Preservation Decisions: Terms and Conditions Apply

Challenges, Misperceptions and Lessons Learned in Preservation Planning

Christoph Becker, Andreas Rauber

ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011)

Ottawa, ON, Canada

June 14, 2011

SCAPE

SCAlable Preservation Environments

FACULTY OF !NFORMATICS

Digital Preservation decisions

- Digital Preservation arises from change
  - organizational, users, technical, legal, contextual…
- Alignment of technology and business
  - Continuum between business and technology
  - User requirements vs. IT operations
  - Technology obsolescence vs. technological opportunities
- Reconciling Conflicts
  - between ends and means
  - between strategy and tactics
- Core decision: How to preserve content information
  - *Preservation action*: A concrete action (usually implemented by a software tool) performed on content in order to achieve preservation goals.

Preservation Planning

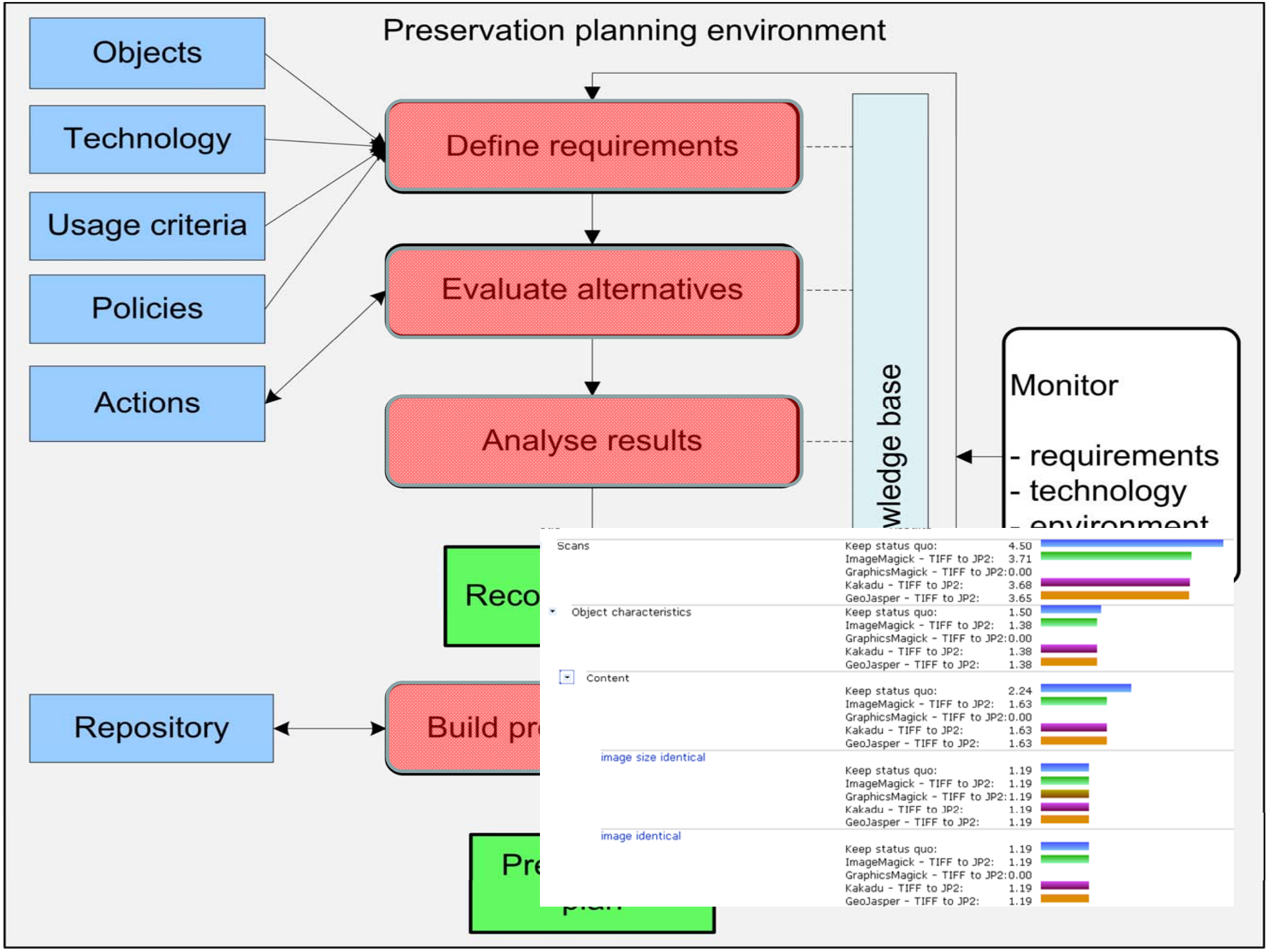
- ***Preservation Planning:*** is the ability to monitor, steer and control the preservation operation to meet preservation goals and manage obsolescence threats
- Systematic evaluation of candidate actions against scenario-specific requirements in a standardized, repeatable workflow using controlled experimentation on sample content
- 'A ***preservation plan*** defines a series of preservation actions to be taken by a responsible institution to address an identified risk for a given set of digital objects or records (called collection).'
- Plato: The Planning Tool - www.ifs.tuwien.ac.at/dp/plato
  - Growing user community
  - Series of case studies and productive decisions
- From planets to SCAPE SCAlable Preservation Environments …. 2011-2014

