Digital Preservation: A systems perspective

goals and capabilities

March 11, 2013

Christoph Becker

Vienna University of Technology

www.ifs.tuwien.ac.at/~becker

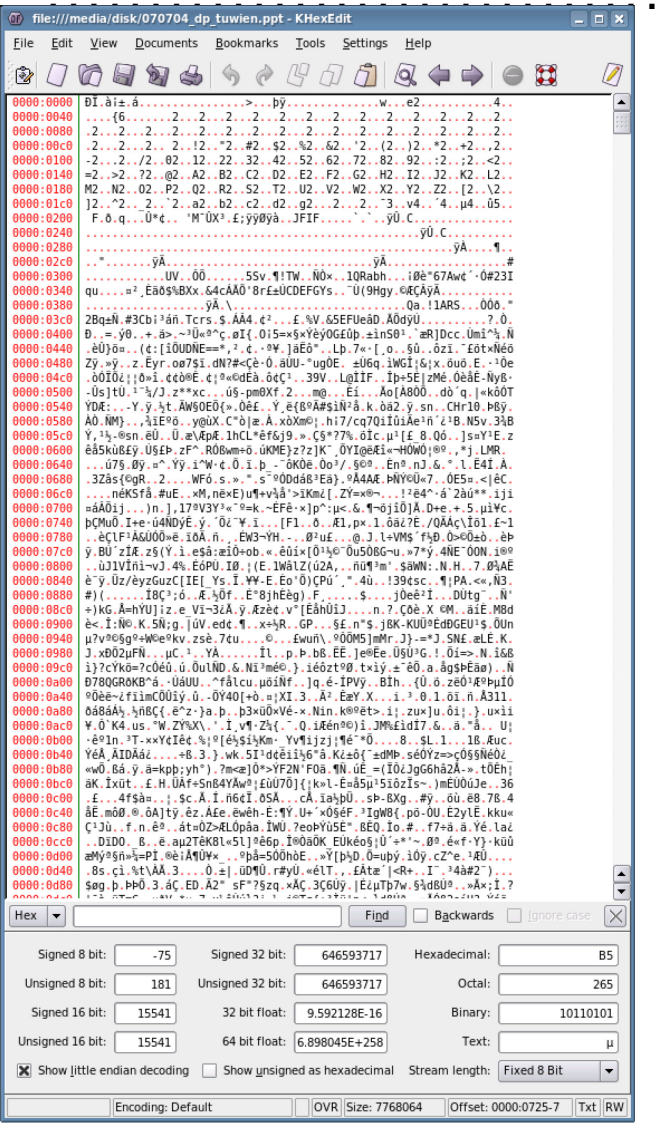
Why do we need Digital Preservation?

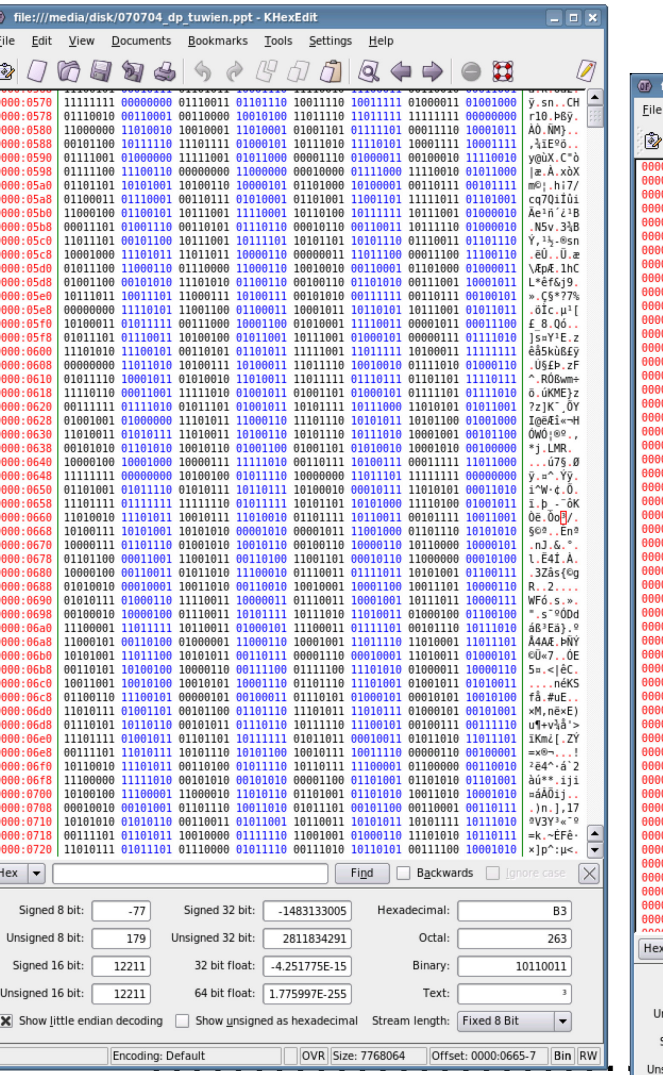
- .
  - .
  - .
  - .
- . .
  - .
  - .
  - rograms won‘t
  - .
- . .

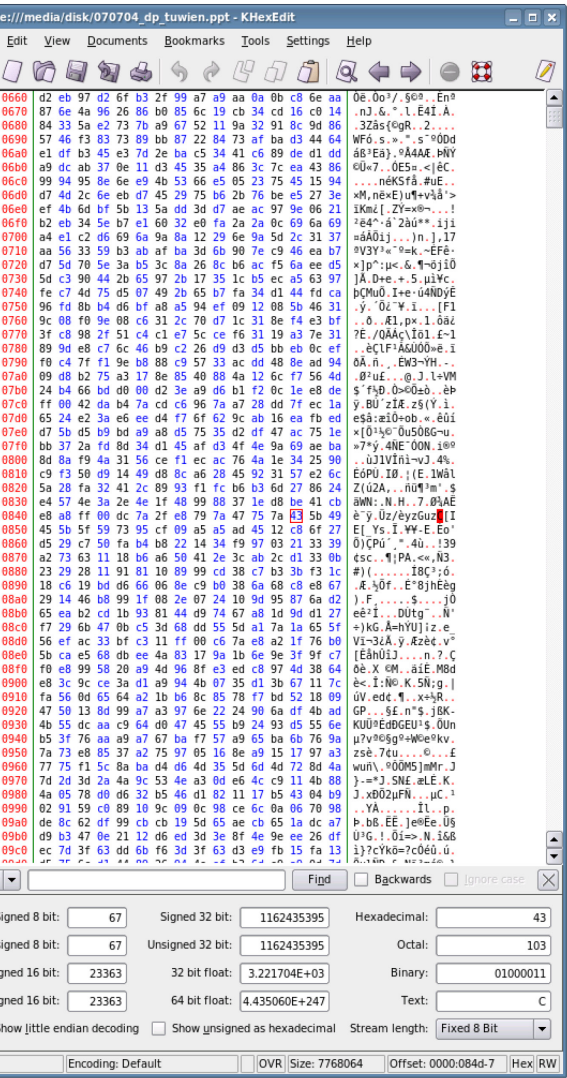
Why do we need Digital Preservation?

- 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.

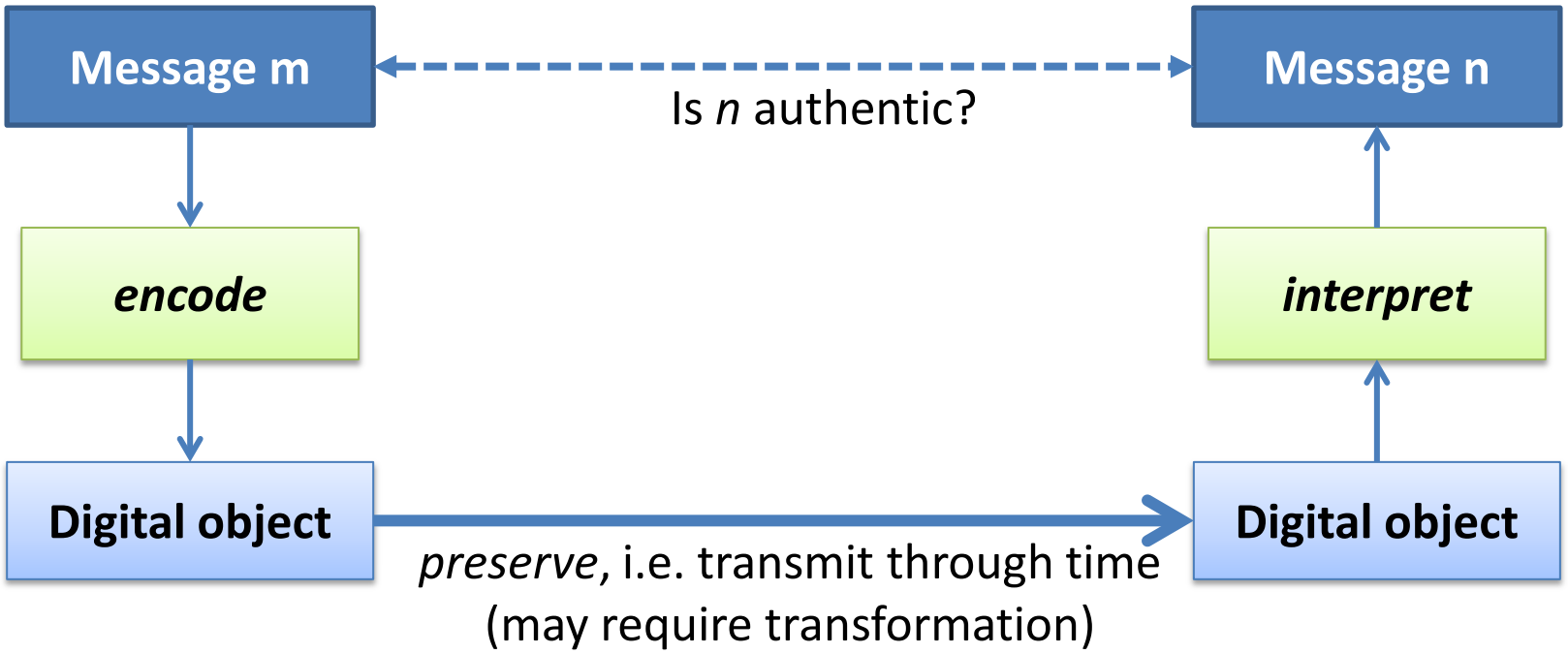
Why do we need Digital Preservation?







Digital preservation is communication.

