Digital Preservation Introduction

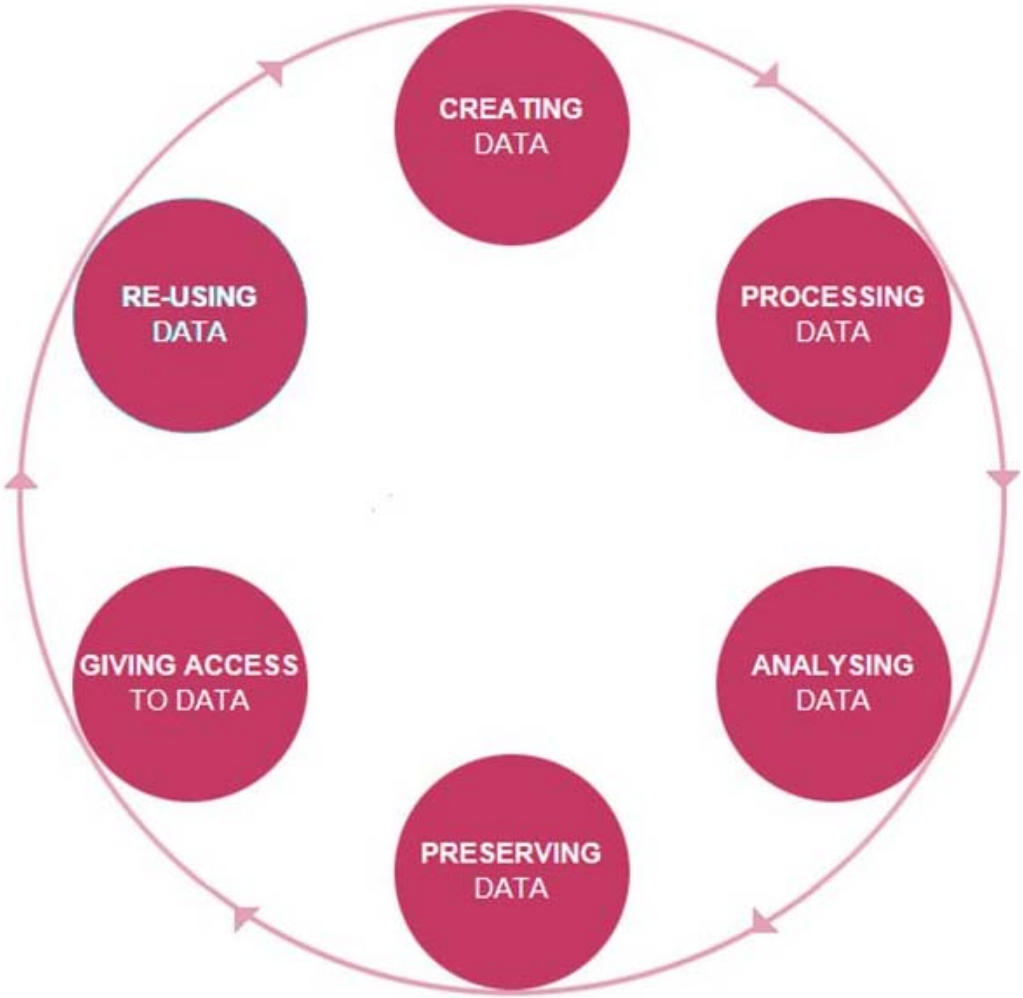
Andreas Rauber

Department of Software Technology and Interactive Systems

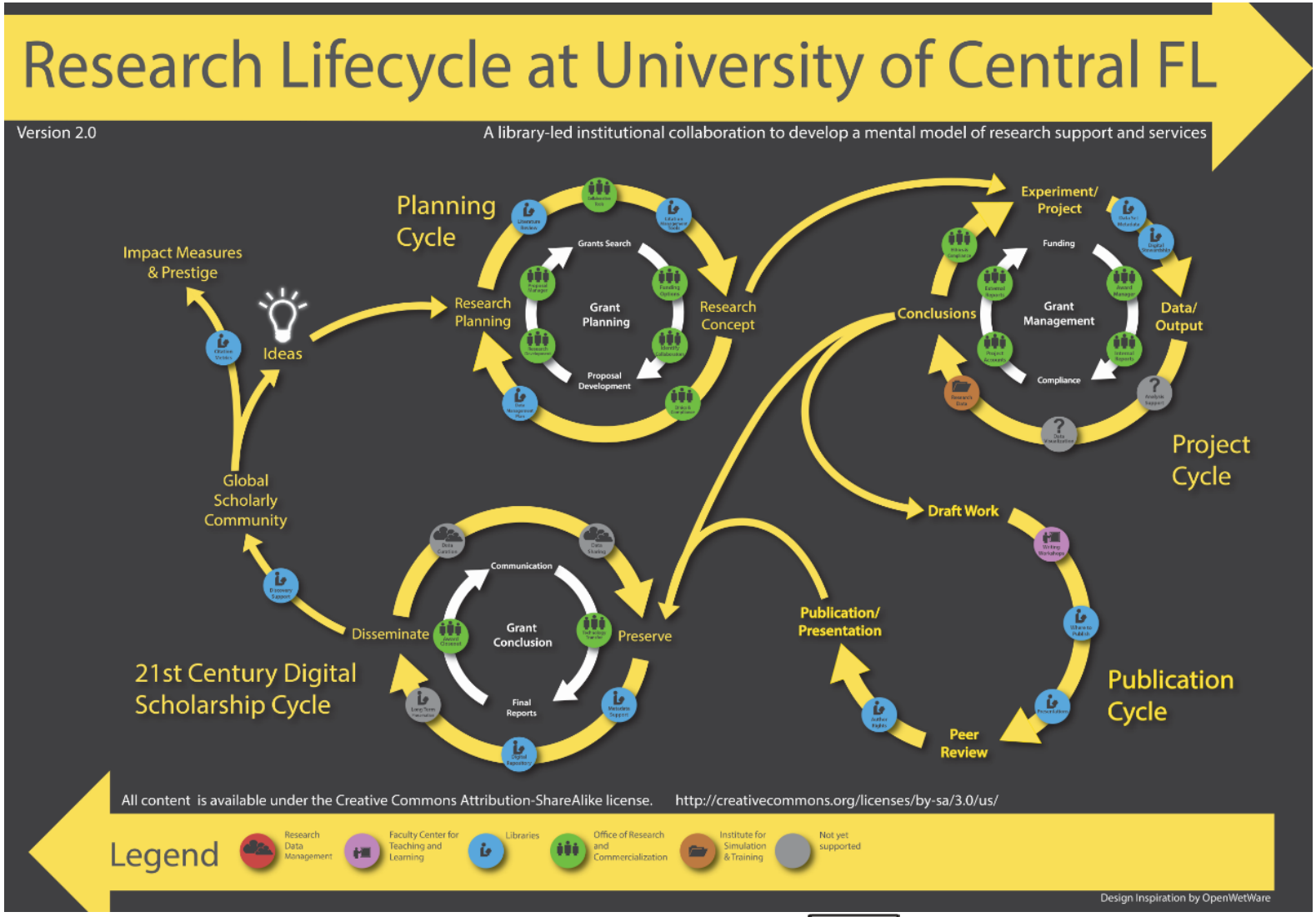
Vienna University of Technology

http://www.ifs.tuwien.ac.at/~andi

UK Data Archive Lifecycle model



University of Central Florida Lifecycle Model



http://guides.ucf.edu/ScholarlyCommunication/ResearchLifecycle

Overview

- What are the challenges in Digital Preservation?
- What methods can we use to counter them?
- What's the issue with file formats?
- What do we have in our repository?

Why do we need Digital Preservation?

Questions / discussion:

- What is *Digital Preservation?*

Why do we need Digital Preservation?