Outline

- Preservation Planning
  - Planning method and Plato
  - Case studies
  - Decision criteria: What to measure and how
- Lessons Learned
  - Necessity, Scope, Costs, Benefits
  - Prerequisites and Critical Success Factors
  - Common misperceptions
- Observations and Future Challenges

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 (ePrints; RODA, eSciDoc,…)
- Plato guides, validates and documents planning
- Automation: Reduce manual effort


Preservation planning environment
Objects
Technology
Usage criteria
Policies
Actions
Define requirements
Evaluate alternatives
Analyse results
wledge base
Monitor
- requirements
- technology
Repository
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
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
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

Case studies

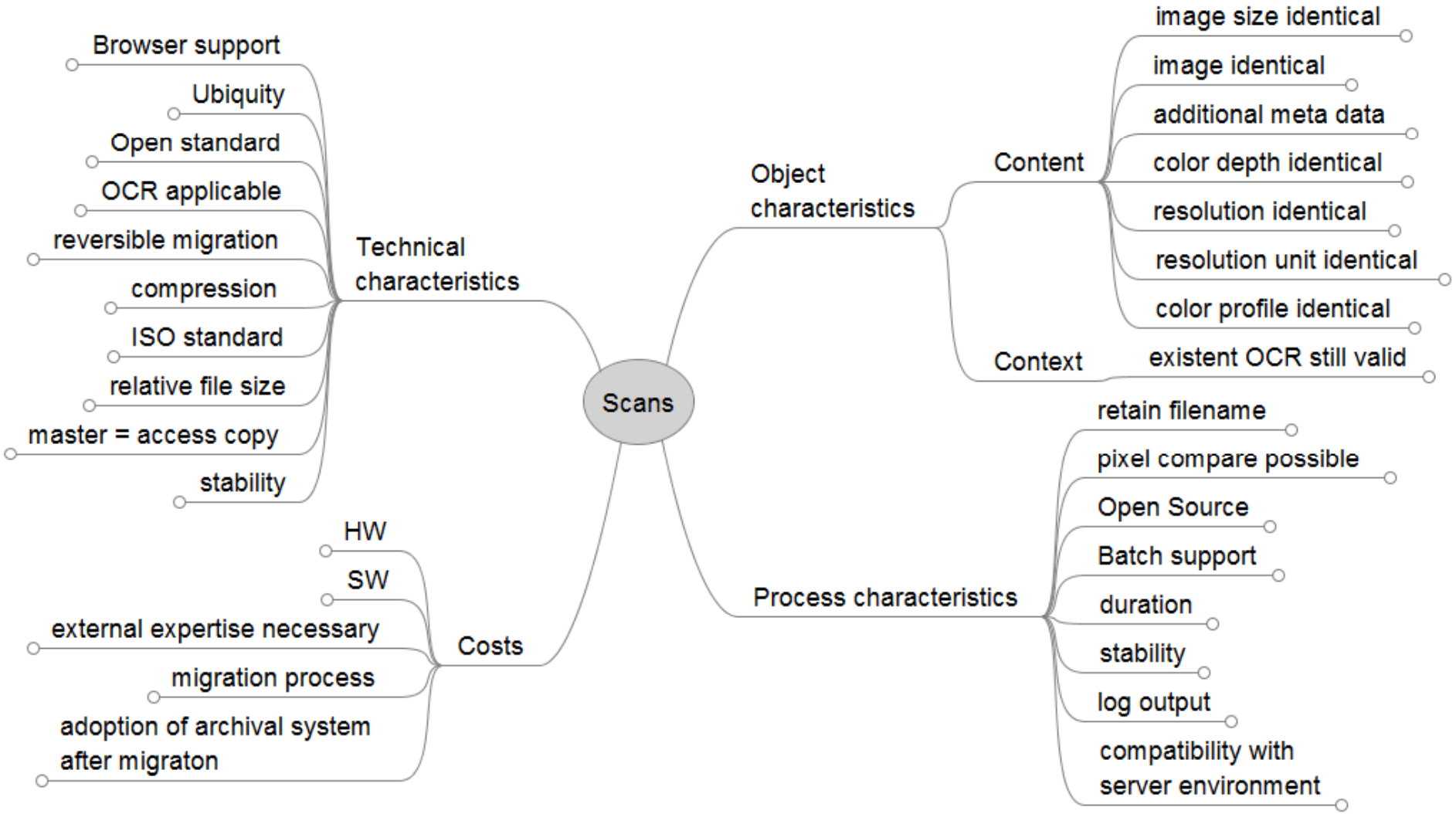
- Case studies conducted with Plato
  - Electronic documents
  - Interactive art
  - Console video games
  - **Scanned images**
  - Relational databases
  - Interactive art
  - Computer games
  - Born-digital photographs
  - Documents
  - Emails
  - …
  - And: Bitstream preservation (Zierau et al., IPRES 2010)