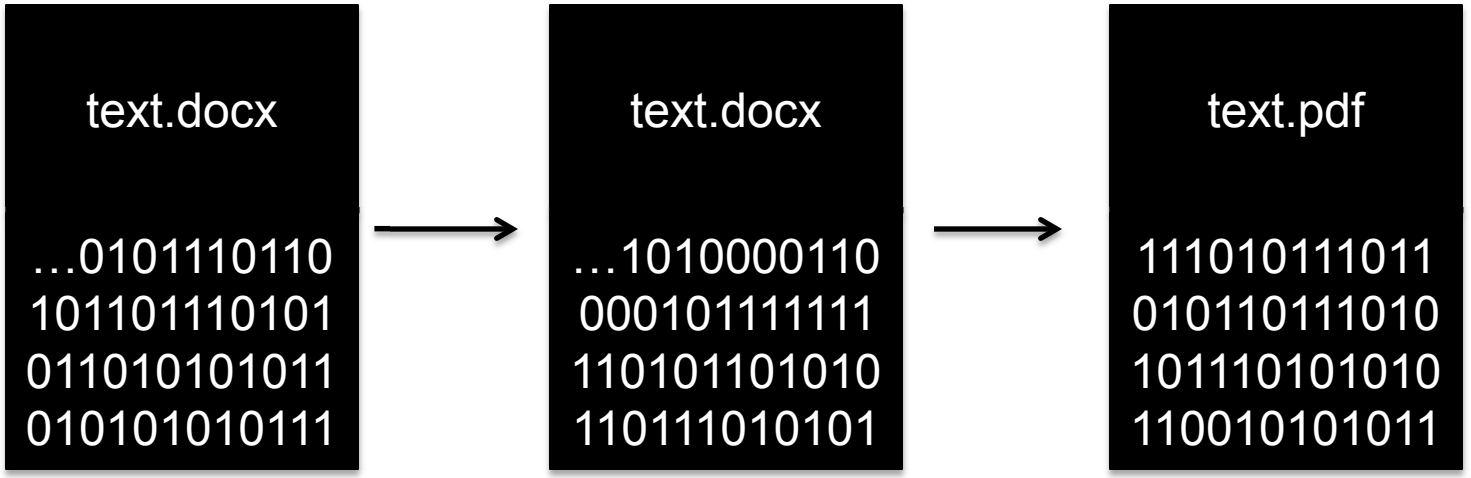


… But at the time of reception

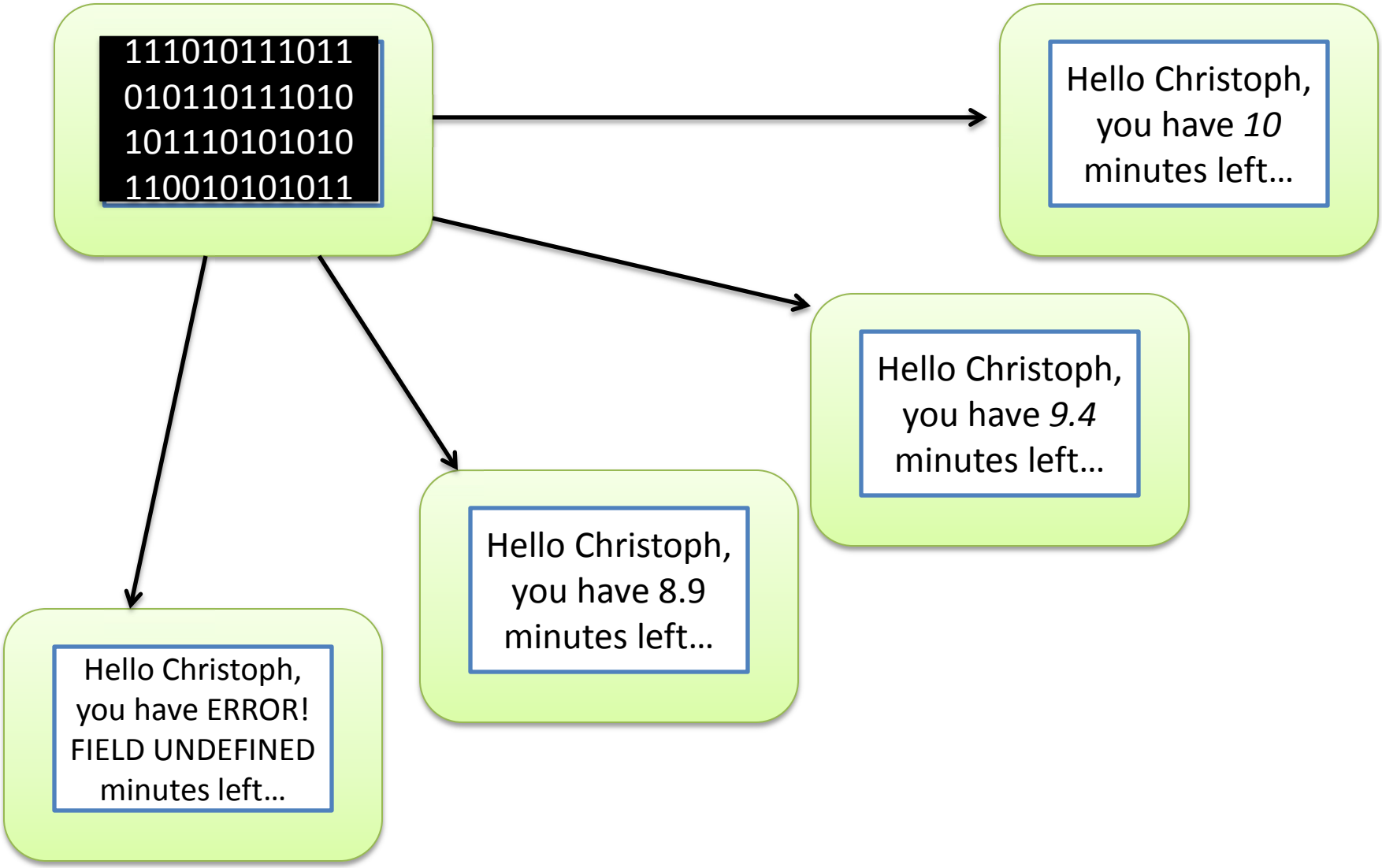
- there is no message *m* any more
- there may be no sender (any more)
- there may be no encoder to check against
- there may be no decoder
- the recipient may not be the original addressee

Digital content and understandability

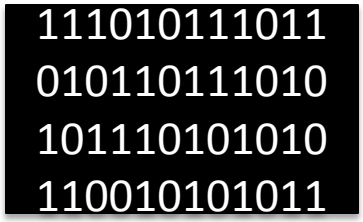
- Digital content is great, but…
- Content and environments
- ‘Documents cannot be edited’