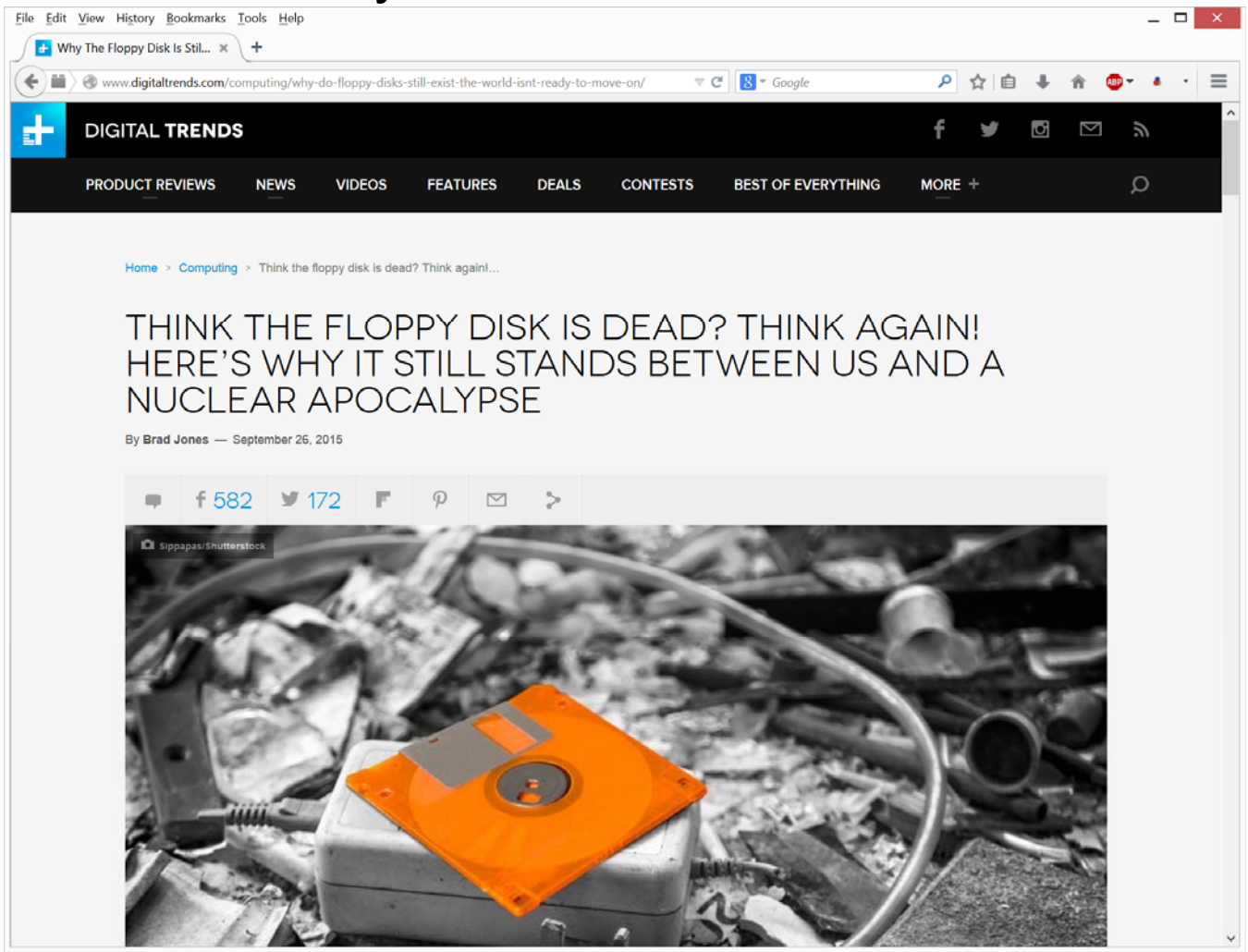
Why do we need Digital Preservation?

1. Physical Preservation (Bit-stream preservation)

- Transfering to current storage systems
  - note: transfer may not be trivial
    (file systems, encodings, relative references, copy protection,...)
- Ensure redundancy
  - technologically
  - geographic spread
- Access, security
- Error detection, recovery, disaster planning

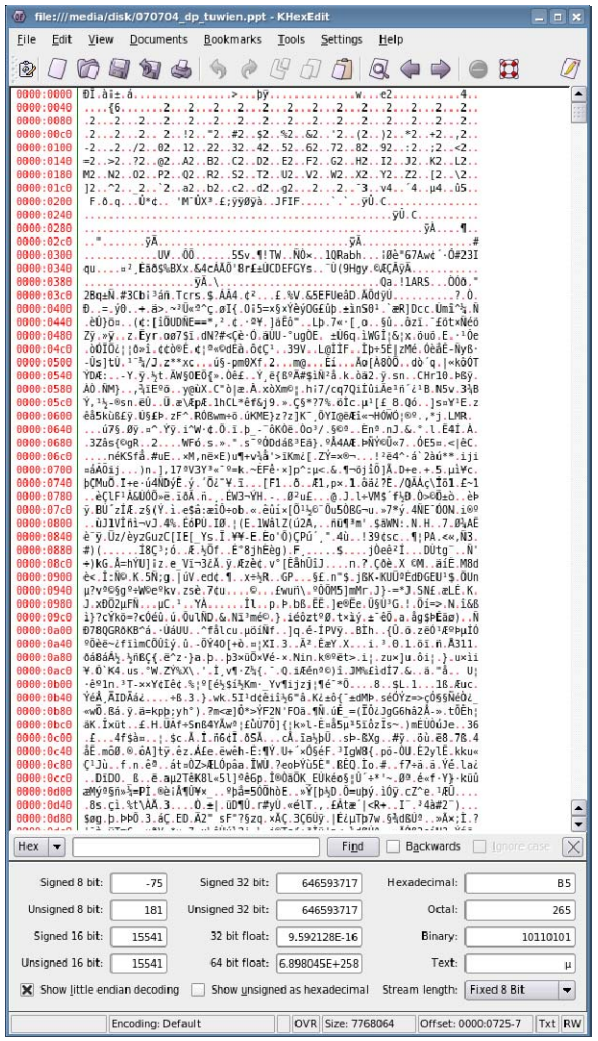
Why do we need Digital Preservation?

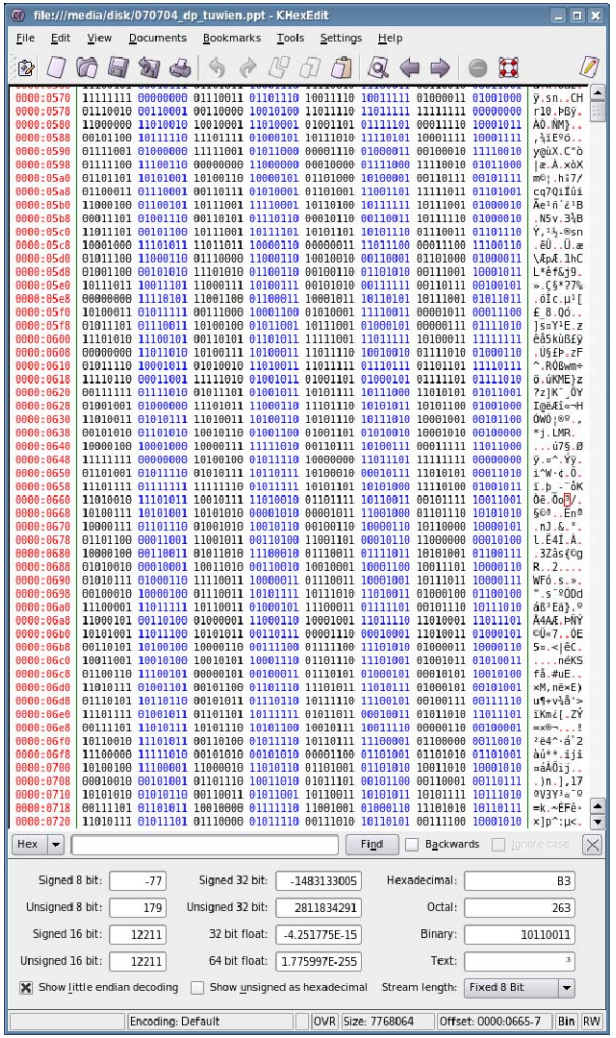
Just as a curiosity:

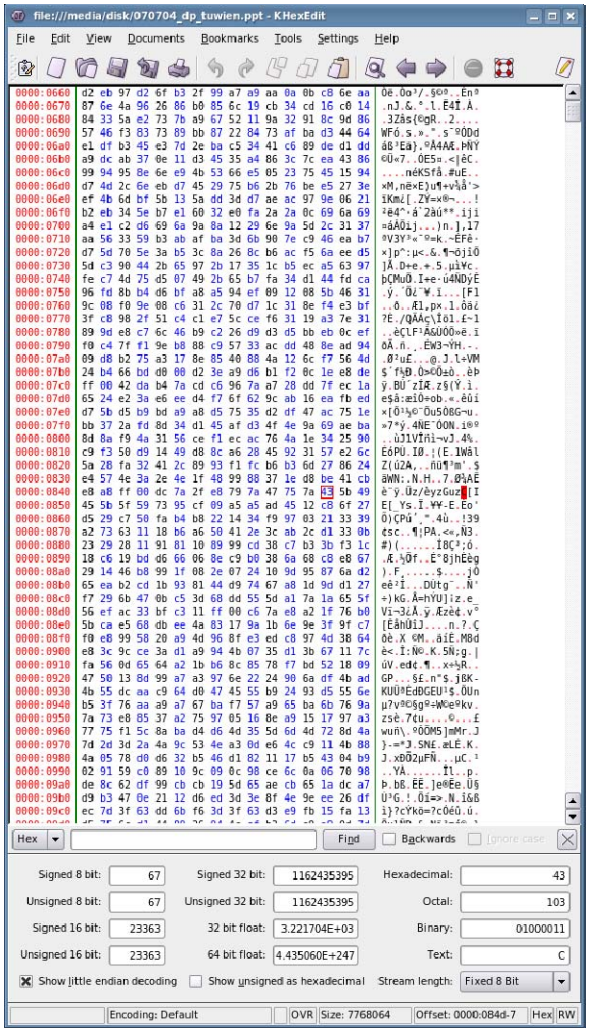


http://www.digitaltrends.com/computing/why-do-floppy-disks-still-exist-the-world-isnt-ready-to-move-on/

Why do we need Digital Preservation?







