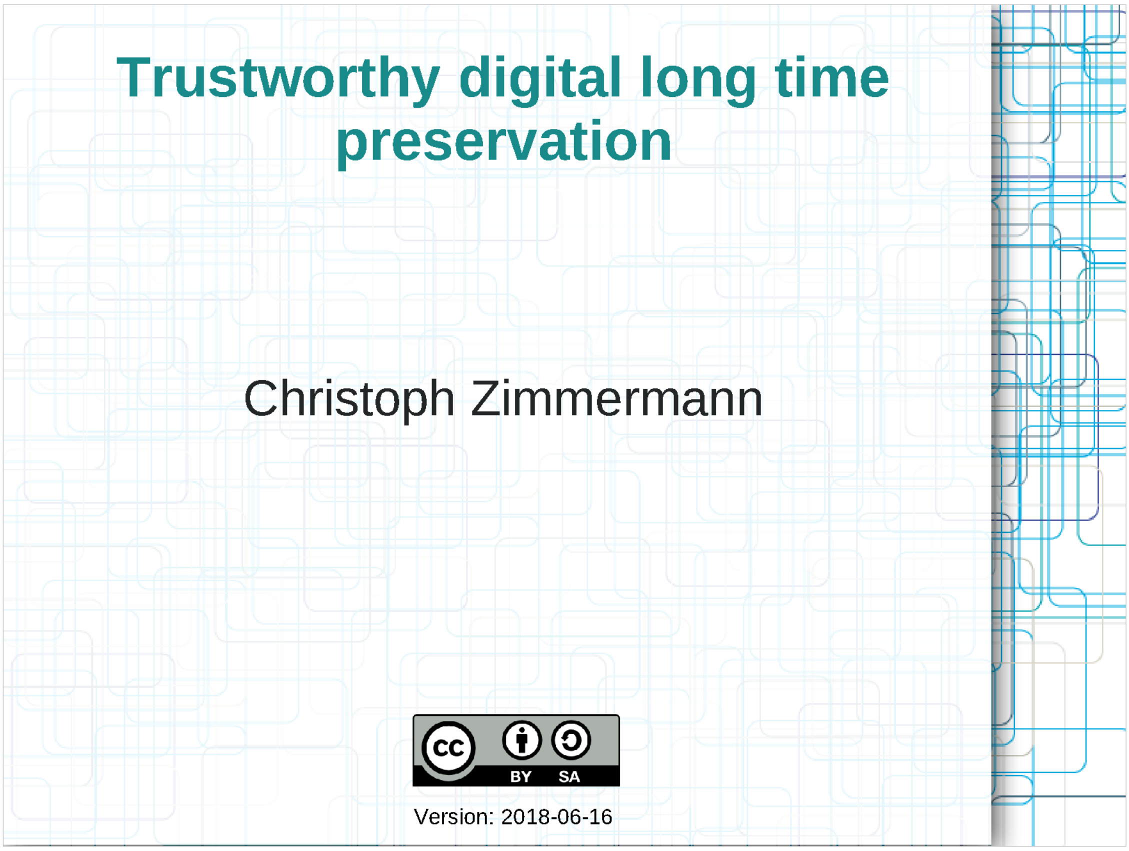
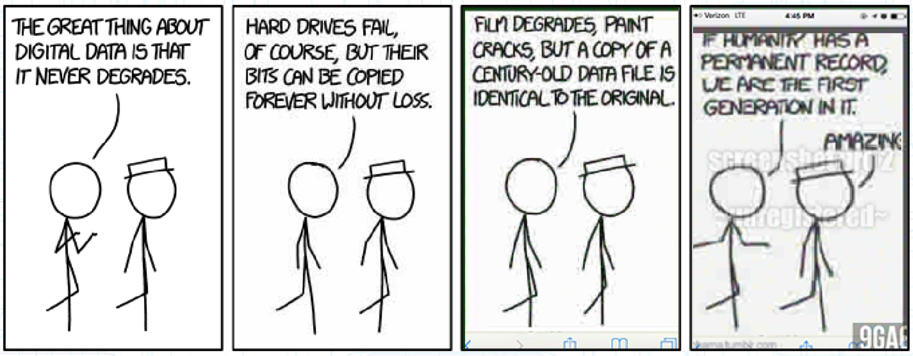


Trustworthy digital long time preservation

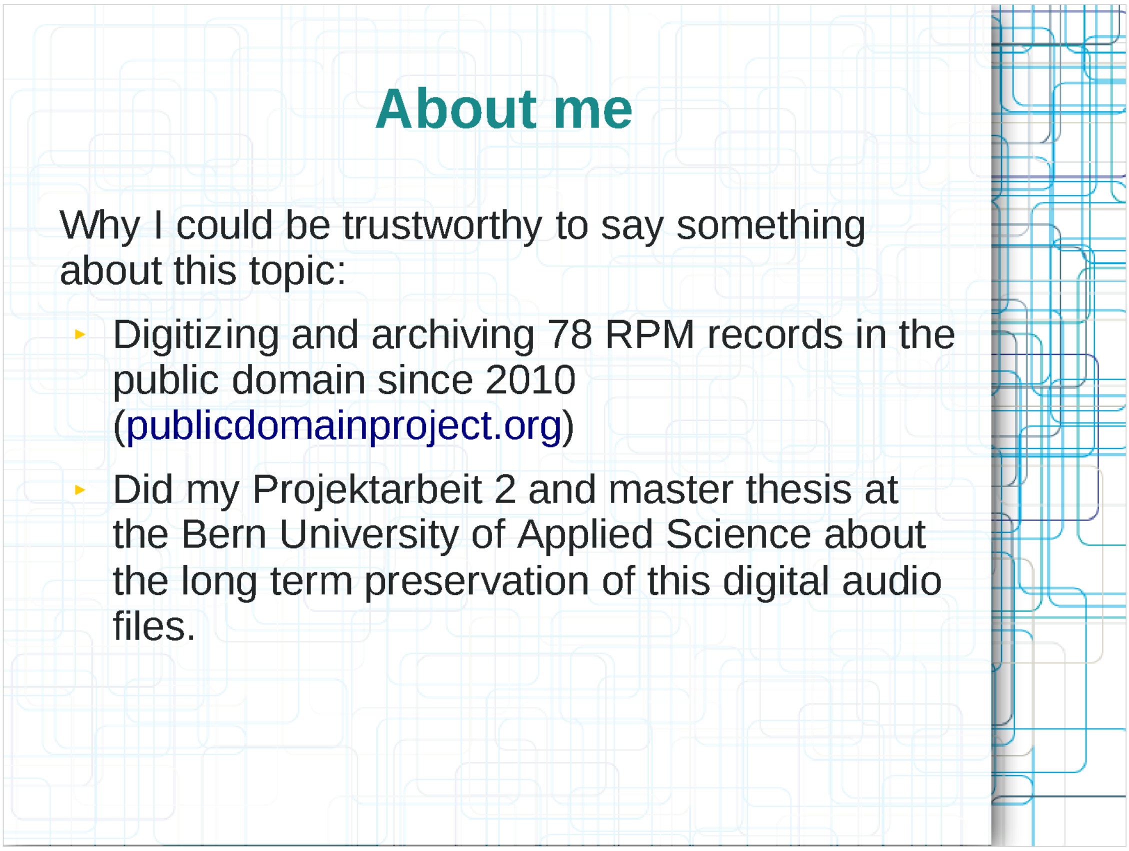
Christoph Zimmermann

Version: 2018-06-16

Introduction


THE GREAT THING ABOUT DIGITAL DATA IS THAT IT NEVER DEGRADES.
HARD DRIVES FAIL, OF COURSE, BUT THEIR BITS CAN BE COPIED FOREVER WITHOUT LOSS.
FILM DEGRADES, PAINT CRACKS, BUT A COPY OF A CENTURY-OLD DATA FILE IS IDENTICAL TO THE ORIGINAL.
Verizon LTE 4:45 PM
IF HUMANITY HAS A PERMANENT RECORD, WE ARE THE FIRST GENERATION IN IT.
AMAZING
9GAG

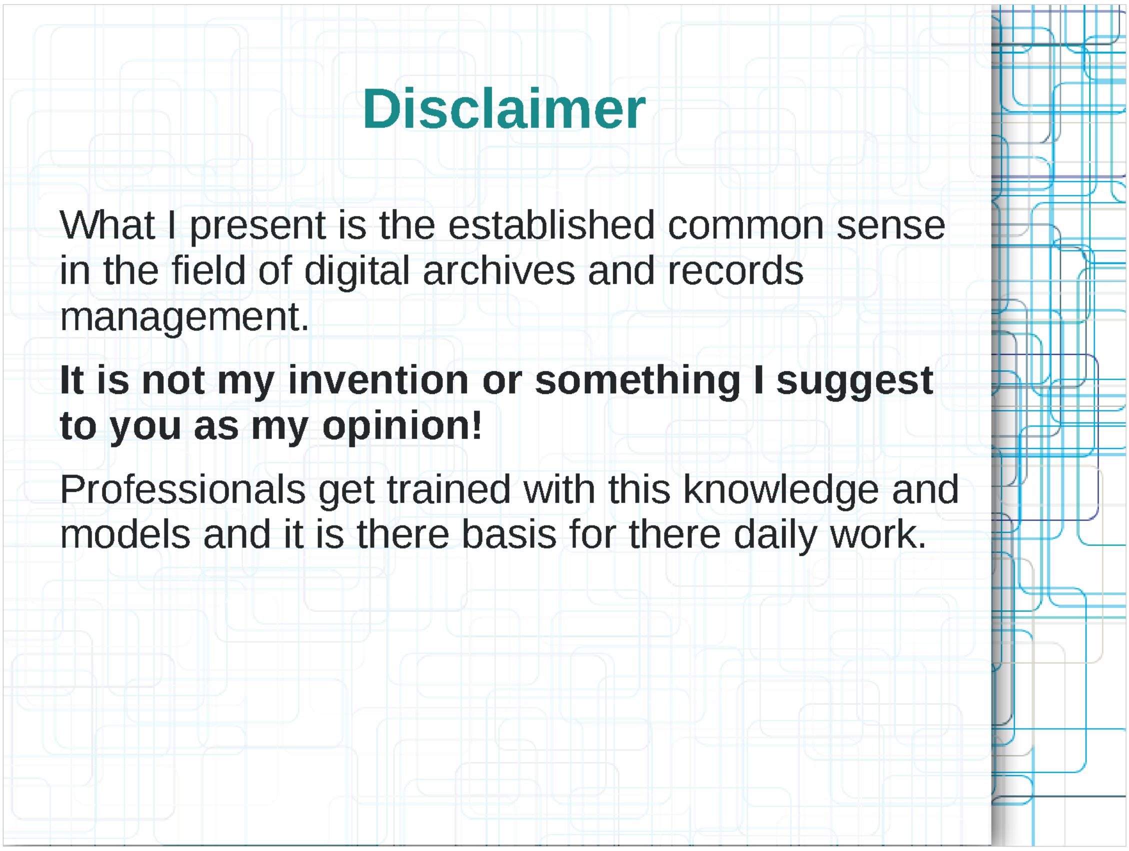
Let's start with a XKCD comic that leads us in very short time to the essence of this talk and the topic in general



About me

Why I could be trustworthy to say something about this topic:

- Digitizing and archiving 78 RPM records in the public domain since 2010 (publicdomainproject.org)
- Did my Projektarbeit 2 and master thesis at the Bern University of Applied Science about the long term preservation of this digital audio files.



Disclaimer

What I present is the established common sense in the field of digital archives and records management.