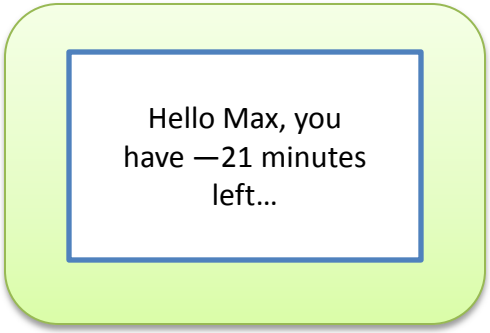


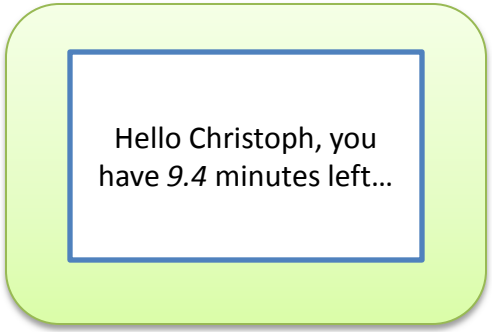
The black box problem

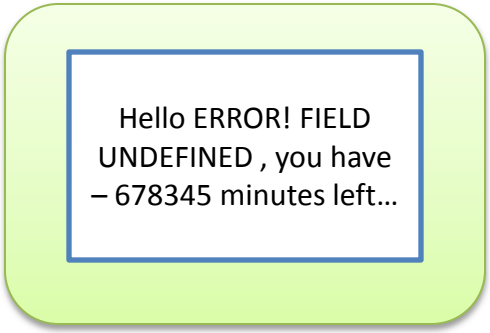


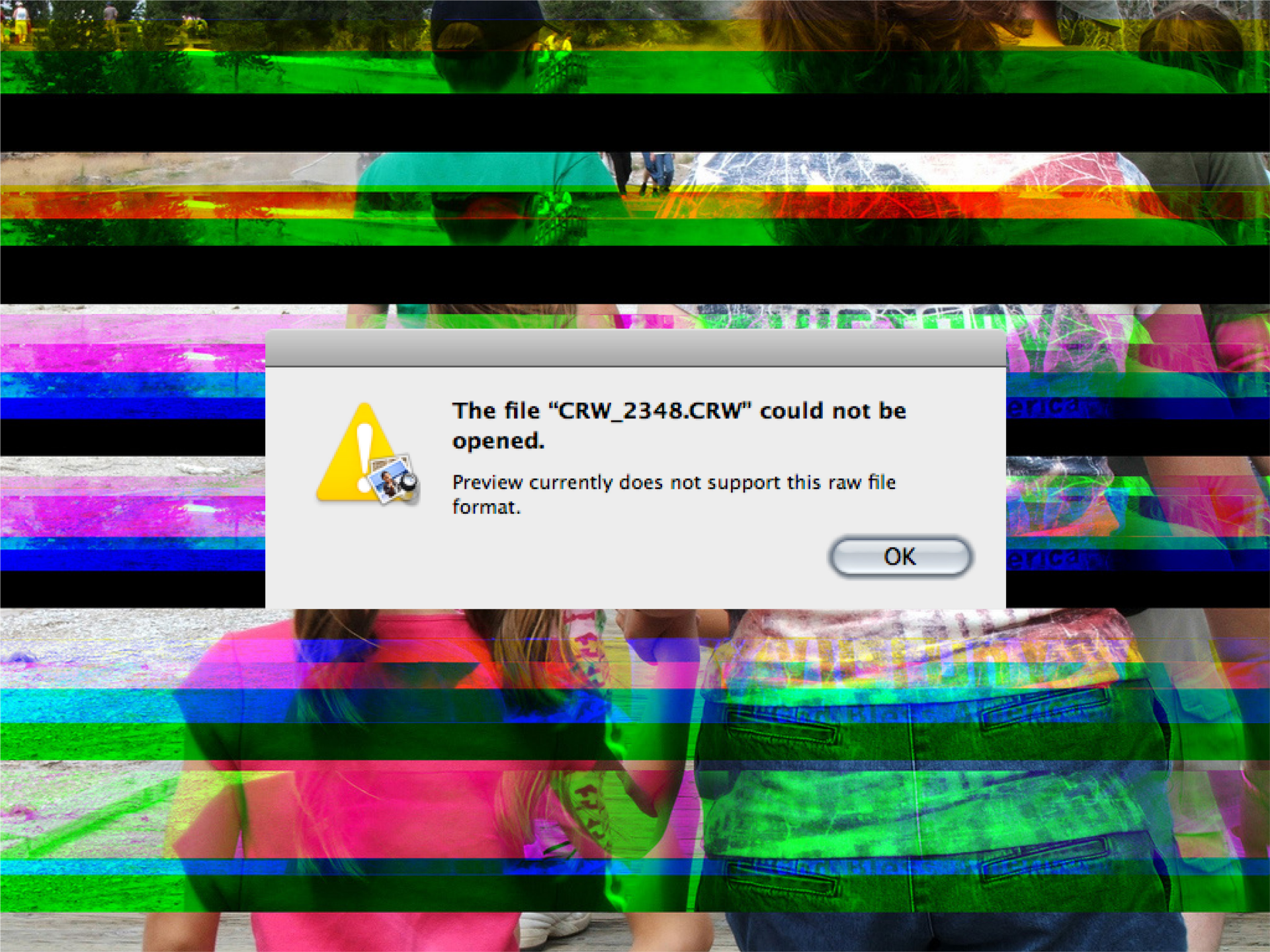
Five years later…



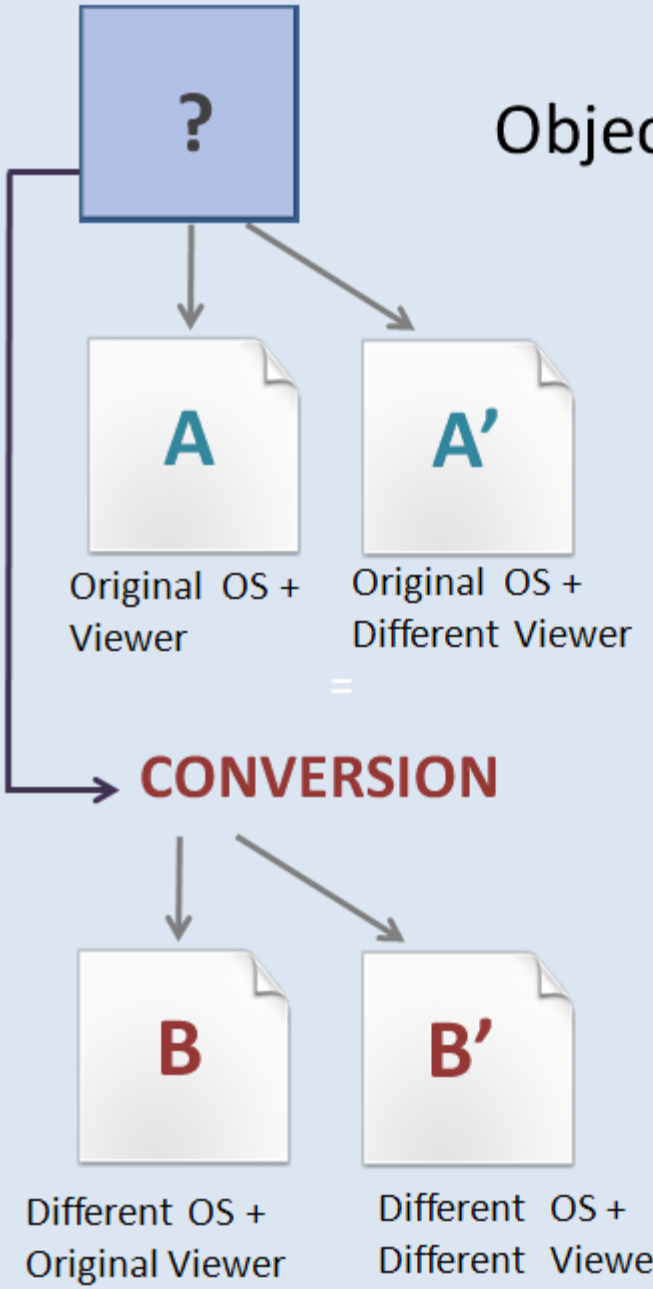


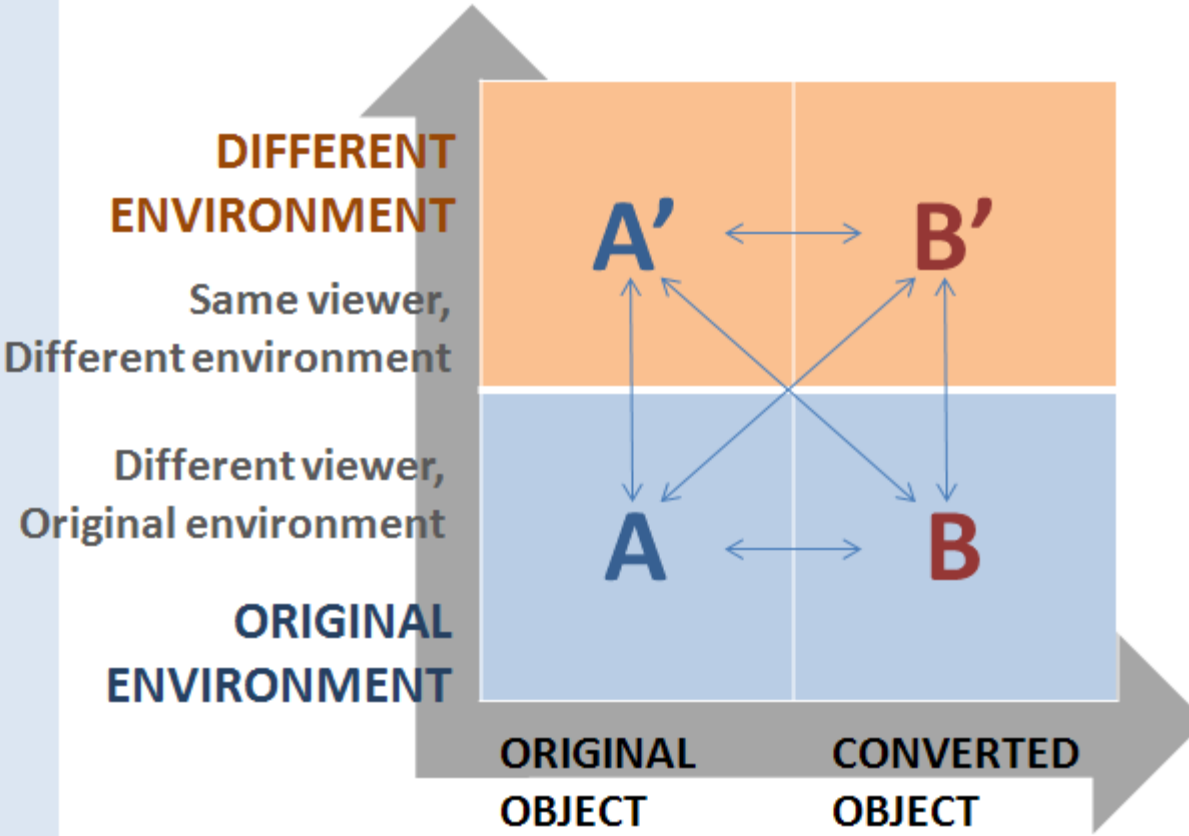





The file “CRW_2348.CRW" could not be opened.
Preview currently does not support this raw file format.
OK

Objects, environments and dependencies





...networks of objects

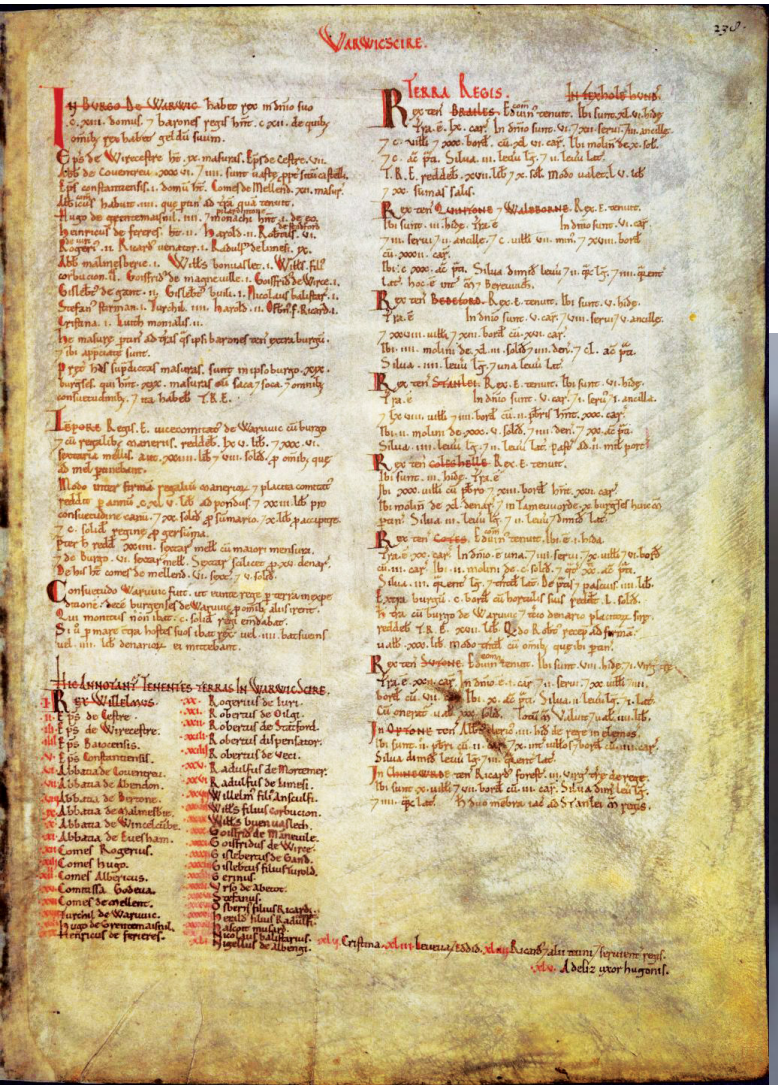
(Diagram by Natasa Milic-Frayling, MSRC)

Digital Longevity

- The mission of Digital Preservation is to keep content authentic and understandable for a user community over time
- Three levels
  - Physical
  - **Logical**
  - Semantic
- From Cultural heritage and space data systems to HEP, the web, business-critical information, and people
- **How long is long-term?**
  - From 7 years of taxes to 7000 years of nuclear waste documentation

