Storing Vs Preserving Data

Homes in Ancient Egypt were made mainly out of mud bricks, papyrus and little wood.
They were build in few months/weeks by a crew of 10-20 men.
Houses built around 2500 b.c. protected their owners for a generation or more, serving their scope for a limited amount of time, but none of them remains fully intact today.

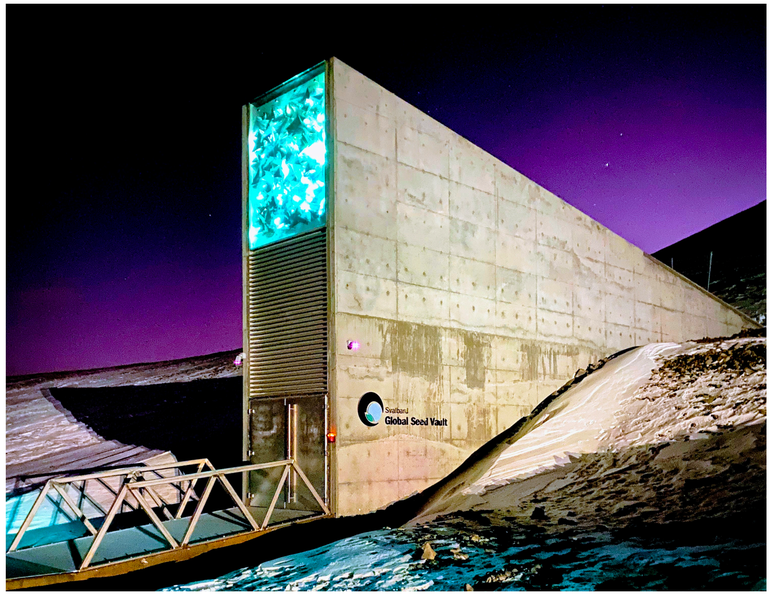
The Great Pyramid of Giza is made out of stone. It has been Estimated that roughly 40000 men worked to complete the Great Pyramid in 27 years.
The Great Pyramid, at today, remain intact.

Can you tell the difference between the house built with mud bricks and the pyramid? Certain things are made to serve their purpose for a very limited amount of time, but if we want something to endure time, then we need to make a lot of effort into solving long-term maintenance problems.

How does your data repository look like?

A “homy”, yet well organised, seed collection in a portable hermetic plastic container.

A seed storage room. A highly organised system of seed storage with controlled temperature and humidity.

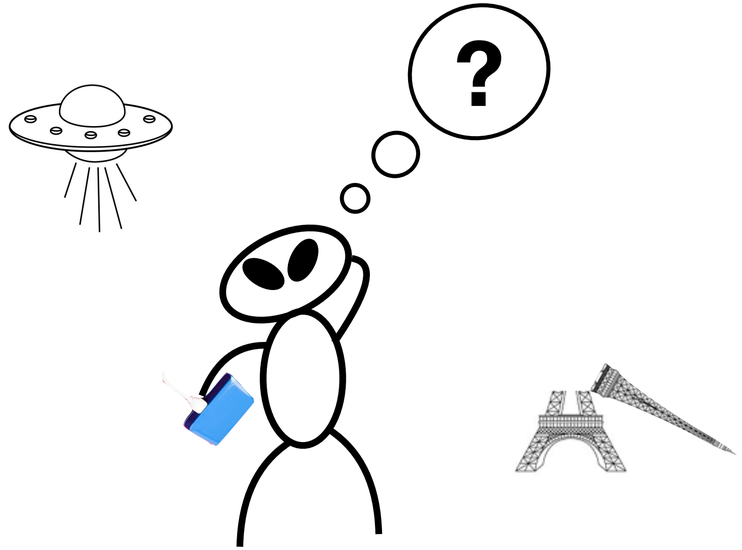


The Svalbard Seed Vault, a backup for the world’s 1750 seed banks that is supposed to last for 1000 years

Months **Years** **Centuries**

Time

Storing Vs Preserving Data



During our research we use up to date technologies and data formats to store our data. Simple **data storing** works just fine if we do not intend to use or make usable the data for a long time period. However, because of the nature of our research (a multi-generational cohort study, for example), our university regulations, or our will to make research products accessible, we may need to guarantee access and usability of the data for a long period of time.

