Schedule

(1) Introduction

Who are we?

(2) Digital Preservation and Preservation Planning:

What is a Preservation Plan and why do we need it?

(3) Preservation Planning:

How do we build a Preservation Plan? How does Plato help?

(4) Exercise (& Coffee Break):

Which objectives should we meet for preserving scanned images?

(5) Decision Criteria in Digital Preservation:

How do we make goals and objectives measurable?

(6) Lessons learned:

What are the most common misconceptions?

First: Some case studies

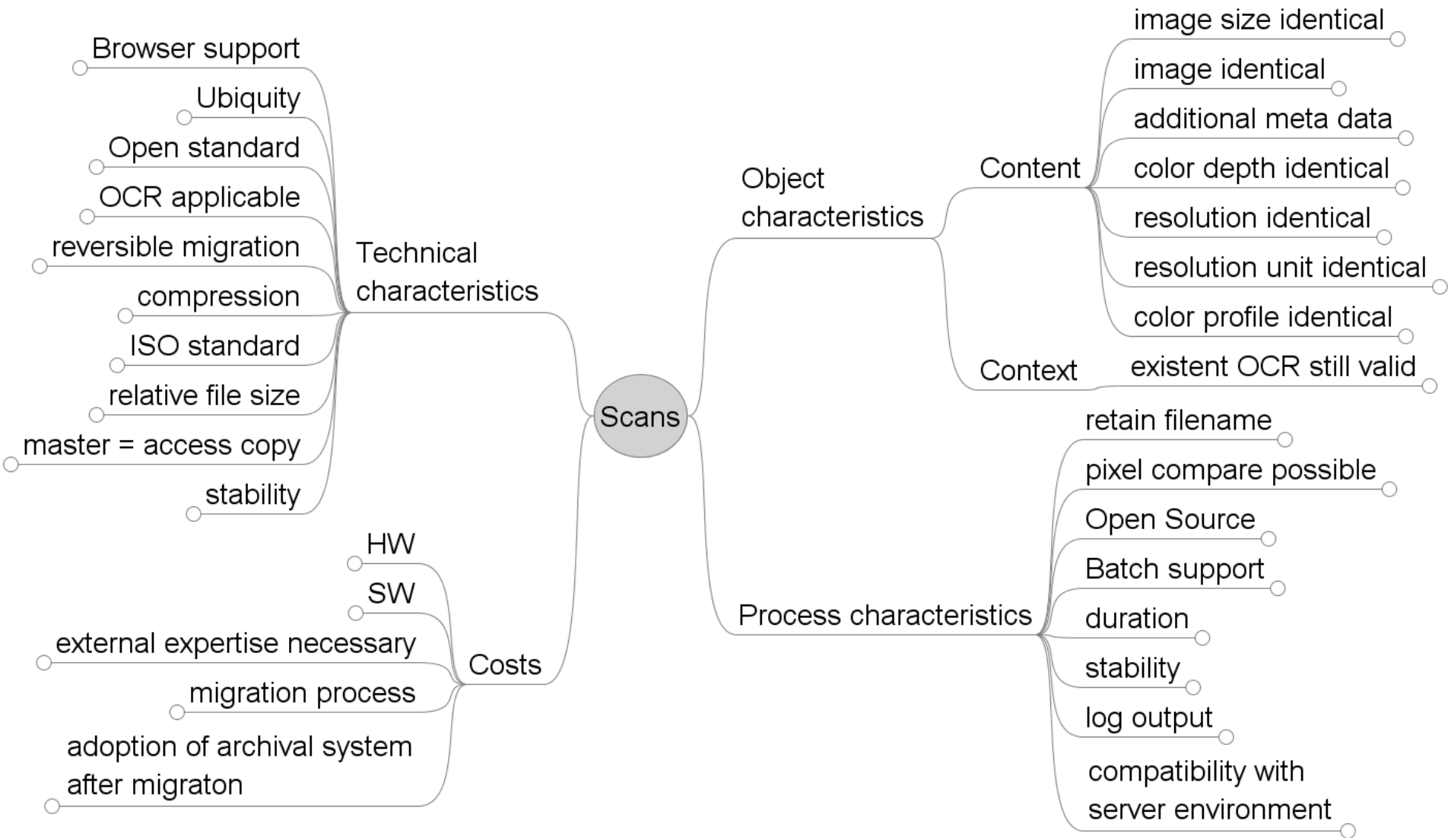
- Case studies conducted with Plato
  - Electronic documents
  - Interactive art
  - Console video games
  - **Scanned images**
  - Relational databases
  - Interactive art
  - Computer games
  - Born-digital photographs
  - Documents
  - Emails
  - and: Bitstream preservation
- *… see www.ifs.tuwien.ac.at/dp/plato for publications*