The Domesday book

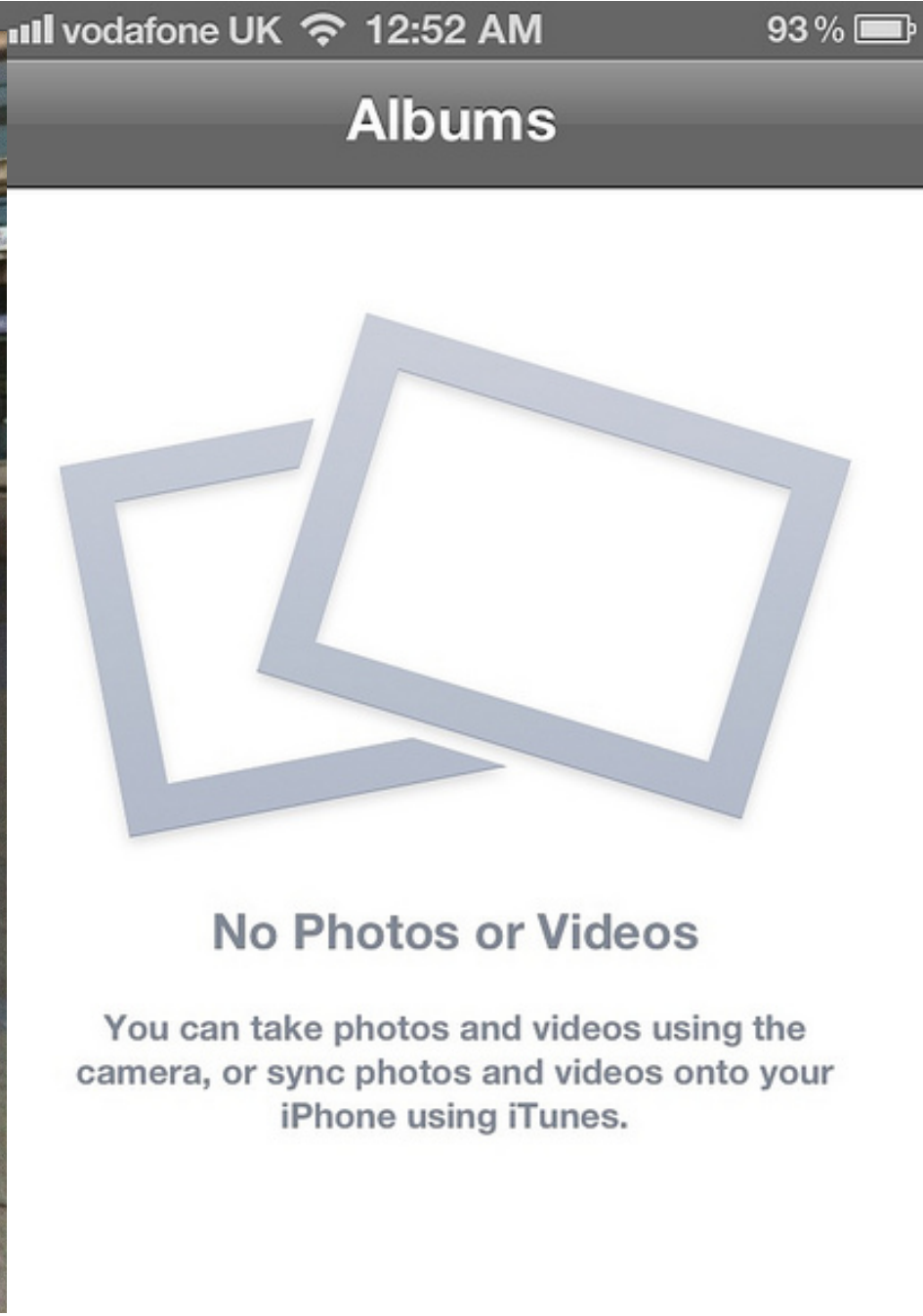
- http://www.ariadne.ac.uk/issue36/tna


WARWICSCIRE.

TERRA REGIS.

Tevatron (particle physics)

- 20 PB of data accumulated over 26 years
- *Both D0 and the CDF expect to lose their dedicated computing infrastructure over the next five years. A gradual loss of knowledge about how to deal with the complex data, which includes raw detector readouts, reconstructed particle trajectories and higher-level analyses, could also present a serious hurdle to exploiting the data in the future.*
- http://www.nature.com/news/2011/110527/full/474016a.html



http://www.flickr.com/groups/2121762@N23/pool/

Digital damages...

- When the National Archives received data in the mid-seventies from the Census Bureau, it was in a 1960's then state-of-the-art UNIVAC format. At the time "**there were only two UNIVAC computers left in the world**: one in Japan and the other housed in the Smithsonian Institute as a museum piece**. Heroic and costly rescue efforts recovered much, but not all, of the data**." [1]
  - *http://www.atlasofdigitaldamages.info/v1/stories/census-bureau-us/*

- … an attempt to obtain an article beginning on page 415 of a scientific journal revealed that the online version, available via Science Direct, only shows articles in that volume up to page 389. The response to a query to Science Direct was that at least 2% of its **electronic journal content is missing**. [1]

- *[1] Warner, Dorothy and Buschman, John. Studying the Reader/Researcher Without the Artifact: Digital Problems in the Future History of Books. Library Philosophy and Practice Vol. 7, No. 1 (Fall 2004) http://www.webpages.uidaho.edu/~mbolin/warner-buschman.htm*

A Digital Archive Disappears

- In 1991, … [an] ambitious digital preservation project sought to capture and archive the content of numerous nascent electronic journals. The project attracted considerable attention among authors and editors of e-journals… “to create a significant collection of electronic journals on the Internet, which scholars, libraries, and individuals around the world can access via the Web.” (info.lib.uh.edu/pr/v7/n4/mace7n4.html).
  Alas, essential funding never appeared, and CICNet itself ceased operations in 1997. The CICNet Journal Archive vanished with it. **Ironic, indeed, to lose not a mere collection but an archive whose purpose was to prevent loss of electronic content**. How many pioneering e-journals, many of them hosted on now-defunct Gopher servers, were lost for eternity? [2]

Lost Bush emails

- Under the Presidential Records Act, White House is legally obliged to keep copies of all communication, including emails
- In 2007, 22 million emails were declared missing
- With tremendous effort emails could be recovered
- Melanie Sloan (CREW executive director): We may never discover the full story of what happened here. It seems like they just didn't want the e-mails preserved."
- *[http://www.guardian.co.uk/technology/2009/dec/15/bush-emails-recovered]*

Digital Longevity

- The mission of Digital Preservation is to keep content authentic and understandable for a user community over time
- Levels of challenges
  - Physical
  - Logical
  - Semantic
- Dimensions
  - Organisational
  - Technological
  - Economic
- **Waiting for the outcome? Not an option…**

In the absence of that option…

- design principles
  - Standardisation
  - Openness
  - Transparency
  - Reversibility and traceability
  - Well-documented decision making
  - Governance, Risk and Compliance
- secondary indicators
  - Trust (what is trust?)
  - Proof of trustworthiness
  - Certification
  - Predictive measures of likely success
  - “Best practices”
  - Internal process metrics

Authenticity and understandability

- Understandability *for whom?*
- We need to understand the *Designated user community*
  - *What technology do they have?*
  - *What requirements for access do they have?*
  - *What do they want to do with the content?*
  - *What do they care about in the content?*
- Authenticity: An object
  - Is what it purports to be
  - Has created by the purported author
  - Has not been tampered with
  - Contains what it is supposed to contain...

- Are we talking about the bitstream here?

Authenticity and *significant properties*

- Significant properties are a way of expressing what is considered the essence of an object (a *performance)*
  http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf
  http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops
- This needs an understanding of the purpose (or the *intent*)
  - http://www.dlib.org/dlib/january13/webb/01webb.html
- If we understand what is essential about the content
  - we can devise measures to quantify that essence
  - we can systematically evaluate
- How to structure significant properties? What are the significant properties of this presentation? How to measure them?
  - Content
  - Context
  - Appearance
  - Structure
  - Behaviour

Digital preservation goals

- (for an archive, for example)
  1. **Acquire content** from producers in accordance to the mandate
  2. **Deliver** authentic, complete, usable and understandable **objects** to designated user community
  3. Faithfully **preserve provenance** of all objects
  4. Authentically **preserve objects** for the specified time horizon, securing their integrity and protecting them from threats
  5. **React to changes** in the environment timely in order to keep objects accessible and understandable
  6. Ensure repository **sustainability**: mandate, technical, financial, operational, communities
  7. Build **trust** in the depositors, the designated community and other stakeholders
  8. Maximize **efficiency** in all operations

DP requires certain capabilities

- A capability is an ability that an organization, person, or system possesses. Capabilities are typically expressed in general and high-level terms and typically require a combination of organization, people, processes, and technology to achieve.
  - Capabilities are not just tools
  - DP capabilities will require tools
  - Needs solid understanding of the socio-technical environment and its change processes

Key DP capabilities

- Keep bitstreams safe
- Provide access to content
- Monitor the outside world and the inside of the organisation
- Execute operations
  to keep content authentic, understandable, and accessible
- Plan those operations
- Manage the context: users, technologies, economy, legal…
- Governance, Risk and Compliance
- Infrastructure and support
- Sustainability

Key DP capabilities

- Keep bitstreams safe
- Provide access to content
- Monitor the outside world and the inside of the organisation
- **Execute operations**
  to keep content authentic, understandable, and accessible
- Plan those operations
- Manage the context: users, technologies, economy, legal…
- Governance, Risk and Compliance
- Infrastructure and support
- Sustainability

Preservation operations

- Analysis (“characterisation”)
  - Identification
  - Feature extraction
  - Validation
- Actions
  - Migration
  - Emulation
  - Others
- Quality Assurance
  - Authenticity
  - Significant properties
- Metadata
- Reporting

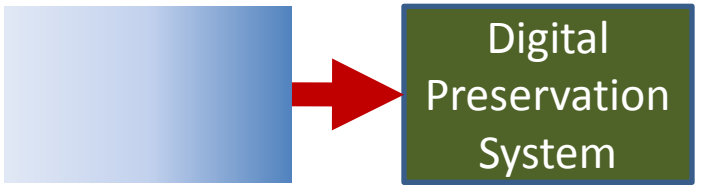
Capabilities and processes

- Capabilities are abilities to achieve goals
  - Delivered by processes
  - Supported by services and tools
  - Dependent on infrastructure
  - Managed by people
  - Measured by internal and external metrics
  - Assessed according to their maturity

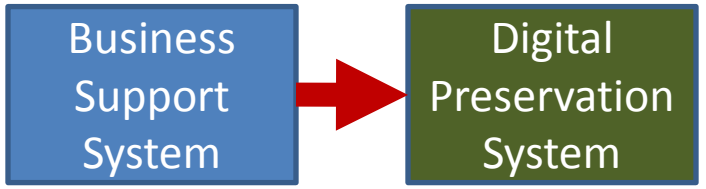
Information systems and preservation

Scenarios of systems
and their perceived relevance of digital preservation requirements

The Digital Preservation System:
DP as *functional requirements*



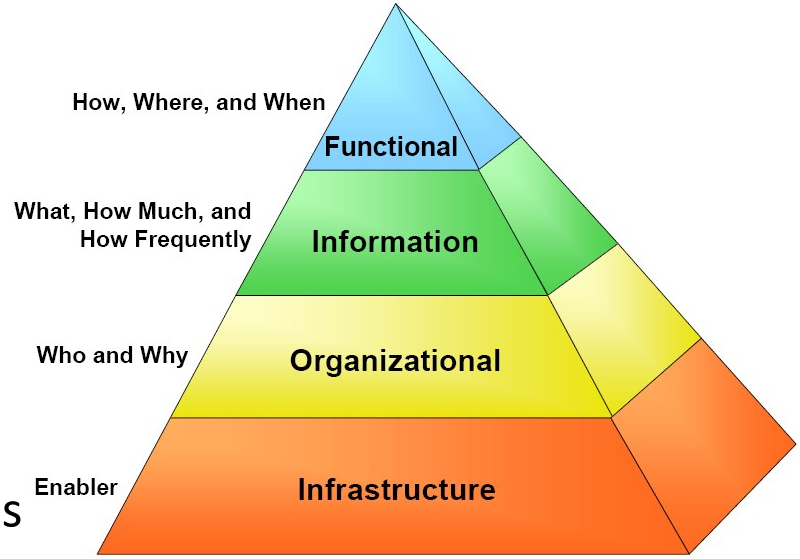
The Systems of Systems: Business system delegates DP responsibility to a DPS



The “Digital Preservation Capable” System:
Longevity as a *non-functional requirement*!