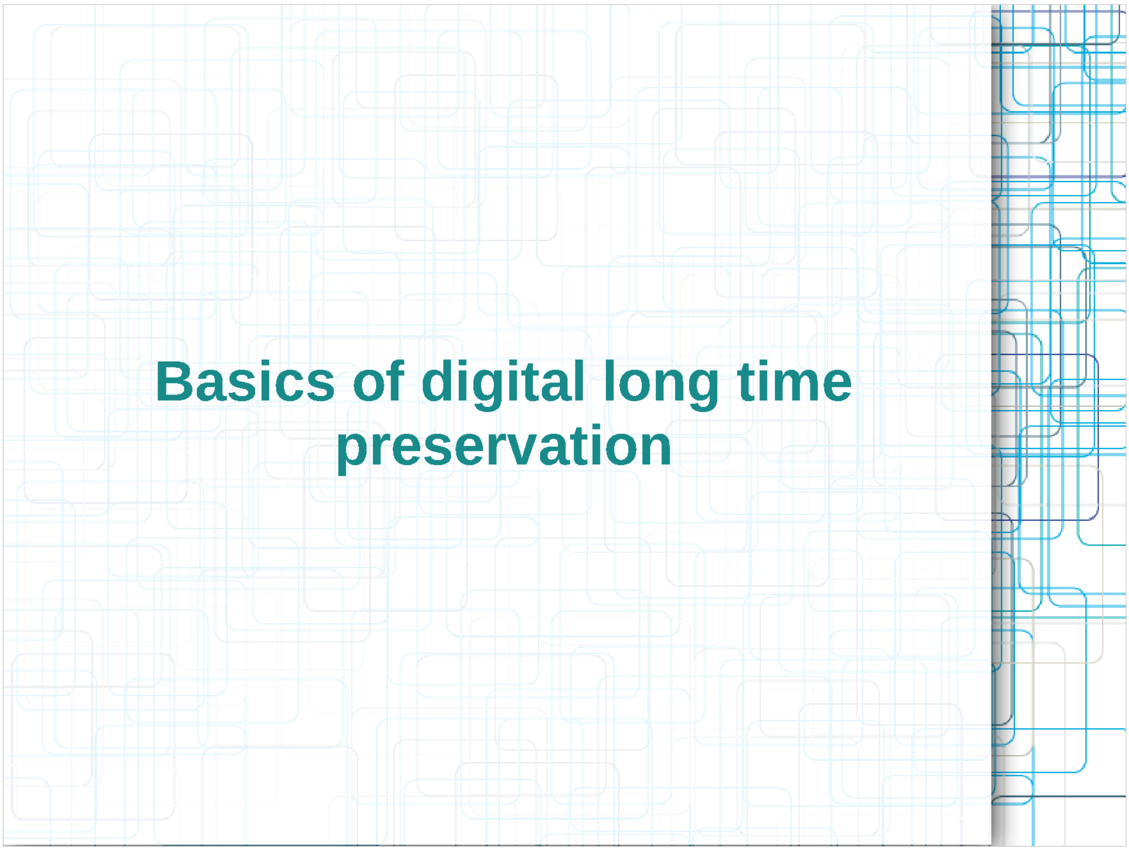
It is not my invention or something I suggest to you as my opinion!

Professionals get trained with this knowledge and models and it is there basis for there daily work.

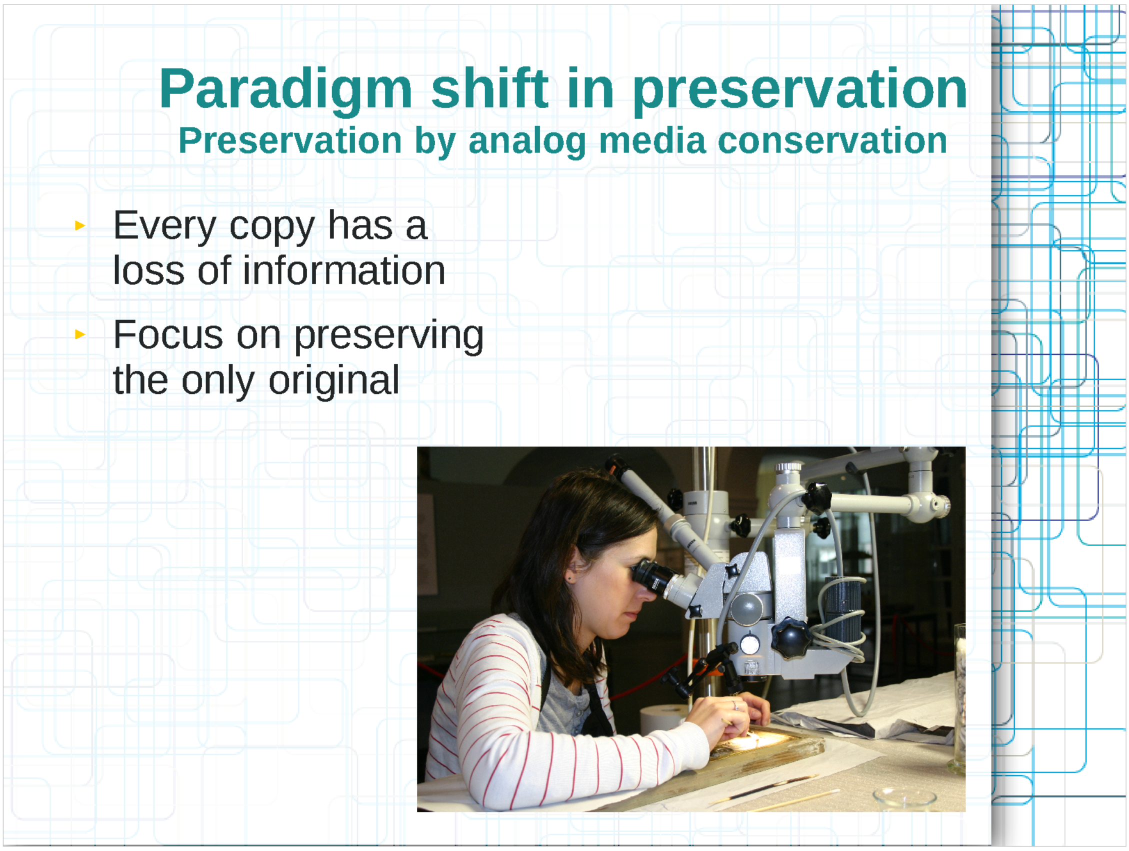
Don't shoot the messenger. It will not help you :-)

Structure of this presentation

- Basics of digital long time preservation
- History of digital preservation standards
- The OAIS model
- What does all that mean in practical terms?
- Questions, answers and discussions



Basics of digital long time preservation



Paradigm shift in preservation

Preservation by analog media conservation

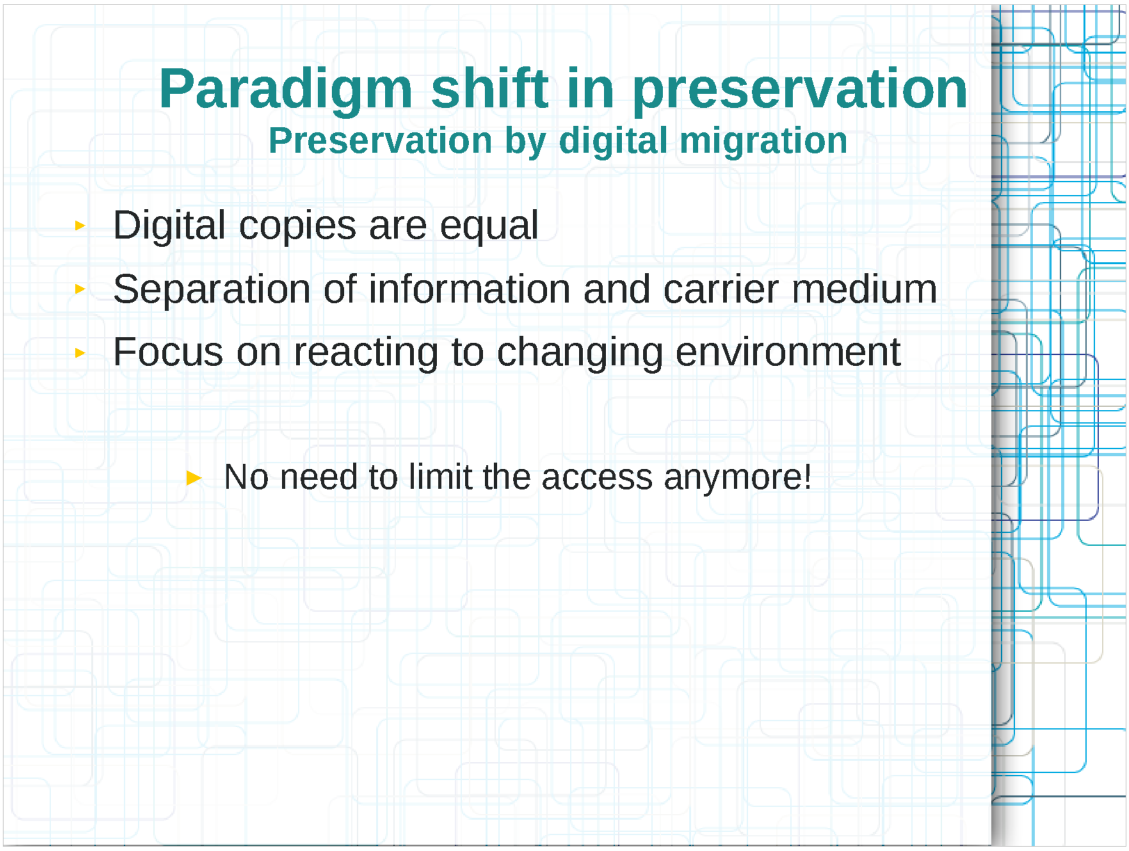
- Every copy has a loss of information
- Focus on preserving the only original

In the analog world there is always only one (or a few) originals and there is always a loss when this original is copied.

All the effort goes into the preservation of this only original that is in an archive. Means protected access, climate control, keep away from sunlight/fingerprints etc.

Often only copies are available for public access or no access at all.

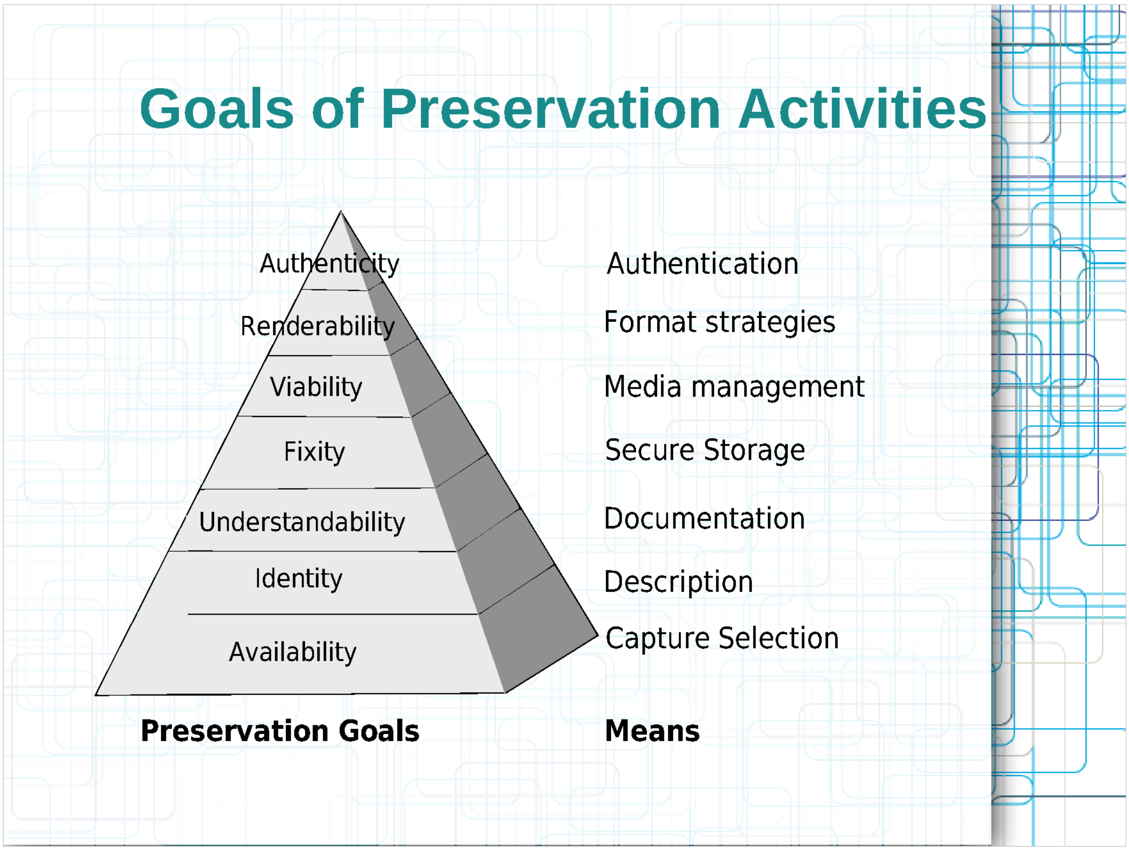
Image Source: Museum für Hamburgische Geschichte, CC BY SA
https://commons.wikimedia.org/wiki/File:MHG_Arbeitsfotos_Holz_3.jpg



Paradigm shift in preservation

Preservation by digital migration

- Digital copies are equal
- Separation of information and carrier medium
- Focus on reacting to changing environment
  - No need to limit the access anymore!