Four cases, three solutions: Scanned images

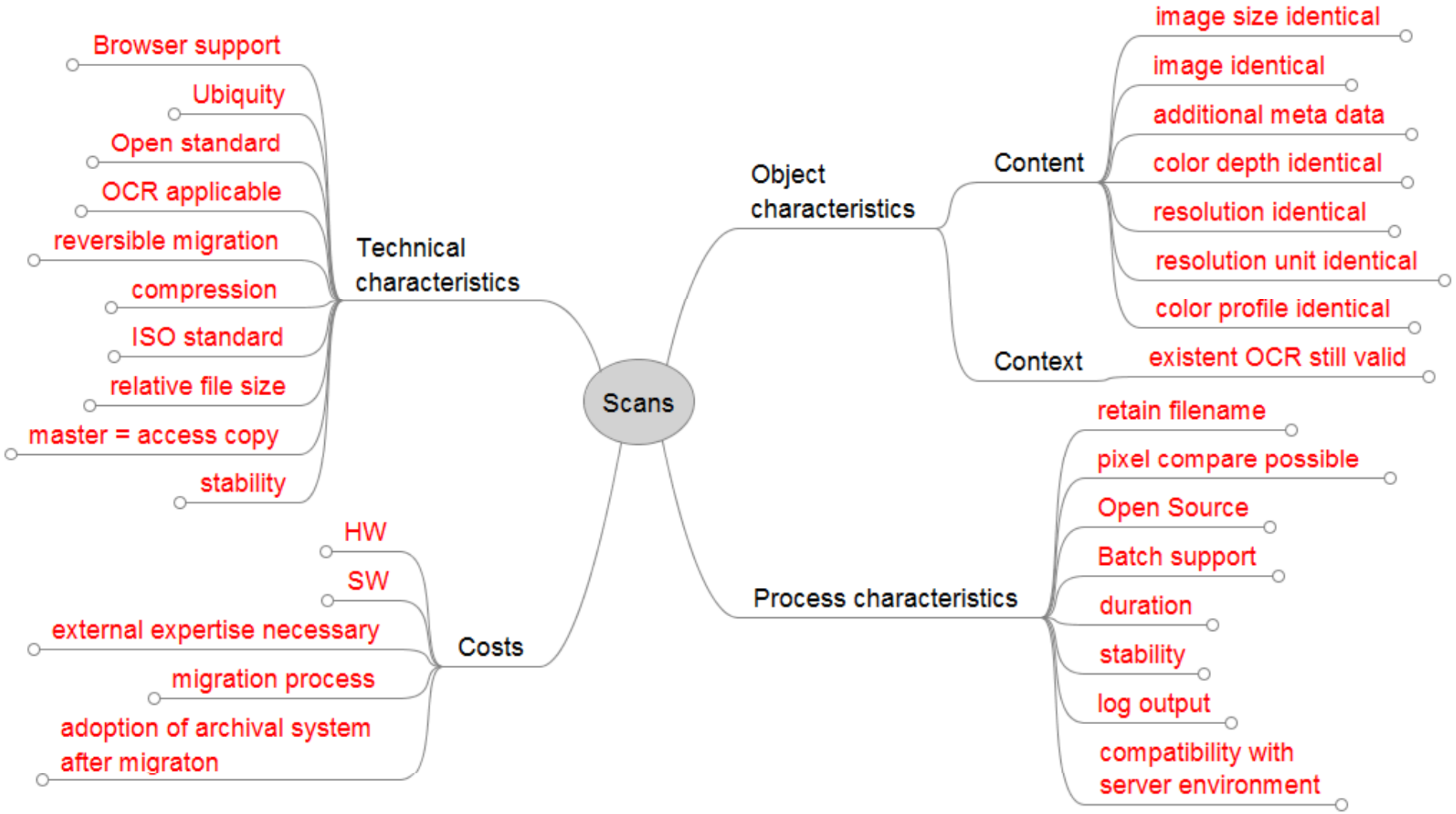
- Bavarian State Library, 72TB TIFF6: *Leave and monitor*
- British Library, 80TB TIFF5: *Migrate to JP2 (ImageMagick)*
- Royal Library of Denmark, ~10.000 aerial photographs in TIFF6: *Leave and monitor*
- State and University Library Denmark, scanned yearbooks in GIF: *Migrate to TIFF 6*