Enterprise Architecture

- Enterprises are systems
- Enterprise architecture (EA)
  - models the role of information systems and technology on organizations in a system architecture approach
  - in order to align enterprise-wide concepts, business processes and information with information systems
  - The core driver is planning for change and providing self-awareness to the organization



- The Zachman Framework
  - Very influential early EA approach
  - describes the elements of an enterprise's systems architecture
  - Each cell: a set of models, principles, services, standards
  - Rows: viewpoints
  - Columns: Focus

	DATA What	**FUNCTION** How	**NETWORK** Where	**PEOPLE** Who	**TIME** When	MOTIVATION Why
SCOPE (contextual)	List of things important in the business	List of business processes	List of business locations	List of important organizations	List of events	List of business goals and strategies
ENTERPRISE (business model)	Conceptual data/object model	Business process model	Business logistics system	Work flow model	Master schedule	Business plan
SYSTEM (logical model)	Logical data model	System architecture model	Distributed systems architecture	Human interface architecture	Processing structure	Business rule model
TECHNOLOGY (physical model)	Physical data/class model	Technology design model	Technology architecture	Presentation architecture	Control structure	Rule design
COMPONENTS (detailed)	Data definition	Program	Network architecture	Security architecture	Timing definition	Rule specification
INSTANCES (functioning enterprise)	Usable data	Working function	Usable network	Functioning organization	Implement ed schedule	Working strategy

Everything needs to fit together: Zachman

	DATA What	**FUNCTION** How	**NETWORK** Where	**PEOPLE** Who	**TIME** When	MOTIVATION Why
SCOPE (contextual)	List of things important in the business	List of business processes	List of business locations	List of important organizations	List of events	List of business goals and strategies
ENTERPRISE (business model)	Conceptual data/object model	Business process model	Business logistics system	Work flow model	Master schedule	Business plan
SYSTEM (logical model)	Logical data model	System architecture model	Distributed systems architecture	Human interface architecture	Processing structure	Business rule model
TECHNOLOGY (physical model)	Physical data/class model	Technology design model	Technology architecture	Presentation architecture	Control structure	Rule design
COMPONENTS (detailed)	Data definition	Program	Network architecture	Security architecture	Timing definition	Rule specification
INSTANCES (functioning enterprise)	Usable data	Working function	Usable network	Functioning organization	Implement ed schedule	Working strategy

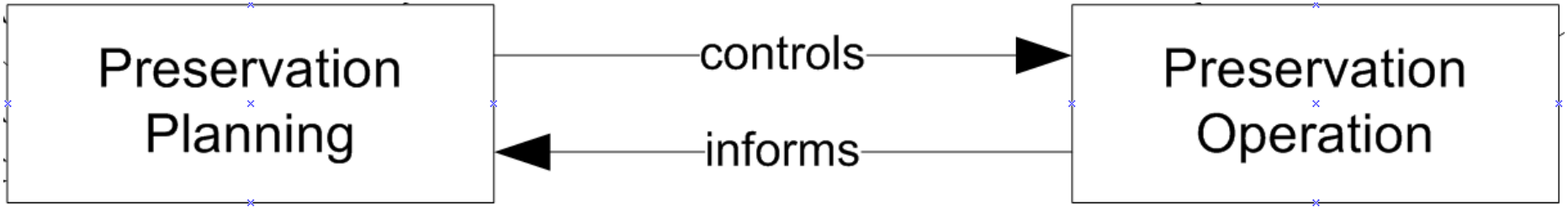
IT Governance

- Digital Preservation vs. Information Management
  - Information Management needs IT
  - Information Technology needs controls
- What is IT Governance?
  - Expectations, goals, responsibility, performance, control
  - Systems, Organizations and Goals
  - Strategic alignment of business and technology
  - Enterprise Architecture
- Transparency, compliance, trust: Governance

IT Governance and COBIT

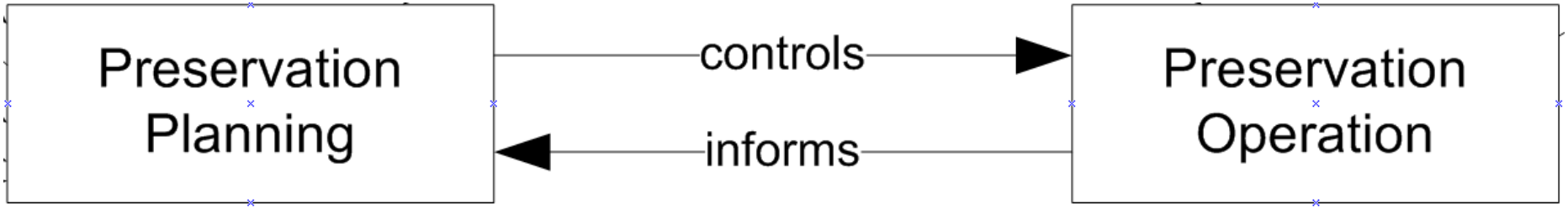
- IT Governance: decision making and communication within IT-supported organizations
- COBIT: Control Objectives for Information Technology
  - "the leadership, organisational structures and processes that ensure that the enterprise's IT sustains and extends the organisation's strategies and objectives"
  - goal-driven, process-oriented and control-based
  - How to leverage resources to achieve desired ends?
  - Goals – processes - activities
    - *Ensure systems security*
    - *Acquire and maintain application software*, ….
  - Sophisticated, adaptable process model

Core Preservation Capabilities



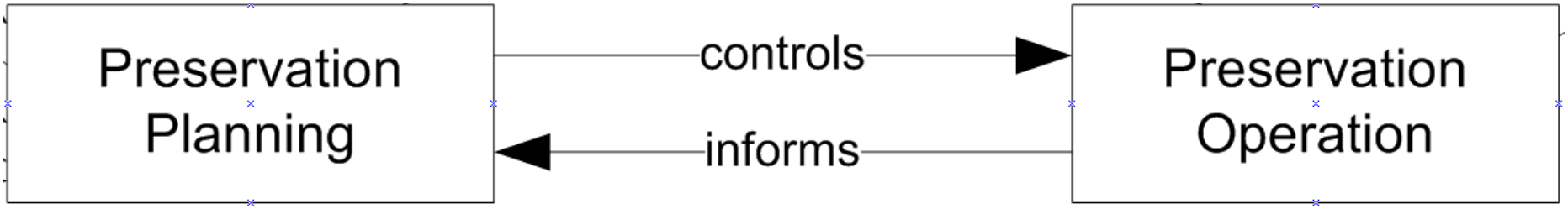
Preservation Planning	Preservation Operation
Monitor, steer and control the preservation operation of content	Control the deployment and execution of preservation plans.

Core Preservation Capabilities



Preservation Planning	Preservation Operation
Monitor, steer and control the preservation operation of content	Control the deployment and execution of preservation plans.
•Drivers and constraints •Decision making •Options diagnosis •Specification and delivery •Monitoring	•Analyze content •Execute preservation actions •Ensure adequate provenance trail •Handle preservation metadata •Conduct Quality Assurance •Provide reports and statistics