Goals of Preservation Activities

Preservation Goals	Means
Authenticity	Authentication
Renderability	Format strategies
Viability	Media management
Fixity	Secure Storage
Understandability	Documentation
Identity	Description
Availability	Capture Selection

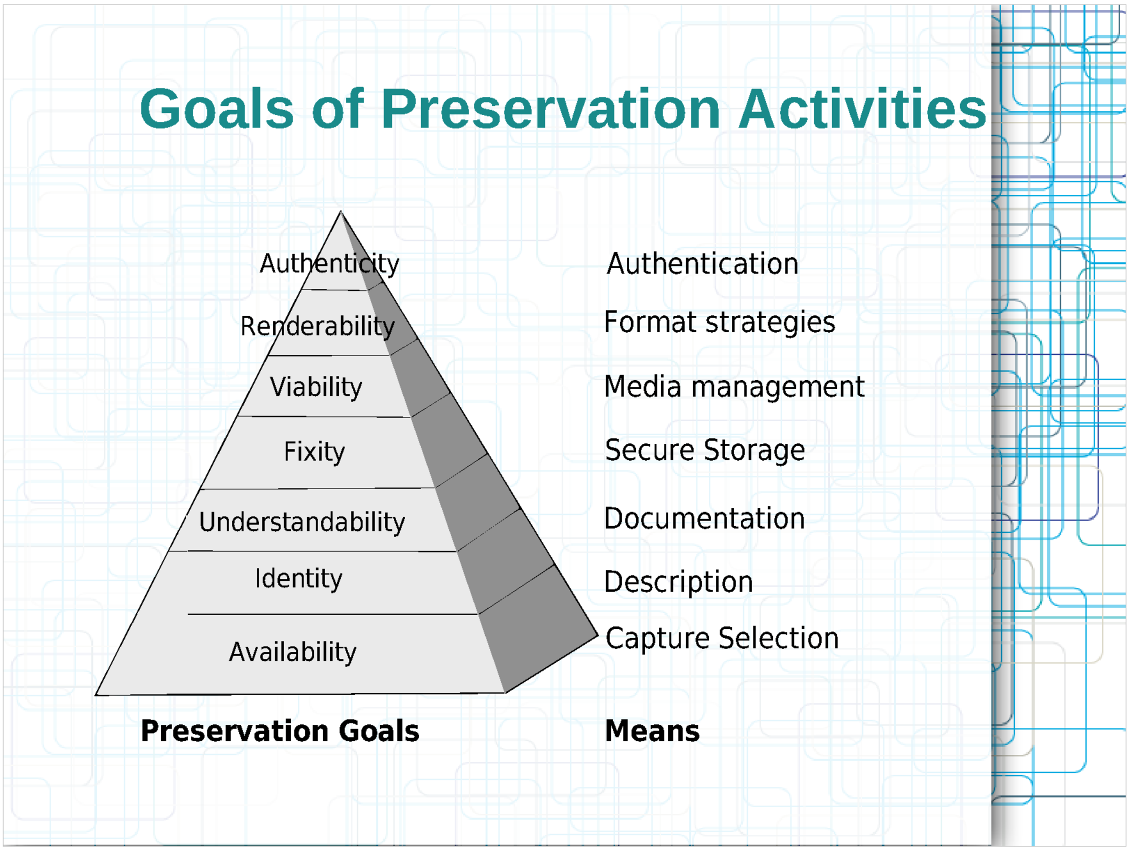
The preservation pyramid from Priscilla Caplan is a compact summarization of the whole range of properties of a digital object that have to be preserved to achieve the long term preservation. Source: Priscilla Caplan, What Is Digital Preservation?, Library Technology Reports, https://journals.ala.org/ltr/article/view/4224/4809, 2008

Since digital preservation is defined as a set of activities, it is most easily approached by asking what these activities are intended to accomplish:

Availability: It is a truism that you cannot preserve digital objects that you do not control. Depending on the materials and the circumstances, getting a copy of the objects may be trivial or quite difficult.

Identity: The creation of descriptive metadata is usually thought of in the context of discovery and access, but it is also a preservation activity. If the end of digital preservation is long-term access and/or usability, the digital object must be described in sufficient detail to allow future access and/or use. Ideally, digital objects should be self-describing; that is, they should carry descriptive metadata within them. [...] However, there is universal agreement that persistent identifiers are a critical element of descriptive metadata for preservation.

Understandability: […] The repository is responsible for providing and preserving enough information, as metadata, documentation, and/or related objects, to enable future users to understand the preserved objects.



Goals of Preservation Activities

Preservation Goals	Means
Authenticity	Authentication
Renderability	Format strategies
Viability	Media management
Fixity	Secure Storage
Understandability	Documentation
Identity	Description
Availability	Capture Selection

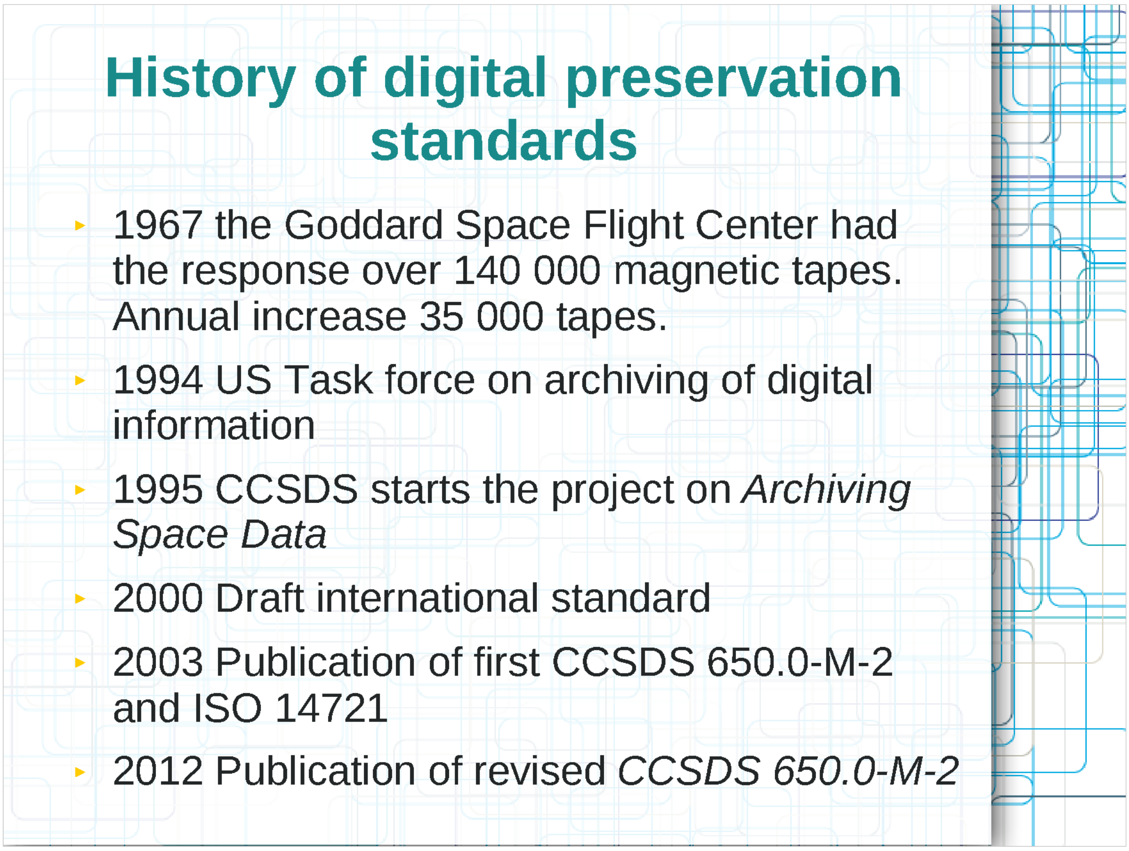
Fixity: Preservation systems must protect digital objects from unauthorized changes, whether deliberate or inadvertent. Industry-standard computer security regimes are the best defense against both malicious and careless behavior.[...] Media degradation can also cause bitstream corruption and is prevented by sound storage management practices[...] Fixity errors are detected by comparing message digests (more commonly called “checksums”) calculated over the same file at different times.

Viability: is the quality of being readable from media. Media deterioration and media obsolescence are threats to viability[...]Files should be copied periodically to new media, and backup copies should be kept on different physical devices. However, ensuring viability for content that has been neglected can be a serious problem. Archives that once may have received cartons of personal papers may now be given shoeboxes full of obsolete floppy disks.

Renderability: Ensuring that a digital file is renderable (displayable, playable, or otherwise usable as appropriate) may be the heart of the digital preservation process. A file may be authentic, uncorrupted, and perfectly viable, but if the hardware or software required to render it is no longer available, the file is essentially unusable.

Authenticity: Often defined anthropomorphically as “the quality that an object is what it purports to be,” authenticity means that the integrity of both the source and the content of an object can be verified. (It does not refer to the veracity of the object—an authentic document could be totally untrue.) [...] and to ensure that the chain of custody and all authorized changes are documented. The event history pertaining to a digital object is known as its “digital provenance” and is a critical part of preservation metadata.