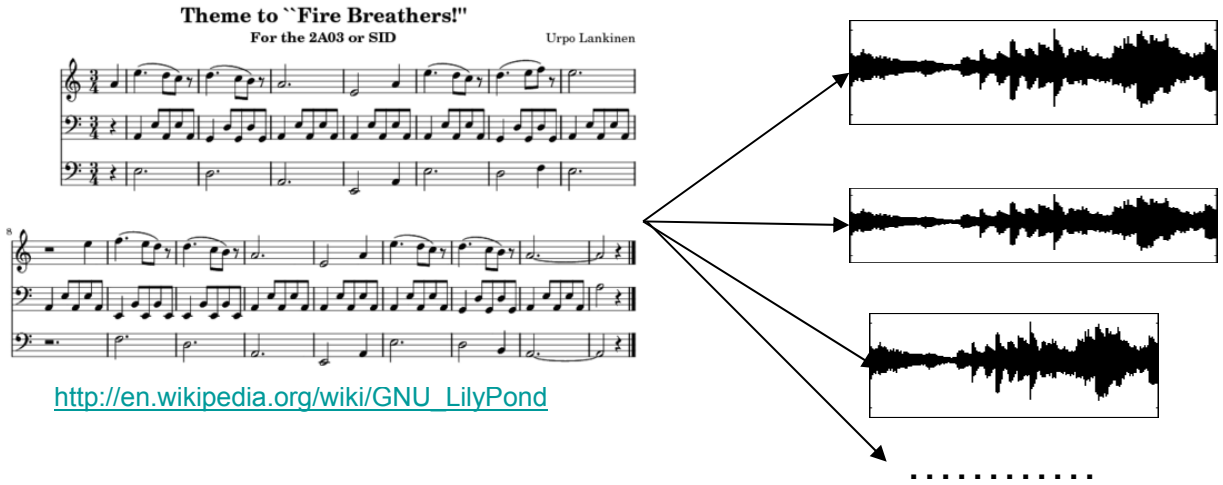
Why do we need Digital Preservation?

2. Logical Preservation

- Digital Objects require specific environment to be accessible :
  - Files need specific programs
  - Programs need specific operating systems (-versions)
  - Operating systems need specific hardware components
- SW/HW environment is not stable:
  - Files cannot be opened anymore
  - Embedded objects are no longer accessible/linked
  - Programs won‘t run
  - Information in digital form is lost
    (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

Strategies for Logical Preservation

Another way of viewing this…



http://en.wikipedia.org/wiki/GNU_LilyPond

.....

https://en.wikipedia.org/wiki/Odeon_of_Herodes_Atticus

https://commons.wikimedia.org/wiki/File:Stereoanlage_Vision_2000.jpg

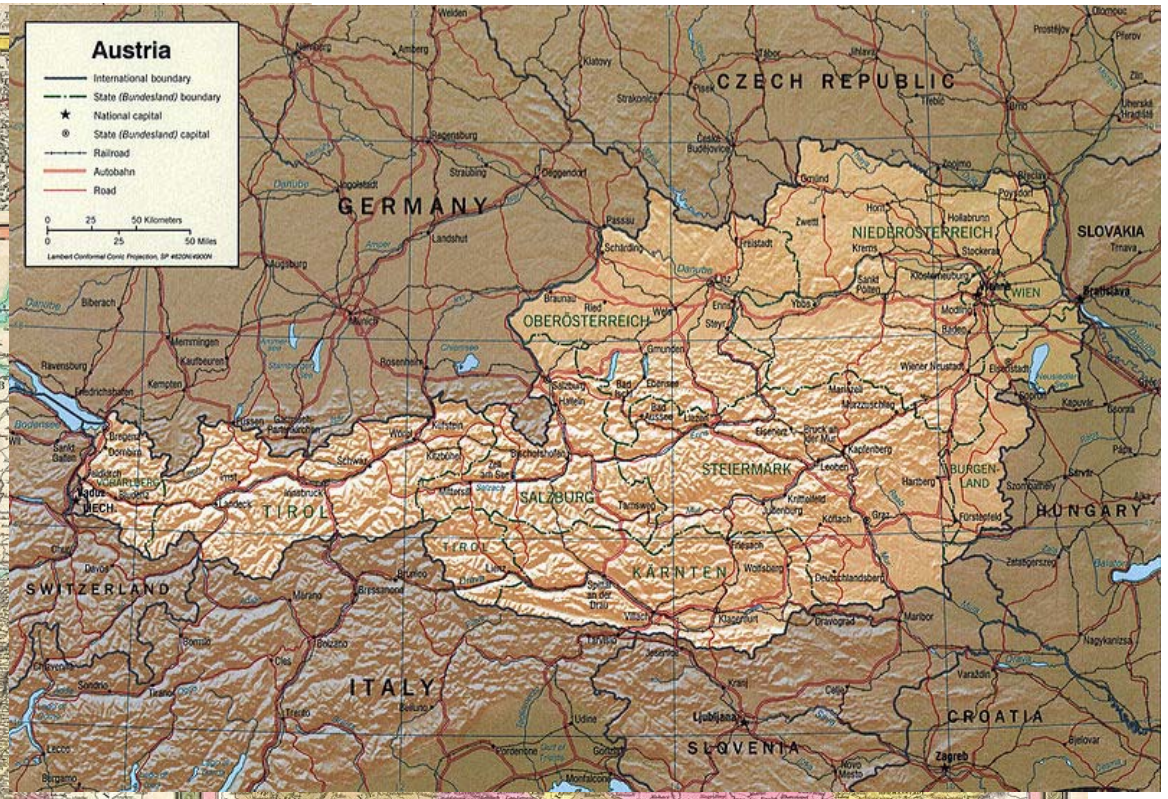
https://en.wikipedia.org/wiki/Konzerthaus,_Vienna

.....

Why do we need Digital Preservation?

Homann Heirs Ma
Wikimedia 1747

Cary Map of Au
Wikimedia 1801



Austria 1999 CIA map
Wikimedia, 1999

Mitchell Map of Austria, Hungary and Transylvania
Wikimedia 1850

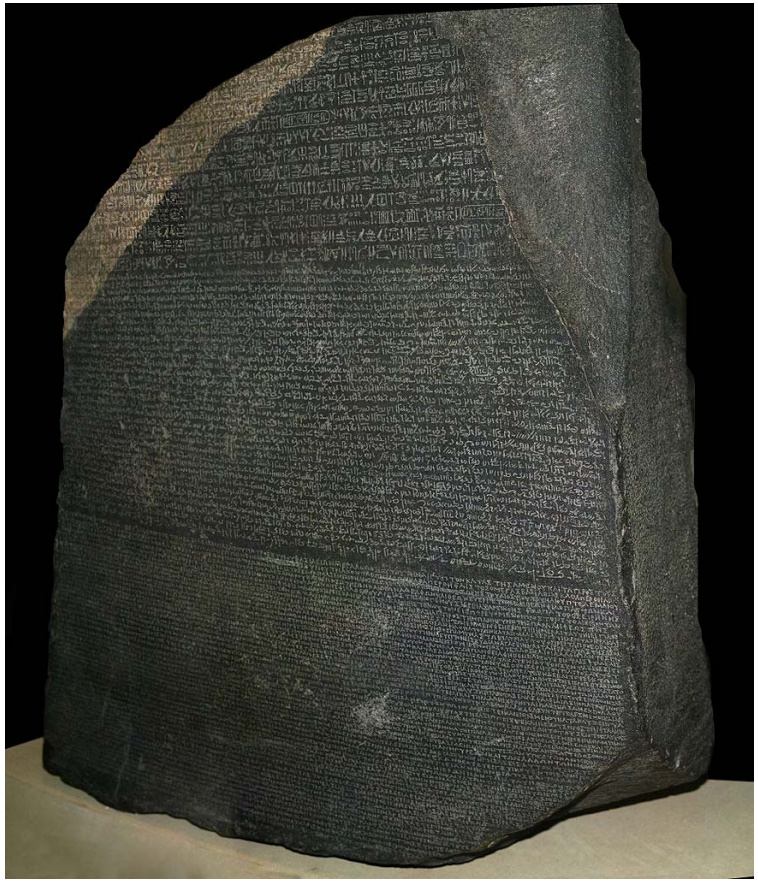
Why do we need Digital Preservation?

3. Semantic Layer: information object

- How to interpret the data (information?) in the objects?
  - terminology changes:
    changes in country names, borders, connotation of words,...
  - concept changes:
    drunk driving: before 1998: 0.8‰ , afterwards 0.5‰
  - transformations: currencies/exchange rates, sensor resolutions,
  - provenance: actions applied to objects
    sources: who? / which sensor?, transformations, post-processing
  - context of objects:
    understanding the context of decisions, side-effects, quotations, calibration timestamps
- For preserving digital information, all 3 layers need to be addressed

Why do we need Digital Preservation?

One of the most famous examples…



https://commons.wikimedia.org/wiki/File:Rosetta_Stone.JPG

Why do we need Digital Preservation

- The goal of Digital Preservation is to **maintain digital objects accessible and usable in an authentic manner for a long term** into the future.

Why do we need Digital Preservation?

Questions / discussion:

- What is *digital data*?
- What is *digital storage*?
- What do we mean by
  - *accessible*?
  - *authentic*?
  - *long-term*?

Why do we need Digital Preservation?

- Essential for all digital objects
  - Office documents, accounting, emails, …
  - Scientific datasets, sensor data, metadata, …
  - Applications, simulations, business processes, …
- All application domains
  - Cultural heritage data
  - eGovernment, public administration
  - Science / Research
  - Industry
  - Health, pharmaceutical industry
  - Aviation, control systems, construction, …
  - Private data
  - …

Why do we need Digital Preservation

3 levels of threat / preservation

1. Bit rot – physical preservation / bit preservation
   Physical Layer: how to keep the 0's and 1's
2. Object formats – logical preservation
   Logical Layer: how to remain able to open a file, run a program
3. Authenticity, interpretability – semantic preservation
   Semantic Layer: how to ensure we can understand/interpret data correctly

- What can we do?

Overview

- What are the challenges in Digital Preservation? (recap)
- What methods can we use to counter them?
- What's the issue with file formats?
- What do we have in our repository?