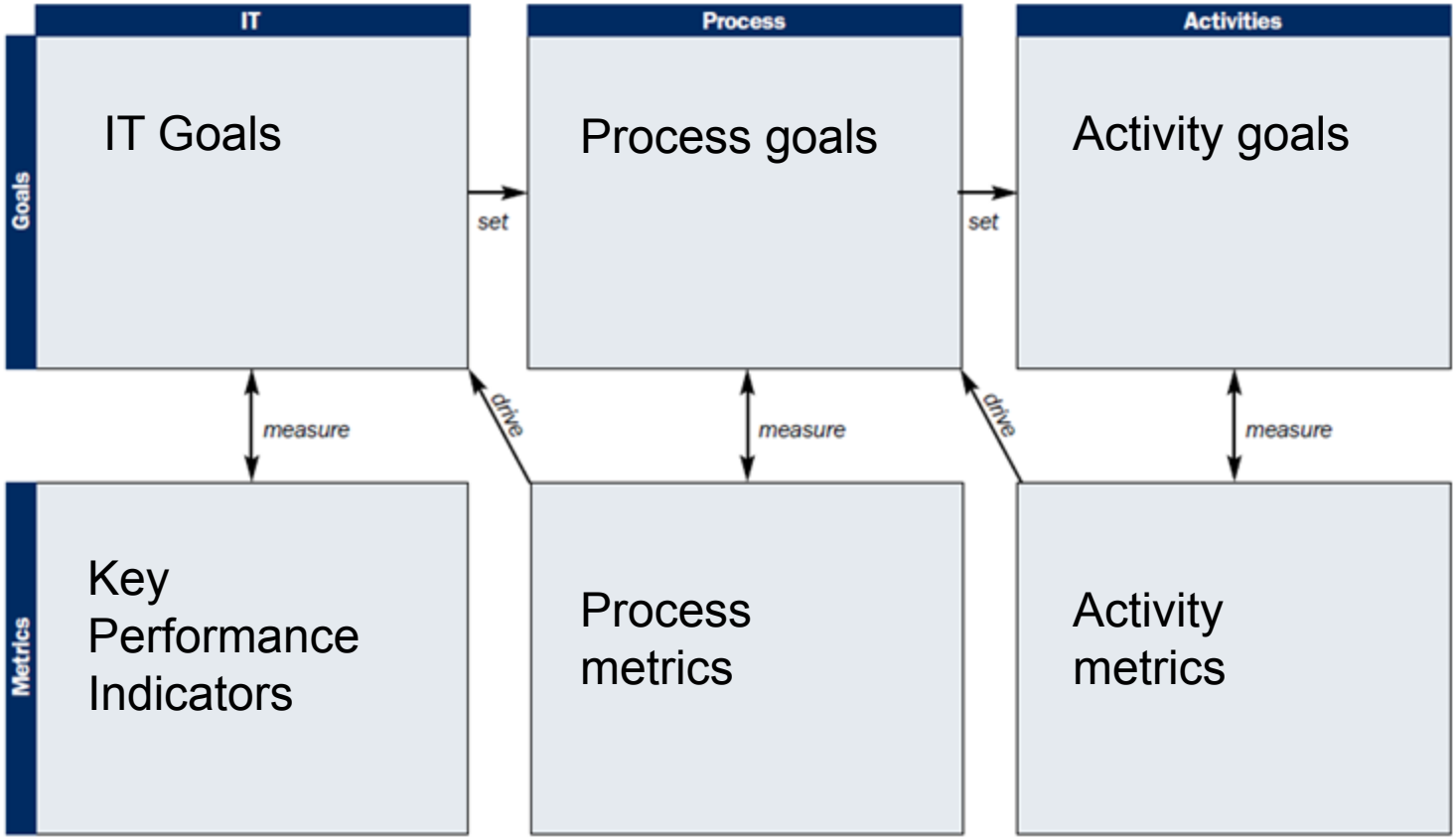
Core Preservation Capabilities



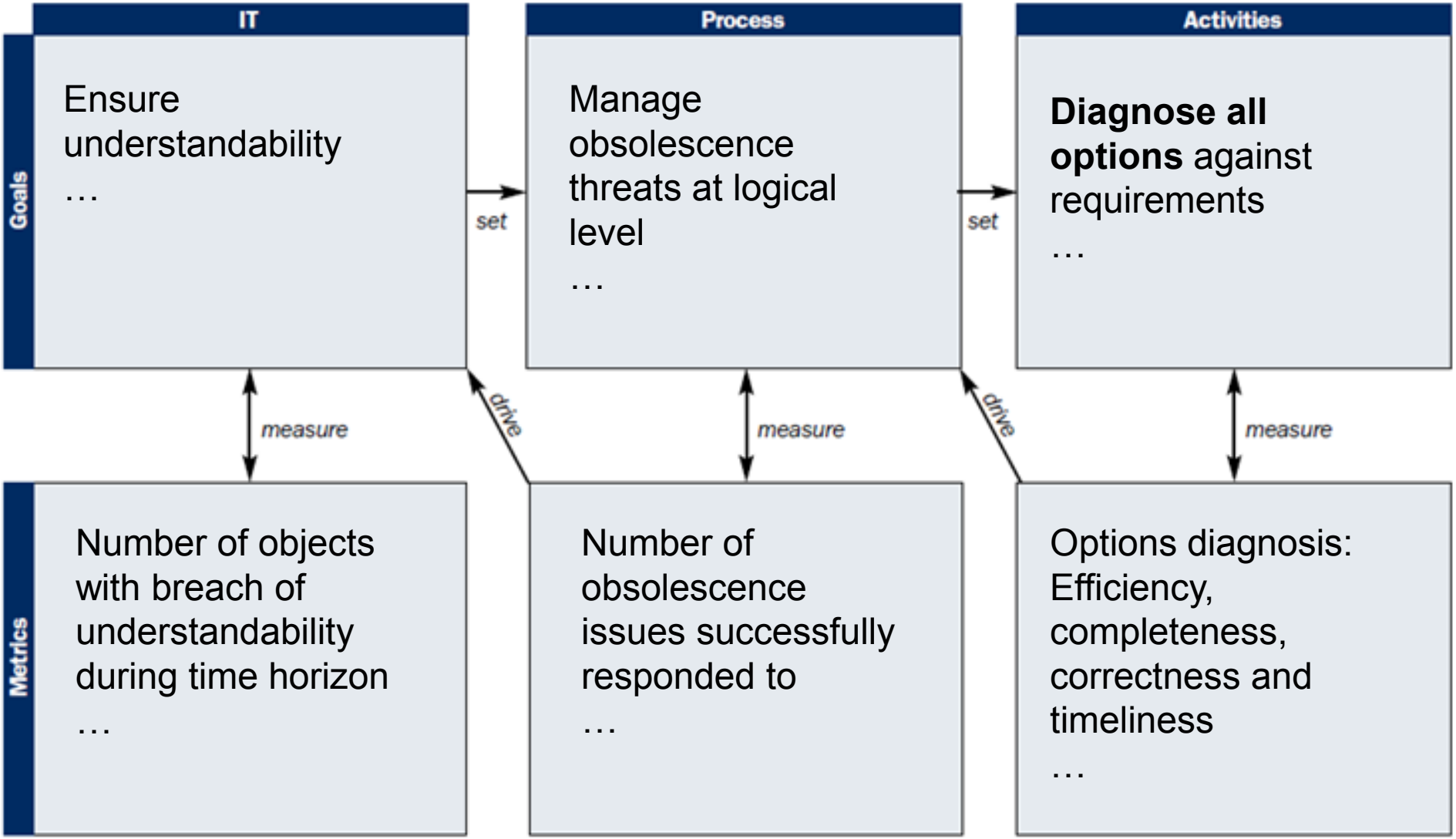
Preservation Planning	Preservation Operation
Monitor, steer and control the preservation operation of content	Control the deployment and execution of preservation plans.
•Drivers and constraints •Decision making •Options diagnosis •Specification and delivery •Monitoring	•Analyze content •Execute preservation actions •Ensure adequate provenance trail •Handle preservation metadata •Conduct Quality Assurance •Provide reports and statistics
"Migrate this set of images (in TIFF-5) to JP2 using ImageMagick 6.3 with parameters a,b,c"	*•Analyse original* *•Migrate, analyse output* *•Conduct quality assurance* *•Provenance, metadata, Reporting*

COBIT processes...

- Driven by specific goals and controls
- Organized into activities with assigned responsibilities
- Related to other processes
- Measured on all levels: Internal vs. external goals and metrics



How to measure

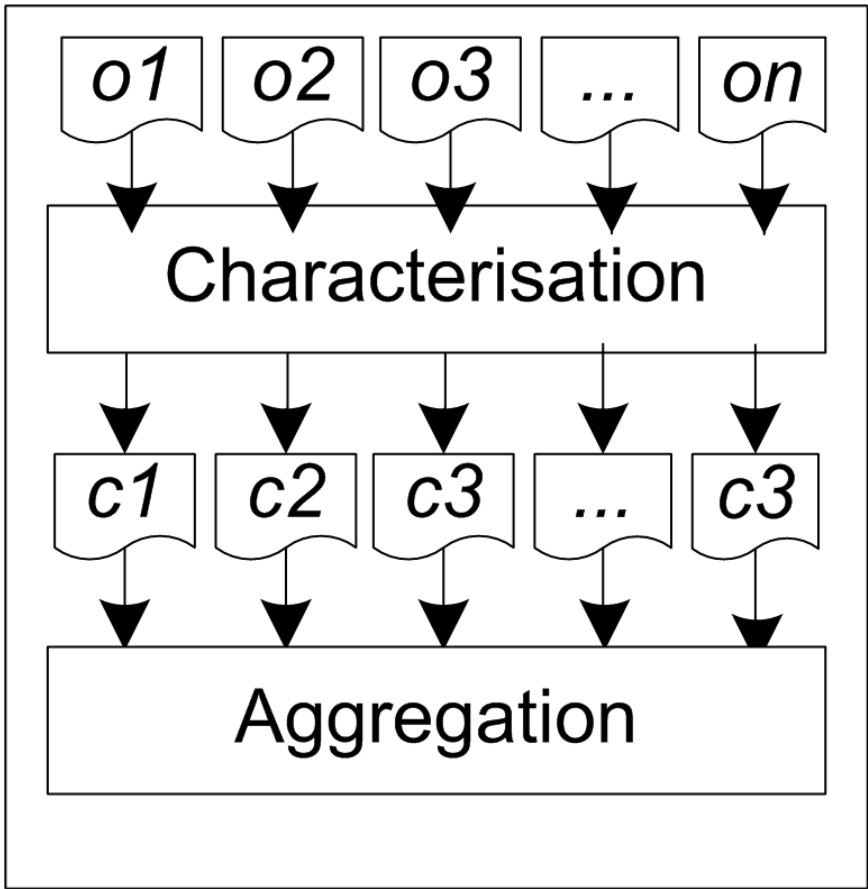


Digital Longevity

- Levels of challenges
  - Physical
  - Logical
  - Semantic
- Dimensions
  - Organisational
  - Technological
  - Economic
- Key capabilities
  - How to measure organisational processes?
- **How to start?**
  - Analysis: Find out what we have

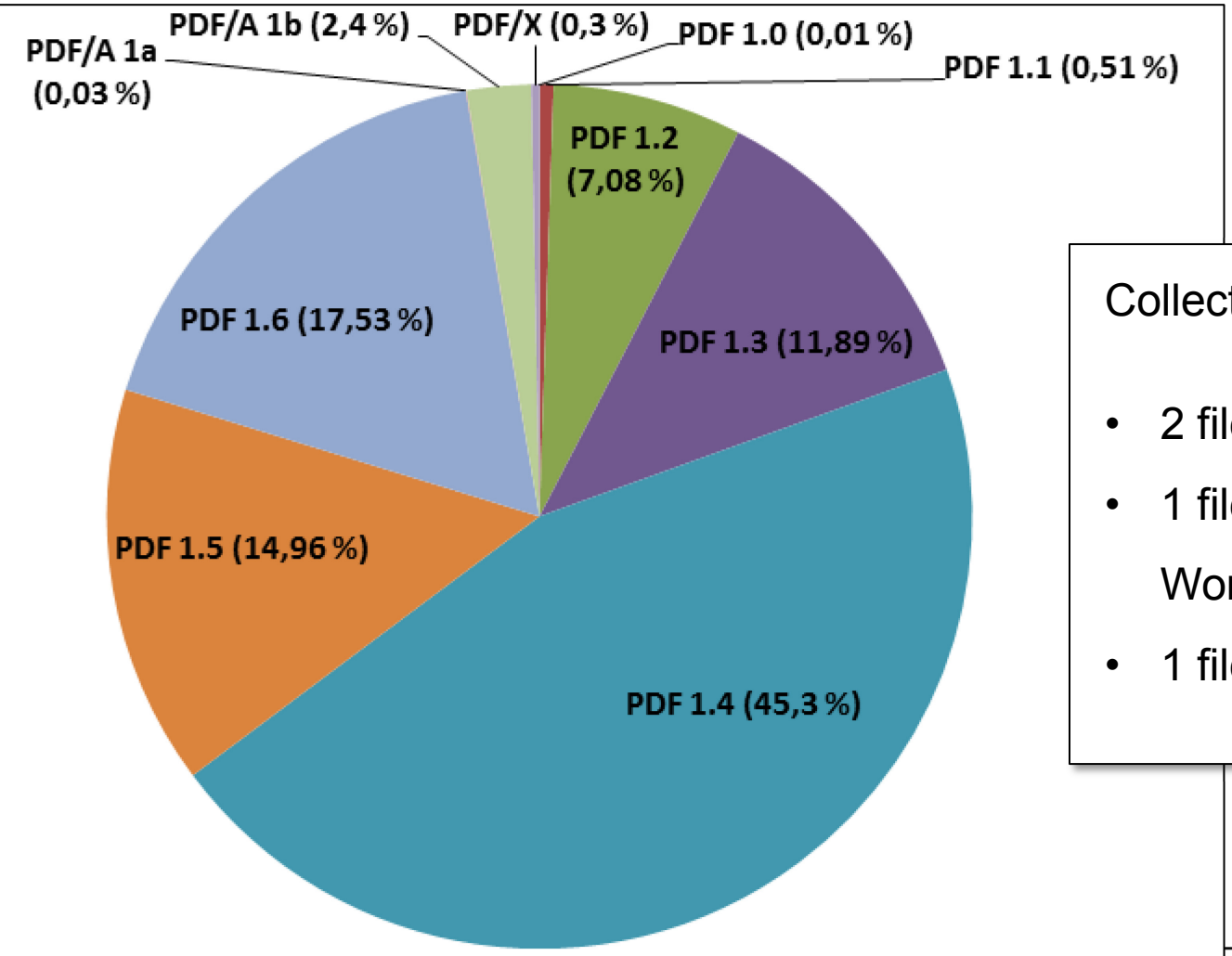
Content Profiling

- Multi-dimensional view on the features and feature distribution of a set of digital objects
  - Distribution of file formats
  - Distribution of characteristics
  - Representative data sets
- Stages
  0. *Characterise*
  1. Collect metadata
  2. Combine and filter
  3. Reason on the result
- Aggregated analysis of *characterisation results*



Content Profiling

- An example collection of ~42.000 documents



Collection size: 42.038 files

- 2 files „Unknown Binary“
- 1 file exposed as MS Word DOC
- 1 file with size 0

Some combined measures

- Number of PDF files with
  - `wellformed=false`: 604 (1,4%)
  - `valid=false`: 2494 (5,9%)
  - `wellformed=false` and `valid=false`: 604 (1,4%)
  - `has.forms=true` and `valid=false`: 705 (1,7%)
  - `has.forms=true` and `wellformed=false`: 153 (0,4%)

- For PDF/A and PDF/X validity and wellformedness could not be determined for all files
  - Challenge: Deal with sparsity

Ok, so I have a lot of *X*. What now?