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, Standardisation
Aerial photographs in TIFF6	Leave unchanged and monitor	Lack of JP2 browser support, Process costs

Scanned books requirements



…. and in Plato?

Scanned books results

Results: Weighted multiplication

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

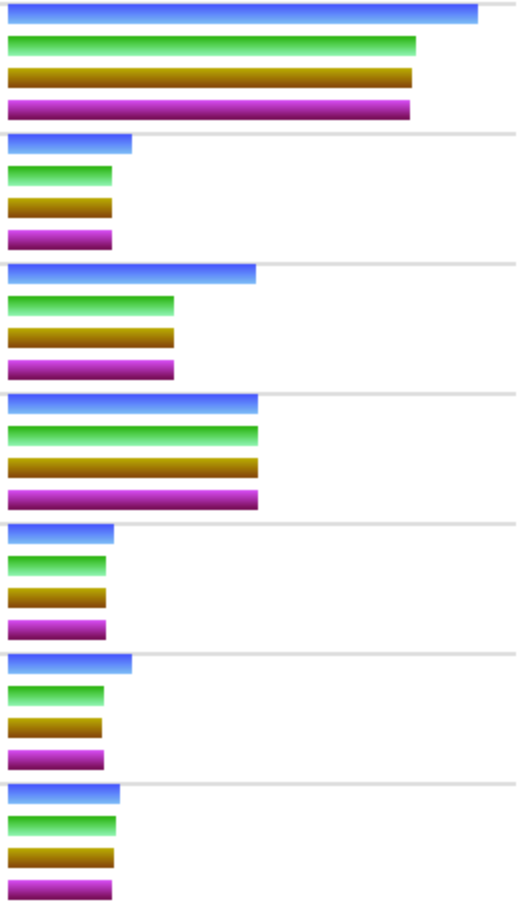
Node	Results	
Scans	Keep status quo:	4.50
	ImageMagick - TIFF to JP2:	3.71
	GraphicsMagick - TIFF to JP2:	0.00
	Kakadu - TIFF to JP2:	3.68
	GeoJasper - TIFF to JP2:	3.65
Object characteristics	Keep status quo:	1.50
	ImageMagick - TIFF to JP2:	1.38
	GraphicsMagick - TIFF to JP2:	0.00
	Kakadu - TIFF to 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
additional meta data	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	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

Scanned books results WS

Results: Weighted sum

Result-Tree with all Alternatives, Aggregation method: Weighted sum.
This tree contains only strategies that do not have knock-out evaluation criteria; see above

Node	Results	
Scans	Keep status quo:	4.70
	ImageMagick - TIFF to JP2:	4.09
	Kakadu - TIFF to JP2:	4.06
	GeoJasper - TIFF to JP2:	4.03
Object characteristics	Keep status quo:	1.25
	ImageMagick - TIFF to JP2:	1.04
	Kakadu - TIFF to JP2:	1.04
	GeoJasper - TIFF to JP2:	1.04
Content	Keep status quo:	2.50
	ImageMagick - TIFF to JP2:	1.68
	Kakadu - TIFF to JP2:	1.68
	GeoJasper - TIFF to JP2:	1.68
Context	Keep status quo:	2.50
	ImageMagick - TIFF to JP2:	2.50
	Kakadu - TIFF to JP2:	2.50
	GeoJasper - TIFF to JP2:	2.50
Technical characteristics	Keep status quo:	1.06
	ImageMagick - TIFF to JP2:	0.98
	Kakadu - TIFF to JP2:	0.98
	GeoJasper - TIFF to JP2:	0.98
Costs	Keep status quo:	1.25
	ImageMagick - TIFF to JP2:	0.97
	Kakadu - TIFF to JP2:	0.95
	GeoJasper - TIFF to JP2:	0.97
Process characteristics	Keep status quo:	1.14
	ImageMagick - TIFF to JP2:	1.10
	Kakadu - TIFF to JP2:	1.08
	GeoJasper - TIFF to JP2:	1.04



