

# **Digital Preservation Decisions and Governance**

An IT Perspective

Christoph Becker

Vienna University of Technology

Vienna, Austria

[www.ifs.tuwien.ac.at/~becker](http://www.ifs.tuwien.ac.at/~becker)

# Why do we need Digital Preservation?

- ...
- ...
- ...
- ...
- ...
- programs 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.
- .....













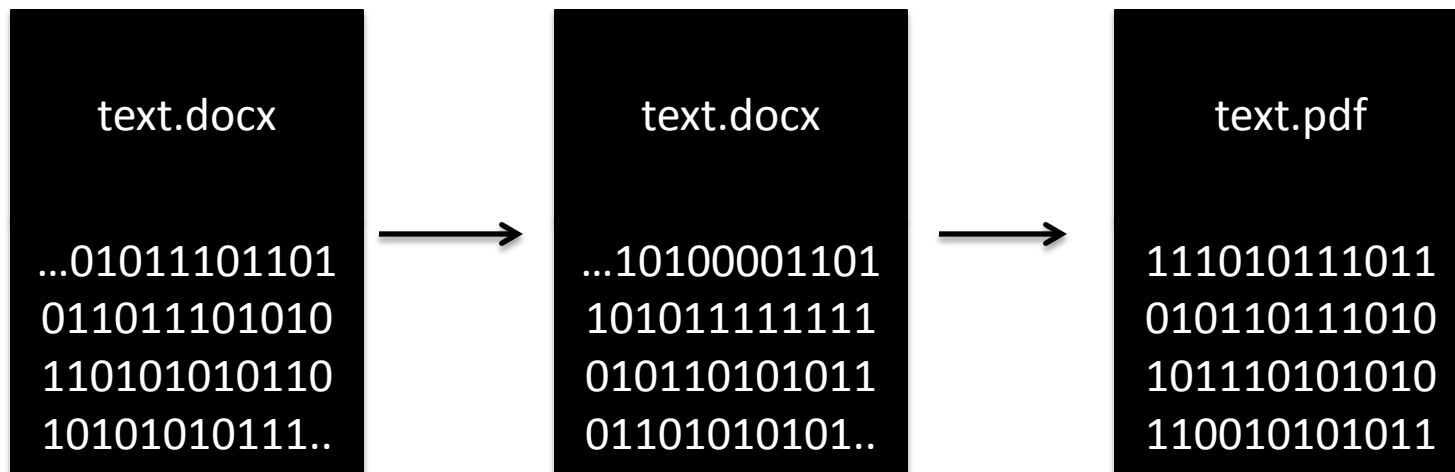
**The file "CRW\_2348.CRW" could not be opened.**

Preview currently does not support this raw file format.

OK

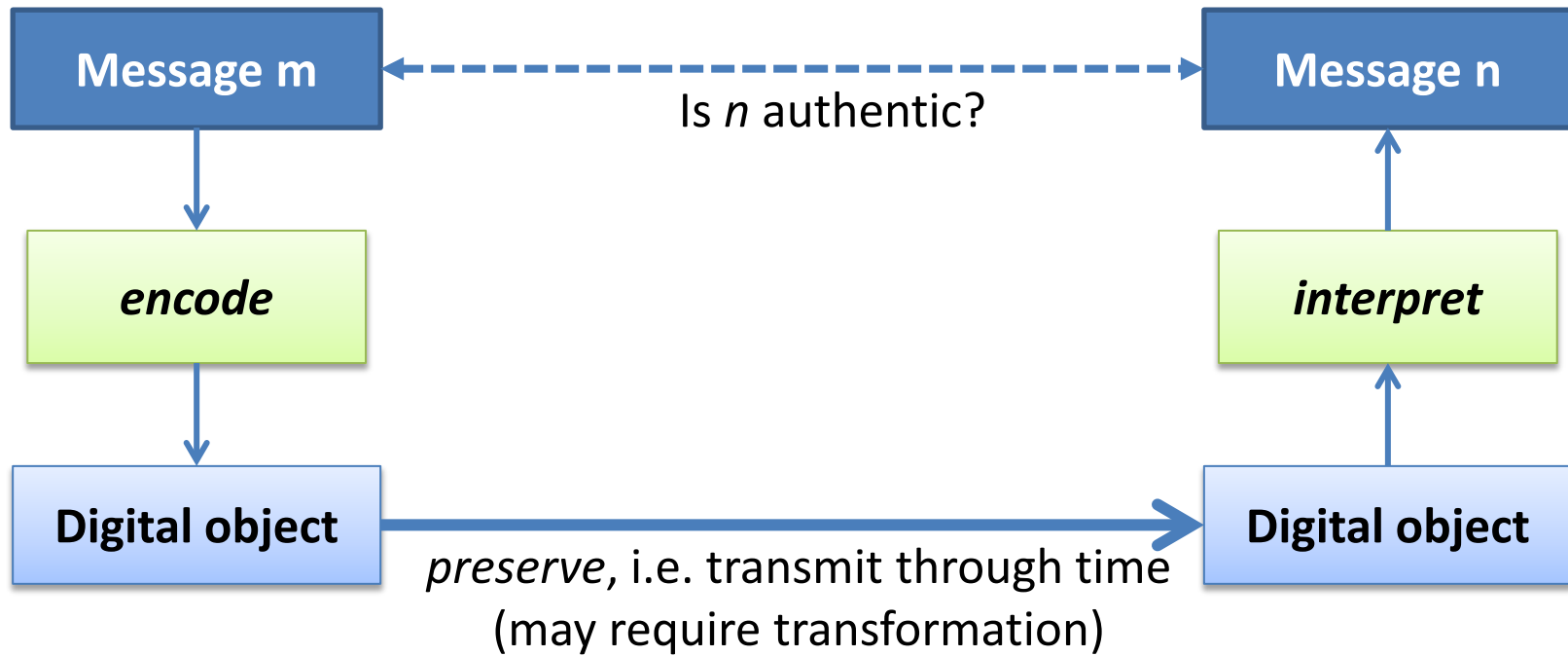
# Digital content and understandability

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





# Digital preservation is communication.



... But at the time of reception

there is no message  $m$  any more

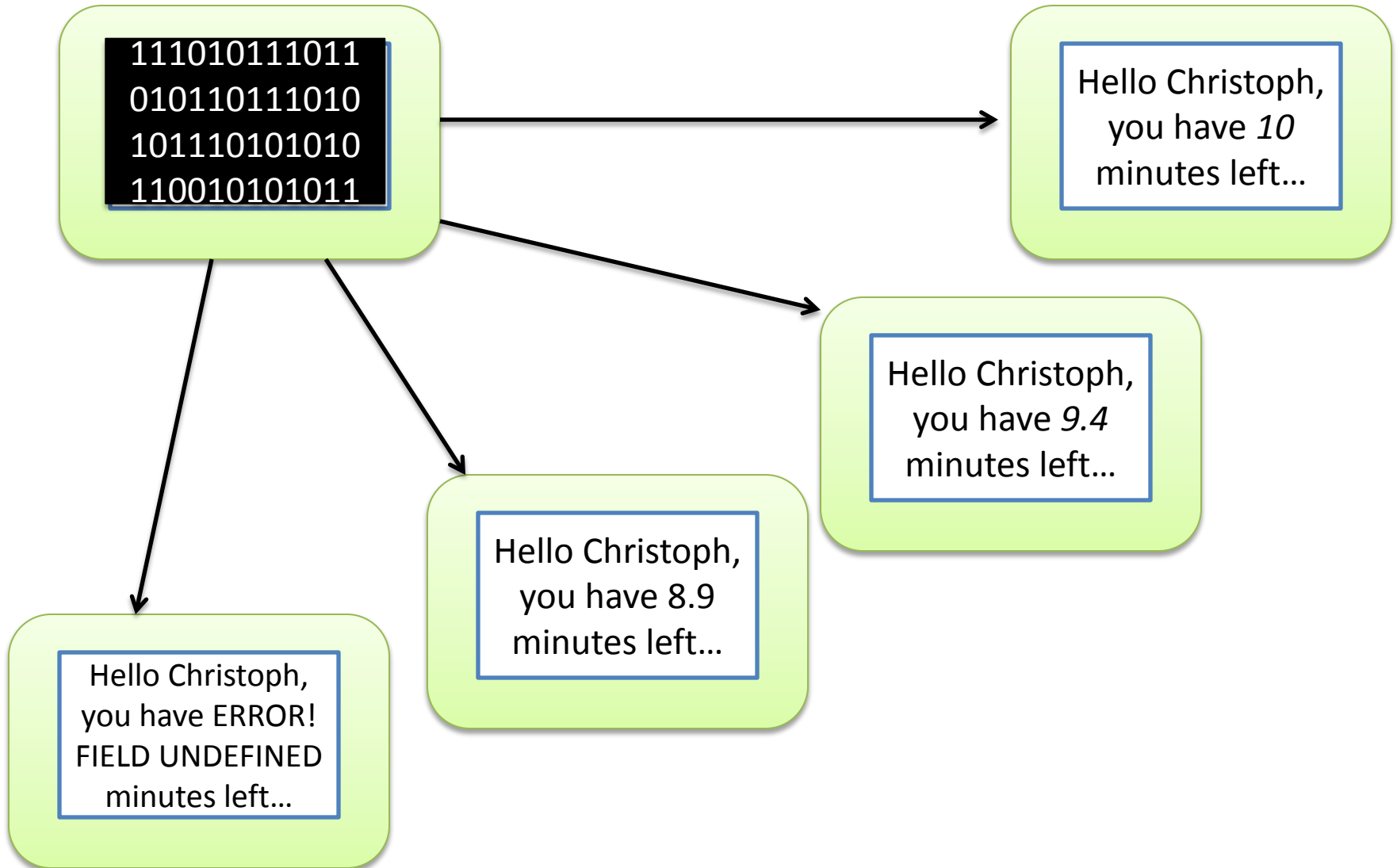
there may be no sender (any more)

there may be no encoder to check against

there may be no decoder

the recipient may not be the original addressee

# The black box problem



# Five years later...

```
111010111011  
010110111010  
101110101010  
110010101011
```

text.pdf

Hello Max, you  
have -21 minutes  
left...