History of digital preservation standards

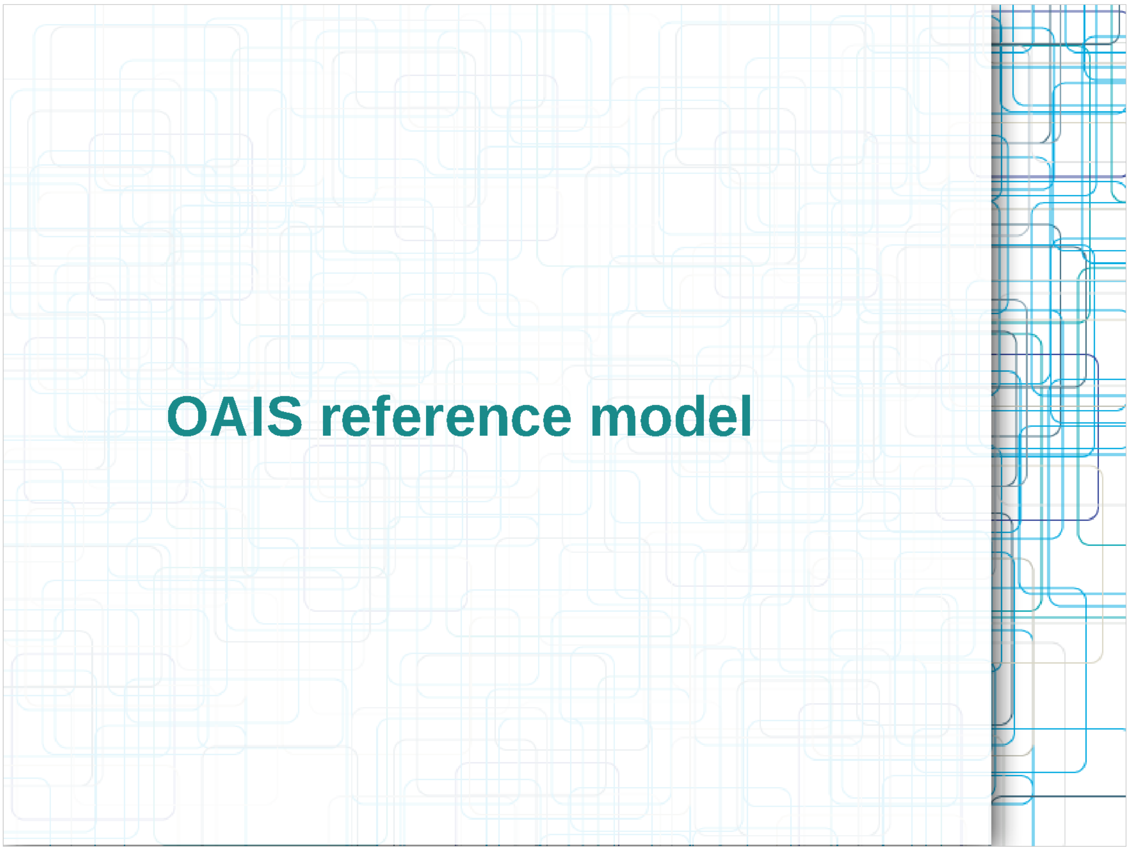


History of digital preservation standards

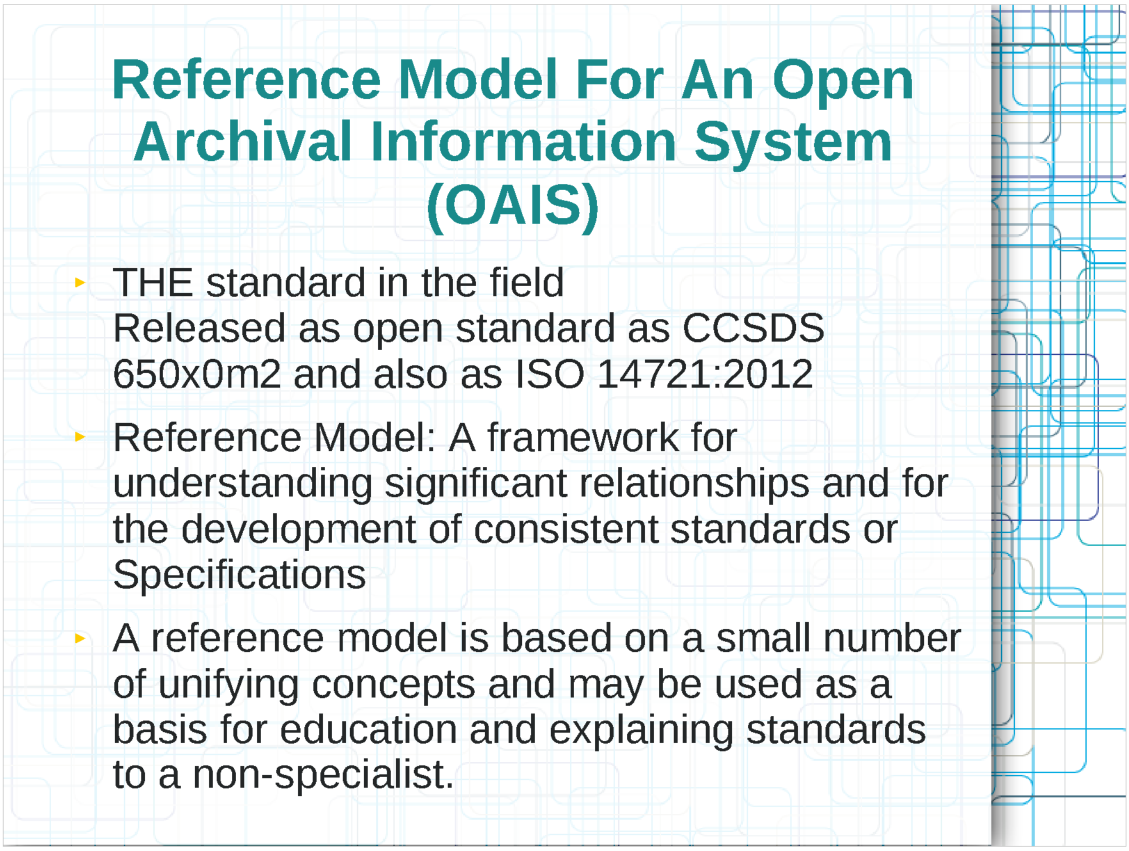
- 1967 the Goddard Space Flight Center had the response over 140 000 magnetic tapes. Annual increase 35 000 tapes.
- 1994 US Task force on archiving of digital information
- 1995 CCSDS starts the project on *Archiving Space Data*
- 2000 Draft international standard
- 2003 Publication of first CCSDS 650.0-M-2 and ISO 14721
- 2012 Publication of revised *CCSDS 650.0-M-2*

1994 US Task force on archiving of digital information by the US *Comission on Preservation and Access* and the *Research Library Group*

The Consultative Committee for Space Data Systems (CCSDS)



OAIS reference model



Reference Model For An Open Archival Information System (OAIS)

- THE standard in the field
  Released as open standard as CCSDS 650x0m2 and also as ISO 14721:2012
- Reference Model: A framework for understanding significant relationships and for the development of consistent standards or Specifications
- A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist.

REFERENCE MODEL FOR AN OPEN ARCHIVALINFORMATION SYSTEM (OAIS), CCSDS, CCSDS 650.0-M-2, 2012
http://public.ccsds.org/publications/archive/650x0m2.pdf

nestor-Arbeitsgruppe OAIS-Übersetzung / Terminologie, Referenzmodell für ein Offenes Archiv-Informations-System - Deutsche Übersetzung, Version 2, nestor, urn:nbn:de:0008-2013082706, 2013
http://nbn-resolving.de/urn:nbn:de:0008-2013082706

Definitions

What means long term?

A period of time long enough for:

- Concern about the impacts of changing technologies, including support for new media and data formats
- Changing Designated Community

This period extends into the indefinite future.

Long Term Archive

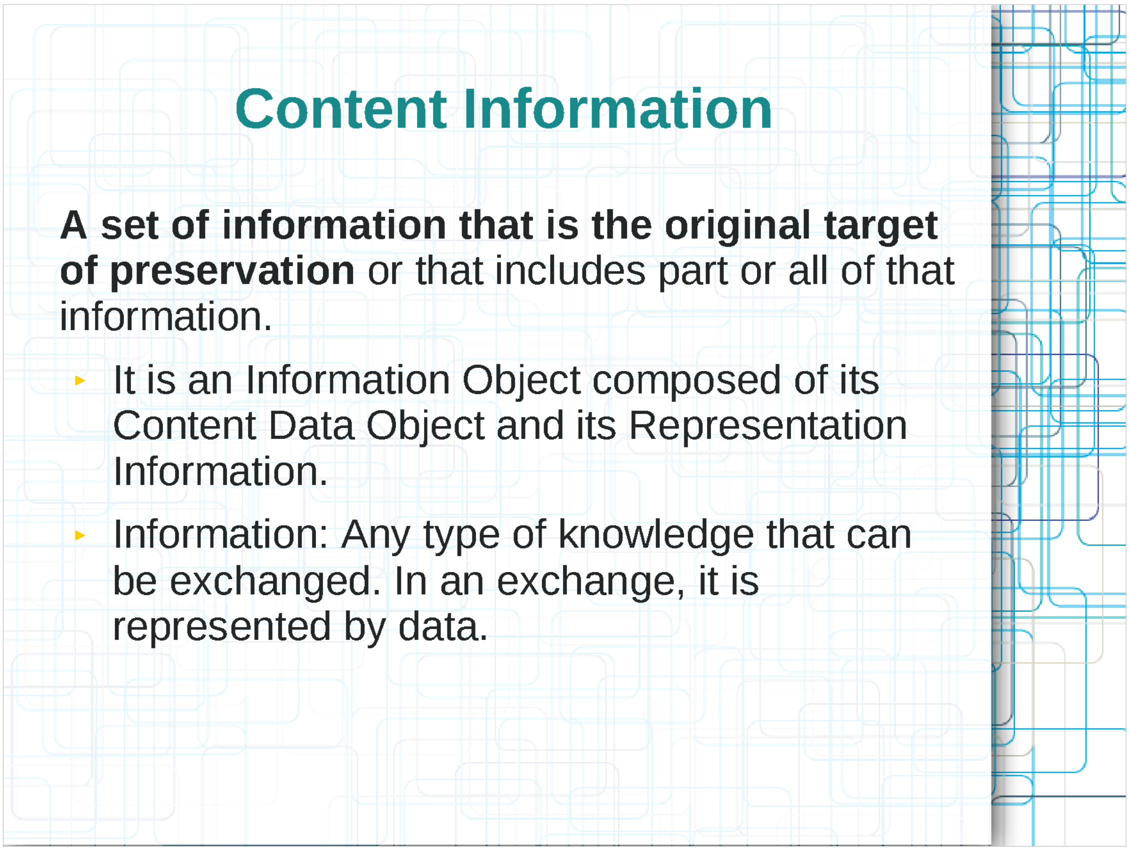
[...] an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community.

- **Remember: People and systems!**

Long Term Preservation

The act of maintaining information

- Independently Understandable by a Designated Community
- and with evidence supporting its Authenticity, over the Long Term.



Content Information

A set of information that is the original target of preservation or that includes part or all of that information.

- It is an Information Object composed of its Content Data Object and its Representation Information.
- Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by data.

Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by data. An example is a string of bits (the data) accompanied by a description of how to interpret the string of bits as numbers representing temperature observations measured in degrees Celsius (the Representation Information)

Designated communities

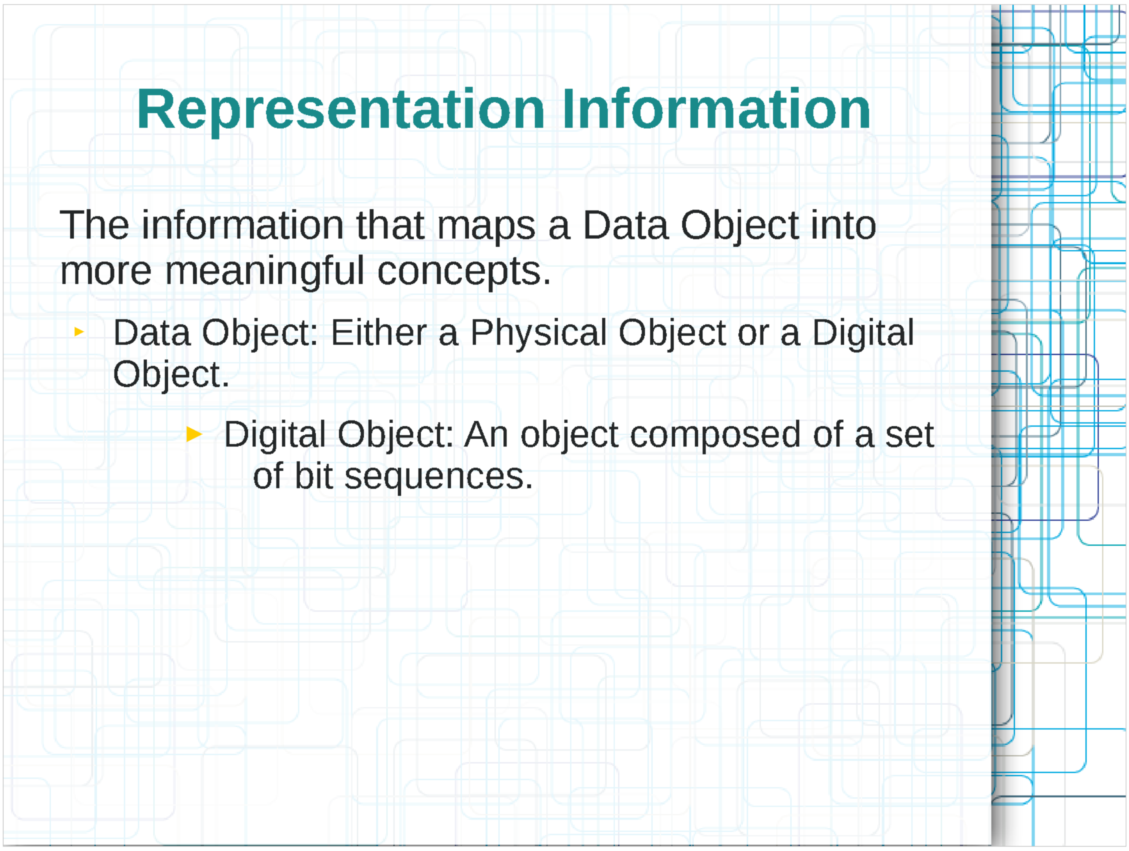
An identified group of potential Consumers who should be able to understand a particular set of information.