In this section you will learn that time may affect our data storage in a large variety of ways. **Preserving data** means fighting against the effects of time on our data repository and the longer the preserving time, the greater the effort.

What may happen to data?

Bit-Rot

Data degradation affecting SSD, magnetic media (hard disks) and optical media (CD, DVD)

Defective media

Data saved on media that became defective or get damaged over time

Data repository oblivion

Data saved on a repository that stops to be maintained, its location and content is forgotten, falling in this way into oblivion.

Outdated media

Data saved on media that, over time, became unreadable by modern computers

Outdated file format

Data saved on file formats that become outdated, and so unreadable, over time

Data storage loss

Data saved on media that gets damaged, lost, stolen, or destroyed by accident.

What may happen to data?

Multiple backups and **checksum**

A checksum is a sequence of numbers and letters uniquely associated with a file. Comparing checksum of two data copies that are supposed to be identical reveal bit rot effects.

Defective media

Data saved on media that became defective or get damaged over time

Data repository oblivion

Data saved on a repository that stops to be maintained, its location and content is forgotten, falling in this way into oblivion.

Outdated media

Data saved on media that, over time, became unreadable by modern computers

Outdated file format

Data saved on file formats that become outdated, and so unreadable, over time

Data storage loss

Data saved on media that gets damaged, lost, stolen, or destroyed by accident.

What may happen to data?

Multiple backups and **checksum**

A checksum is a sequence of numbers and letters uniquely associated with a file. Comparing checksum of two data copies that are supposed to be identical reveal bit rot effects.

Defective media

Data saved on media that became defective or get damaged over time

Data repository oblivion

Data saved on a repository that stops to be maintained, its location and content is forgotten, falling in this way into oblivion.

Outdated media

Data saved on media that, over time, became unreadable by modern computers

Choosing common and up to date file formats

Give priority to open formats instead of formats maintained by private. Stay up to date about the most common file formats and update your data repository accordingly.

Data storage loss

Data saved on media that gets damaged, lost, stolen, or destroyed by accident.

What may happen to data?

Multiple backups and **checksum**

A checksum is a sequence of numbers and letters uniquely associated with a file. Comparing checksum of two data copies that are supposed to be identical reveal bit rot effects.

Defective media

Data saved on media that became defective or get damaged over time

Hiring a data manager

A data manager/specialist will be in change of organising and maintaining your data repository.

Outdated media

Data saved on media that, over time, became unreadable by modern computers

Choosing common and up to date file formats

Give priority to open formats instead of formats maintained by private. Stay up to date about the most common file formats and update your data repository accordingly.

Data storage loss

Data saved on media that gets damaged, lost, stolen, or destroyed by accident.

What may happen to data?

Multiple backups and **checksum**

A checksum is a sequence of numbers and letters uniquely associated with a file. Comparing checksum of two data copies that are supposed to be identical reveal bit rot effects.

Defective media

Data saved on media that became defective or get damaged over time

Hiring a data manager

A data manager/specialist will be in change of organising and maintaining your data repository.

Outdated media

Data saved on media that, over time, became unreadable by modern computers

Choosing common and up to date file formats

Give priority to open formats instead of formats maintained by private. Stay up to date about the most common file formats and update your data repository accordingly.

Multiple backups

Making multiple backups of your data in different locations is an effective measure against data loss.

What may happen to data?

Multiple backups and **checksum**

A checksum is a sequence of numbers and letters uniquely associated with a file. Comparing checksum of two data copies that are supposed to be identical reveal bit rot effects.

Refreshing data

The concept is simple: make a copy of your data in new storage media before the old one stops operating correctly.

Hiring a data manager

A data manager/specialist will be in change of organising and maintaining your data repository.

Outdated media

Data saved on media that, over time, became unreadable by modern computers

Choosing common and up to date file formats

Give priority to open formats instead of formats maintained by private. Stay up to date about the most common file formats and update your data repository accordingly.

Multiple backups

Making multiple backups of your data in different locations is an effective measure against data loss.

What may happen to data?

Multiple backups and **checksum**

A checksum is a sequence of numbers and letters uniquely associated with a file. Comparing checksum of two data copies that are supposed to be identical reveal bit rot effects.

Refreshing data

The concept is simple: make a copy of your data in new storage media before the old one stops operating correctly.