Bit-level preservation

- Maintain bit-sequence
- Redundant storage:
  - Lockss: lots of copies keeps stuff safe
  - Cloud
- Distributed storage – physically separated
- Different technologies / platforms / production batches
- Controlled storage conditions
- Regular maintenance: tape rewinding, disc spinning, …
- Maintain devices for accessing storage!
- Trade-off capacity, energy, effort

Bit-level preservation

Questions / discussion:

- How long do tapes / CDs / DVDs / HDDs / SSD last?
- What are the costs of bit-level preservation?
- What are the logistic challenges?
- Is a DVD that lasts for 200 years a solution?
- What would be the most durable storage technologies?
- What is "digital storage"?
- Distribution and Trust?
- Are we allowed to store redundantly? in the cloud?
  - Copyright
  - Copy protection
  - Distributed objects, referenced via URL? DOI?

Why do we need Digital Preservation

3 levels of threat / preservation

1. Bit rot – physical preservation / bit preservation
   Physical Layer: how to keep the 0's and 1's
2. Object formats – logical preservation
   Logical Layer: how to remain able to open a file, run a program
3. Authenticity, interpretability – semantic preservation
   Semantic Layer: how to ensure we can understand/interpret data correctly

- How? What can we do?

Strategies for Logical Preservation

Technology Museum

- ▪ Keep the hardware (drives, computer,…)
- \+ Maintains full functionality
- \+ Creates time buffer to develop more permanent strategies
- \+ Requires detailed documentation of HW and SW, but this also helps
- \+ Only strategy for some types of objects? (which?)
- \- Economically and technically infeasible to maintain spare parts forever
- \- Requires huge "museum"
- \- Requires highly specialized know-how for all platforms and software

Strategies for Logical Preservation

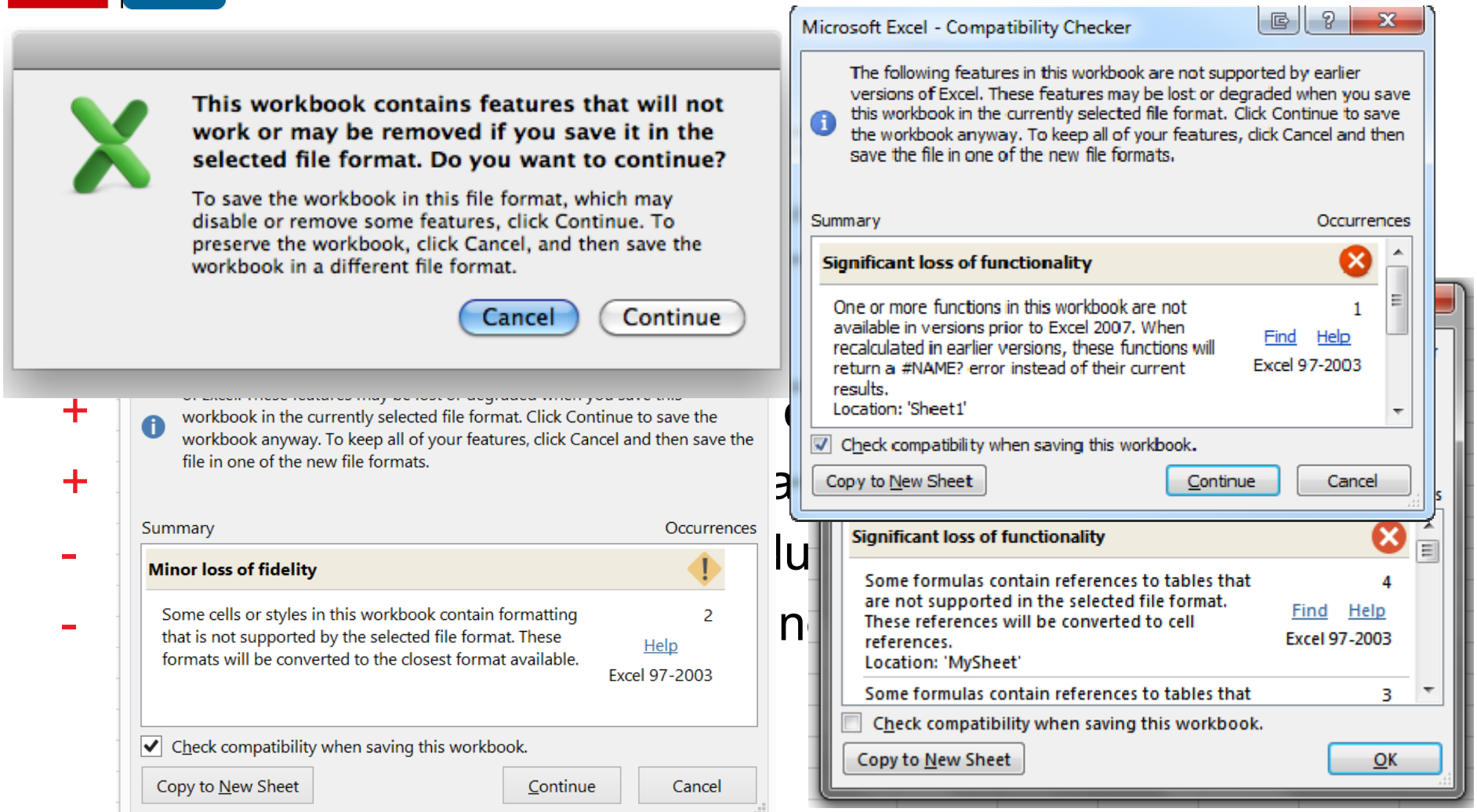
Migration

- Transform into different format
- Continually or on demand (Viewer)

+ Widely used
+ Possibility to compare at time of migration
+ Resulting objects are always accessible

- Possibly undesired changes during migration
- Needs to be repeated again and again

Strategies for Logical Preservation



https://support.office.com/en-us/article/Use-Office-Excel-2010-with-earlier-versions-of-Excel-2fd9ffcb-6fce-485b-85af-fecfd651a5ac

Strategies for Logical Preservation

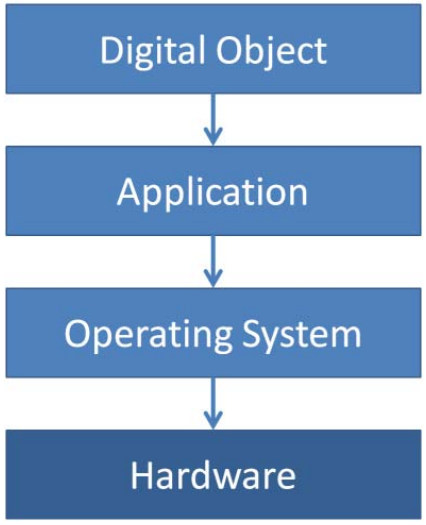
Emulation

- Emulation of Hardware or Software (OS, application)
+ Widely used principle
+ Many emulators available
+ Potentially preserving complete functionality
+ *Document is unchanged*
- *Document is unchanged*
- Complex technology, lot of research required
- Requires detailed documentation of the system
- Requires experience how to interact with emulated historic system in the future
- Emulators must be migrated as well
- Emulators potentially erroneous (Complexity)

Strategies for Logical Preservation

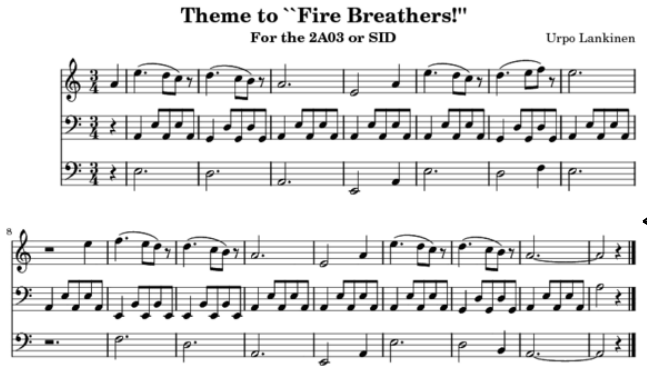
Excursion: Emulation vs. Migration

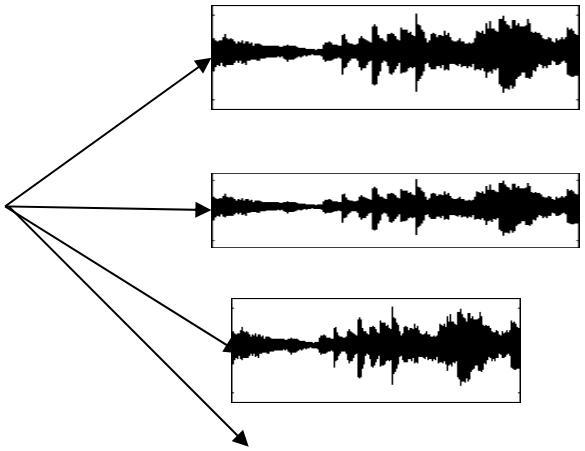
- Different on the pragmatic level, but conceptually identical
- Change occurs somewhere in the viewpath
- Have basically the same advantages/disadvantages and characteristics
- None of them guarantees identical rendering/performance of digital objects
- Many variants (e.g. viewer, virtualization)
- Need to be evaluated the same way



Strategies for Logical Preservation

Emulation vs. Migration – remember this?





http://en.wikipedia.org/wiki/GNU_LilyPond

.....

https://en.wikipedia.org/wiki/Konzerthaus,_Vienna

https://en.wikipedia.org/wiki/Odeon_of_Herodes_Atticus

https://commons.wikimedia.org/wiki/File:Stereoanlage_Vision_2000.jpg

.....