Hello Christoph, you  
have 9.4 minutes left...

Hello ERROR! FIELD  
UNDEFINED , you have  
- 678345 minutes left...

text.docx

Hello Christoph,  
you have 10  
minutes left

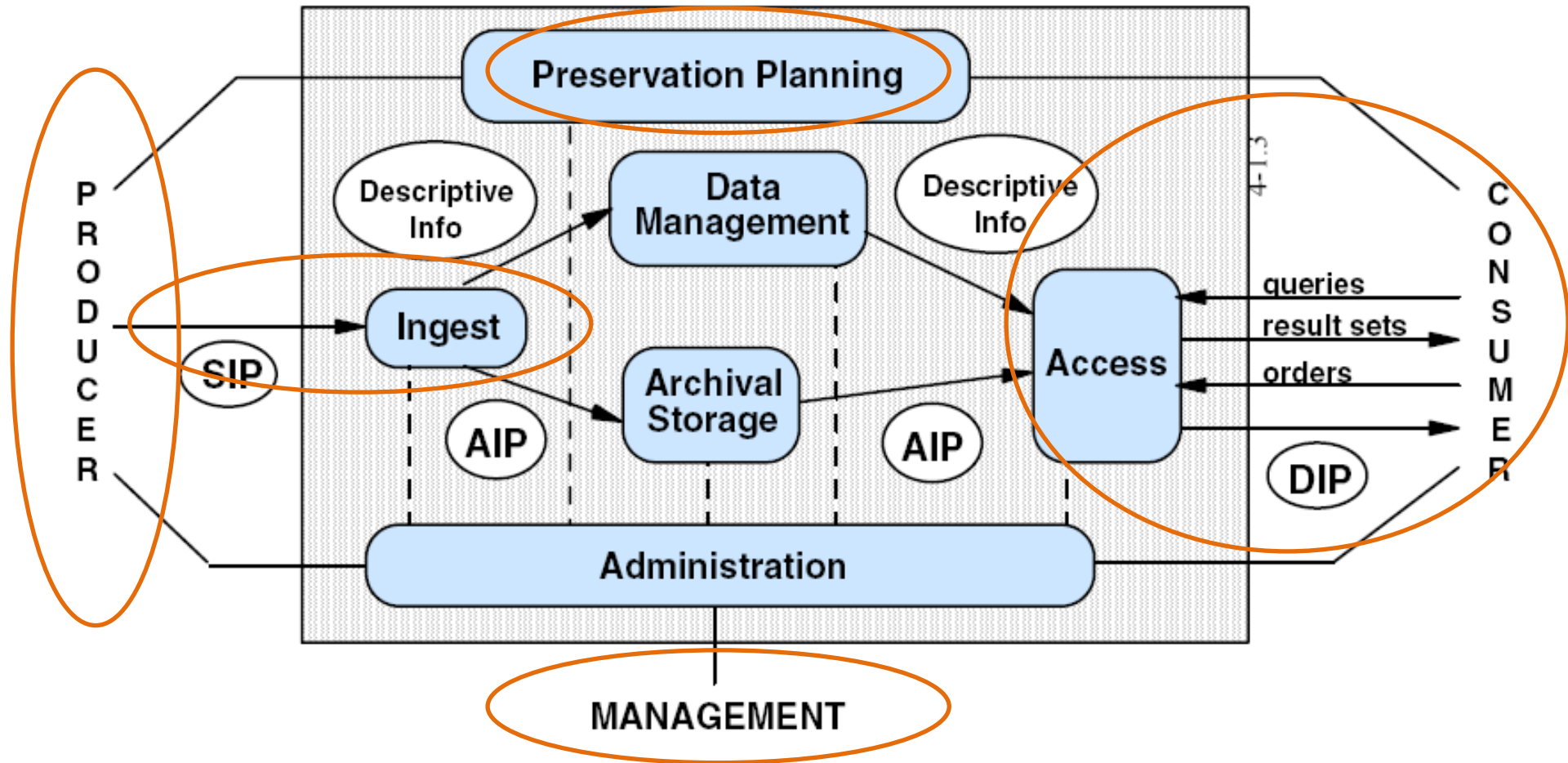


# Digital Longevity

- 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
- Focus on a repository institution responsible for safeguarding cultural heritage



# A repository



- ... What to do with the Word files?



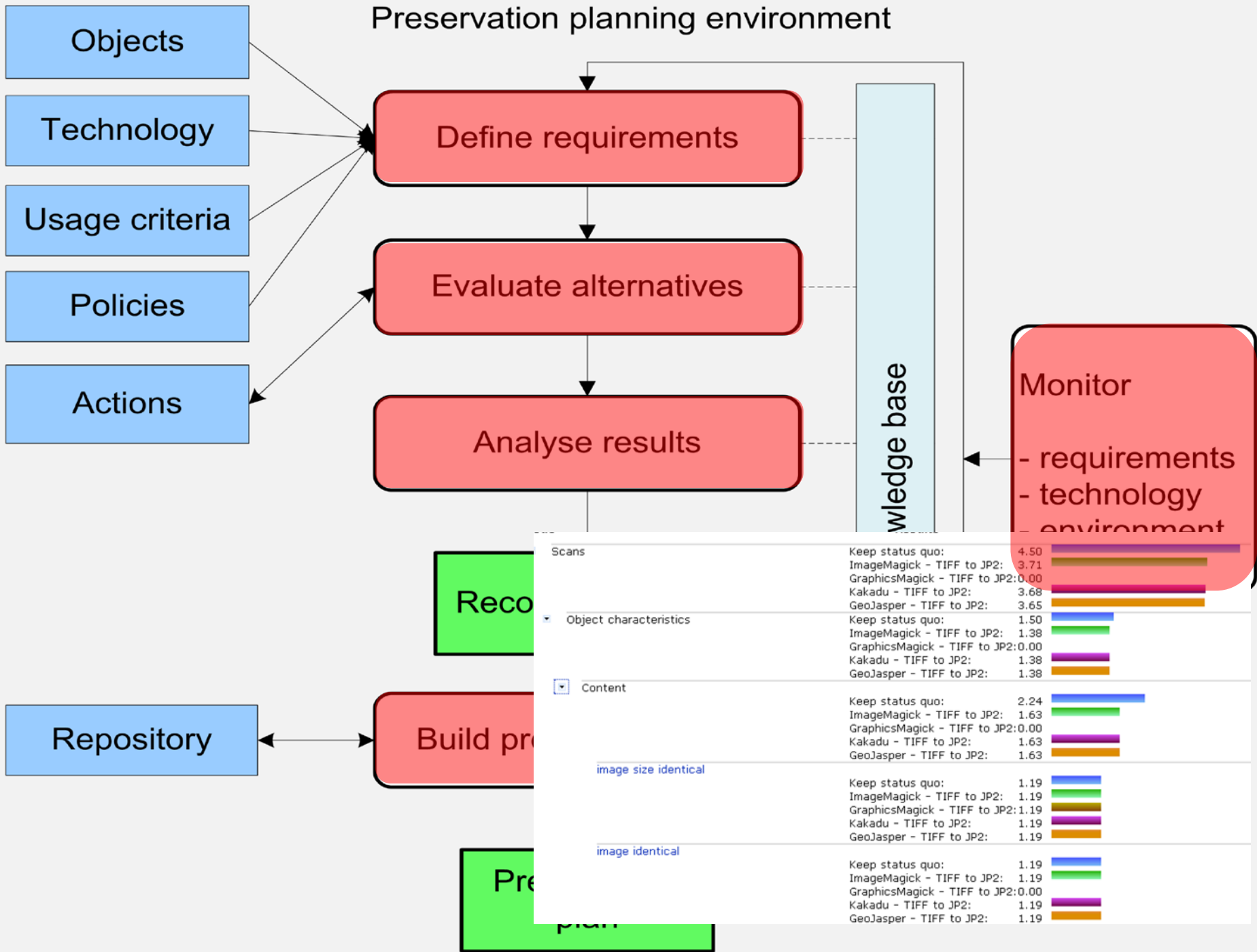
# The problem

- Challenges in evaluating preservation actions
  - Quality varies across tools
  - Properties vary across content
  - Usage varies across communities
  - Requirements vary across scenarios
  - Risk tolerance varies across collections
  - Preferences and constraints vary across organisations
  - Cost structures and compatibility varies across environments
  - Constraints, priorities and requirements shift constantly