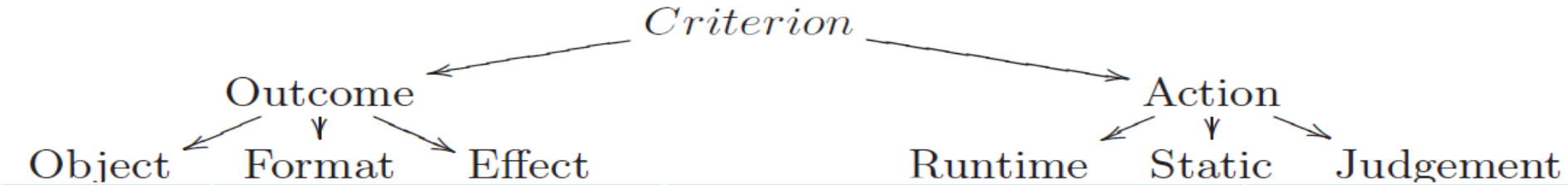
Addressing the evaluation gap

- Manual evaluation is very effort intensive
- Analysis of criteria specified in ten case studies
- Creation of a taxonomy of criteria
- Creation of measurement devices for each category
- Integration with Plato through an extensible measurement framework
- Quantitative analysis of measurement coverage

Decision criteria

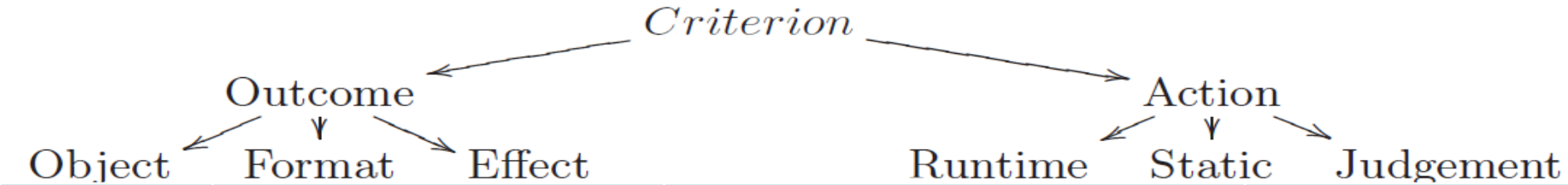
- Each criterion concerns either the action or its outcome
- Outcome
  - Object (authenticity, editability, …)
  - Format (licensing, standardisation, complexity…)
  - Effect (Costs…)
- Action
  - Runtime properties (performance, stability, logging…)
  - Static (price, license…)
  - Judgement (configuration interface usability…)

How to measure?



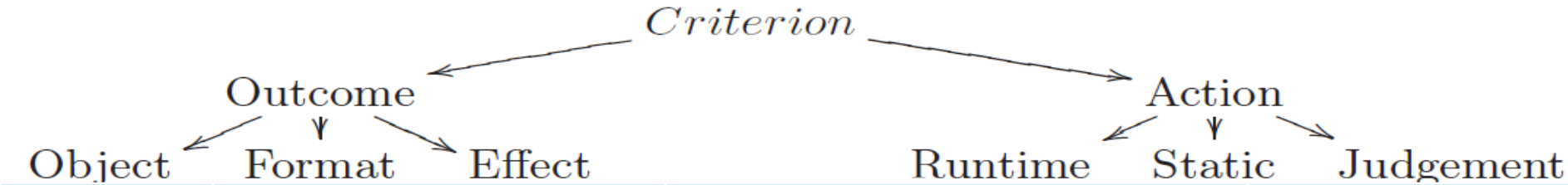
Category	Example	Data collection and measurement	Tools

How to measure?



