This is no less true for digital collections than for traditional collections. Yet over the very long term, the inherent instability and transient nature of digital content makes preservation a significant challenge, one faced not just by the library but by institutions worldwide.

The main objective of the study is protection of data, so that libraries can function when computer system fails. For that strong backup policy must be practice in libraries.

There are three important points to consider to protect data for short and long term.

- Backing up the data on to a removable storage medium
- Storing the data in a safe, secure location.
➢ Being able to recover the data and reinstall it onto the existing or recreated databases.

Libraries must have such backup policy which is multipurpose mechanism that will safeguard the material digital library holds, provide quick, easy and continues access as well as ensure that human errors, disasters and technological advances do not undo the digital library efforts.
References:


http://sony.storagesupport.com/files/SONY%20AIT%20EOL%20second%20notice%202010-12-09_Fin.pdf

http://www.computerworld.com/s/article/9134468/Review_Hard_disk_vs._solid_state_drive_is_an_SSD_worth_the_money_

http://www.networkcomputing.com/storage-networking-management/are-ssd-based-storagearrays-doomed/240000637


Watson, J. Preservation Concerns in the E-resources Environment. Managing the transition from print to electronic journals and resources (pp. 45-61). Retrieved from INFLIBNET