# Trustworthy preservation planning

- Preservation planning:
  - the ability to assess the impact of influencers and specify actionable preservation plans that define concrete courses of actions and the directives governing their execution
  - the operative management of obsolescence to maximize expected value with minimal costs
- A preservation plan specifies actions
  - scope and what, how, when, who, why
- Trust requires evidence
  - Trust has to be evaluated in a realistic context
    - Documented evidence
    - Controlled experimentation
    - scenario-specific requirements assessment

# Preservation planning environment



Reco

Pre  
plan

Category	Tool	Value
Scans	Keep status quo:	4.50
	ImageMagick - TIFF to JP2:	3.71
	GraphicsMagick - TIFF to JP2:	0.00
	Kakadu - TIFF to JP2:	3.68
	GeoJasper - TIFF to JP2:	3.65
Object characteristics	Keep status quo:	1.50
	ImageMagick - TIFF to JP2:	1.38
	GraphicsMagick - TIFF to JP2:	0.00
	Kakadu - TIFF to JP2:	1.38
	GeoJasper - TIFF to JP2:	1.38
Content	Keep status quo:	2.24
	ImageMagick - TIFF to JP2:	1.63
	GraphicsMagick - TIFF to JP2:	0.00
	Kakadu - TIFF to JP2:	1.63
	GeoJasper - TIFF to JP2:	1.63
image size identical		
image identical	Keep status quo:	1.19
	ImageMagick - TIFF to JP2:	1.19
	GraphicsMagick - TIFF to JP2:	1.19
	Kakadu - TIFF to JP2:	1.19
	GeoJasper - TIFF to JP2:	1.19



# Preservation Planning: Key concepts

- Repeatable, standardized planning workflow
- A weighted hierarchy of objectives
  - Measurable criteria on the leaf level of the tree
  - Utility functions make criteria comparable
- Controlled experimentation on sample content
  - Evidence-based decision making
- Standardized structure for plan specification
  - Transparency and documentation
  - Comparability across scenarios
  - Integration with repository systems
- Planning tool Plato guides, validates, documents
- Automation: Reduce manual effort

# Case studies

- Case studies conducted with Plato
  - Scanned images
  - Interactive art
  - Computer games
  - Born-digital photographs
  - Relational databases
  - Electronic documents
  - Console video games
  - Emails
  - ...
- <http://www.ifs.tuwien.ac.at/dp/plato>
- Plato is free

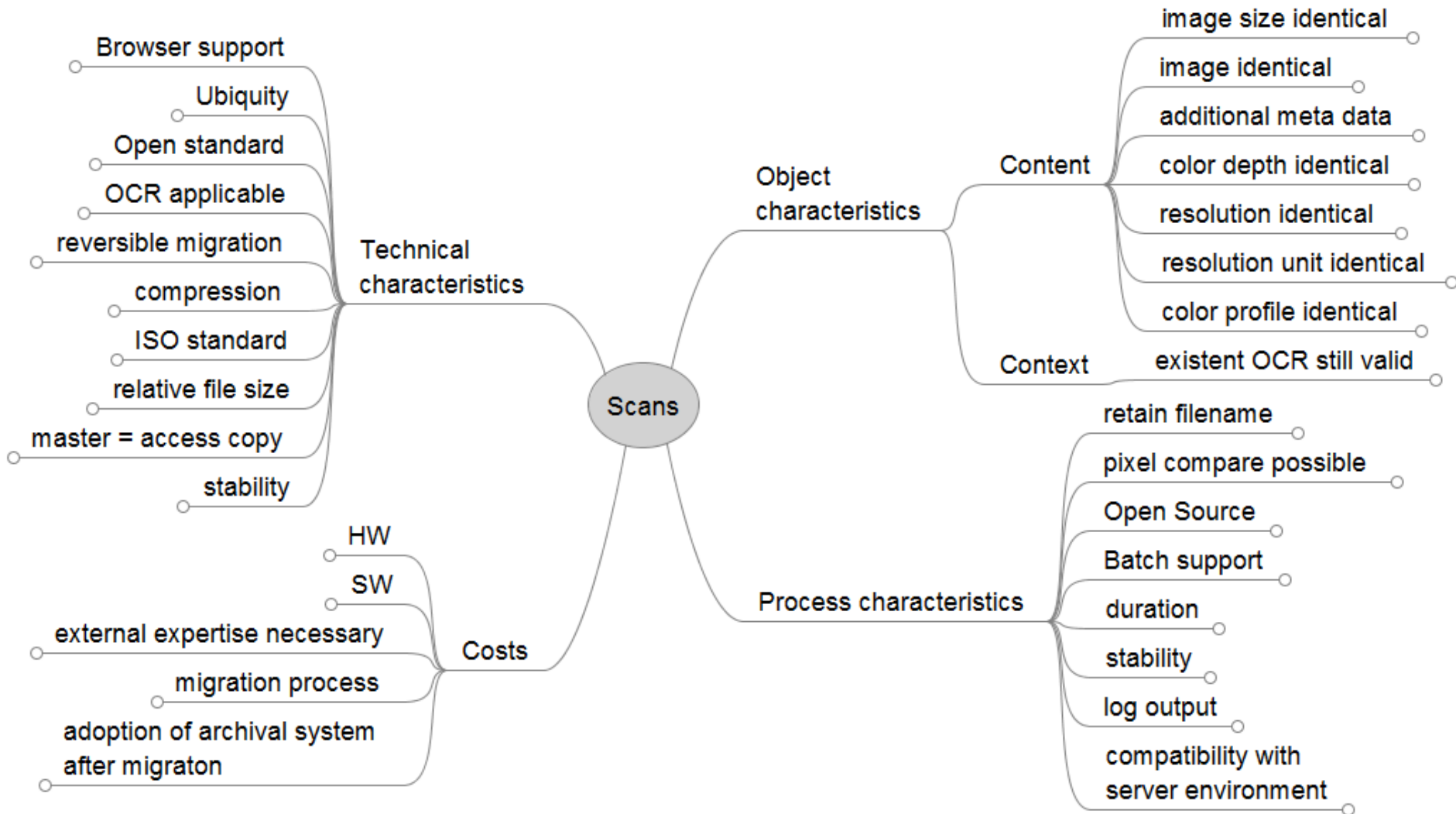
# Four cases, three solutions: Scanned images

- Bavarian State Library, 72TB TIFF6: *Leave and monitor*
- British Library, 80TB TIFF5: *Migrate to JP2 (ImageMagick)*
- Royal Library of Denmark, ~10.000 aerial photographs in TIFF6: *Leave and monitor*
- State and University Library Denmark, scanned yearbooks in GIF: *Migrate to TIFF 6*

Scenario	Chosen action	Main reasons
72 TB scanned book pages in TIFF6	Leave unchanged and monitor	Color profile complications, lack of JP2 browser support, Process costs
80 TB scanned newspapers in TIFF5	Migrate to JP2	Storage costs, Standardization
Aerial photographs in TIFF6	Leave unchanged and monitor	Lack of JP2 browser support, Process costs
































































