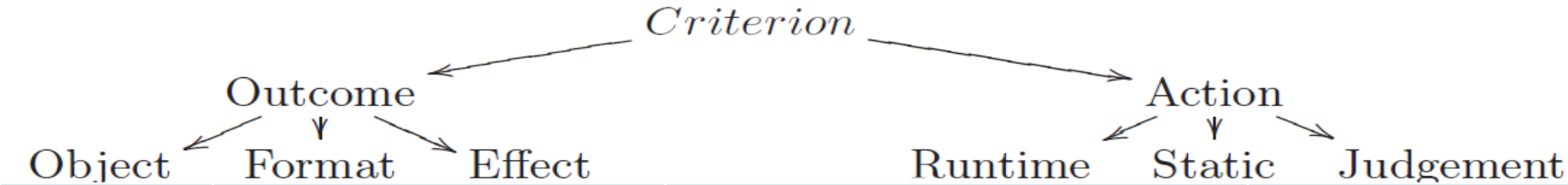
Category	Example	Data collection and measurement	Tools
Outcome Object	**Image pixelwise identical Footnotes preserved**	**Measurements of output and input, comparison**	**XCL, FITS, JHove, ImageMagick...**

How to measure?



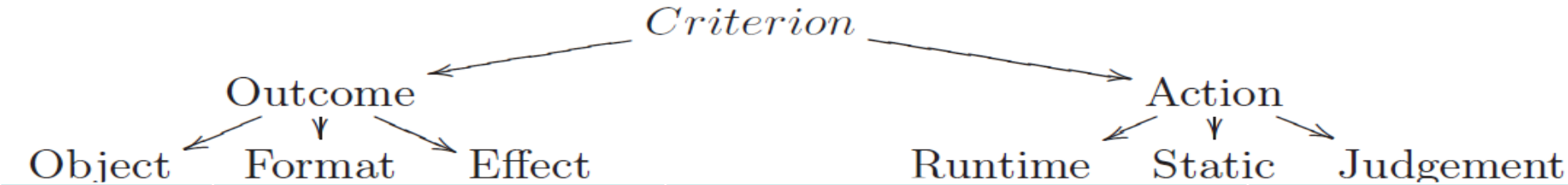
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	XCL, FITS, JHove, ImageMagick...
Outcome Format	**Format is ISO standardised**	**Measurements of the output, Trusted external data sources**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

How to measure?



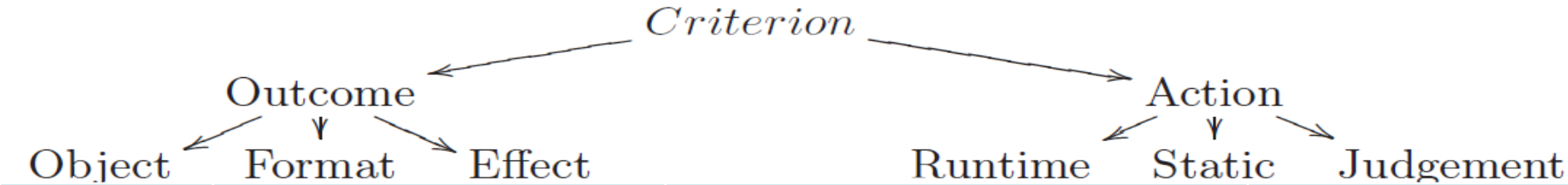
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	XCL, FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID,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

How to measure?