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



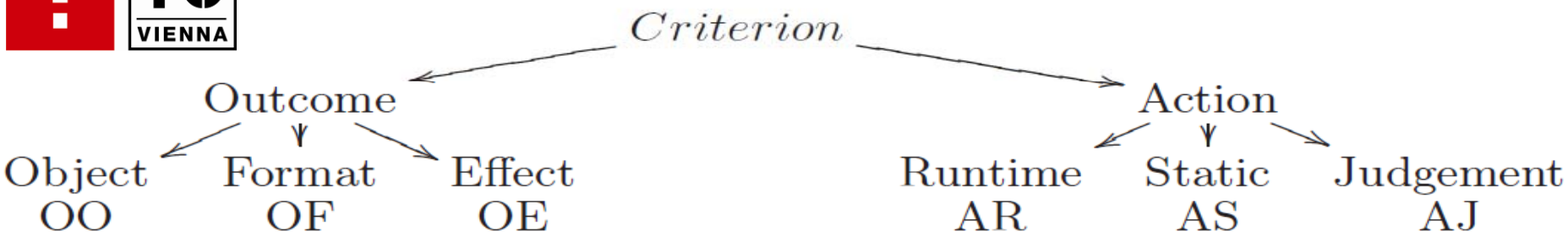
Scanned books requirements



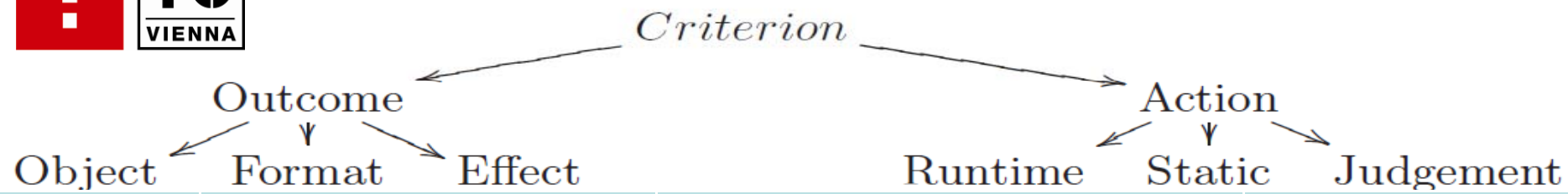
Addressing the evaluation gap

- 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
- Quantitative analysis of measurement coverage

What to measure?