Strategies for Logical Preservation

Standardization

- Using open or de-facto standards
- \+ Simplifies DP process
- \+ Many tools available
- \+ Tools for standards are easier to build also in the future
- \- Significant effort required for standardization
- \- Loss at converting into standard
  (who is responsible?)
- \- Some object types cannot be standardized

Strategies for Logical Preservation

Standardization - Excursion into file formats
Proprietary vs. Open

- Proprietary
  - Documentation mostly not available
  - License and patent rules
  - License agreements subject to change
  - Restrictions for use and modifications may apply
- Open
  - Documentation available!
  - Unlimited use
  - No license fee
  - Open for modifications
  - No patent owners
- But: sometimes proprietary may better than open - **why**?
- Is the concept of "file formats" still useful?

Strategies for Logical Preservation

Limiting Accepted Formats

- Similar to standardization
- \+ Reduces challenge to smaller number of formats
- \- Does not solve the problem
- \- Limits the type of objects that can be accepted
- \- Potential loss at conversion
- \- Requires strict control of formats (and what's in them!)

Strategies for Logical Preservation

Data/Information Extraction

- Create abstract representation of information (e.g. databases or documents -> XML)
- \+ Independent of specific infrastructure
- \+ Many tools available
- \+ Easier to develop tools in the future
- \- High effort to develop tools for specific abstraction scenario
- \- Limited functionality of tools designed to interpret information, many aspects not preservable
- \- Cannot be applied to all types of objects

Strategies for Logical Preservation

Encapsulation

- Add metadata, software,... (representation information) to object („onion“)
- \+ Simplifies search for preservation solution on demand, offering several potential layers
- \+ Always allows for the application of several other strategies at different levels
- \- Does not solve the problem
- \- Even with all information encapsulated we may not be able to find a solution

Strategies for Logical Preservation

Universal Computing Platform

- Example: UVC: Universal Virtual Computer (IBM)
- Abstract virtual machine, intermediate platform that can be implemented on many other platforms

\+ Works for documents and software

\+ A kind of standardization for platform, reduces development effort

\+ Can test solution at time when being developed

\- Pretty complex (cf. Java, but that's still simple)

\- High effort at time of preservation

\- Requires cooperation of the producers of information

\- High risk of loosing aspects of information

Strategies for Logical Preservation

Backwards Compatibility and Version Migration

- current SW reads old versions and performs migration
- \+ Usually available
- \+ Creates time buffer for more permanent solutions
- \+ sometimes equal or better functionality
- \- Doubtful whether this will work for a long time (why?)
- \- Each change might lead to unwanted changes
- \- No guarantee from part of the producer of the SW

Strategies for Logical Preservation

Viewer

- Migration on demand, interpretation by Viewer software
+ Original datastream unchanged, interpreted directly
+ No continuous migration
+ No cummulative errors
- Viewer sometimes cannot process all (parts of) objects
- Time delay when developing viewers, increasing
- Viewer SW must be carried along with technology changes
- Hard to evaluate whether viewer is correct

Strategies for Logical Preservation

Non-digital Strategies

- Printing to paper, microfilm, ...
- \+ Reqires transformation to readable form -> stable
- \+ Coding of digital data is possible
- \+ Lots of experience in handling analog data carriers
- \+ High stability -> Bit-stream Preservation
- \- Loosing functionality, loosing advantage of digital technology
- \- Not applicable for all objects
- \- High costs for preserving some of the analog data carrier material, low storage density, ...
- \- Even this can be “buggy” (Xerox bug, manipulation)

Strategies for Logical Preservation

Data Recovery, Data Archeology

- Analysis of bit-stream to interpret data, digital forensics
- \+ Probably only approach to recover "lost" information
- \- No guarantee that it works
- \- Without sufficient documentation close to "guessing"
- \- Extremely high costs per object
- \- Hard to guess whether it may be successful for a given object

Logical Preservation

Summary

- Changing object, environment
- Loss upon migration / emulation
- Decision of what to preserve → **Significant Properties**!
- How to detect/document what you lost?
- Range of strategies available, none is perfect
- Combination of strategies
- No solution forever -> DP is a process!

Logical Preservation

- Preservation Planning
- Identify objects at risk
- Standardization reduces risk (why?)
- Apply preservation actions such as migration / emulation / HW-museum
- Identify what you need to preserve (significant properties)
- Identify suitability of tools
- Find out what you can preserve / what you loose
- Do it, document it, verify it, monitor it

Logical Preservation

Questions / Discussion:

- What are the problems of logical preservation?
- What is the optimal strategy?
- What is the optimal strategy for a specific object?
- What is a good format / platform (e.g. to migrate to)?
- What are characteristics of good formats/platforms/… ?
- How can we identify objects at risk?
- When is a format "more/less risky"?
- What is a file format?
- How can we find out what we loose with a strategy?

Logical Preservation

Questions / Discussion (2):

- What is the difference between emulation and migration? Are they different? Are they not different?
- What are the significant properties of an object / process?
- “I want to preserve everything” – (how) can we do this?
- What is the “original object”?
- What is the complexity of each strategy? Costs? Effort?
- What know-how do we need to decide on a strategy?
- What would be potential risks/difficulties e.g. for construction plans? Medical imaging (DICOM)?

Logical Preservation

Questions / discussion (3):

- Which objects are most at risk?
- Which objects are most difficult to preserve?
- How do we preserve entire business processes?
- If we loose significant properties with a strategy, what is the impact on authenticity? Can we use a “changed” object?
- What is the difference to systems engineering?

Why do we need Digital Preservation

3 levels of threat / preservation

1. Bit rot – physical preservation / bit preservation
   Physical Layer: how to keep the 0's and 1's
2. Object formats – logical preservation
   Logical Layer: how to remain able to open a file, run a program
3. Authenticity, interpretability – semantic preservation
   Semantic Layer: how to ensure we can understand/interpret data correctly

- How? What can we do?

Semantic preservation

- Threats at semantic level
  - meaning of terms change: city names, …
  - measurement scales, sensor sensitivity, …change
  - interpretation of facts change: alcohol levels, …
- Rather long-term, but subtle to notice
- Consider context of objects
  - purpose, setting, limitations, cultural context, related objects, …

Semantic preservation

- Approaches / solutions:
  - Semantic enrichment
  - Metadata
  - Migration at semantic level
  - Documentation of context
  - Tracing of metadata
  - Document intended meaning / interpretation

Semantic preservation

Questions / discussion:

- How do we identify need for action?
- What is the risk of missing timely action?
- How do we solidly identify and document context?
- How can we implement semantic enrichment / semantic migration, …?
- What about security issues?
- Is PDF save? PDF/A?
- Who is allowed to have access to which documents? Who had access to them?
- Are differences in the communication protocol at an API level a problem of logical or semantic preservation?

From Data to Processes

- Assume we know how to preserve data - **Is this sufficient?**
- Preserving data: Data Management Plans
  - describing data and context: provenance, authenticity, representation information,...
  - range of (ambiguous) definitions of context
  - But: mostly not actionable, not enforcable,...
  - BUT: data are (just) results of processes!
- Processes may be needed to
  - verify data
  - understand provenance
  - re-use process on new data
  - integrate data over time
- **Process curation instead of data curation!**

What can go wrong?

- A lot……
- Many times: trivial mistakes!
- But also more serious / conceptual issues
- From mistakes to actual fraud
- Overlap with security research, digital forensics, …
- Roles and Responsibilities, Policies, …

What can go wrong?

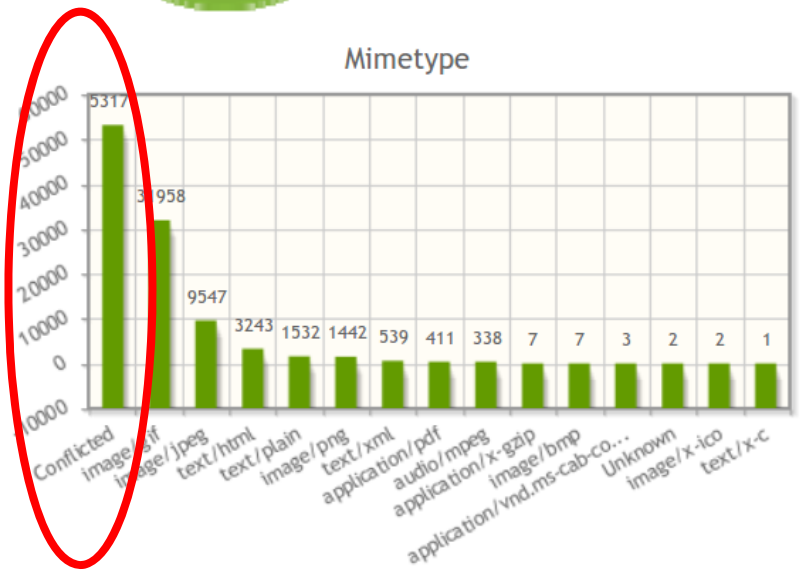
- Ingest / Standardization:
  - Who is performing the initial migration?
  - Who is liable?
  - Who will need to manage any problems subsequently?

- Migration
  - Something added? E.g. Word -> TXT, Excel -> TXT
  - Something lost?

- What is a PDF file?
  - A malicious invoice…
  - A multi-purpose paper: https://bigdata.uni-saarland.de/publications/p1972-dittrich.html

What can go wrong?

- Collection Profiling: what is in the repository?



JHOVE2

Apache Tika

ffident

Droid

TrID

METADATA XTRACTOR

Overview

- What are the challenges in Digital Preservation? (recap)
- What methods can we use to counter them?
- What's the issue with file formats?
- What do we have in our repository?

What File is This?

- „Clever software“
  - inspects files to decide how to process them
- Format registries

What Kind of File is This?

- By external characteristics (file extensions)
- By internal characteristics („magic number“, „signature“).