# Scanned books requirements



# Scanned books results

## Results: Weighted multiplication

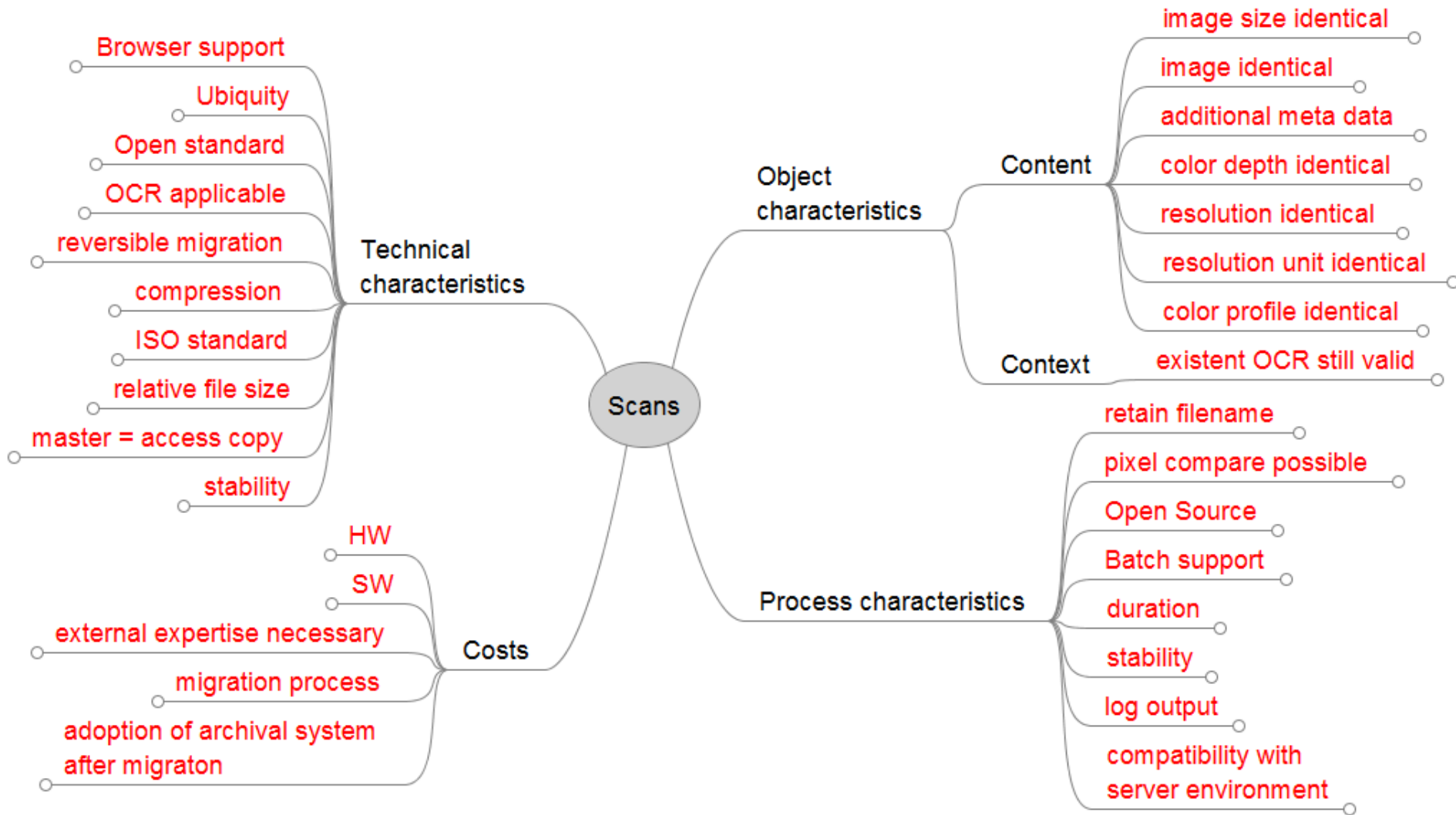
Result-Tree with all Alternatives, Aggregation method: Weighted multiplication

Node	Results															
Scans	<table><tr><td>Keep status quo:</td><td>4.50</td><td></td></tr><tr><td>ImageMagick - TIFF to JP2:</td><td>3.71</td><td></td></tr><tr><td>GraphicsMagick - TIFF to JP2:</td><td>0.00</td><td></td></tr><tr><td>Kakadu - TIFF to JP2:</td><td>3.68</td><td></td></tr><tr><td>GeoJasper - TIFF to JP2:</td><td>3.65</td><td></td></tr></table>	Keep status quo:	4.50		ImageMagick - TIFF to JP2:	3.71		GraphicsMagick - TIFF to JP2:	0.00		Kakadu - TIFF to JP2:	3.68		GeoJasper - TIFF to JP2:	3.65	
Keep status quo:	4.50															
ImageMagick - TIFF to JP2:	3.71															
GraphicsMagick - TIFF to JP2:	0.00															
Kakadu - TIFF to JP2:	3.68															
GeoJasper - TIFF to JP2:	3.65															
Object characteristics	<table><tr><td>Keep status quo:</td><td>1.50</td><td></td></tr><tr><td>ImageMagick - TIFF to JP2:</td><td>1.38</td><td></td></tr><tr><td>GraphicsMagick - TIFF to JP2:</td><td>0.00</td><td></td></tr><tr><td>Kakadu - TIFF to JP2:</td><td>1.38</td><td></td></tr><tr><td>GeoJasper - TIFF to JP2:</td><td>1.38</td><td></td></tr></table>	Keep status quo:	1.50		ImageMagick - TIFF to JP2:	1.38		GraphicsMagick - TIFF to JP2:	0.00		Kakadu - TIFF to JP2:	1.38		GeoJasper - TIFF to JP2:	1.38	
Keep status quo:	1.50															
ImageMagick - TIFF to JP2:	1.38															
GraphicsMagick - TIFF to JP2:	0.00															
Kakadu - TIFF to JP2:	1.38															
GeoJasper - TIFF to JP2:	1.38															
Content	<table><tr><td>Keep status quo:</td><td>2.24</td><td></td></tr><tr><td>ImageMagick - TIFF to JP2:</td><td>1.63</td><td></td></tr><tr><td>GraphicsMagick - TIFF to JP2:</td><td>0.00</td><td></td></tr><tr><td>Kakadu - TIFF to JP2:</td><td>1.63</td><td></td></tr><tr><td>GeoJasper - TIFF to JP2:</td><td>1.63</td><td></td></tr></table>	Keep status quo:	2.24		ImageMagick - TIFF to JP2:	1.63		GraphicsMagick - TIFF to JP2:	0.00		Kakadu - TIFF to JP2:	1.63		GeoJasper - TIFF to JP2:	1.63	
Keep status quo:	2.24															
ImageMagick - TIFF to JP2:	1.63															
GraphicsMagick - TIFF to JP2:	0.00															
Kakadu - TIFF to JP2:	1.63															
GeoJasper - TIFF to JP2:	1.63															
image size identical	<table><tr><td>Keep status quo:</td><td>1.19</td><td></td></tr><tr><td>ImageMagick - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>GraphicsMagick - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>Kakadu - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>GeoJasper - TIFF to JP2:</td><td>1.19</td><td></td></tr></table>	Keep status quo:	1.19		ImageMagick - TIFF to JP2:	1.19		GraphicsMagick - TIFF to JP2:	1.19		Kakadu - TIFF to JP2:	1.19		GeoJasper - TIFF to JP2:	1.19	
Keep status quo:	1.19															
ImageMagick - TIFF to JP2:	1.19															
GraphicsMagick - TIFF to JP2:	1.19															
Kakadu - TIFF to JP2:	1.19															
GeoJasper - TIFF to JP2:	1.19															
image identical	<table><tr><td>Keep status quo:</td><td>1.19</td><td></td></tr><tr><td>ImageMagick - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>GraphicsMagick - TIFF to JP2:</td><td>0.00</td><td></td></tr><tr><td>Kakadu - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>GeoJasper - TIFF to JP2:</td><td>1.19</td><td></td></tr></table>	Keep status quo:	1.19		ImageMagick - TIFF to JP2:	1.19		GraphicsMagick - TIFF to JP2:	0.00		Kakadu - TIFF to JP2:	1.19		GeoJasper - TIFF to JP2:	1.19	
Keep status quo:	1.19															
ImageMagick - TIFF to JP2:	1.19															
GraphicsMagick - TIFF to JP2:	0.00															
Kakadu - TIFF to JP2:	1.19															
GeoJasper - TIFF to JP2:	1.19															
additional meta data	<table><tr><td>Keep status quo:</td><td>1.19</td><td></td></tr><tr><td>ImageMagick - TIFF to JP2:</td><td>1.08</td><td></td></tr><tr><td>GraphicsMagick - TIFF to JP2:</td><td>1.08</td><td></td></tr><tr><td>Kakadu - TIFF to JP2:</td><td>1.08</td><td></td></tr><tr><td>GeoJasper - TIFF to JP2:</td><td>1.08</td><td></td></tr></table>	Keep status quo:	1.19		ImageMagick - TIFF to JP2:	1.08		GraphicsMagick - TIFF to JP2:	1.08		Kakadu - TIFF to JP2:	1.08		GeoJasper - TIFF to JP2:	1.08	
Keep status quo:	1.19															
ImageMagick - TIFF to JP2:	1.08															
GraphicsMagick - TIFF to JP2:	1.08															
Kakadu - TIFF to JP2:	1.08															
GeoJasper - TIFF to JP2:	1.08															
color depth identical	<table><tr><td>Keep status quo:</td><td>1.19</td><td></td></tr><tr><td>ImageMagick - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>GraphicsMagick - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>Kakadu - TIFF to JP2:</td><td>1.19</td><td></td></tr><tr><td>GeoJasper - TIFF to JP2:</td><td>1.19</td><td></td></tr></table>	Keep status quo:	1.19		ImageMagick - TIFF to JP2:	1.19		GraphicsMagick - TIFF to JP2:	1.19		Kakadu - TIFF to JP2:	1.19		GeoJasper - TIFF to JP2:	1.19	
Keep status quo:	1.19															
ImageMagick - TIFF to JP2:	1.19															
GraphicsMagick - TIFF to JP2:	1.19															
Kakadu - TIFF to JP2:	1.19															
GeoJasper - TIFF to JP2:	1.19															

Take a look...