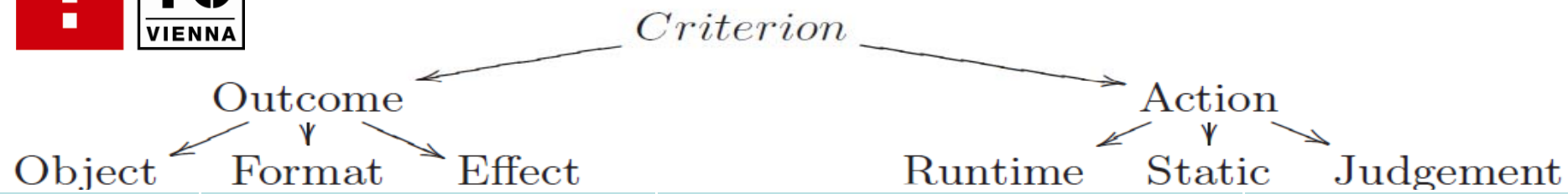


How to measure?



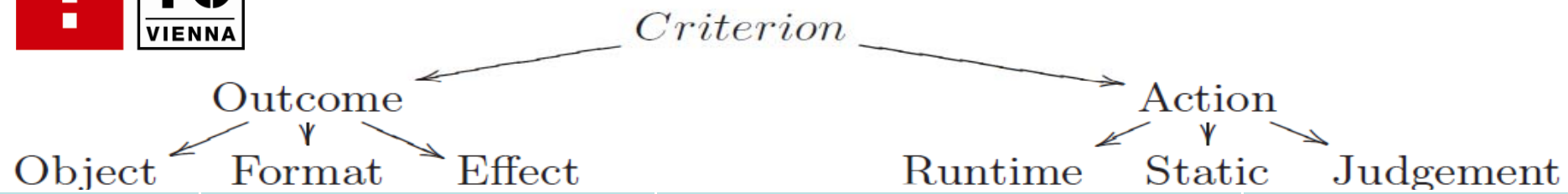
Category	Example	Data collection and measurement	Tools

How to measure?



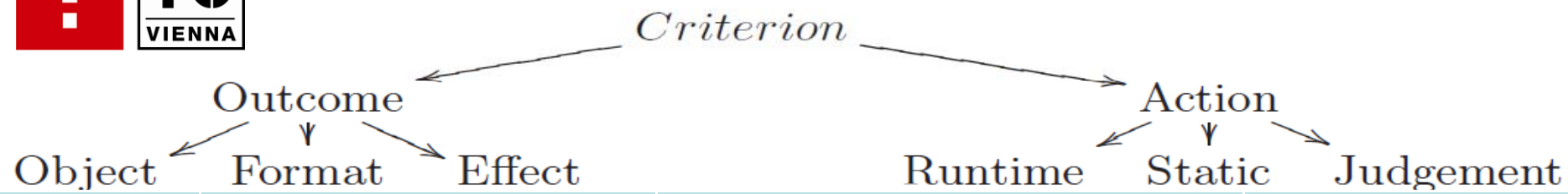
Category	Example	Data collection and measurement	Tools
Outcome Object	**Image pixelwise identical Footnotes preserved**	**Measurements of output and input, comparison**	**FITS, JHove, ImageMagick...**

How to measure?



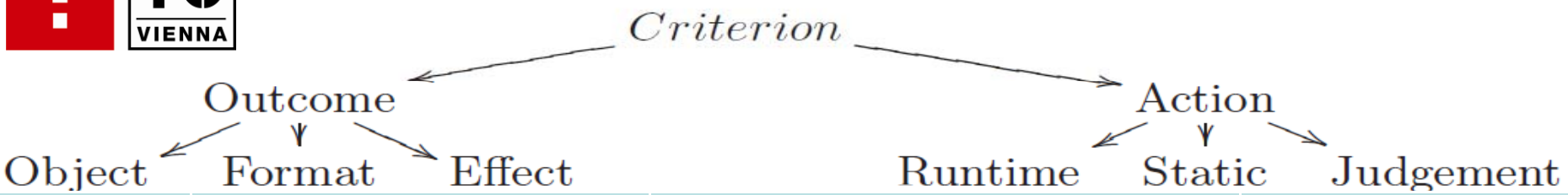
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	**Format is ISO standardised**	**Measurements of the output, Trusted external data sources**	**DROID,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

How to measure?