E.g. A TIFF file begins with …

- Bytes 0-1
  - The byte order used within the file. Legal values are: “II” (4949.H) / “MM” (4D4D.H)
- Bytes 2-3
  - An arbitrary but carefully chosen number (**42**) that further identifies the file as a TIFF file.

What Kind of File is This?

- What's wrong with file extensions?
  - Not necessarily unique (e.g. wks)
  - Granularity not sufficient
  - Can be altered by users
- Formats vs. Format profiles
  - PDF is not **one** format
  - DOC is not **one** format
  - TIFF is not **one** format
- A lot of things can go wrong: by coincidence or maliciously!
  - Word -> TXT, Excel -> TXT
  - Standardization: "We only use PDFs, we've no problem"
    - Not all PDFs are created equal…
    - A PDF file? https://bigdata.uni-saarland.de/publications/p1972-dittrich.html

What's Wrong with MIME Types?

- Insufficient depth of detail
  - No requirements regarding syntax and semantic description
  - No requirement for complete disclosure, especially of proprietary formats
- Insufficient granularity
  - Both tiled RGB GeoTIFF with LZW and striped bi-tonal TIFF-FX with Group 4 are typed as "image/tiff"
  - All of PDF 1.0 – 1.4, PDF/X-1, X-2, X-3, and PDF/A are typed as "application/pdf"
  - These variants might require radically different workflows

Why Do We Need a Format Registry?

- Knowledge base of file format representation information
  - properties,
    - what do they mean?
    - how to read them?
  - supporting software
- Unification of vocabulary (entity names and mappings)
- A (single?) access point to various information about formats though a common API

File Format Registries

- PRONOM
  - http://www.nationalarchives.gov.uk/pronom/
- Global Digital Format Registry (defunct)
  - http://library.harvard.edu/preservation/digital-preservation_gdfr.html
- Unified Digital Format Registry (UDFR) (defunct)
  - http://www.udfr.org/
- Sustainability of Digital Formats Planning for Library of Congress Collections
  - https://www.loc.gov/preservation/digital/formats/index.shtml
- FileExt
  - http://filext.com

PRONOM

You are here: Home > Services for professionals > Preservation > PRONOM > Search by format > Details: Summary

The technical registry
PRONOM

Welcome : About Add an entry
Search Help Information resources

Details: File format summary

Help : detailed report on file format

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles

Details for: Microsoft Word for Windows Document 97-2003

Save as... XML | CSV Print

Go to: Summary | Documentation > | Signatures > | Compression > | Character encoding > | Rights > | Reference files > Properties >

Summary

Name	Microsoft Word for Windows Document
Version	97-2003
Other names	Microsoft Word for Windows Document (97-XP)
Identifiers	MIME: application/msword Apple Uniform Type Identifier: com.microsoft.word.doc PUID: fmt/40
Family	
Classification	Text (Wordprocessed)
Disclosure	None
Description	With the release of Word 97, Microsoft revised the native binary word processing format, which is based on its generic OLE2 Compound Document Format. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily from OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Word document is stored as a 'WordDocument' stream within a Compound Document Format file. The format remained unchanged with the releases of Word 2000, 2002 and 2003.
Orientation	Binary
Byte order	Little-endian (Intel)
Related file formats	Has priority over OLE2 Compound Document Format Is subsequent version of Microsoft Word for Windows Document (6.0/95) Is subtype of OLE2 Compound Document Format

Registry Content

- Descriptive information
- Identifiers
  - MIME
  - Pronom Unique Identifier (PUID)
- Relationships to formats
- Technical environment
- References and links
- Risk factors

Registry Use Cases

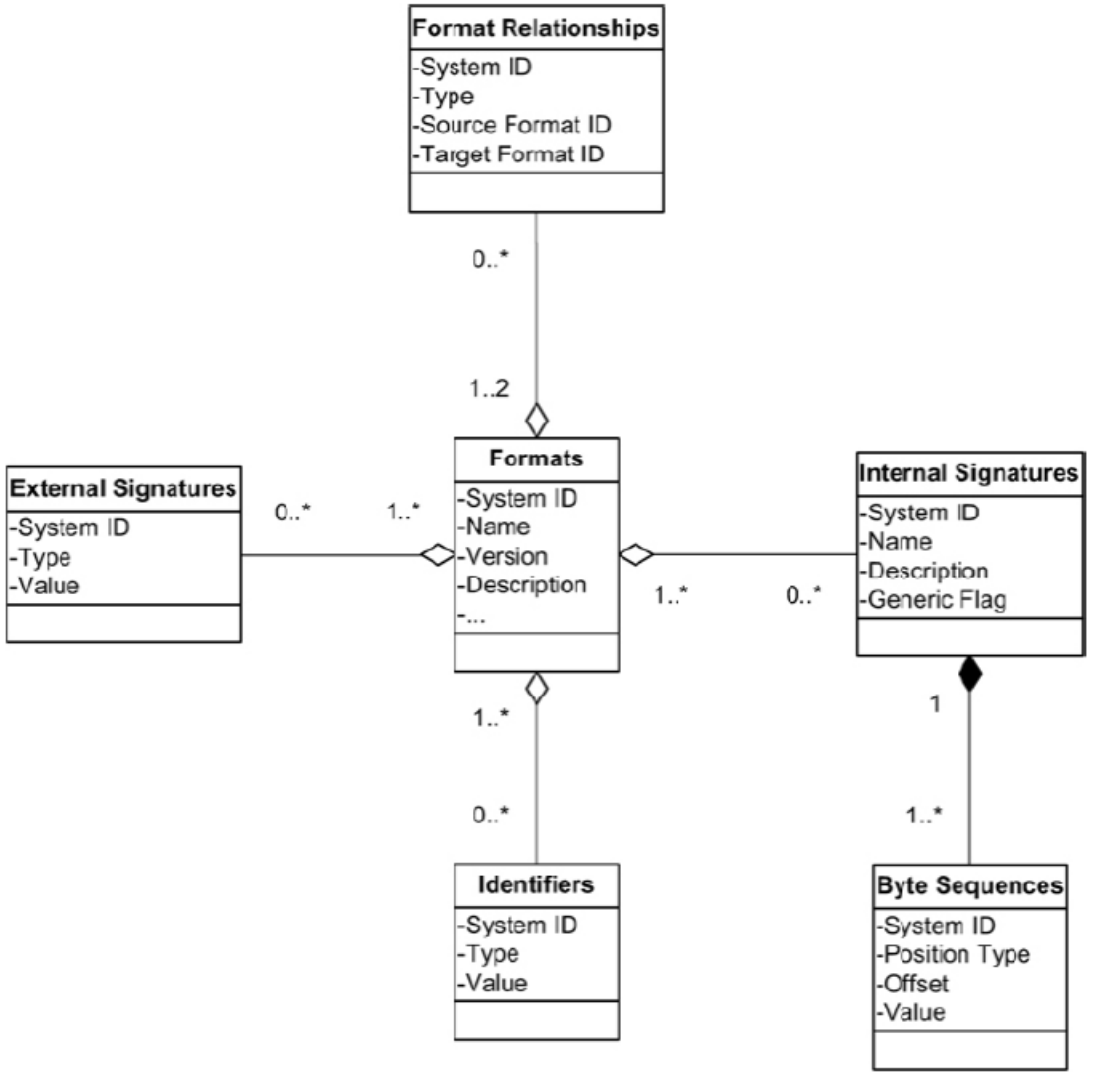
- Identification
  - "I have a digital object; what format is it?"
- Validation
  - "I have an object purportedly of format *F*; is it?"
- Transformation
  - "I have an object of format *F*, but need *G*; how can I produce it?"
- Characterization
  - "I have an object of format *F*; what are its significant properties?"
- Risk assessment
  - "I have an object of format *F*; is it at risk of obsolescence?"
- Delivery
  - "I have an object of format *F*; how can I render it?"

Identification Tools

- DROID (Digital Record Object Identification)
  - relies on PRONOM
  - The National Archives, UK
- JHOVE
  - JSTOR/Harvard Object Validation Environment
  - **Validation and characterisation**
- FITS (File Information Tool Set)
- veraPDF: PDF/A validation
  http://verapdf.org/home/

Signatures in DROID

- External signatures
  - -File extensions
- Internal signatures
  - -Format indicators in the bitstream
  - -Byte sequences



JHove

- JSTOR/Harvard Object Validation Environment
- Modular and extensible Java-based architecture
  - Image modules: GIF, JPEG, JPEG2000, TIFF
  - Document modules: ASCII,HTML,PDF, UTF-8, XML
  - ...
- Three functions
  - Identification
  - Validation
  - Characterisation
- JHove2
  - Identification and validation
  - Feature extraction
  - Policy based assessment
  - Able to handle complex objects

The TIFF Module…

- Tagged Image File Format (TIFF) raster images TIFF 4.0, 5.0, and 6.0 [TIFF 4.0, TIFF 5.0, TIFF 6.0]
- Baseline 6.0 Class B, G, P, and R [TIFF 6.0]
- Extension Class Y [TIFF 6.0]
- TIFF/IT (ISO 12639:2003) [TIFF/IT] File types CT, LW, HC, MP, BP, BL, and FP, and conformance levels P1 and P2
- TIFF/EP (ISO 12234-2:2001) [TIFF/EP]
- Exif 2.0, 2.1 (JEIDA-49-1998), and 2.2 (JEITA CP-3451) [Exif 2.1, Exif 2.2]
- GeoTIFF 1.0 [GeoTIFF]
- TIFF-FX (RFC 2301) [TIFF-FX]
- Profiles C, F, J, L, M, and S
- Class F (RFC 2306) [Class F, RFC 2306]
- RFC 1314 [RFC 1314]
- DNG (Adobe Digital Negative) [DNG]

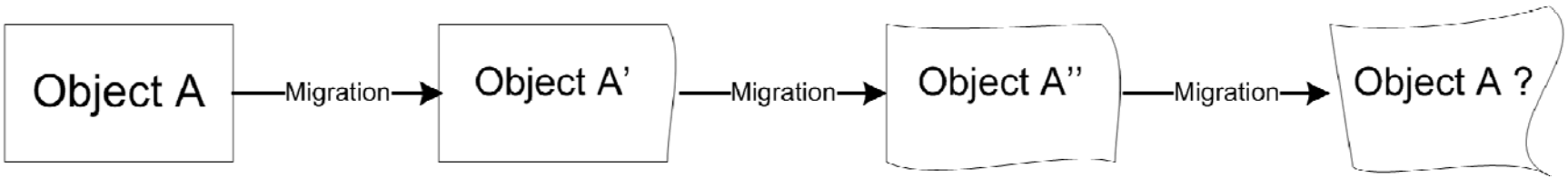
Validation

- A digital object is **well-formed** if it meets the purely syntactic requirements for its format.
- An object is **valid** if it is well-formed and it meets additional semantic-level requirements.