[www.ifs.tuwien.ac.at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)

# Scanned books requirements



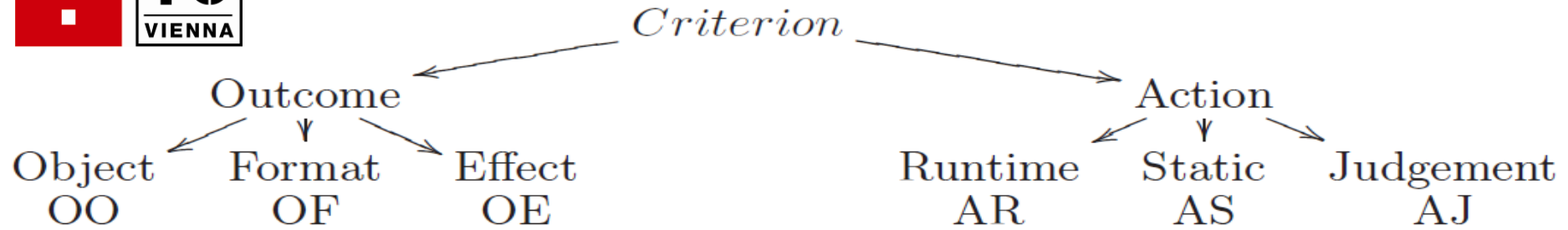


# Decision criteria and evaluation

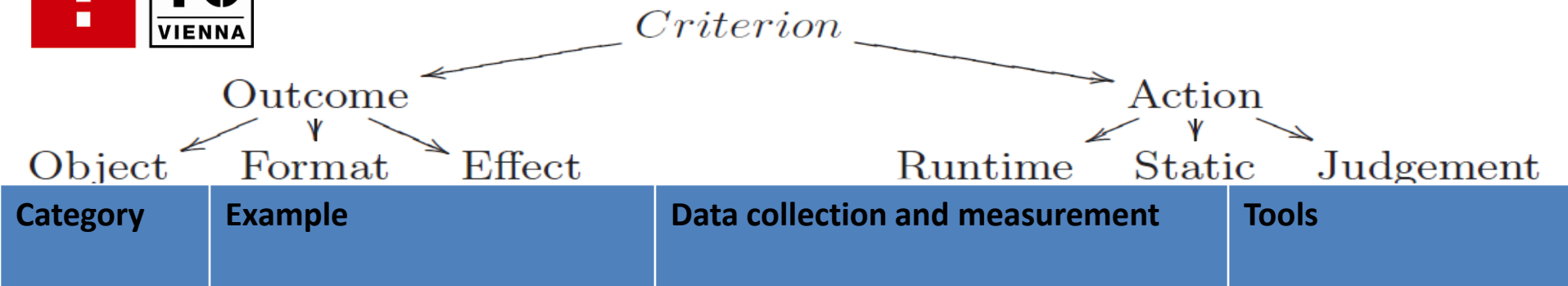
- Problems
  - Manual evaluation is very effort intensive
  - Need for sharing knowledge and comparing experiences
- Decision criteria
  - Analysis of >600 criteria specified in 12 case studies
  - A taxonomy of criteria
  - Measurement devices for each category
  - Integration with Plato through an extensible measurement framework
- Types of criteria
- Quantitative analysis of measurement coverage
- Quantitative analysis of decision criteria *impact*



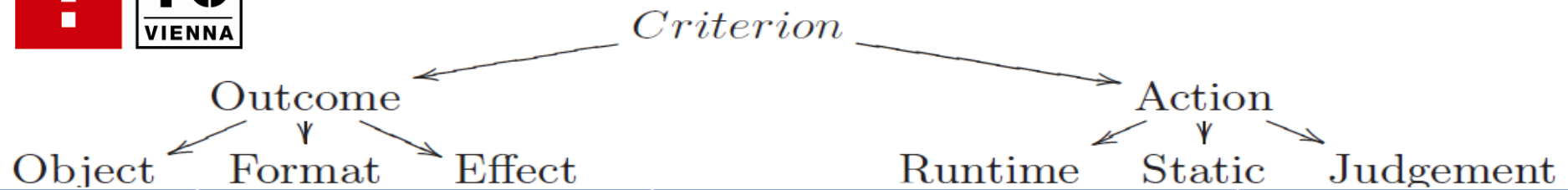
# What to measure?



# How to measure?

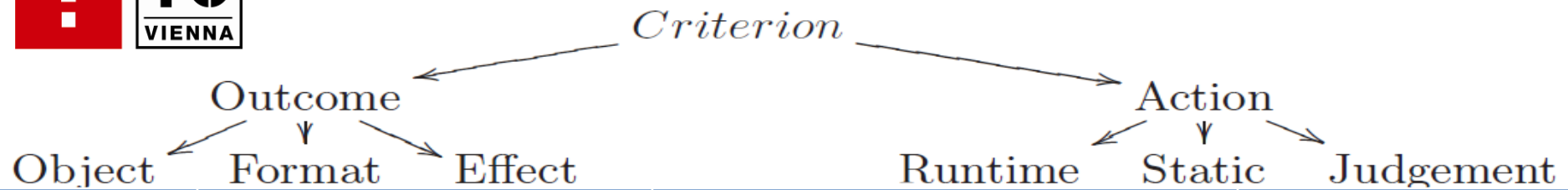


# How to measure?



Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, image comparison...

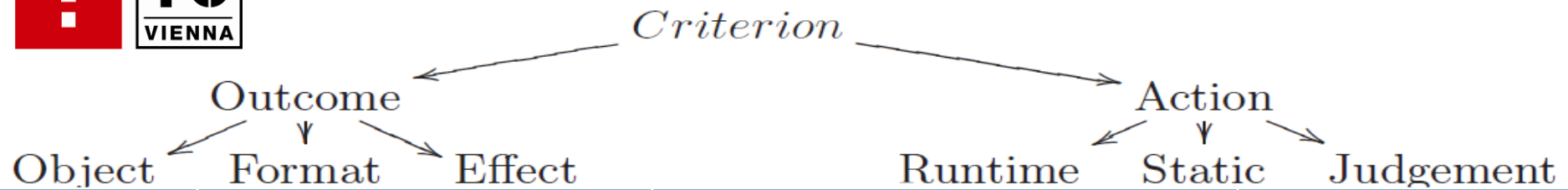
# How to measure?



Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	<b>Format is ISO standardised</b>	<b>Measurements of the output, Trusted external data sources</b>	<b>DROID, PRONOM, UDFR, P2</b>

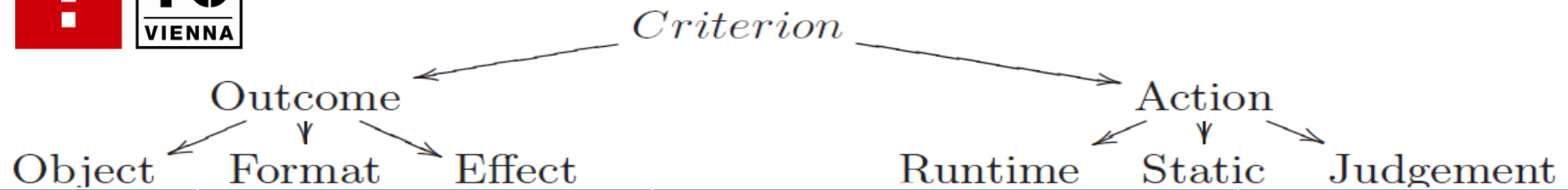


# How to measure?



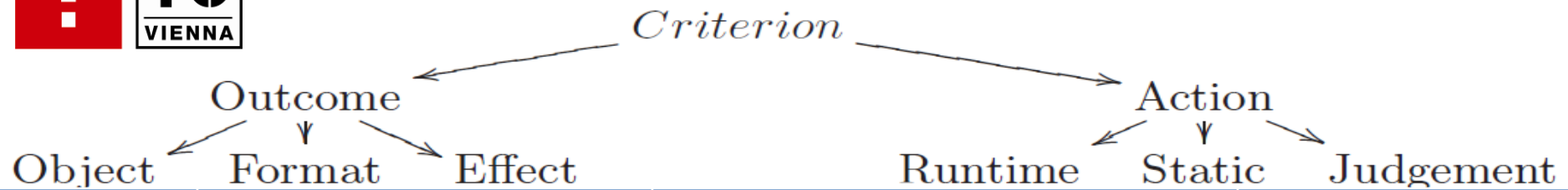
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM, UDFR, P2
Outcome effect	<b>Annual bitstream preservation costs (€)</b>	<b>Measurements of the output, external data sources, models (LIFE)...</b>	<b>LIFE model</b>