- A Designated Community is defined by the Archive and this definition may change over time.
- Knowledge Base: A set of information, incorporated by a person or system, that allows that person or system to understand received information.

Independently Understandable

A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

- To achieve this we need so called representation information



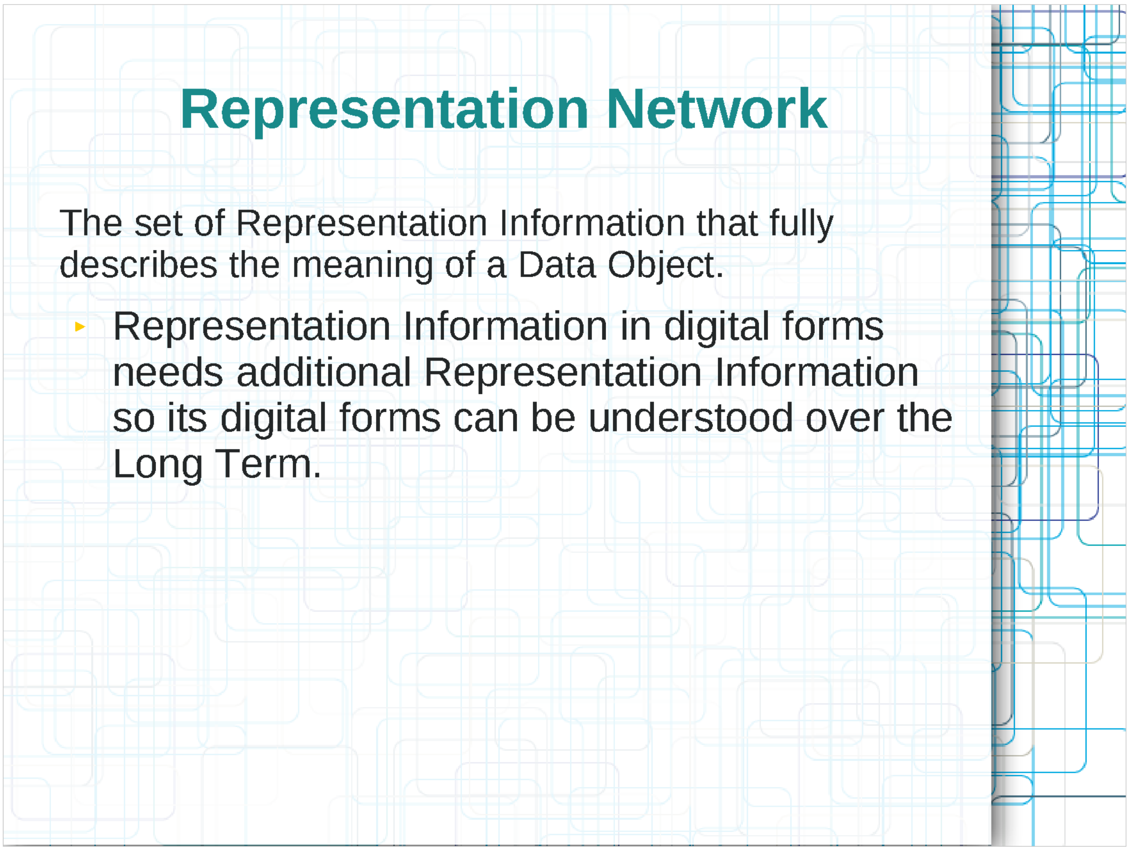
Representation Information

The information that maps a Data Object into more meaningful concepts.

- Data Object: Either a Physical Object or a Digital Object.
    - Digital Object: An object composed of a set of bit sequences.

An example of Representation Information for a bit sequence which is a FITS file might consist of the FITS standard which defines the format plus a dictionary which defines the meaning in the file of keywords which are not part of the standard.

Another example is JPEG software which is used to render a JPEG file; rendering the JPEG file as bits is not very meaningful to humans but the software, which embodies an understanding of the JPEG standard, maps the bits into pixels which can then be rendered as an image for human viewing.



Representation Network

The set of Representation Information that fully describes the meaning of a Data Object.

- Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term.

Structure Information: The Representation Information that imparts meaning about how other information is organized. For example, it maps bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays.

Semantic Information: The Representation Information that further describes the meaning beyond that provided by the Structure Information.

Other Representation Information: Representation Information which cannot easily be classified as Semantic or Structural. For example software, algorithms, encryption, written instructions and many other things may be needed to understand the Content Data Object, all of which therefore would be, by definition, Representation Information, yet would not obviously be either Structure or Semantics. Information defining how the Structure and the Semantic Information relate to each other, or software needed to process a database file would also be regarded as Other Representation Information.

Provenance Information

The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated.

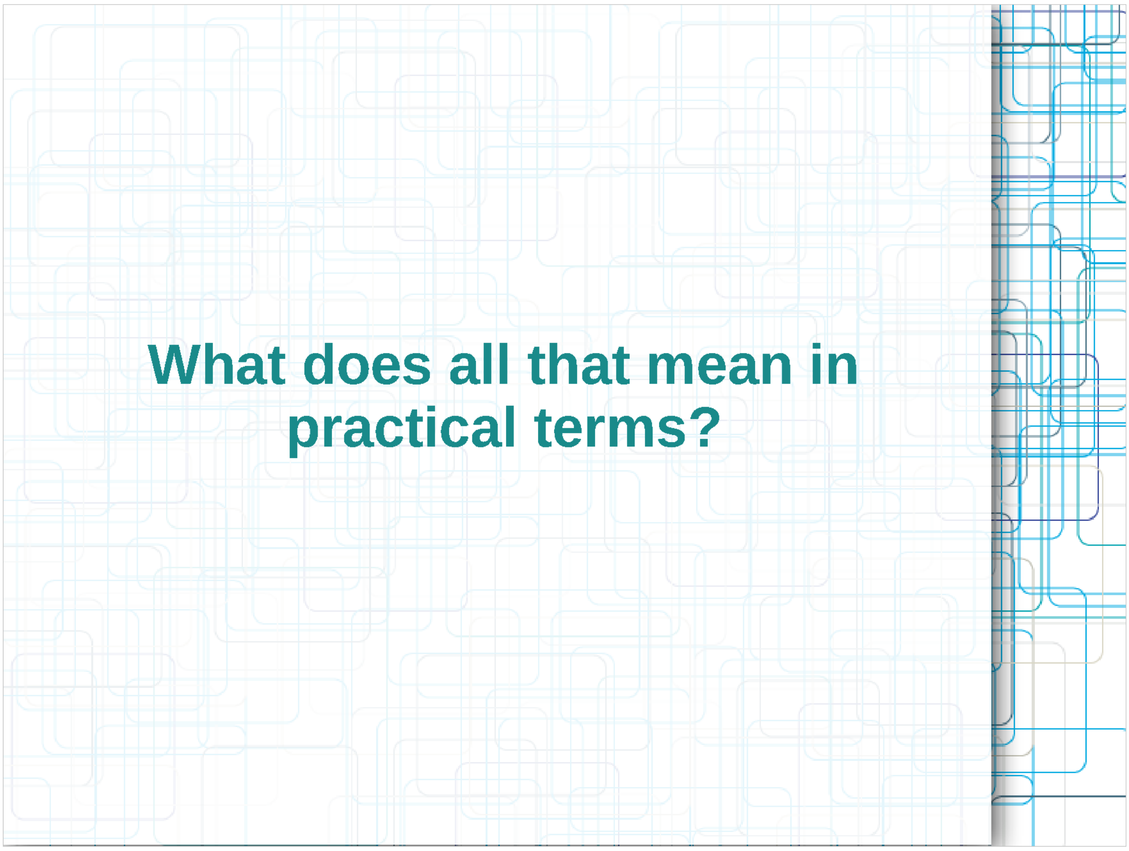
- The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer.
- Provenance Information adds to the evidence to support Authenticity.

Digital Migration

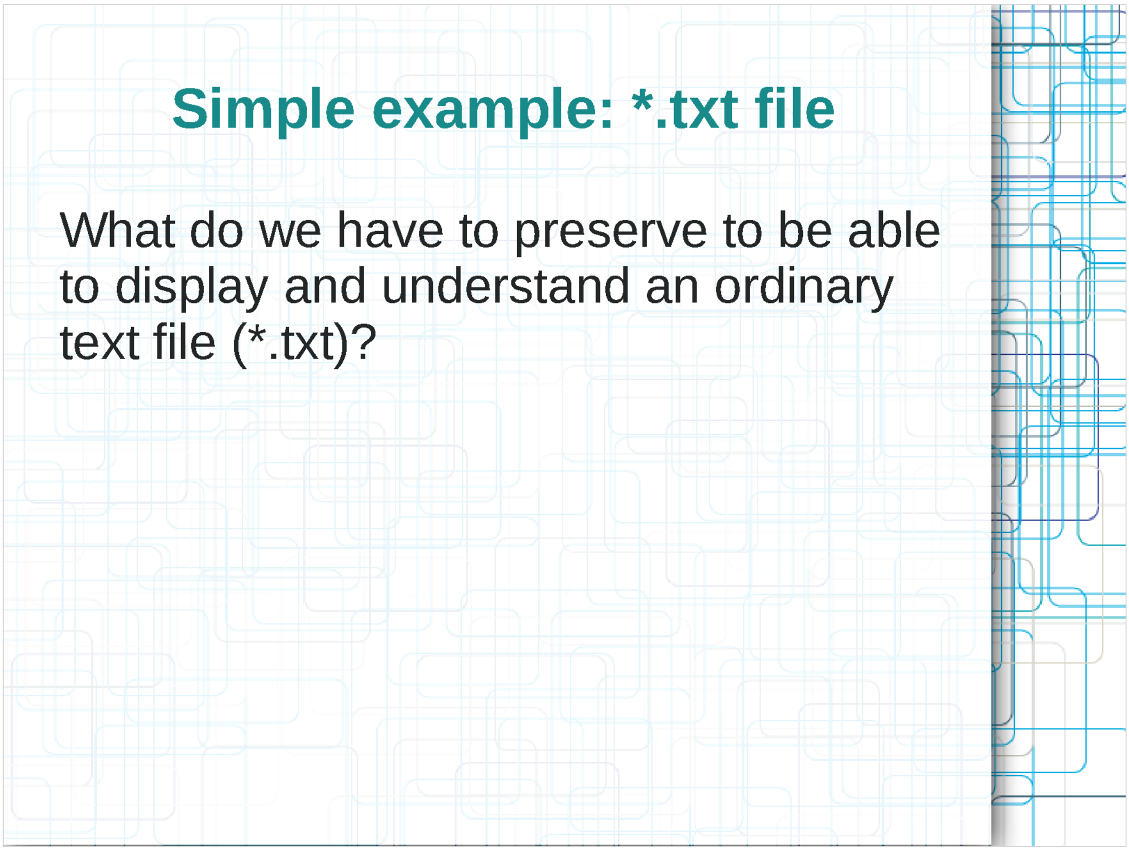
The transfer of digital information, while intending to preserve it, within the OAIS.

It is distinguished from transfers in general by three attributes:

- a focus on the preservation of the full information content that needs preservation
- a perspective that the new archival implementation of the information is a replacement for the old
- an understanding that full control and responsibility over all aspects of the transfer resides with the OAIS.



What does all that mean in practical terms?