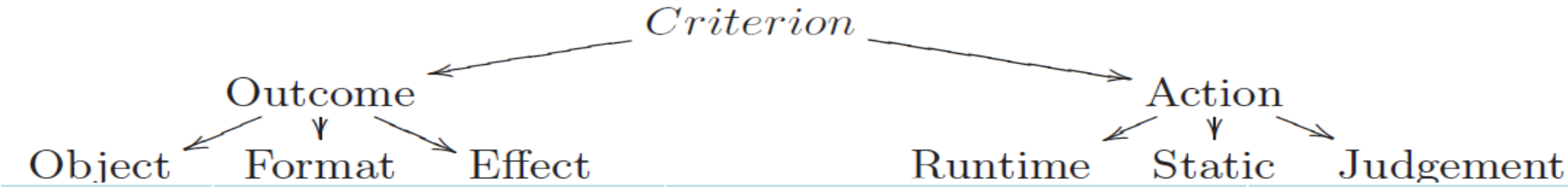
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	XCL, FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID,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



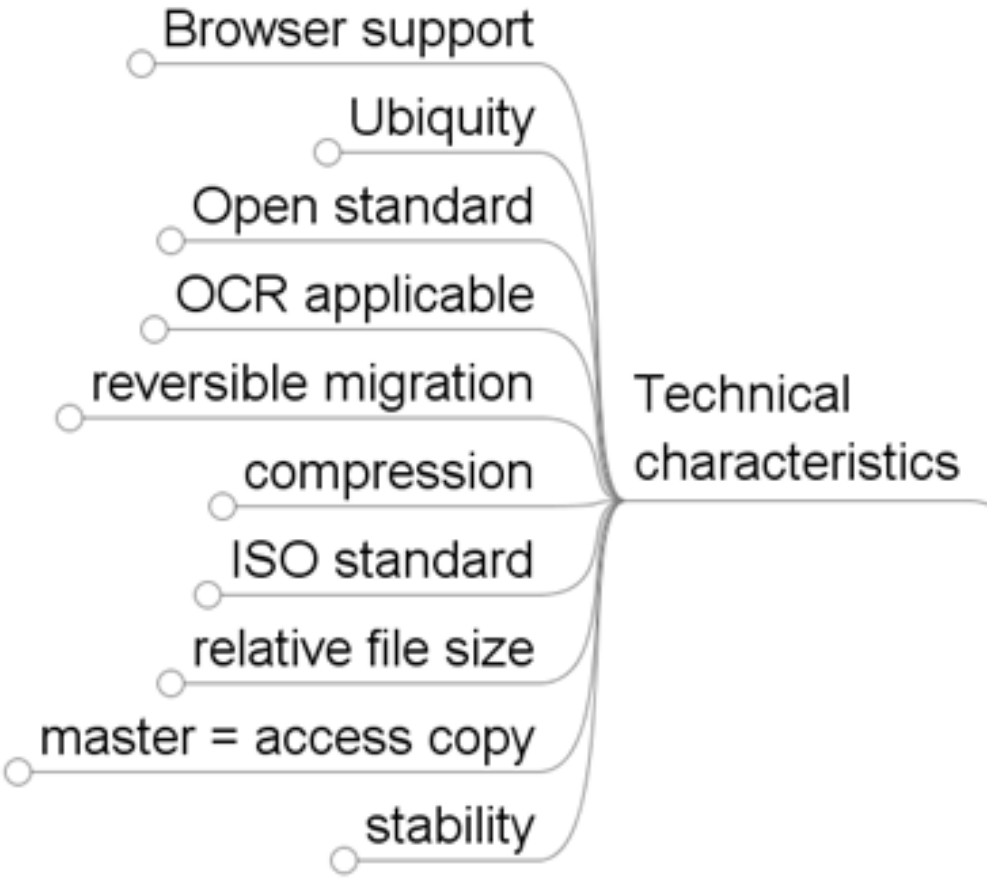
Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	XCL, FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID,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