- Validation use cases:
  - "I have an object that purports to be of format *F*; is it?"
  - "I have an object of format *F*; does it meet profile *P* of *F*?"
  - "I have an object of format *F* and external metadata about *F* in schema *S*; are they consistent?"

Digital Preservation

- Core requirement: Keep object “intact” - authentically
- Essential object characteristics
  - Content
  - Appearance
  - Structure
  - Behaviour
  - Context

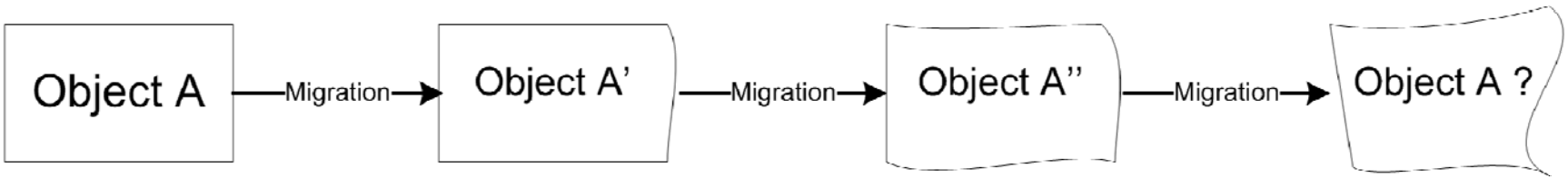


Validating a migrated image

- Yes, it's in JPEG 2000 format

- Yes, it's well-formed
- Yes, it's valid
- Yes, it still has the same dimensions
- …. But is it still the same image?

- We need more characterisation.



Characterisation

- Characteristics describing properties of the content in focus
- 3 aspects of characterisation:
  - Identification
  - Format validation
  - **Feature extraction**
- Depth of characterisation:
  - Narrow (e.g. Format profiles used in ROAR)
  - Deep, depends on context and tasks
    - **Significant properties**

JHOVE2

Apache Tika

ffident

Droid

TrID

METADATA XTRACTOR

FITS - File Information Tool Set

- Main features:
  - Consolidates output
  - Can include raw output
  - Configurable/Extendable
- FITS includes:
  - Droid
  - Metadata Extra
  - Jhove
  - Exiftool
  - FFident
  - File Utility

Conflicts

3 types of conflicts:

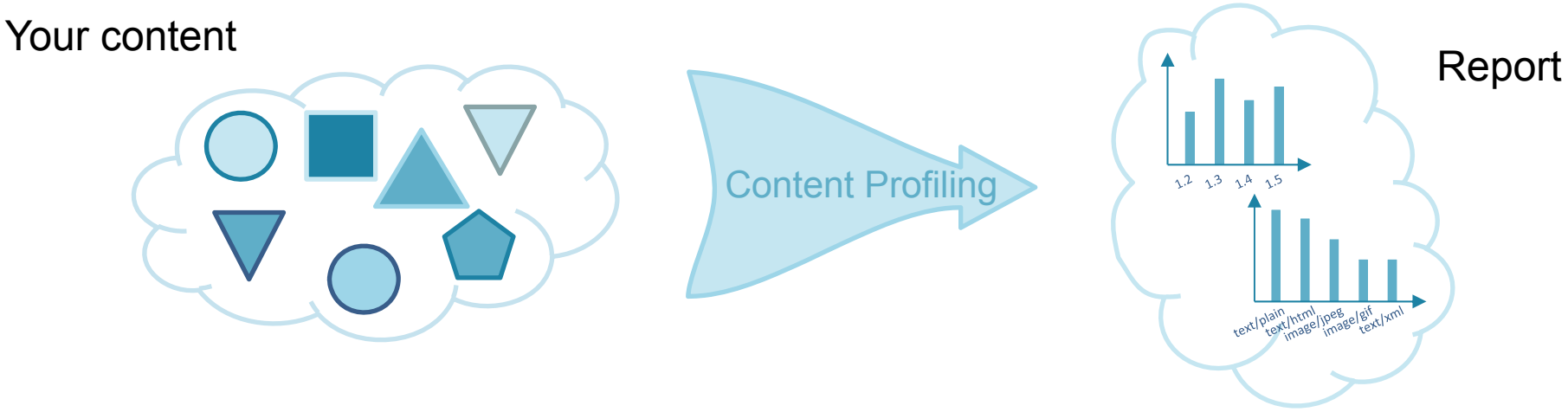
1. Inconsistent property naming, e.g: *image_width* and *imagewidth*
2. Competing characterisation results, e.g: tool1 identifies a file as *plain text*, but tool2 identifies the file as *PDF*
3. Close, but not the same property values, e.g: *application/xhtml+xml* vs. *application/xml*.

Overview

- What are the challenges in Digital Preservation? (recap)
- What methods can we use to counter them?
- What's the issue with file formats?
- What do we have in our repository?

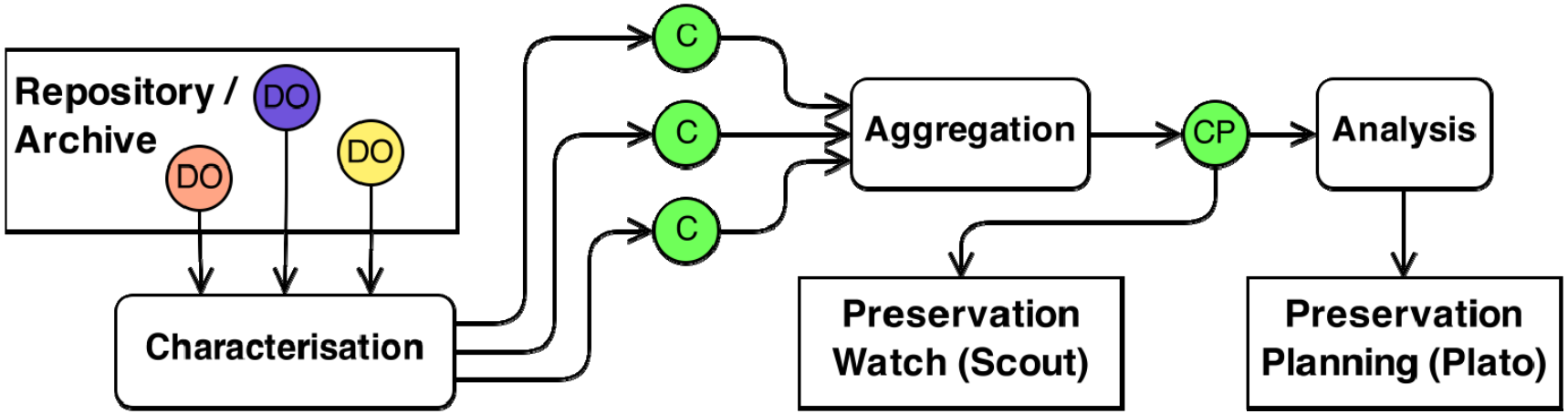
What is Content Profiling

- Better understand your content
- Reveal risks and opportunities
- Part of Preservation Planning
  - Planning and Watch, http://bit.ly/scape-suite



Content Profiling in Details

- A way of getting control over data
  - Decision support
- Consists of:
  - Characterization
  - Aggregation
  - Analysis
  - Reporting / Use for decision making



Aggregation

- Provides an overview of the content
  - Distributions of characteristics
  - Statistics (size, min, max, avg...)
- Data sizes grow dramatically
- Heterogeneity of data
- No universal characterisation tool
  - Combination of such tools