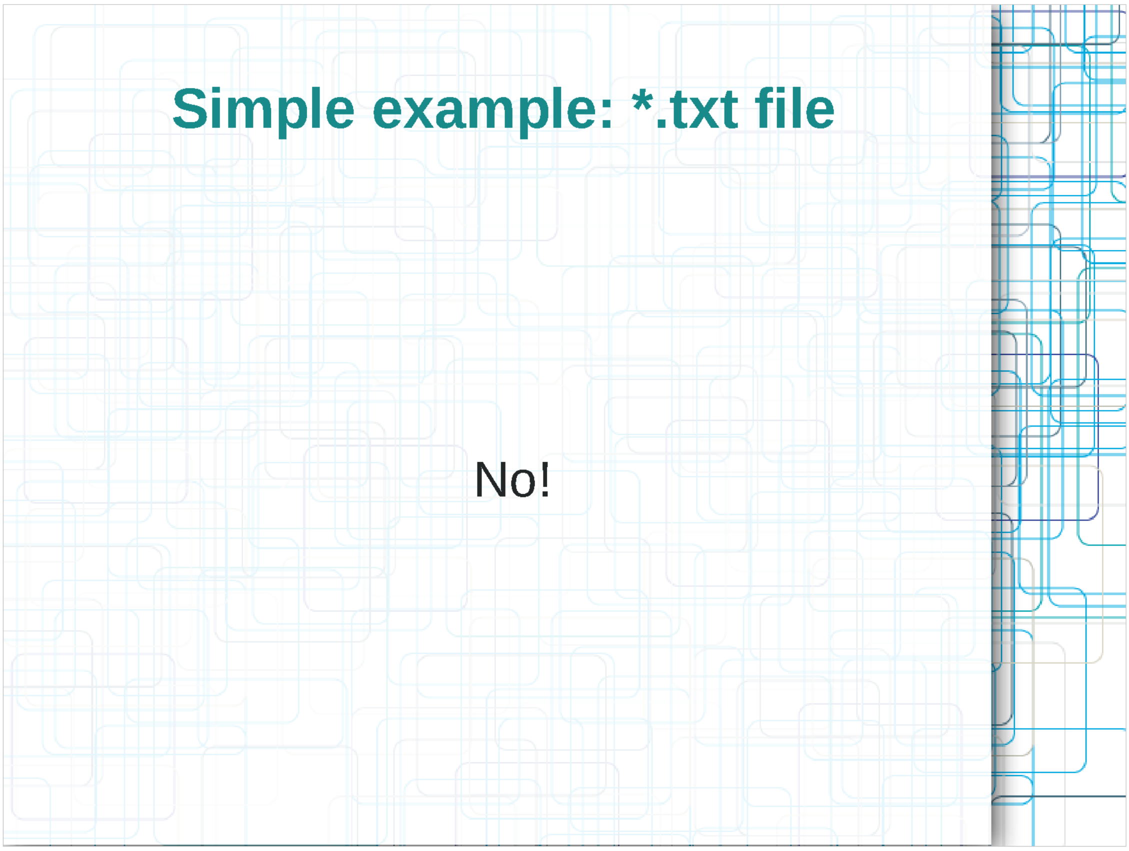


Simple example: *.txt file

What do we have to preserve to be able to display and understand an ordinary text file (*.txt)?

Simple example: *.txt file

- An ASCII Table (Nowadays a Unicode table)
- Is that enough?

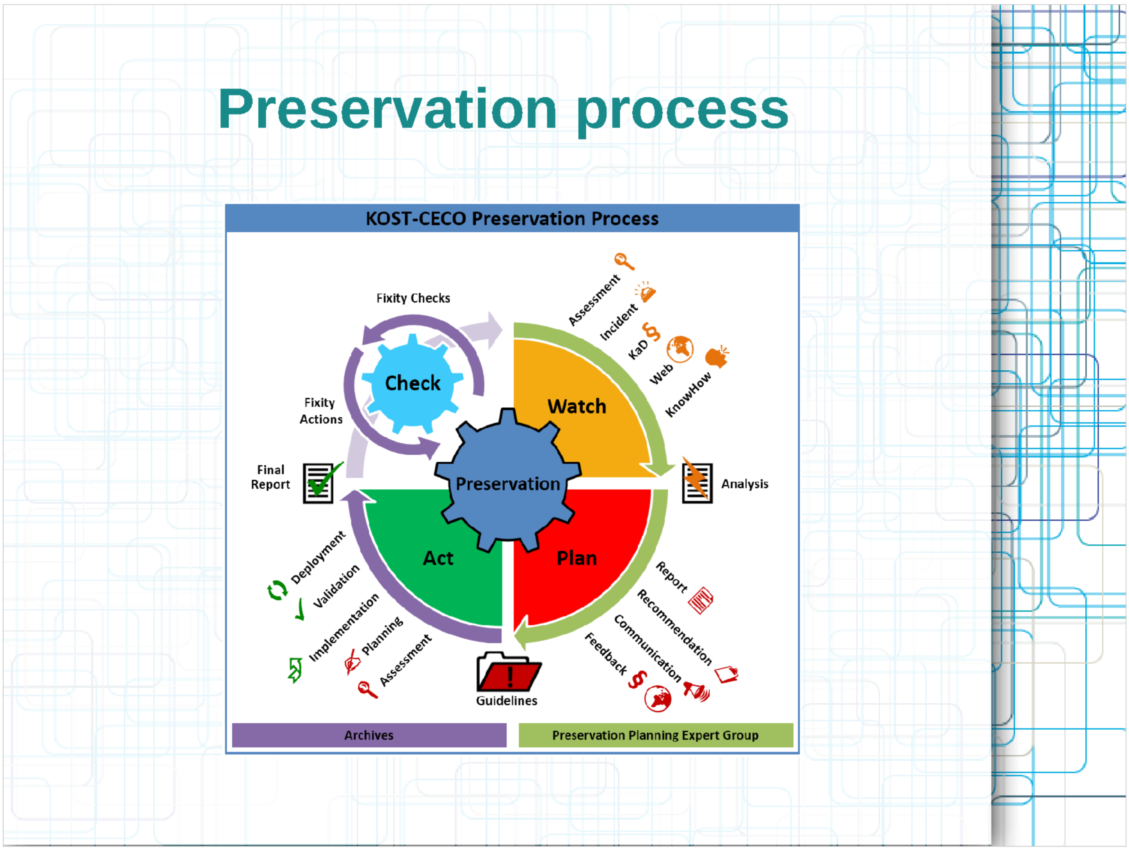


Simple example: *.txt file

No!

Simple example: *.txt file

- Representation information:
  - An ASCII and Unicode table
  - Technical specifications about the file system, storage media, hardware and software interfaces etc. down to the detail level of every bit.
  - Dictionary for the used (human) language
- Metadata about:
  - What, who, why, when, where
  - The history since it is in the repository



Preservation process

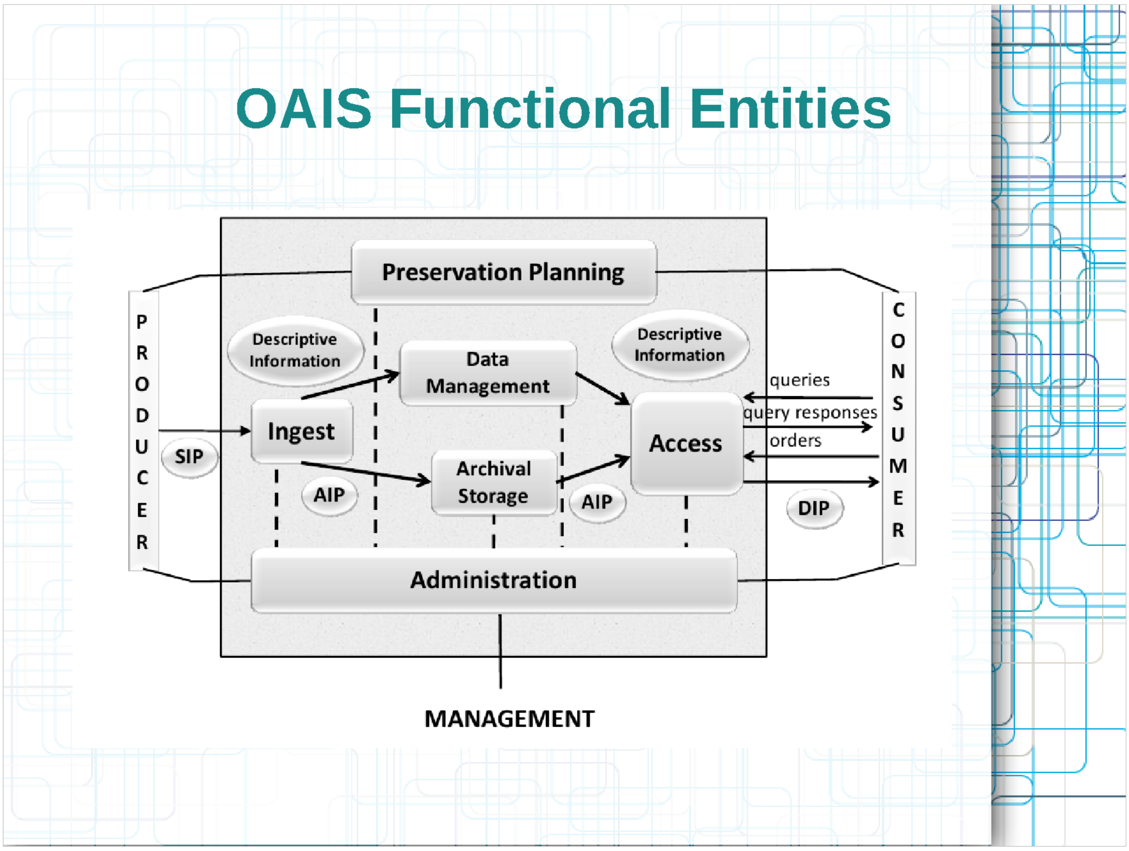
KOST-CECO Preservation Process

Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (kost-ceco), Preservation Process der KOST, 2015, http://kost-ceco.ch/cms/index.php?id=238,422,0,0,1,0

Example:
Designated communities

The Public Domain Project has following designated communities:

- A global community with access to a web browser, supporting HTML 4.0, finished school education, language level for English: A2
- Music scientist, historians with access to a web browser, supporting HTML 4.0, school education: undergraduate level or higher, basic knowledge of Dublin Core, language level for English: B2
- Search engines, meta archives, automated data analysis programs (bots) who can make queries with HTTP 1.1 and can handle as answer HTML 4.0 or RDF 1.1 (serialized as RDF/XML).