Hiring a data manager

A data manager/specialist will be in change of organising and maintaining your data repository.

Refreshing data

The concept is simple: make a copy of your data in new storage media before the old one becomes outdated and stops interacting correctly with modern machines.

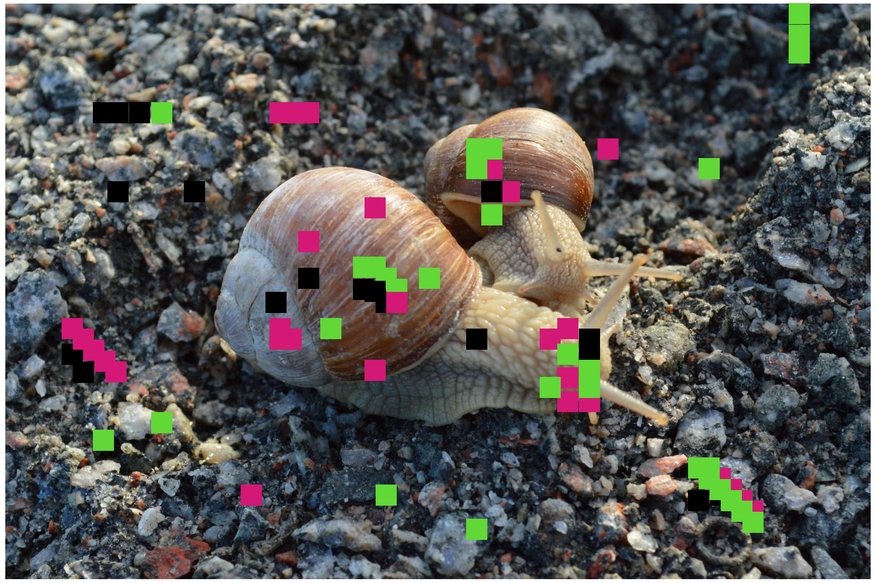
Choosing common and up to date file formats

Give priority to open formats instead of formats maintained by private. Stay up to date about the most common file formats and update your data repository accordingly.

Multiple backups

Making multiple backups of your data in different locations is an effective measure against data loss.

Bit Rot
(semi-fictional)



A researcher spent 15 years taking thousands of pictures (JPEG) of snails in the region of Uppsala to study the morphology of their shell. At the end of the study, she is shocked of finding that hundreds of pictures either appear heavily pixelated or are totally unreadable.

Can you identify the problem?

Outdated media
Incorrect, the researcher can still access her media

Outdated file format
Incorrect, JPEG is still in Use

Defective media
Incorrect, the researcher can still access her media

Bit Rot
Correct, the unridable pixels are a typical sign of bit rot

Data Repository oblivion
Incorrect, the researcher well remember her data location

Media Loss
Incorrect, the data Is still there

Which strategy could you apply to prevent the problem?

Multiple backups
Not enough, bit rot can Happen in all your Backup media

Common File formats
Incorrect, bit rot can Potentially affect any file Format

Checksum
Not enough, checksum Requires at least two Data copies

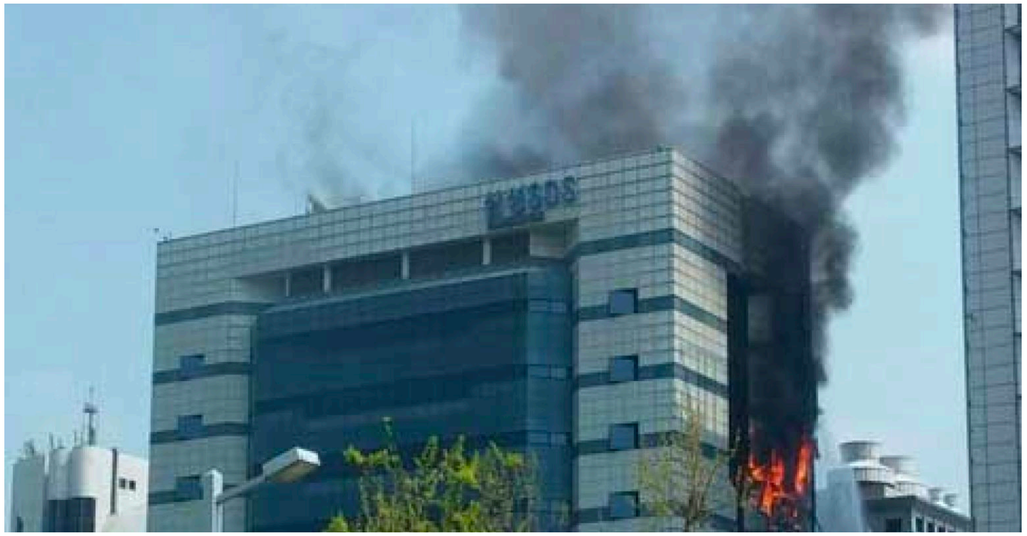
Update File formats
Incorrect, bit rot can Potentially affect any file Format