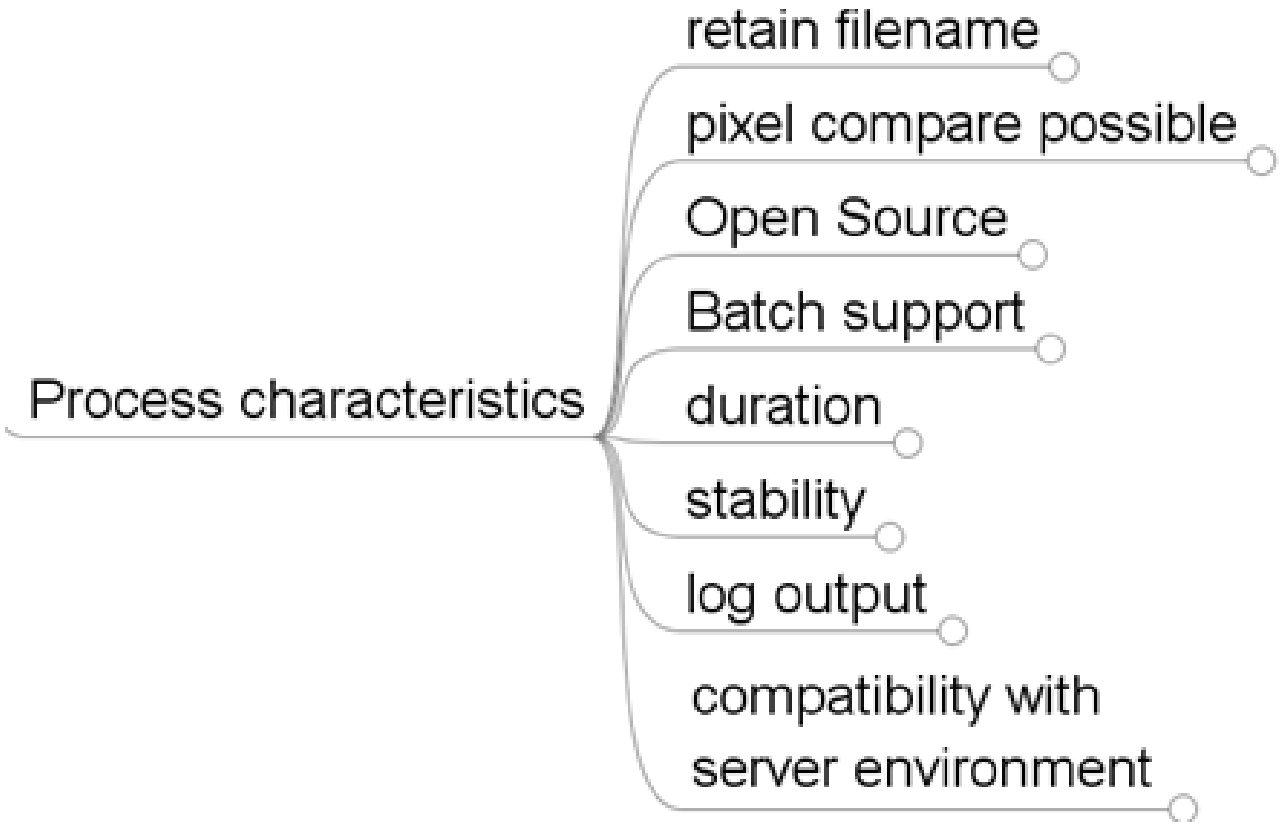


Category	Example	Data collection and measurement	Tools
Outcome Object	Image pixelwise identical Footnotes preserved	Measurements of output and input, comparison	XCL, FITS, JHove, ImageMagick...
Outcome Format	Format is ISO standardised	Measurements of the output, Trusted external data sources	DROID, PRONOM, P2/MiniREEF
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/MiniREEF, manual
Action judgement	**Configuration interface usability**	**Manual judgement, sharing**	