- Having a profile of the content enables first steps of assessing values and risks
  - How valuable are certain parts?
  - How long do we want to keep them?
  - Who needs them?
  - What are their technical dependencies? How can they be performed?
  - Are they standardised? Do they have specific issues? How strongly are they coupled to an environment?
- Who has a digital camera?
  - And who shoots RAW?

Why RAW

digital negative

most authentic version of the image

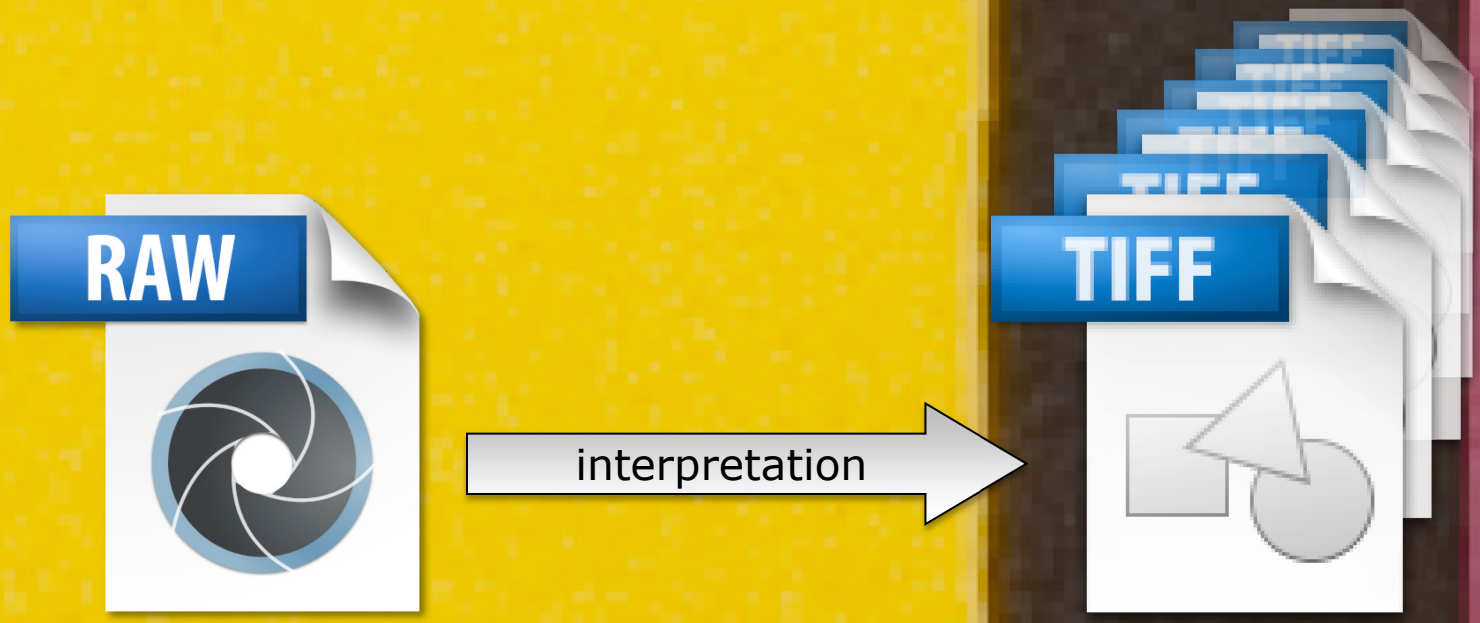
no lossy compression as in jpeg

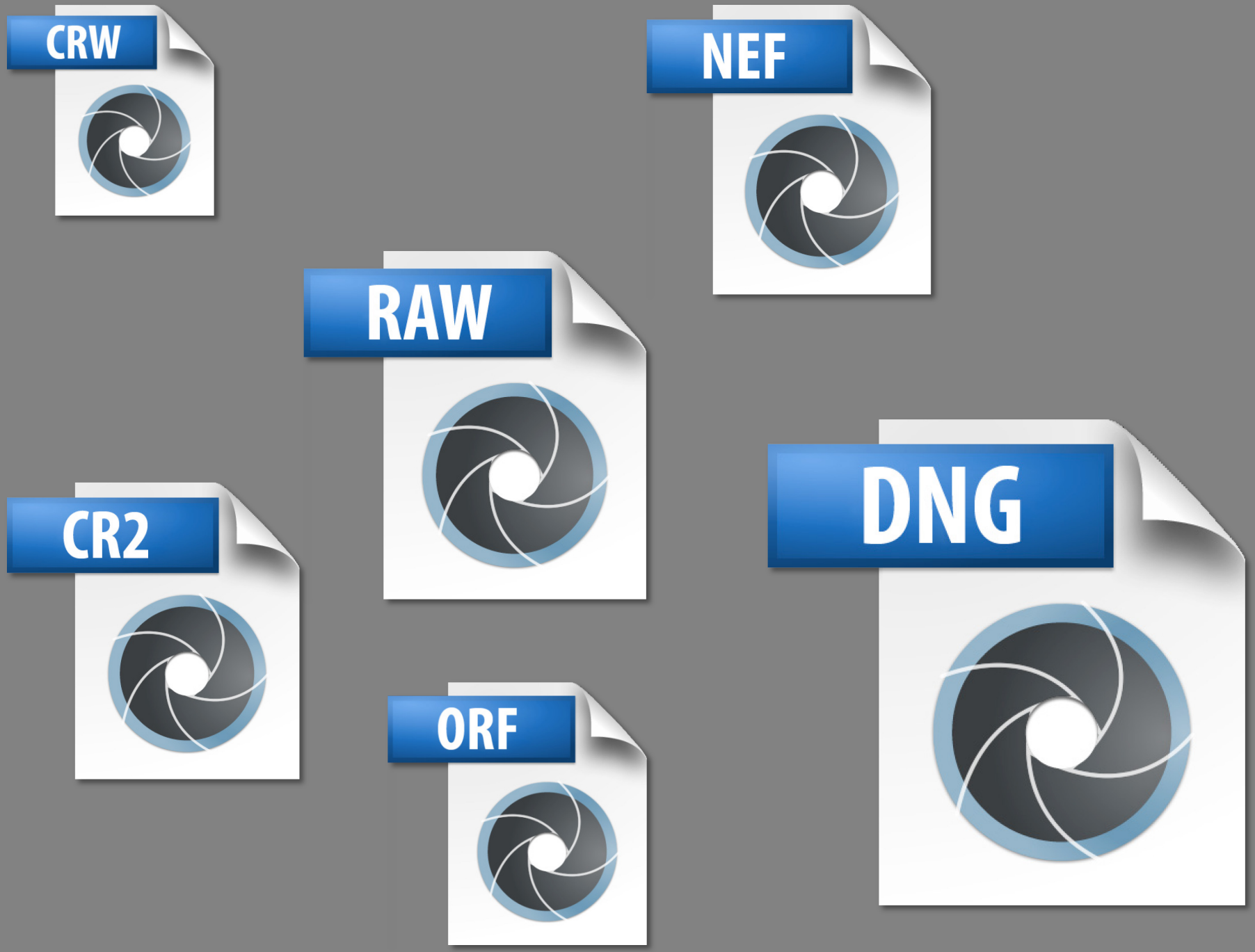
new tools may create better interpretations

not an image: uninterpolated sensor data

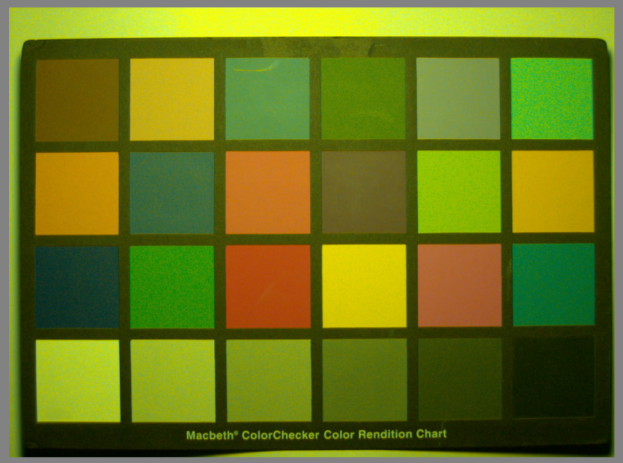
Developing a raw file

- Demosaiquing
- White balance adjustment
- Colorimetric interpretation
- Tone mapping
- Image enhancements