# How to measure?



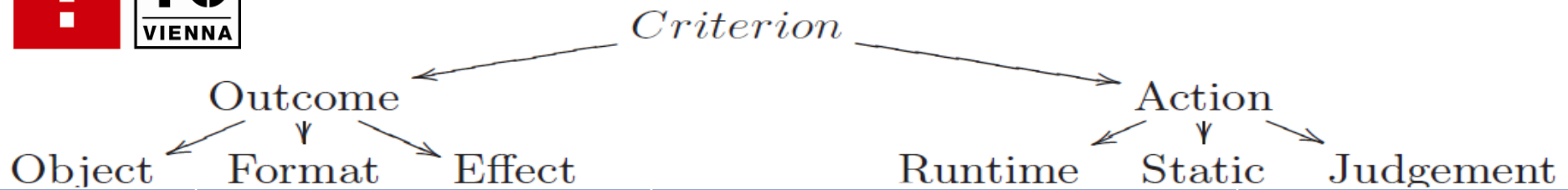
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM, UDFR, P2
Outcome effect	Annual bitstream preservation costs (€)	Measurements of the output, external data sources, models (LIFE)...	LIFE model
Action runtime	<b>Throughput (MB per millisecond), Memory usage</b>	<b>Measurements taken in controlled experimentation</b>	<b>MiniMEE</b>

# How to measure?



Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM, UDFR, P2
Outcome effect	Annual bitstream preservation costs (€)	Measurements of the output, external data sources, models (LIFE)...	LIFE model
Action runtime	Throughput (MB per millisecond), Memory usage	Measurements taken in controlled experimentation	MiniMEE
Action static	<b>License costs per CPU (€), Open Source License</b>	<b>Trusted external data sources, manual evaluation, sharing</b>	<b>UDFR, Pronom, P2, manual</b>

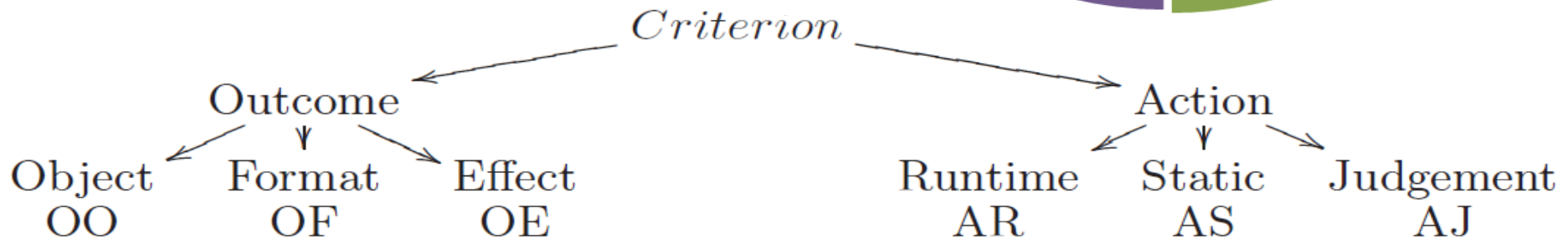
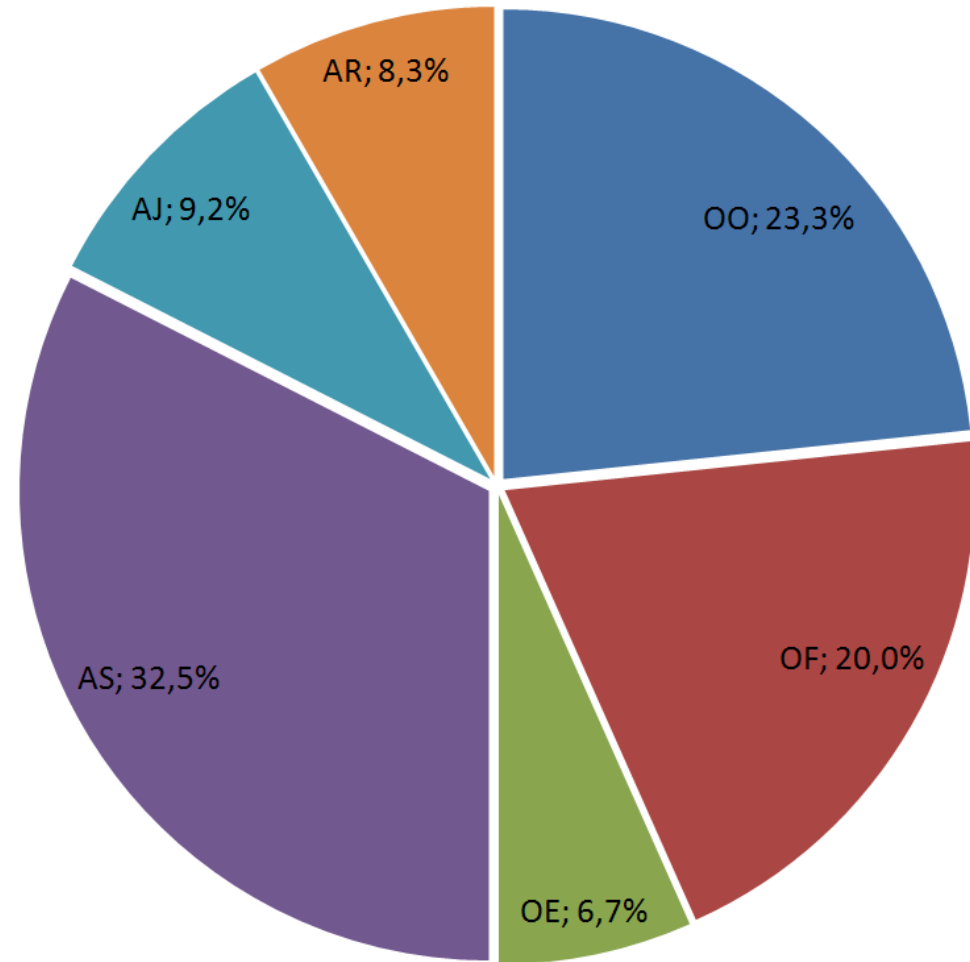
# How to measure?



Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM, LoC format site, UDFR, P2
Outcome effect	Annual bitstream preservation costs (€)	Measurements of the output, external data sources, models (LIFE)...	LIFE model
Action runtime	Throughput (MB per millisecond), Memory usage	Measurements taken in controlled experimentation	MiniMEE
Action static	License costs per CPU (€), Open Source License	Trusted external data sources, manual evaluation, sharing	UDFR, P2, manual
<b>Action judgement</b>	<b>Technical interoperability, configuration flexibility</b>	<b>Manual judgement, sharing</b>	