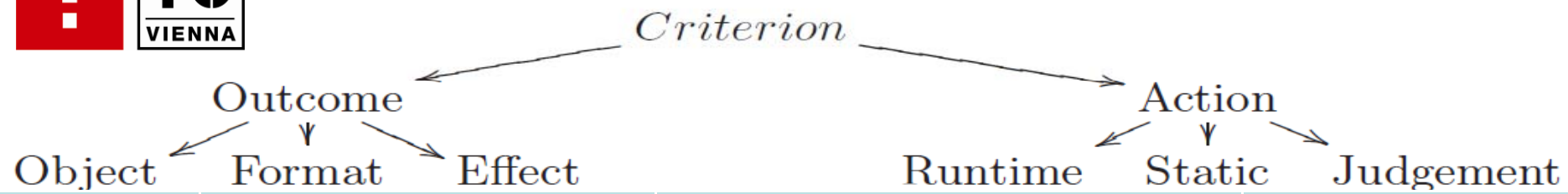
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM,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

How to measure?



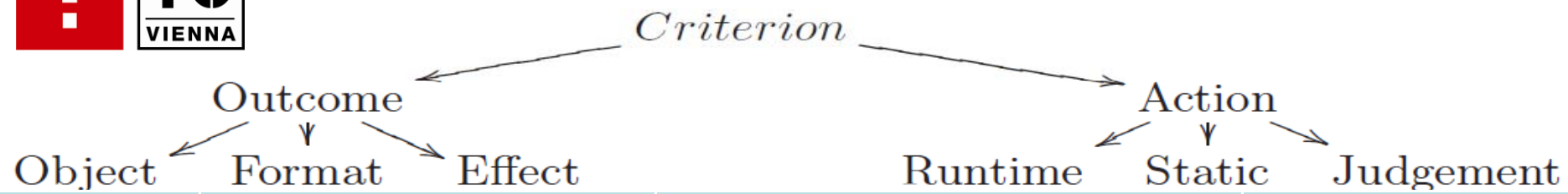
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM,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

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	**License costs per CPU (€), Open Source License**	**Trusted external data sources, manual evaluation, sharing**	**UDFR, Pronom, P2, manual**

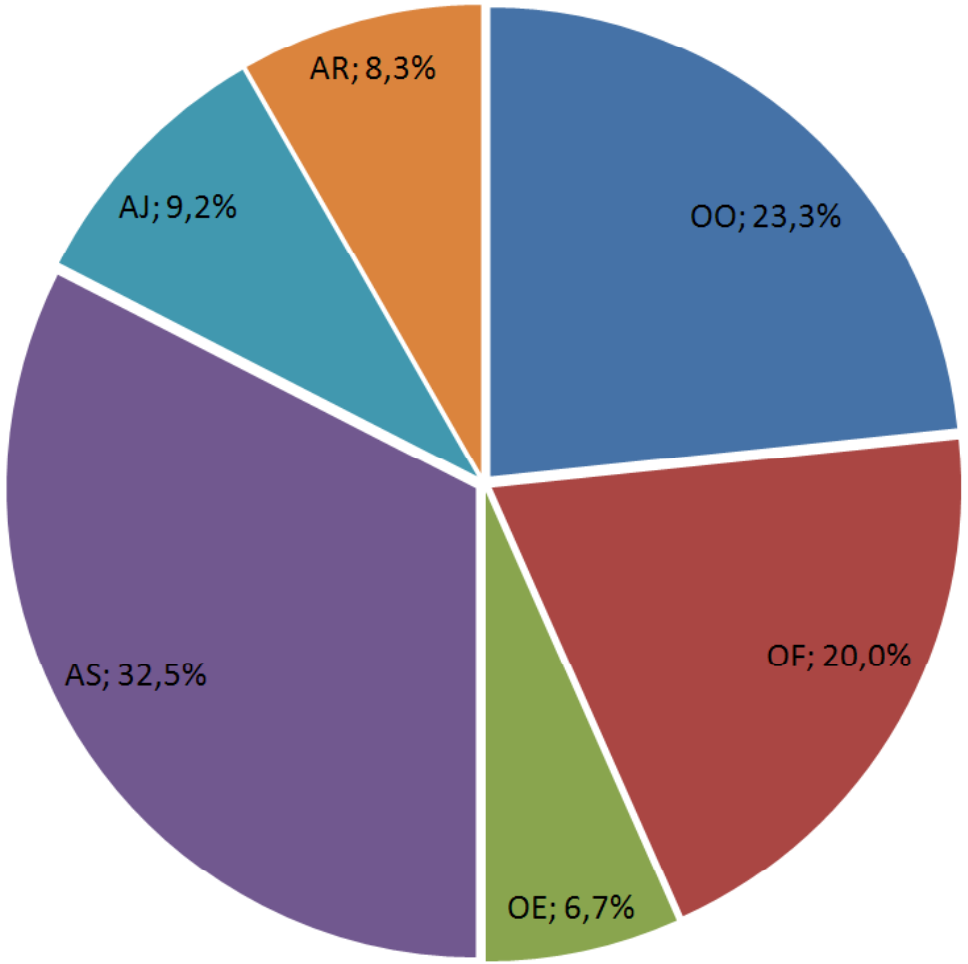
How to measure?