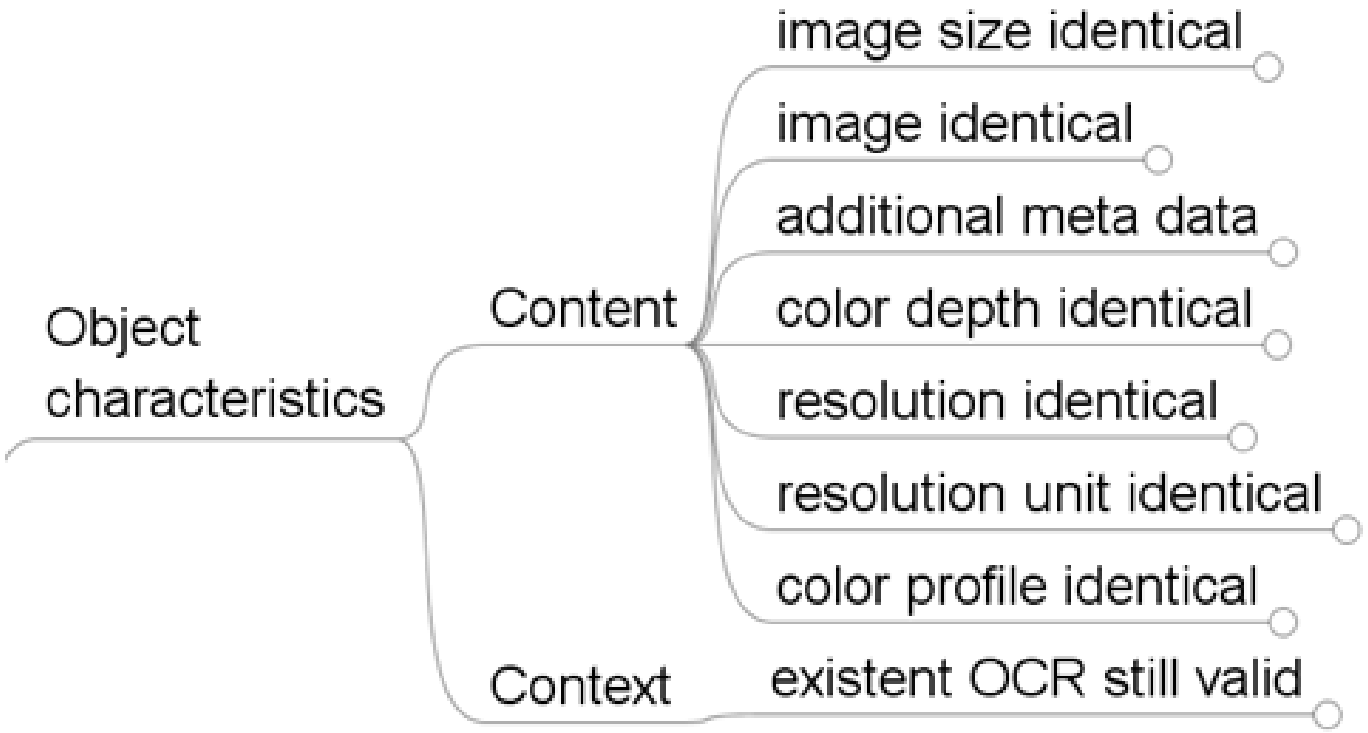
Scanned books requirements again...



Scanned books requirements again...



Scanned books requirements again...



Decision criteria and measurements

- Extraction of structured data
  - FITS
  - Jhove
  - Experiment metadata
- Runtime measurements of actions
  - in a controlled environment
  - MiniMEE: Quality-aware migration
- Comparison of objects
  - Static comparison, interpretation comparison, Observation
  - For images: ImageMagick *compare* and others
- Shared information sources
  - UDFR, PRONOM
  - P2: Semantic web format registry

Decision criteria and measurements

- Extraction of structured data
  - FITS
  - Jhove
  - Experiment metadata
- Runtime measurements of actions
  - in a controlled environment
  - MiniMEE: Quality-aware migration
- **Comparison of objects**
  - **Static comparison, interpretation comparison, Observation**
  - **For images: ImageMagick *compare* and others**
- Shared information sources
  - UDFR, PRONOM
  - P2: Semantic web format registry

Validating a migrated image

- Dimensions, metadata.... easy: extract and compare
- Content... Not always easy
- ImageMagick *compare:* good for simple cases