OAIS Functional Entities

Preservation Planning

PRODUCER

CONSUMER

Descriptive Information

Descriptive Information

Data Management

Ingest

SIP

AIP

Archival Storage

AIP

Access

queries

query responses

orders

DIP

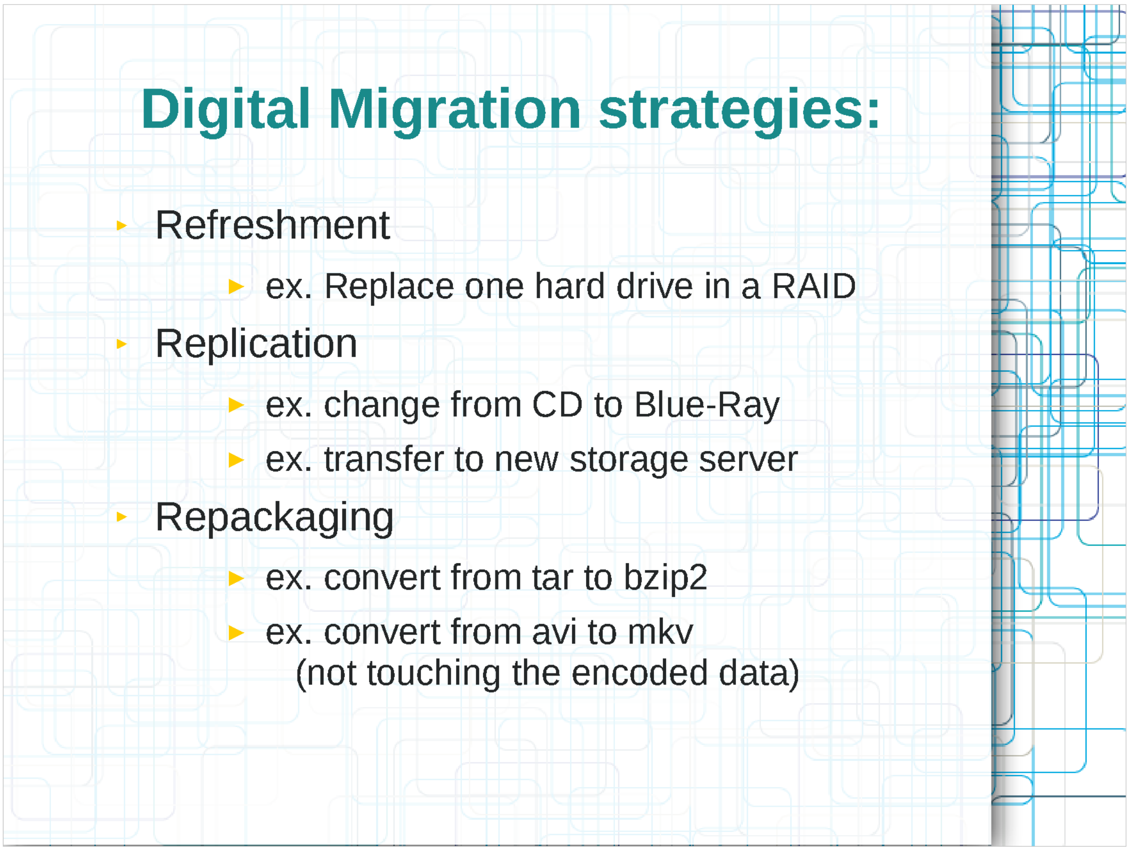
Administration

MANAGEMENT

Submission Information Package (SIP): An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

Dissemination Information Package (DIP): An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.



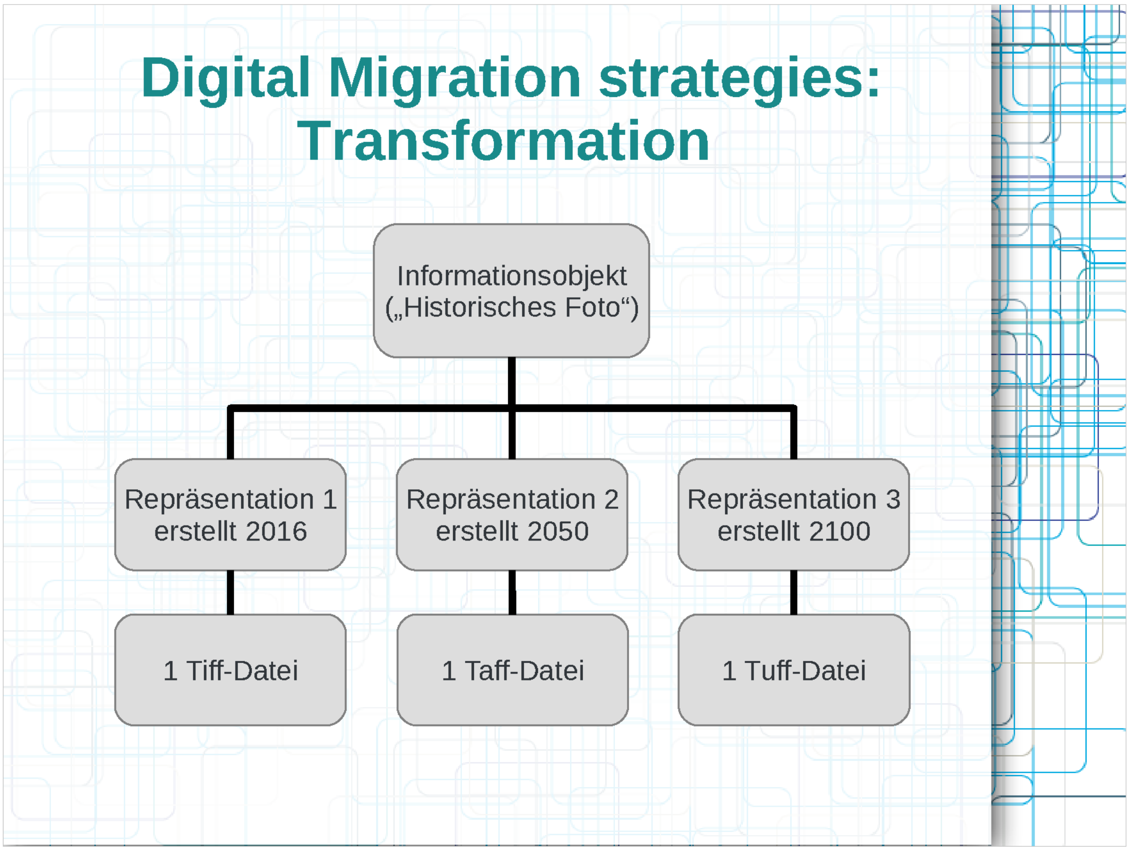
Digital Migration strategies:

- Refreshment
  - ex. Replace one hard drive in a RAID
- Replication
  - ex. change from CD to Blue-Ray
  - ex. transfer to new storage server
- Repackaging
  - ex. convert from tar to bzip2
  - ex. convert from avi to mkv
    (not touching the encoded data)

Refreshment: A Digital Migration where the effect is to replace a media instance with a copy that is sufficiently exact that all Archival Storage hardware and software continues to run as before.

Repackaging: A Digital Migration in which there is an alteration in the Packaging Information of the AIP.

Replication: A Digital Migration where there is no change to the Packaging Information the Content Information, and the PDI. The bits used to represent these Information Objects are preserved in the transfer to the same or new media instance.



Digital Migration strategies: Transformation

Informationsobjekt („Historisches Foto“)

Repräsentation 1 erstellt 2016

Repräsentation 2 erstellt 2050

Repräsentation 3 erstellt 2100

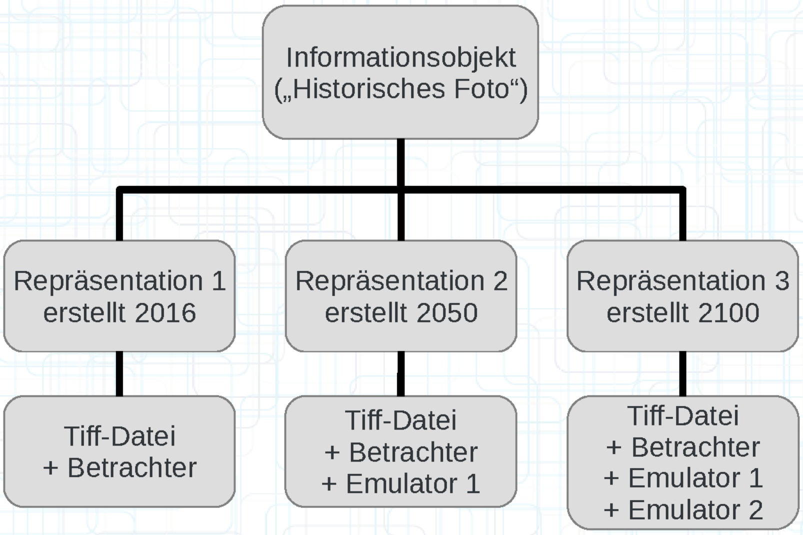
1 Tiff-Datei

1 Taff-Datei

1 Tuff-Datei

Transformation: A Digital Migration in which there is an alteration to the Content Information or PDI of an Archival Information Package. For example, changing ASCII codes to UNICODE in a text document being preserved is a Transformation.

Digital Migration strategies: Emulation



Informationsobjekt („Historisches Foto“)