Refresh Data
Incorrect, you Would copy Corrupted data Into a new media

Hire/Consult a specialist
Maybe a specialist could Have prevented your bitrot Problem, but your database Is very simple: is it worth The cost?

Correct, having multiple backups and checking/correcting frequently for data corruption among identical files could have prevented bit rot

Media Loss
(*real case)



In 2014 a fire temporarily knocked out the Samsung datacenter in Gwacheon (South Korea).
The accident affected connectivity of any Samsung device requiring Samsung servers to operate.

Can you identify the problem?

Outdated media
Incorrect, media was On fire, not outdated

Outdated file format
Incorrect, nothing to do With files, but with their Physical storage

Defective media
Incorrect, the media Was not defective

Bit Rot
Incorrect, here the Problem concerns Physical media

Data Repository oblivion
Incorrect, the data center Was in full operation Before the fire

Media Loss
Correct, the event caused Physical media loss/damage

Which strategy could you apply to prevent the problem?

Multiple backups
Correct, if one of your Storage media gets Damage you can always Use another, and that's what Samsung did, as the server Was up and running in Few hours

Common File formats
Incorrect, the problem Has nothing to do With file formats, but With the physical media

Checksum
Checksum prevents File corruption, but Here our problem Was physical damage Of the media

Update File formats
Incorrect, the problem Has nothing to do With file formats, but With the physical media

Refresh Data
Incorrect, the problem Has nothing to do With outdated media

Hire/Consult a specialist
Incorrect, a specialist Could have not prevented The fire or saved your Media from fire

Outdated media
(*semi-fictional)

A medical study about depression in adolescences started in the early 90’ and concluded in 2020.
All the 90’ data have been stored in Zip drives, floppy disks with the ENORMOUS capacity of 250 MB.
When it’s time to perform the analysis, most of the collaborators have no idea what a Zip drive is
and modern computers don’t recognise the drives.

Can you identify the problem?

Outdated media
Correct, the old media Format cannot Interface with modern Machines

Outdated file format
Incorrect, nothing to do With files, but with their Physical storage

Defective media
Incorrect, the media Was not defective

Bit Rot
Incorrect, here the Problem concerns Physical media

Data Repository oblivion
Incorrect, researchers are Well aware of the dataset, Even if the media is outdated

Media Loss
Incorrect, those Bulky Zip drives Are hard to lose

Which strategy could you apply to prevent the problem?

Multiple backups
Incorrect, in the 90’ they did Indeed multiple backups, but Still in Zip drives!

Common File formats
Incorrect, the problem Has nothing to do With file formats, but With the physical media

Checksum
Checksum prevents File corruption, but Here our problem Is about the physical Media

Update File formats
Incorrect, the problem Has nothing to do With file formats, but With the physical media

Refresh Data
Correct, refreshing, Or copying, Zip drives Data into CDs or other Drives would have Solved the problem.

Hire/Consult a specialist
Maybe a specialist Could have updated the Storage media, but you did not hire him/her, so try again!

Data Repository oblivion (Real case)

According to the New York Times, in 2021, Stefan Thomas, a programmer living in San Francisco, had a small high Security drive (IronKey) containing private keys to a digital wallet. The wallet held 7022 Bitcoin, equivalent, at that time, to about $220 million. The IronKey drive allows a maximum of 10 password input attempts before deleting the entire content of the drive. Stefan forgot the password and never recovered the small fortune.

Can you identify the problem?

Outdated media
Incorrect, media was Not outdated, actually It was the state of art of Security

Outdated file format
Incorrect, nothing to do With files, but with Access to files

Defective media
Incorrect, the media Was not defective, it was working way too well!