Analysis

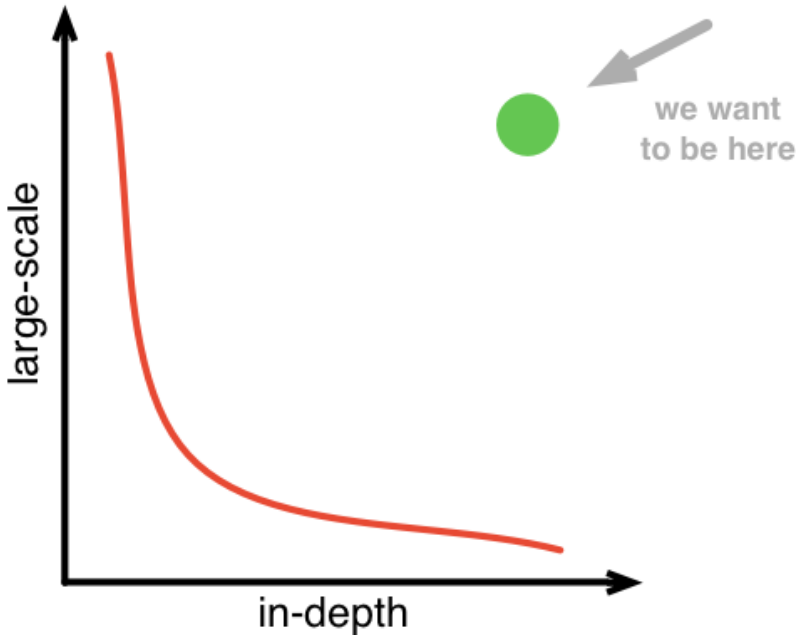
- In-depth research into your content
  - Drilling down
  - Filtering
  - SQL-like queries
- Representative samples generation
  - Based on metadata
  - Outlier detection
  - Stratification across
    - File type,
    - Size,
    - Time, or
    - Any other relevant property

Challenges

- Lack of:
  - Trustworthy tools for characterisation
- Depth
  - Address heterogeneity of data
  - Rise awareness of content properties
  - Combine several characterisation tools
- Quality
  - Conflicts due to combination of characterisation results
  - Resolve conflicting metadata

Challenges

- Scale
  - Effectively analyze substantial amount of metadata
  - Large-scale approaches for content profiling

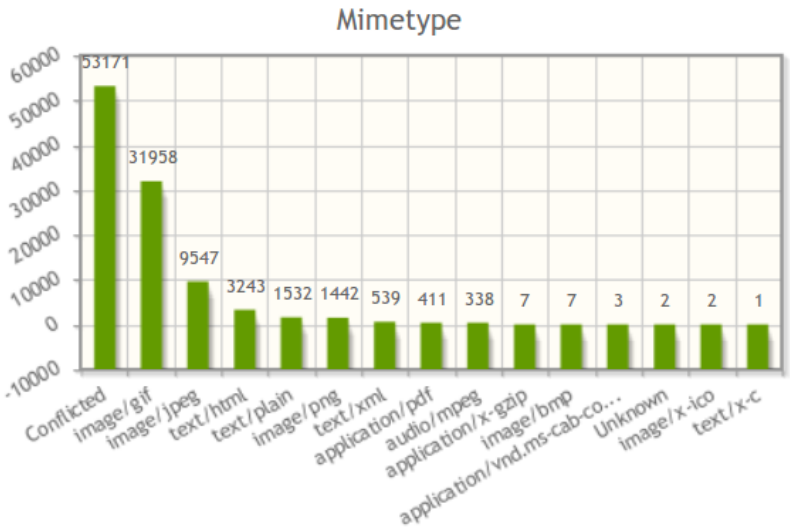


C3PO

- C3PO – Clever, Crafty, Content Profiling of Objects
- Reads and analyzes information from FITS
- Support large-scale database solutions for aggregation and analysis of characterisation metadata
  - MongoDB, http://www.mongodb.org
  - HBase, http://hbase.apache.org
- Aggregation-only mode
  - Useful to fast and explorative generation of a content profile
  - Statistics calculation using predefined filters
  - Single read of data, without computationally expensive ingest and further analysis

C3PO

- Uses characterisation results
- Interface to support other characterisation tools
- Deeper content analysis with interactive visuals through a web-app
- Representative sampling
- Open-source
  - http://ifs.tuwien.ac.at/imp/c3po



C3PO

- Stored metadata property mapping to the existing vocabulary, Planning and Watch Ontology
  - http://purl.org/DP/quality/measures

compression algorithm (Individual)

Definition

Definition

The **URI** of this individual is `http://purl.org/DP/quality/measures#118`

compression algorithm	http://www.w3.org/2004/02/skos/core#prefLabel	compression algorithm
compression algorithm	http://purl.org/DP/quality#scale	http://purl.org/DP/quality/scales#FREETEXT
compression algorithm	http://purl.org/DP/quality#attribute	http://purl.org/DP/quality/attributes#39

C3PO

- Rule-based engine to resolve conflicts in characterization metadata
  - Drools, http://www.jboss.org/drools
- Preservation-specific rules

Target tool	Execution condition	Action
Droid, Exiftool, all	Droid and Exiftool identify a file as “Microsoft Powerpoint Presentation”	Ignore format identifications by other tools
Jhove, all	Jhove reports “text/html” mimetype, other tools report “application/xhtml+xml"	Ignore the “text/html“ mimetype provided by Jhove

Overview

- What are the challenges in Digital Preservation? (recap)
- What methods can we use to counter them?
- What's the issue with file formats?
- What do we have in our repository?

Digital Preservation - Summary

- Is a complex task
- Requires a concise understanding of the objects, their intellectual characteristics, the way they were created and used and how they will most likely be used in the future
- Requires a continuous commitment to preserve objects to avoid the „digital dark hole“
- Requires a solid, trusted infrastructure and workflows to ensure digital objects are not lost
- Is essential to maintain electronic publications & data accessible
- Will become more complex as digital objects become more complex
- Needs to be defined in a preservation plan

Digital Preservation

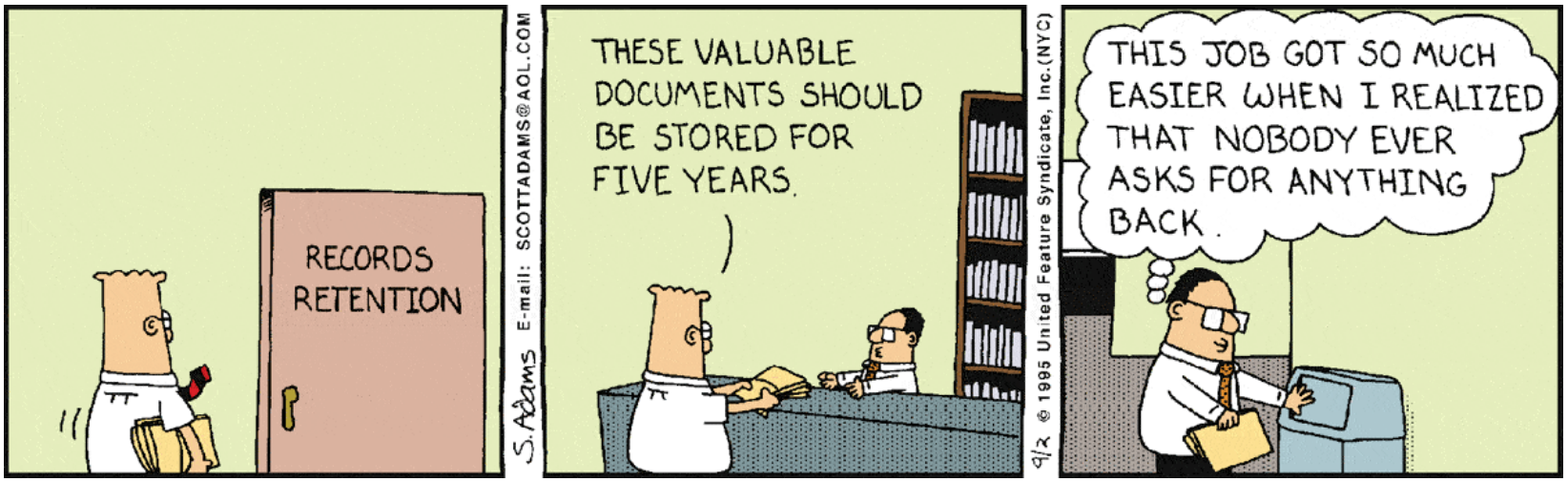
Questions / Discussion:

- At what levels are digital objects threatened?
- What are the time intervals at each level?
- How can we identify objects at risk?
- What can we do to mitigate the risk?
- How can we recover if mitigation fails / is missed?
- What competences do we need?
- How would a training/education program look like?
- How do we know if somebody is doing a good job at DP?

Current Issues

- Atomic file formats, stability of file formats
  - What are the atomic building blocks of information?
  - Can we split information objects?
  - Can we synthesize them? - Help for benchmarking?
- Scalability, Semantics
- Digital forgetting
  - how to decide what to keep and what to forget?
  - keep all? just storage? how to find? utilize? unterstand?
- Sustainable Systems Engineering
  - How can we build preservation-ready systems?
  - How to integrate DP-considerations into software engineering?
- Costs: what does DP cost?
  - cost factors?
  - How to model? evaluate?

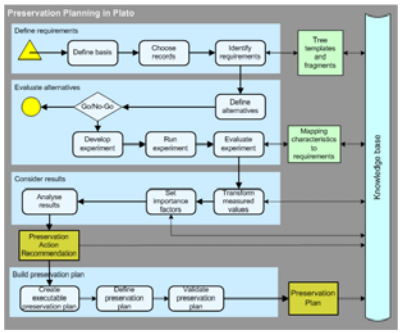
Thank you!

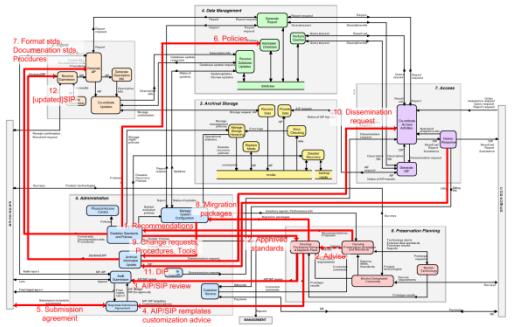


Source: http://dilbert.com/strip/1995-09-02

http://www.ifs.tuwien.ac.at/dp

Thank you!





http://www.ifs.tuwien.ac.at/dp