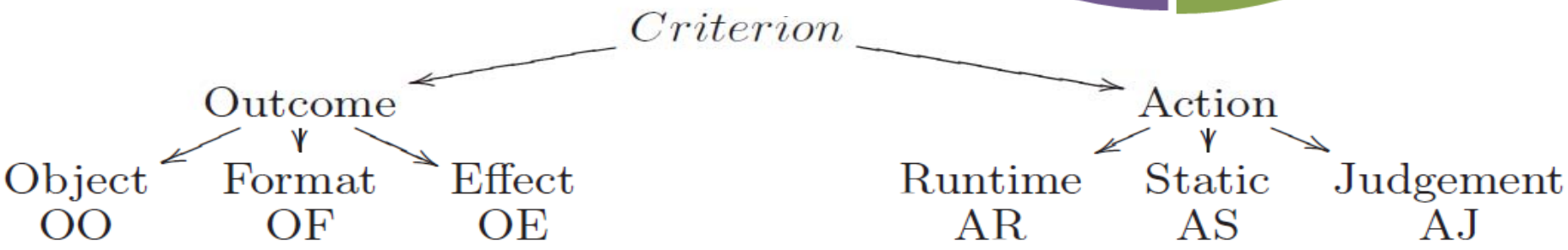


Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM, 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	License costs per CPU (€), Open Source License	Trusted external data sources, manual evaluation, sharing	P2, manual
Action judgement	**Configuration interface usability**	**Manual judgement, sharing**	

Case studies

- Distribution in four case studies on scanned images





Measurement: Where are we now?

- The good news
  - We know the distribution of criteria in the taxonomy
  - We know what we need to measure
  - We have approaches to measuring things
  - We can measure simple properties reliably
- The not so good news
  - Confidence in the measures varies
  - Coverage of measures depends on the objects' formats
  - We do normally not know much about the *impact* of a property
- Bad news
  - Many complex properties cannot be measured yet
  - Universal solutions for QA are not working well
  - Piece by piece, step by step is the way to go

Is all this necessary?

- Challenges when 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
- Trust requires evidence
  - Trust has to be evaluated in a realistic context
  - Controlled experimentation, repeatable documentation, and scenario-specific requirements assessment

Lessons learned

1. Costs and benefits of planning
2. Prerequisites of planning
3. Responsibilities
4. Requirements and Assessment
5. The method, the tools, and the services

What are the costs and benefits of planning?

- Primary cost drivers for the planning activity
  - Maturity of organizational framework:
    Constraints, goals, drivers and responsibilities
  - Degree of familiarity with the planning approach
  - Technical complexity of the content to be preserved
  - Technical proficiency of the staff assigned to do planning
- Effort
  - First run generally effort-intensive: Learning curve, lack of context
  - Subsequent activities significantly easier and faster
- Return on Investment
  - Hard to quantify
  - … but shouldn't we rather ask: What are the costs of NOT planning?
    - This is quite easy to quantify

What are the prerequisites of planning?

- Clear and concise documentation of the organization
  - Constraints
  - Drivers
  - 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

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

What is important?