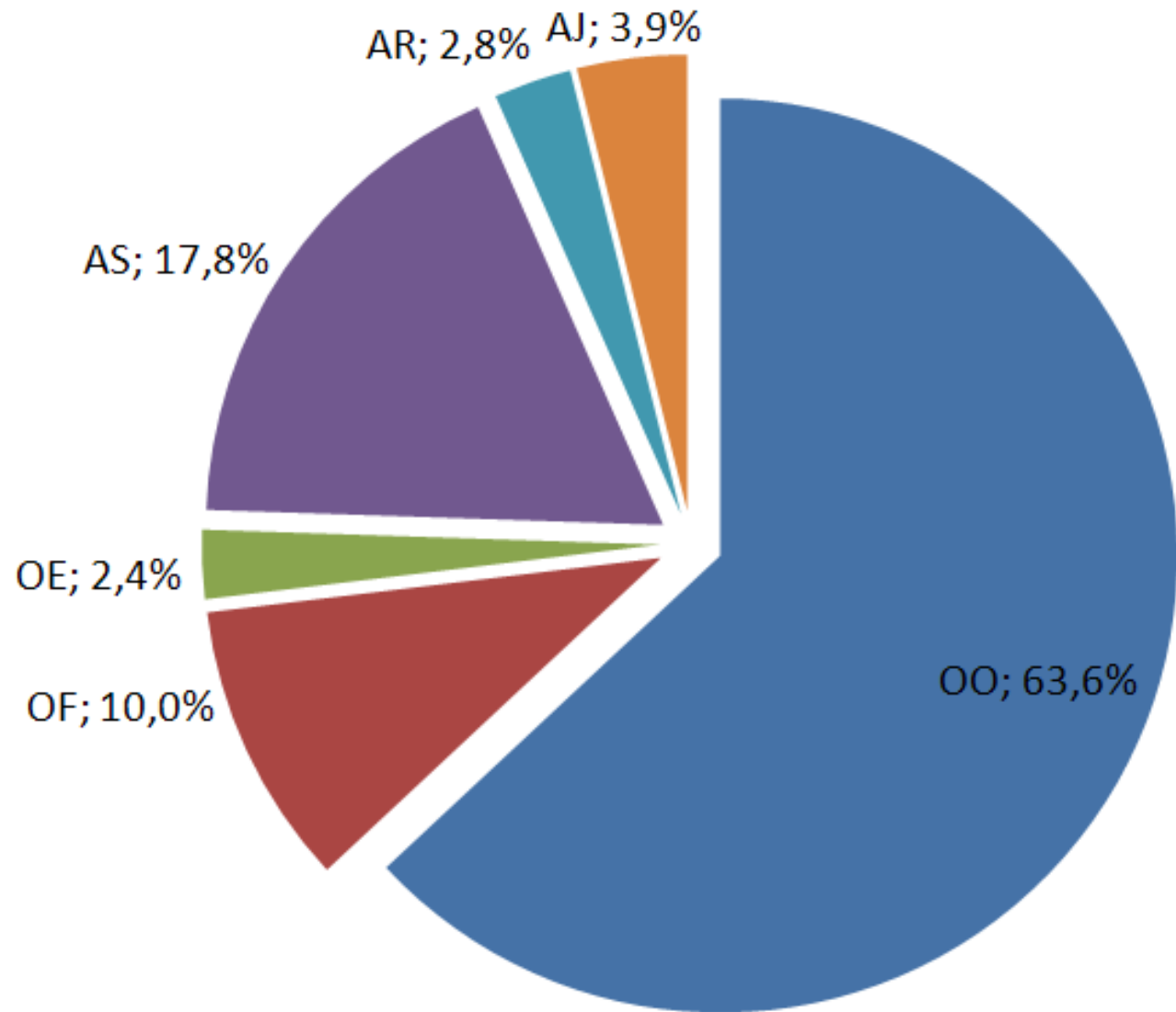
# Case studies

- Distribution in four case studies on scanned images



# Case studies

- Distribution in thirteen cases on various types of content





# Measurement: Where are we now?

- The good news
  - We know what to measure
  - We have tools for measuring aspects
  - We can measure simple properties reliably
  - We can assess the *impact* of a property
- The not so good news
  - Confidence in the measures varies
  - Coverage of measures depends on the objects' formats
- The bad news
  - Many complex properties cannot be measured (yet)
  - Universal solutions for Quality Assurance are not working well
  - Piece by piece, step by step is the way to go

# Collect

- Preservation plans
- Decision criteria

A method and tool  
for decision criteria  
impact  
assessment

# Map

- Significant properties models
- ISO SQUARE Software quality attributes
- Format properties

# Categorise

- Specify uniquely identified criteria
- Categorise all case study decision criteria

# Develop

- Define and implement impact factors
- Visual analysis tools

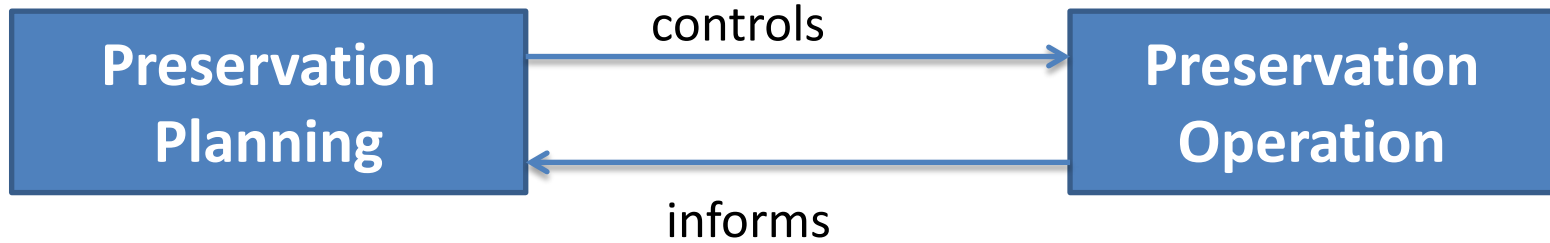
# Analyse

- Impact factors for criteria
- Impact factors for sets of criteria

# The method, the tool, the services

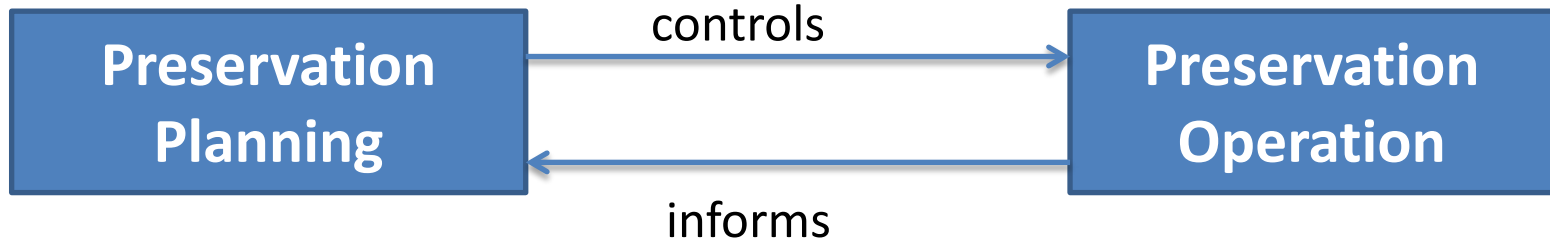
- Method is very generally applicable
  - From computer games to scanned images
  - From databases to born-digital art
  - From private photographs to national heritage institutions
- Tool support varies
  - Degree of automation strongly dependent on content and preservation actions
  - Manual evaluation is always possible
- Integrated services
  - Action services may or may not work on specific content
  - Failure of a service simply means that the service is not suitable
  - Planning and thorough evaluation is important
- How to integrate this into an *organisation*?

# Planning and Operations



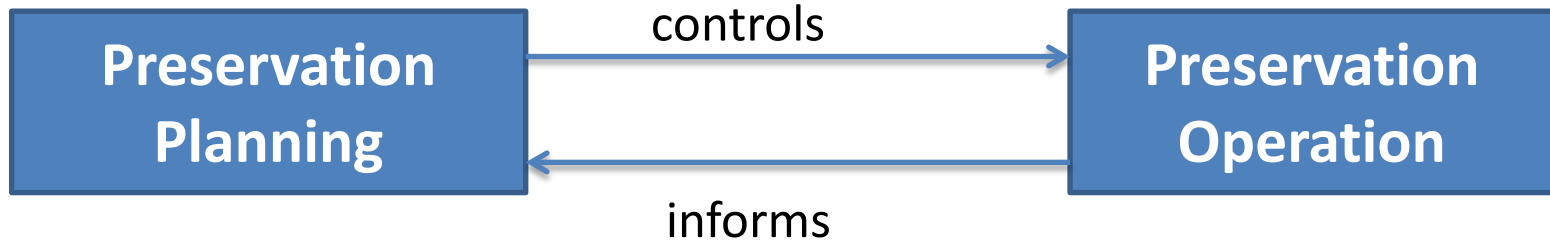
Preservation Planning	Preservation Operation
	Run operations and report on them
	<ul style="list-style-type: none"><li>•Analyze content</li><li>•Execute preservation actions</li><li>•Ensure adequate provenance trail</li><li>•Handle preservation metadata</li><li>•Conduct Quality Assurance</li><li>•Provide reports and statistics</li></ul>

# Planning and Operations



Preservation Planning	Preservation Operation
Monitor and control operations	Run operations and report on them
<ul style="list-style-type: none"><li>•Influencers and Decision making</li><li>•Options diagnosis</li><li>•Specification and delivery</li><li>•Monitoring</li></ul>	<ul style="list-style-type: none"><li>•Analyze content</li><li>•Execute preservation actions</li><li>•Ensure adequate provenance trail</li><li>•Handle preservation metadata</li><li>•Conduct Quality Assurance</li><li>•Provide reports and statistics</li></ul>

# Planning and Operations



Preservation Planning	Preservation Operation
Monitor and control operations	Run operations and report on them
<ul style="list-style-type: none"><li>•Influencers and Decision making</li><li>•Options diagnosis</li><li>•Specification and delivery</li><li>•Monitoring</li></ul>	<ul style="list-style-type: none"><li>•Analyze content</li><li>•Execute preservation actions</li><li>•Ensure adequate provenance trail</li><li>•Handle preservation metadata</li><li>•Conduct Quality Assurance</li><li>•Provide reports and statistics</li></ul>
<i>“Migrate this set of images (in TIFF-5) to JP2 using ImageMagick 6.3 with parameters a,b,c”</i>	<ul style="list-style-type: none"><li>•Analyze original</li><li>•Migrate, analyse output</li><li>•Conduct quality assurance</li><li>•Provenance, metadata, Reporting</li></ul>