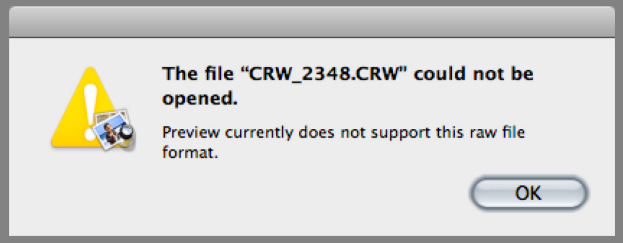

RAW
interpretation
TIFF


CRW
NEF
RAW
CR2
DNG
ORF

Adobe



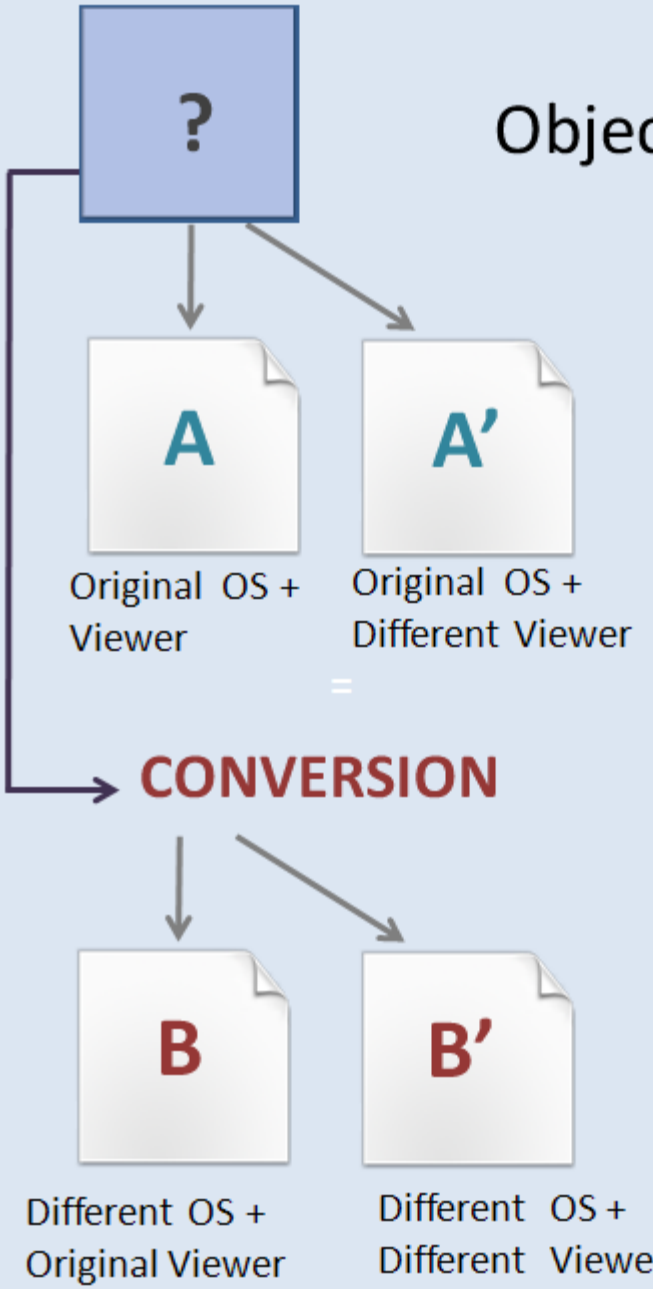
dcraw

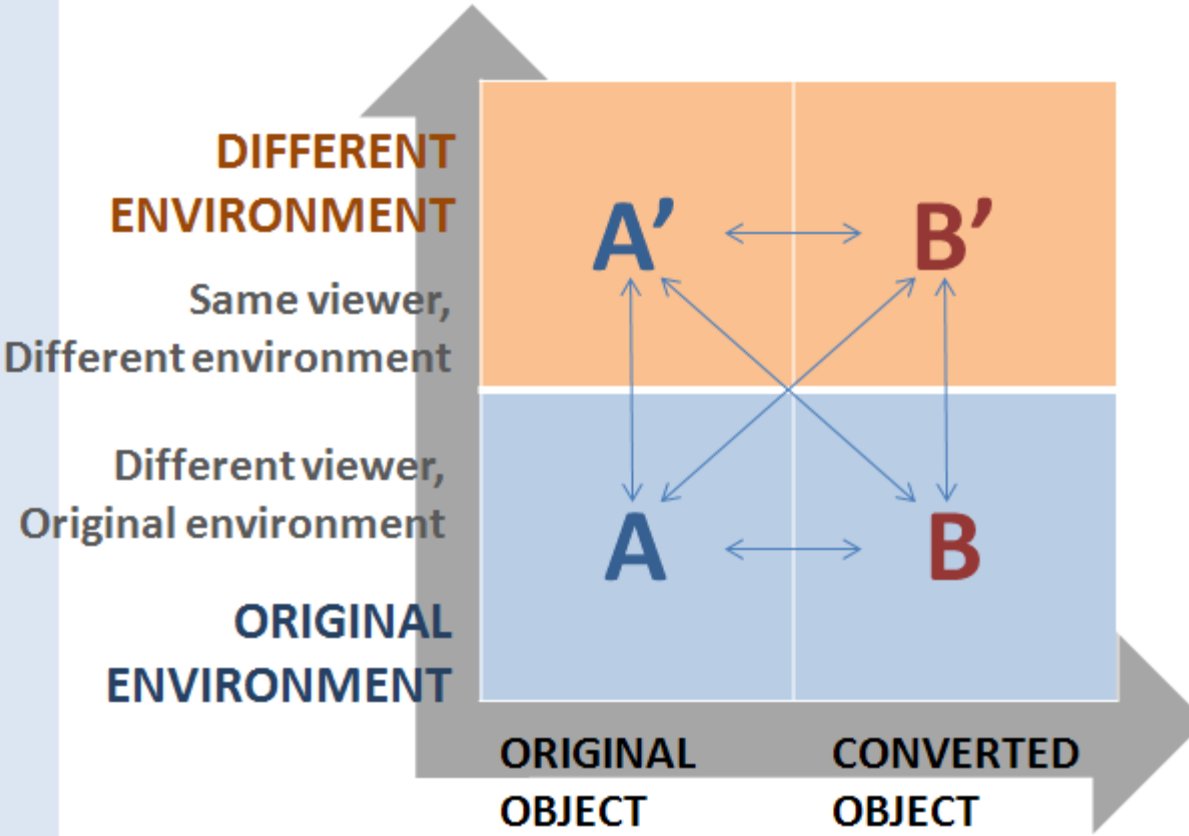


Apple

Adobe DNG Converter

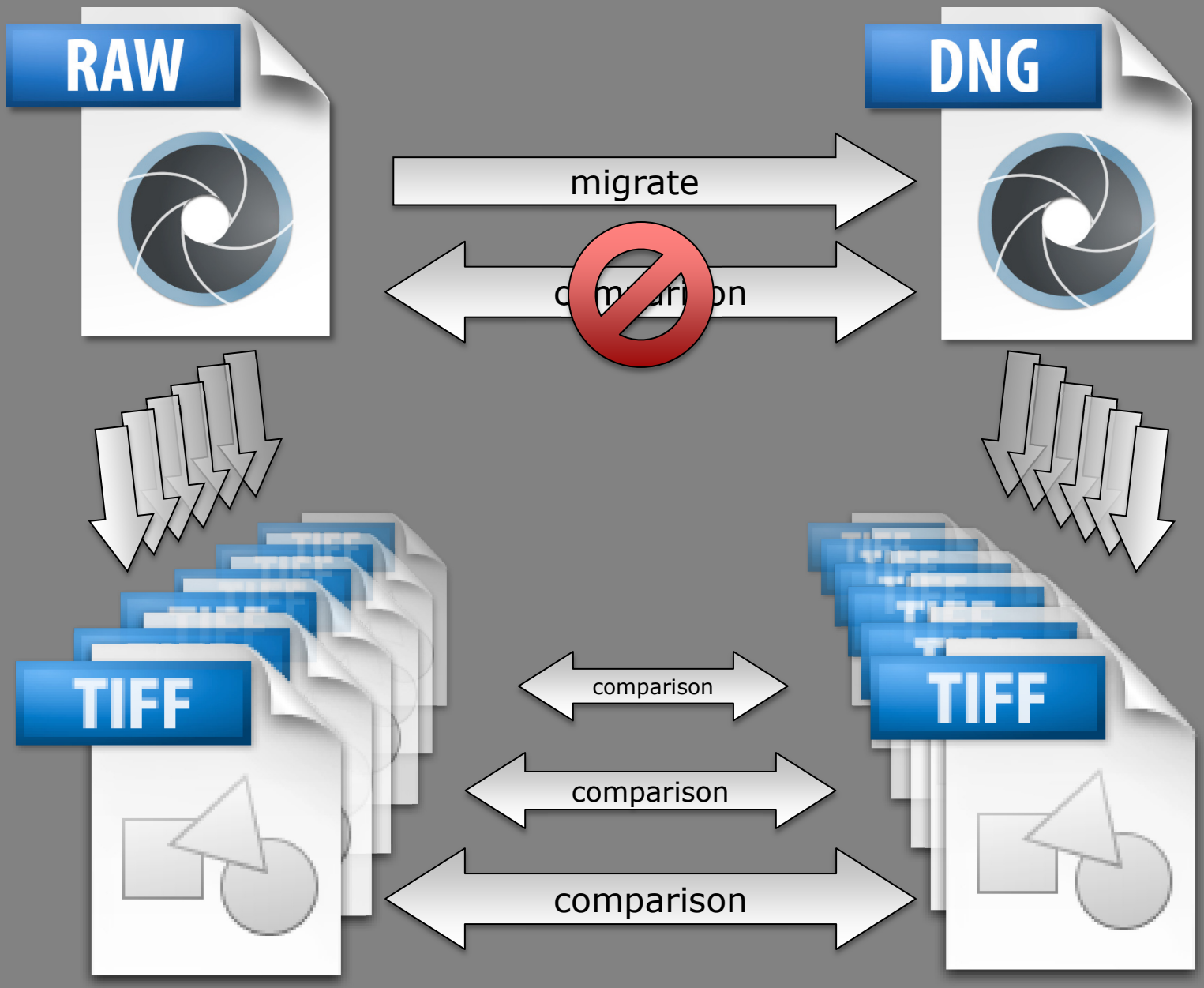
Objects, environments and dependencies



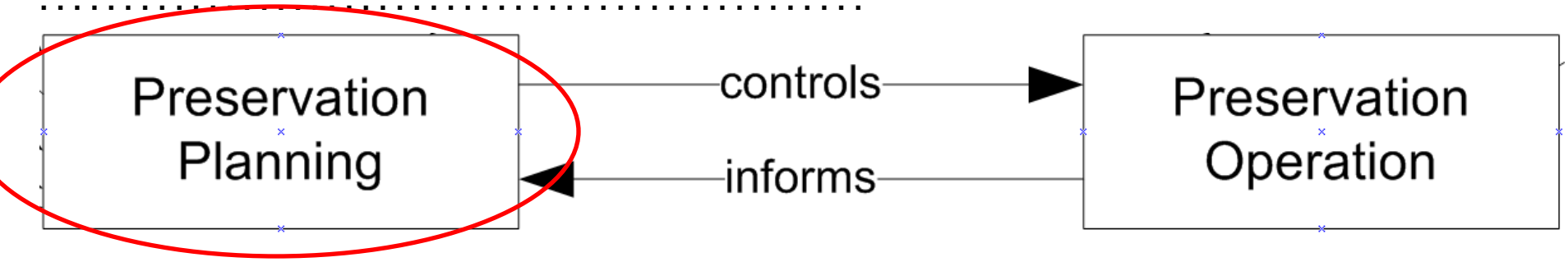


...networks of objects

(Diagram by Natasa Milic-Frayling, MSRC)


RAW
DNG
migrate
TIFF
TIFF
comparison
comparison
comparison

Core Preservation Capabilities



Preservation Planning	Preservation Operation
Monitor, steer and control the preservation operation of content	Control the deployment and execution of preservation plans.
•Drivers and constraints •Decision making •Options diagnosis •Specification and delivery •Monitoring	•Analyze content •Execute preservation actions •Ensure adequate provenance trail •Handle preservation metadata •Conduct Quality Assurance •Provide reports and statistics
"Migrate this set of images (in CRW) to DNG using … with parameters a,b,c and verify significant properties P and Q"	*•Analyse original* *•Migrate, analyse output* *•Conduct quality assurance* *•Provenance, metadata, Reporting*

Summary

- The mission of DP
- Communication over time
- **Authenticity** and **significant properties**
- Understandability for whom?
- Objects and environments: The performance
- Quality assurance is hard
- Content profiling is the first step
- Systems perspective is important
- Core capabilities: Operations and planning
  - But needs much context; governance, risk and compliance; supporting capabilities etc

DP Übung Task 1

- Find out what content you hold yourself and what technical properties it has
- get to know some basic file (format) **analysis tools** for identifying content types and formats and extracting features.
- use this information to **assess basic risks** threatening the longevity of that content.
- think about **how to aggregate such information** to provide a high-level profile view that serves as the starting point for preservation analysis, risk detection, dividing the content into smaller, more easily managed sets, and preservation planning.

DP Übung Task 1

1. Identify and characterise
   - Run *fits* tool on your files
   - http://code.google.com/p/fits/
2. Analyse: Create a content profile
   - Use c3po
   - http://ifs.tuwien.ac.at/imp/c3po
3. Summarise
   - Analyse, write up a short report
4. Upload results until April 3, 23:55
   - PDF report (5 pages)
   - XML profile (exported from c3po)

The DPÜ is managed in TUWEL!

Toy Story 2: The movie vanishes

- http://www.youtube.com/watch?v=EL_g0tyaIeE

Thank you for your attention!

Questions?

www.ifs.tuwien.ac.at/~becker