Bit Rot
Incorrect, here the Problem concerns Accessing files

Data Repository oblivion
Correct, something about the Data was forgotten: the Password to access the keys

Media Loss
Incorrect, the IronKey was Neither lost or Damaged

Which strategy could you apply to prevent the problem?

Multiple backups
Incorrect, Stefan would have Backed up in other IronKeys, Each with its own password Ready to be forgotten

Common File formats
Incorrect, the problem Has nothing to do With file formats, but With accessing the files

Checksum
Checksum prevents File corruption, but Here our problem Is about accessing Files

Update File formats
Incorrect, the problem Has nothing to do With file formats, but Accessing files

Refresh Data
Incorrect, data would Have probably been Refreshed into Another IronKey Device

Hire/Consult a specialist
Correct, assigning a Specialist to the digital Wallet would have been The best way to avoid data Access requirements to Fall into oblivion

Outdated file formats
(*fictional)

In 2025 William Henry Gates III gets hit by an epiphany: everything is meaningless in this world.
He shuts down Microsoft and decides to spend the rest of his life in a Tibetan monastery.
One of the consequences of his decision is that the Windows Media Audio (WMA) format will not longer be supported.
A biology researcher whose work is entirely based on bat audio recordings in WMA format finds herself with quite
A pickle when, four years after Gates decision, she is finalising her PhD thesis.

Can you identify the problem?

Outdated media
Incorrect, media was Not outdated, the Problem concerns File format

Outdated file format
Correct, because of a Random event, a very common File format became now updated

Defective media
Incorrect, the media Was not defective, The problem concerns File format

Bit Rot
Incorrect, here the Problem concerns File format

Data Repository oblivion
Incorrect, the data did not Go into oblivion as it is Currently used to finalise The researcher thesis

Media Loss
Incorrect, the problem concerns Data files not storage

Which strategy could you apply to prevent the problem?

Multiple backups
Incorrect, multiple backups Would have still stored The data in the WMA format

Common File formats
Correct, instead of having Used a format maintained By a private company, the Researcher could have Used an open format

Checksum
Checksum prevents File corruption, but Here our problem Is about file format

Update File formats
Correct, updating file Formats to open file Formats would have Solved the problem

Refresh Data
Incorrect, data would Have probably been Refreshed using the Same outdated file Format

Hire/Consult a specialist
Maybe a specialist could Have been aware of Rumours regarding Gate's Life and he/she could have kept the file formats Updated

Defective media
(*semi-fictional)

Solid State Drives (SSD) have a lifespan of about 10 years, depending on usage.
It's not a matter of **if** a SSD will fail, but **when** it will fail.
SSD lifespan issues have been ignored by Paul, a researcher whose main work was monitoring the shape of Dutch coastlines over a period spanning 15 Years. Paul was storing all his coast pictures in the SSD of his own laptop, until one day, suddenly, his computer OS could not access the SSD anymore.

Can you identify the problem?

Outdated media
Incorrect, media was Not outdated, SSD Is a modern data Storage media

Outdated file format
Incorrect, nothing to do With files, but with Access to files

Defective media
Correct, the media, Over time, got Defective

Bit Rot
Incorrect, here the Problem concerns File format

Data Repository oblivion
Incorrect, the data did not Go into oblivion as it is Currently used to finalise The researcher thesis

Media Loss
Incorrect, the problem concerns Data files not storage

Which strategy could you apply to prevent the problem?

Multiple backups
Maybe multiple backups Could have saved Paul's Data, but SSD as old As the main one would Suffer the same problem Eventually

Common File formats
Incorrect, the problem Has nothing to do With file formats, but With accessing the files

Checksum
Checksum prevents File corruption, but Here our problem Is media

Update File formats
Incorrect, the problem Has nothing to do With file formats, but Accessing files

Refresh Data
Correct, copying The data to new, Fresh, media would Have solved the Problem

Hire/Consult a specialist
Maybe a specialist could Have been aware of The issues concerning SSD Lifespan and could have done something about it.

Wrapping up

- Preserving data requires extra efforts compared to simply storing data;
- Over time data can be lost because a large variety of reasons;
- There are six main causes of data loss over time (excluding human error);
- Each data loss risk may be fought with one or more data preserving strategy.