# What are the prerequisites of planning?

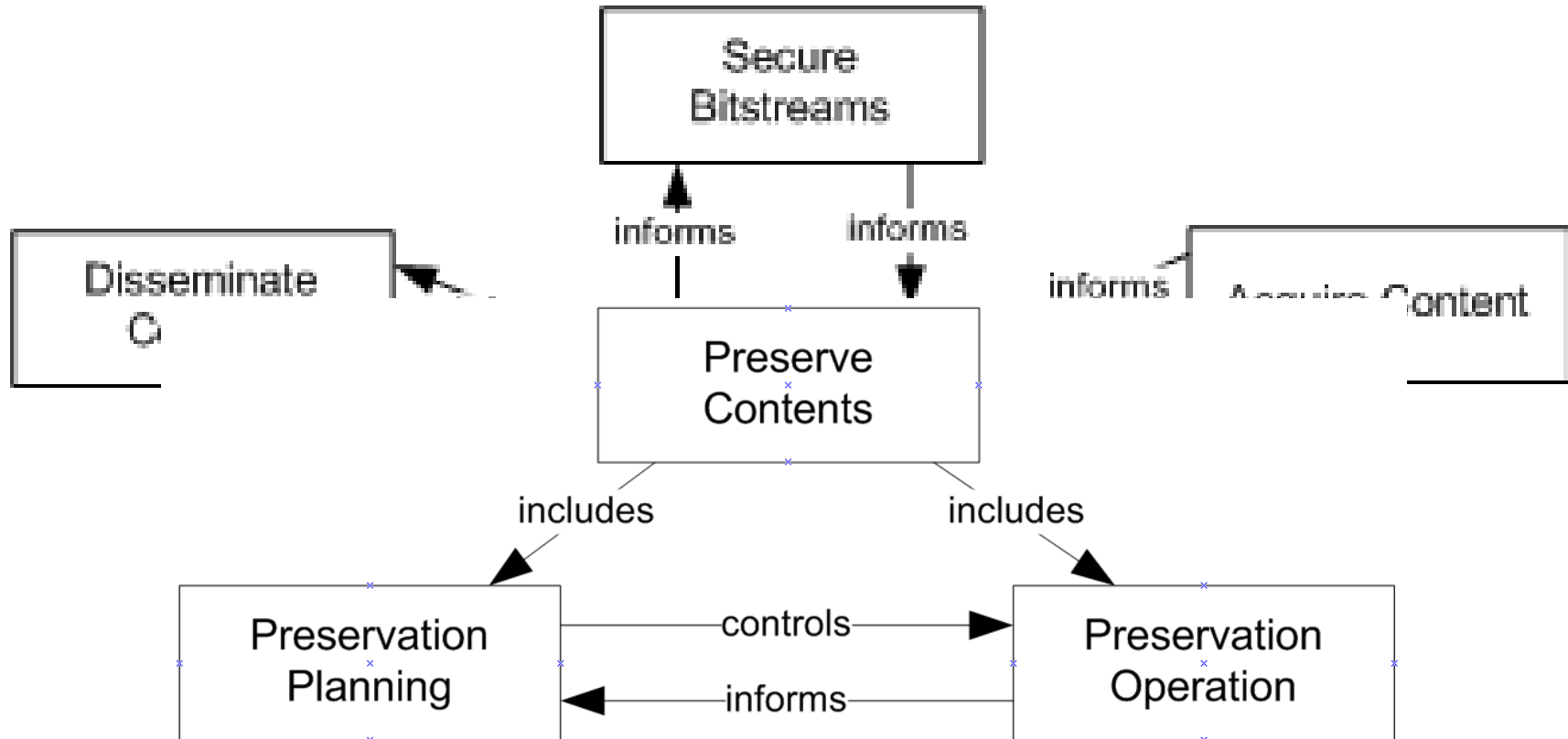
- Clear and concise documentation of the organization
  - Constraints, Drivers and Goals
  - Responsibilities
  - Infrastructure and technical capabilities
  - Cost structures
- Understanding of the decision space
  - Properties of the content
  - Requirements of the stakeholders
  - Available options
  - Relationship between ends and means
  - Relationship between strategies and operations
- Understanding of decision factors and their evaluation

# Who is responsible for planning?

- A full understanding of the planning *role* has yet to be formed
- Combination of expertise and skills required
  - Understanding of business goals to achieve
  - Understanding of organizational environments and processes
  - In-depth knowledge of technical intricacies
- Not all planning activities should be carried out by the same person or role in an organization
- Preservation Planning needs to take place on an operational level
- There are multiple levels of decision making

# Digital Preservation Capabilities

**Preserve Contents** is the ability to maintain content authentic and understandable to the defined user community over time and assure its provenance.



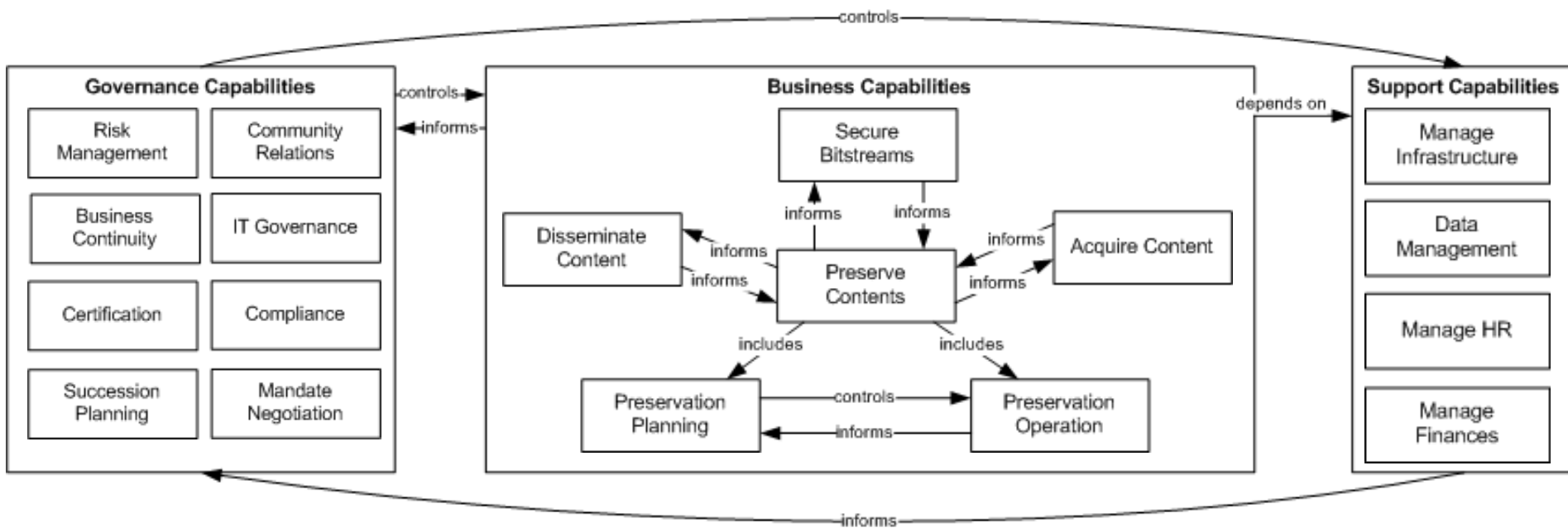
# Digital Preservation Capabilities

**Preserve Contents** is the ability to maintain content authentic and understandable to the defined user community over time and assure its provenance.

Governance  
Capabilities

Business  
Capabilities

Support  
Capabilities



# Some Conclusions

- The planning method and Plato are broadly applicable, but
  - need clear positioning in a well-defined organizational context
  - require clear understanding of the “terms and conditions”
  - Required expertise and skill set needs to be clarified
  - Tool support varies according to content type and action
  - Automation and Scalability
- Integration into an organization's processes
  - understanding of processes, influences, interdependencies



# Current challenges

- From one-off decision making to continuous management
- Scalability for operations
  - Content analysis for 1700TB of web content
  - Quality Assurance for conversion of 100 million Word documents
  - ....
- Scalability for control and monitoring
  - Define preservation plans for a web archive
  - Evaluate preservation actions for 120 different formats
  - Monitor technological shifts and trends in the web
  - ...

# Future Perspectives

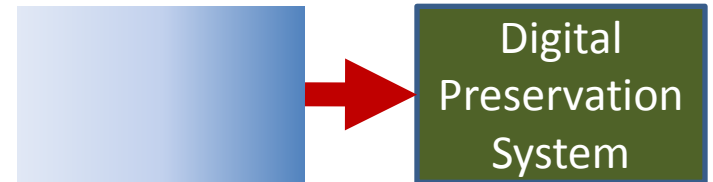
- Shift in stakeholders
  - From Cultural heritage and space data systems to eScience, HEP, big data, businesses and the social web
  - Increasing diversity
  - Entirely new markets, solutions, and jobs
- From information to processes and systems
- We need a holistic view
- Governance, Risk and Compliance



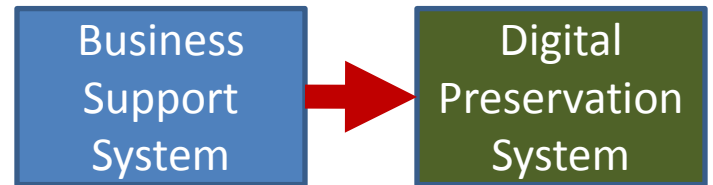
# Information systems and preservation

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!*



# Digital Longevity

- Numerous reference models, frameworks and concepts
  - OAIS and trust: TRAC, RAC (ISO 16363), NESTOR...
  - Records Management: ISO 15489, MoReq 2010...
  - Risk: DRAMBORA...
  - Planning: PLATO, PLATTER
  - Economics: BRTF, LIFE....
- Yet, we still lack a holistic view
  - Maturity of the field is unclear and evolving
  - Integration into Information Systems and Information Technology fields is unclear
  - How does Digital Preservation relate to, e.g., IT Governance? Enterprise Risk Management? Governance, Risk and Compliance?
  - How can we assess and improve organizational capabilities?
  - How can we deal with non-repository scenarios?