Repräsentation 1 erstellt 2016

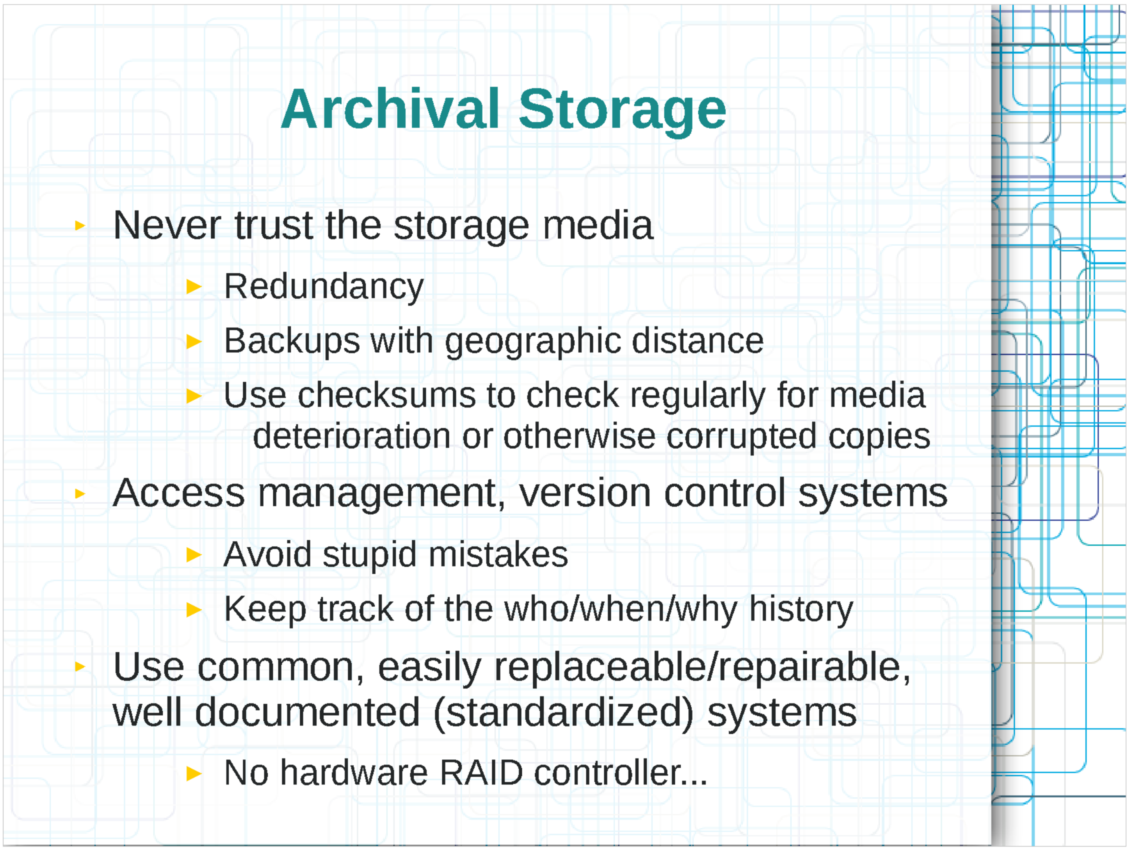
Repräsentation 2 erstellt 2050

Repräsentation 3 erstellt 2100

Tiff-Datei
+ Betrachter

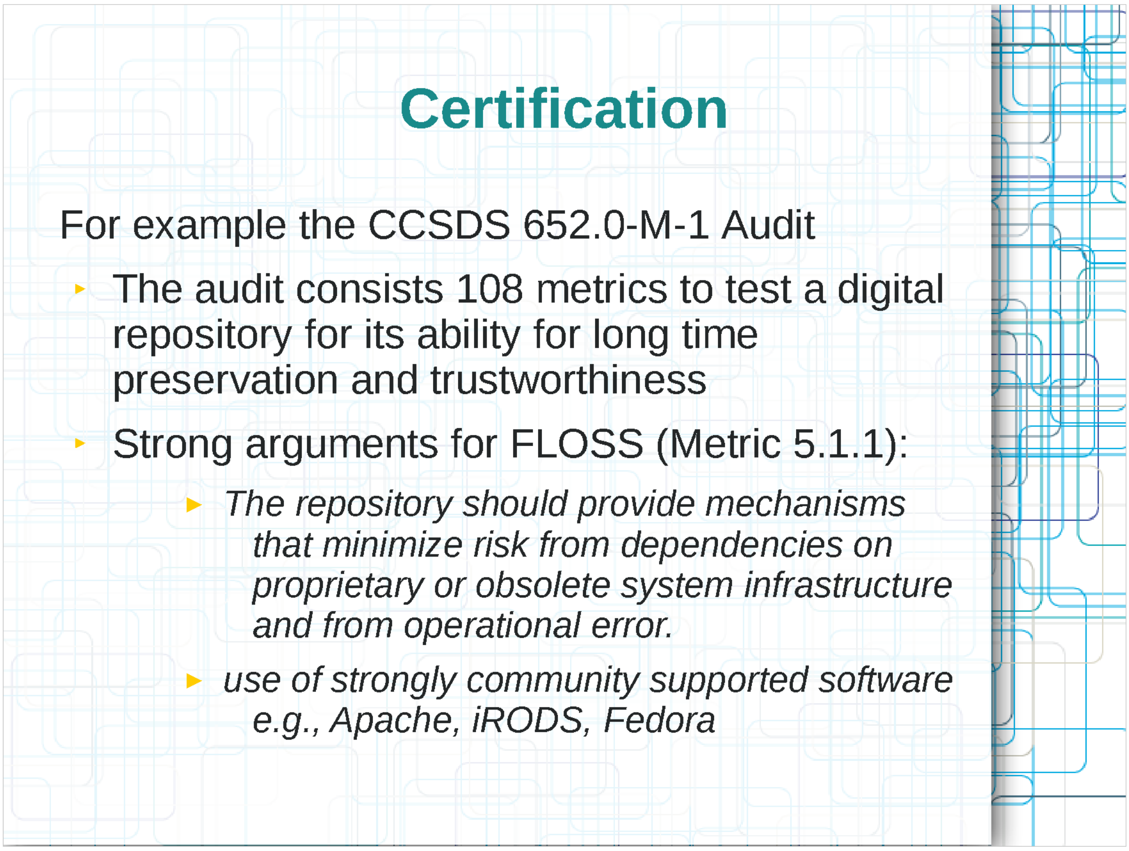
Tiff-Datei
+ Betrachter
+ Emulator 1

Tiff-Datei
+ Betrachter
+ Emulator 1
+ Emulator 2



Archival Storage

- Never trust the storage media
  - Redundancy
  - Backups with geographic distance
  - Use checksums to check regularly for media deterioration or otherwise corrupted copies
- Access management, version control systems
  - Avoid stupid mistakes
  - Keep track of the who/when/why history
- Use common, easily replaceable/repairable, well documented (standardized) systems
  - No hardware RAID controller...



Certification

For example the CCSDS 652.0-M-1 Audit

- The audit consists 108 metrics to test a digital repository for its ability for long time preservation and trustworthiness
- Strong arguments for FLOSS (Metric 5.1.1):
  - *The repository should provide mechanisms that minimize risk from dependencies on proprietary or obsolete system infrastructure and from operational error.*
  - *use of strongly community supported software e.g., Apache, iRODS, Fedora*

https://public.ccsds.org/Pubs/652x0m1.pdf

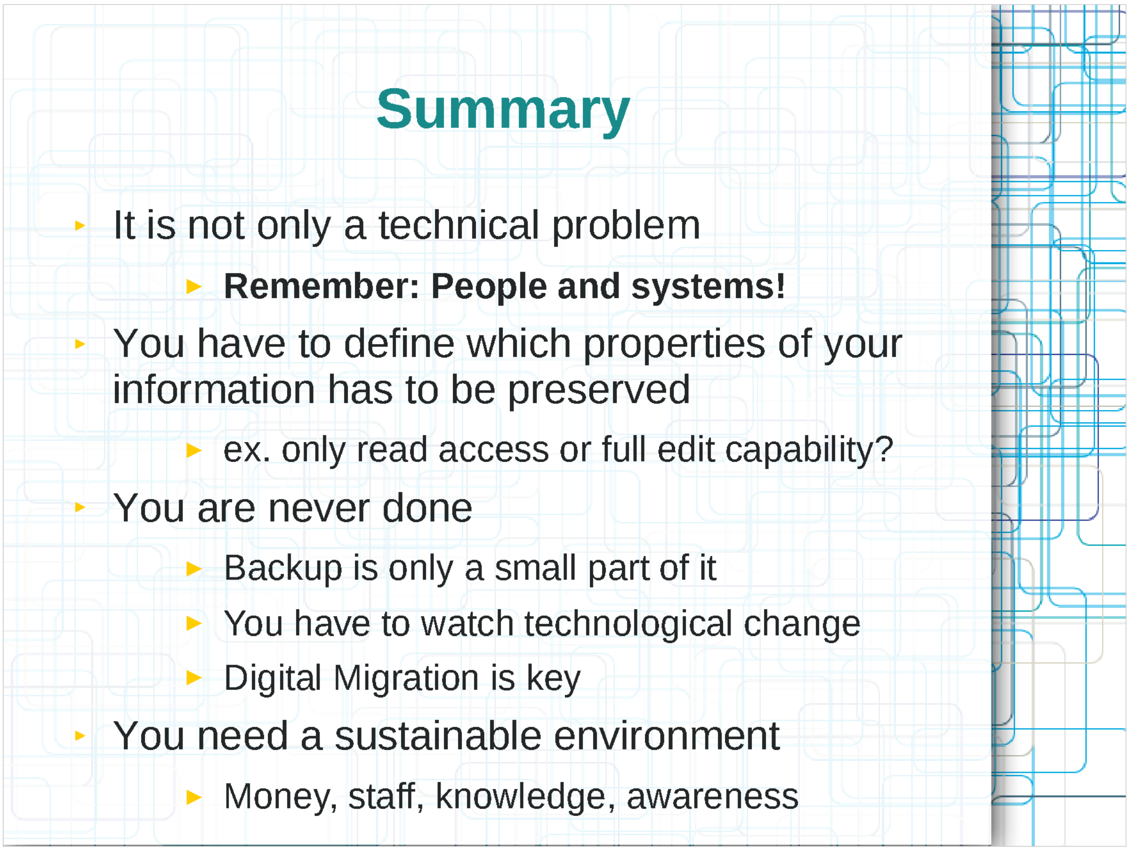
5.1 TECHNICAL INFRASTRUCTURE RISK MANAGEMENT

5.1.1 The repository shall identify and manage the risks to its preservation operations and goals associated with system infrastructure.

Supporting Text: This is necessary to ensure a secure and trustworthy infrastructure.

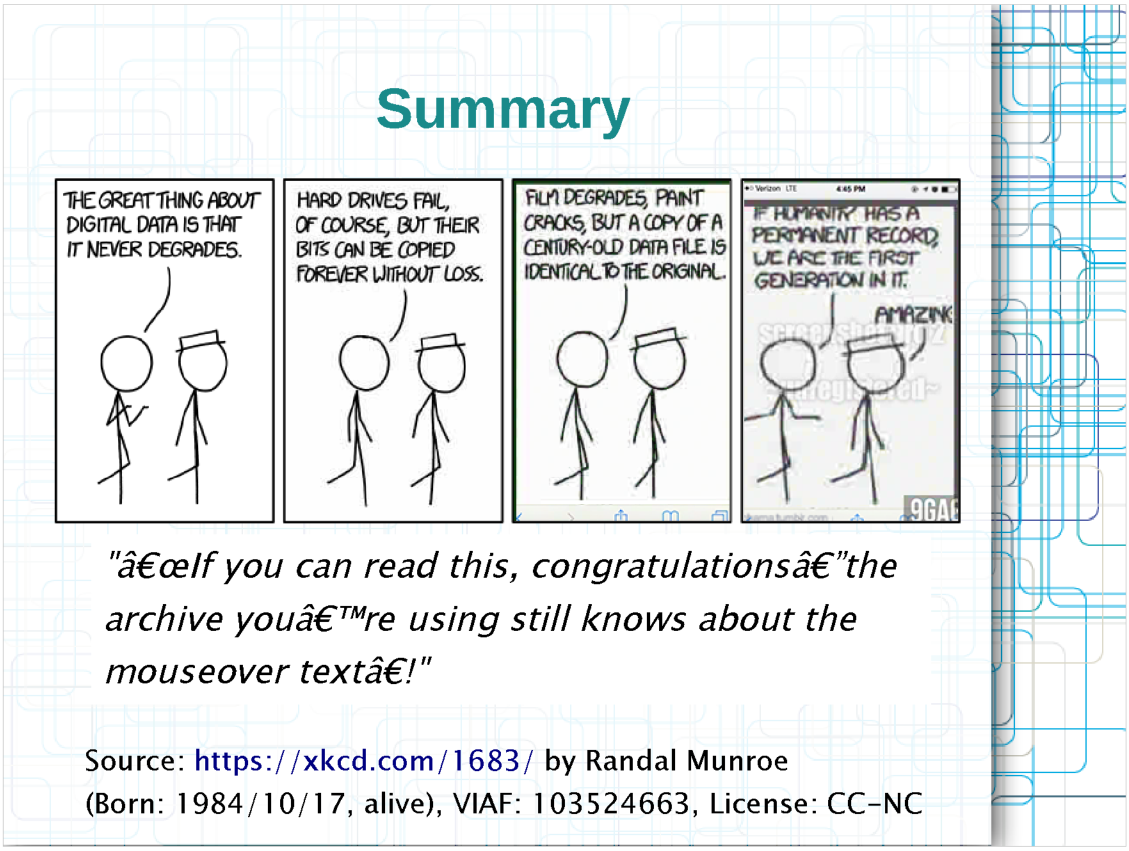
Examples of Ways the Repository Can Demonstrate It Is Meeting This Requirement: Infrastructure inventory of system components; periodic technology assessments; estimates of system component lifetime; export of authentic records to an independent system; use of strongly community supported software e.g., Apache, iRODS, Fedora); re-creation of archives from backups.

Discussion: The repository should conduct or contract assessments of the risks related to hardware and software infrastructure, and operational procedures. The repository should provide mechanisms that minimize risk from dependencies on proprietary or obsolete system infrastructure and from operational error. The degree of support required relates to the criticality of the subsystem(s) involved in long-term preservation. The repository should maintain a system that is scalable (e.g., able to handle anticipated future volumes of both bytes and files) without a major disruption of the system. The repository should maintain a system that is evolvable. That is, the system should be designed in such a way that major components of the system can be replaced with newer technologies without major disruption of the system as a whole. The repository system should be extensible. That is, the system should be designed to accommodate future formats (media and files) without major disruption of the system as a whole. The repository should be able to export its holdings to a future custodian. The repository should be able to re-create the archives after an operational error that overwrites or deletes digital holdings.



Summary

- It is not only a technical problem
  - **Remember: People and systems!**
- You have to define which properties of your information has to be preserved
  - ex. only read access or full edit capability?
- You are never done
  - Backup is only a small part of it
  - You have to watch technological change
  - Digital Migration is key
- You need a sustainable environment
  - Money, staff, knowledge, awareness



Summary

THE GREAT THING ABOUT DIGITAL DATA IS THAT IT NEVER DEGRADES.
HARD DRIVES FAIL, OF COURSE, BUT THEIR BITS CAN BE COPIED FOREVER WITHOUT LOSS.
FILM DEGRADES, PAINT CRACKS, BUT A COPY OF A CENTURY-OLD DATA FILE IS IDENTICAL TO THE ORIGINAL.
IF HUMANITY HAS A PERMANENT RECORD, WE ARE THE FIRST GENERATION IN IT.
AMAZING
9GAG

"â€œIf you can read this, congratulationsâ€”the archive youâ€™re using still knows about the mouseover textâ€!"

Source: https://xkcd.com/1683/ by Randal Munroe
(Born: 1984/10/17, alive), VIAF: 103524663, License: CC-NC

As it is with XKCD, he knows a hell lot what about he is drawing...

This talk is not about backup strategies or if PNG is better than Jpeg. That is far to short sighted.

Q & A, Discussions

Christoph Zimmermann
nuess0r@pdproject.org