- 3 key levers influencing the decision outcome
  1. Requirements definition
  2. Utility functions
  3. Importance weighting
- Weighting requirements
  - Assigns relative importance factors on all level of the tree
  - Low level changes in relative importance have little influence
  - Criteria often have a total weight of 1-5%
- Weighting vs. utility function
  - Key effects of criteria with low weight: Acceptance or rejection
  - Output range of utility function may include 0.0
  - Utility function is much more critical on the level of criteria
- Measurements vs. Assessment

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

Some Conclusions

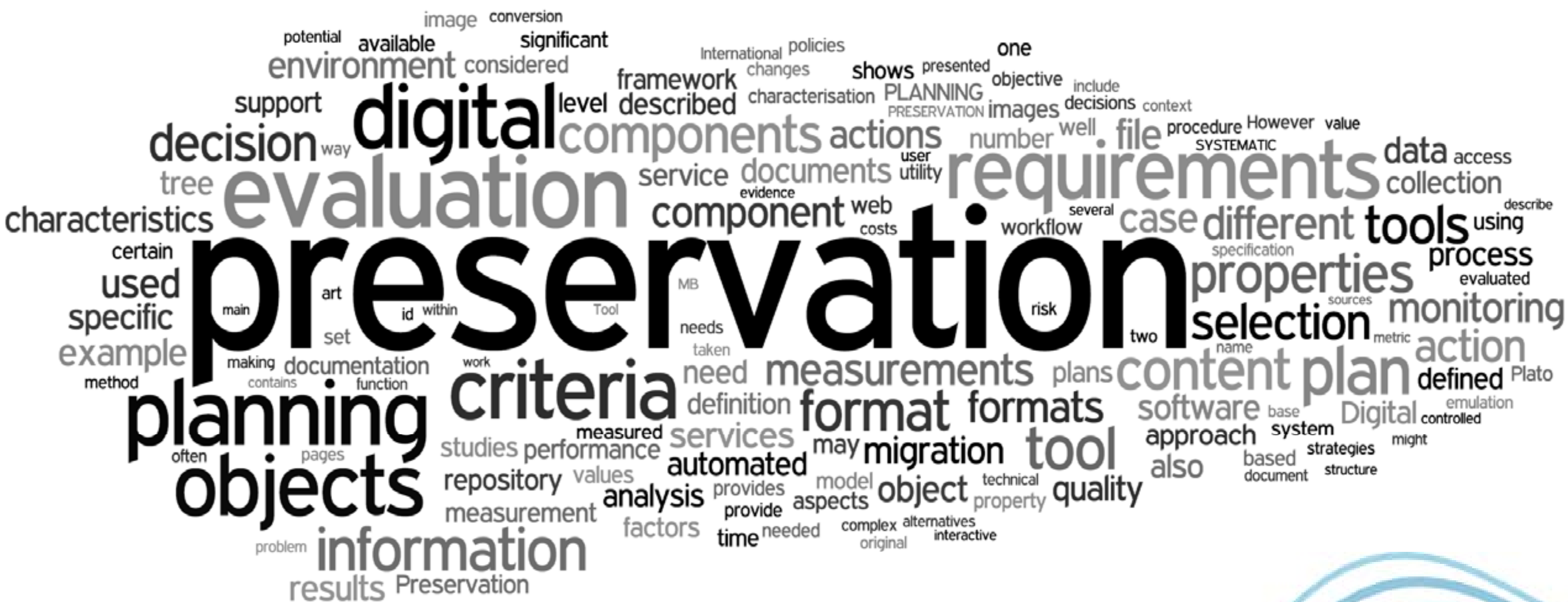
- The planning method and Plato are broadly applicable, but
  - need clear positioning in a well-defined organizational context
  - require clear understanding of the “terms and conditions”
  - Required expertise and skill set needs to be clarified
  - Tool support varies according to content type and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

- Integration into an organization's processes
  - understanding of processes, influences, interdependencies
- Governance, Risk and Compliance: We’d like to see…
  - An integration of DP into IT Governance
  - An integration of DP into Enterprise Risk Management
  - A better understanding of the relationship with Governance, Risk and Compliance

Thank you for your attention!

Questions?

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



SCAPE
SCAlable Preservation Environments

FACULTY OF !INFORMATICS