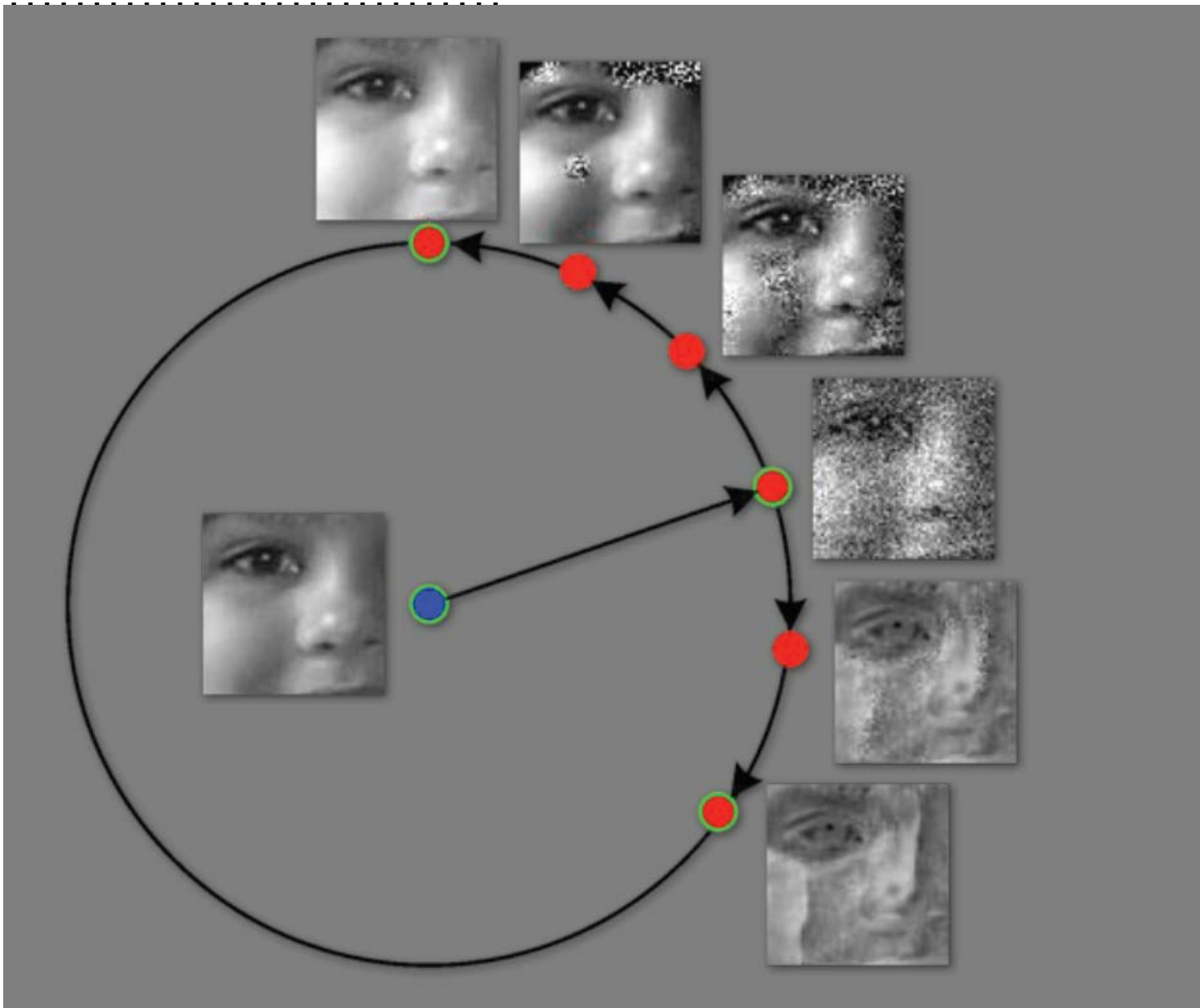
Abbr.	Metric	Description
AE	Absolute Error	The number of different pixels (0 means identical images). This value can be thresholded to only count pixels that have a difference larger then a specified threshold.
PAE	Peak Absolute Error	The highest difference of any single pixel.
PSNR	Peak Signal to Noise Ratio	The ratio of mean square difference to the maximum mean square that can exist between any two images, expressed as a decibel value. The higher the PSNR, the closer the images are, with a maximum difference occurring at 1.
MAE	Mean Absolute Error	Average over all pixels
MSE	Mean Squared Error	Averaged squared error distance
RMSE	Root mean squared error	Identical to *sqrt(MSE)*.

Distance metrics: How meaningful?

One error distance, many images

Anything but “0” is a problematic result

Address QA piece by piece



RAW

Adobe

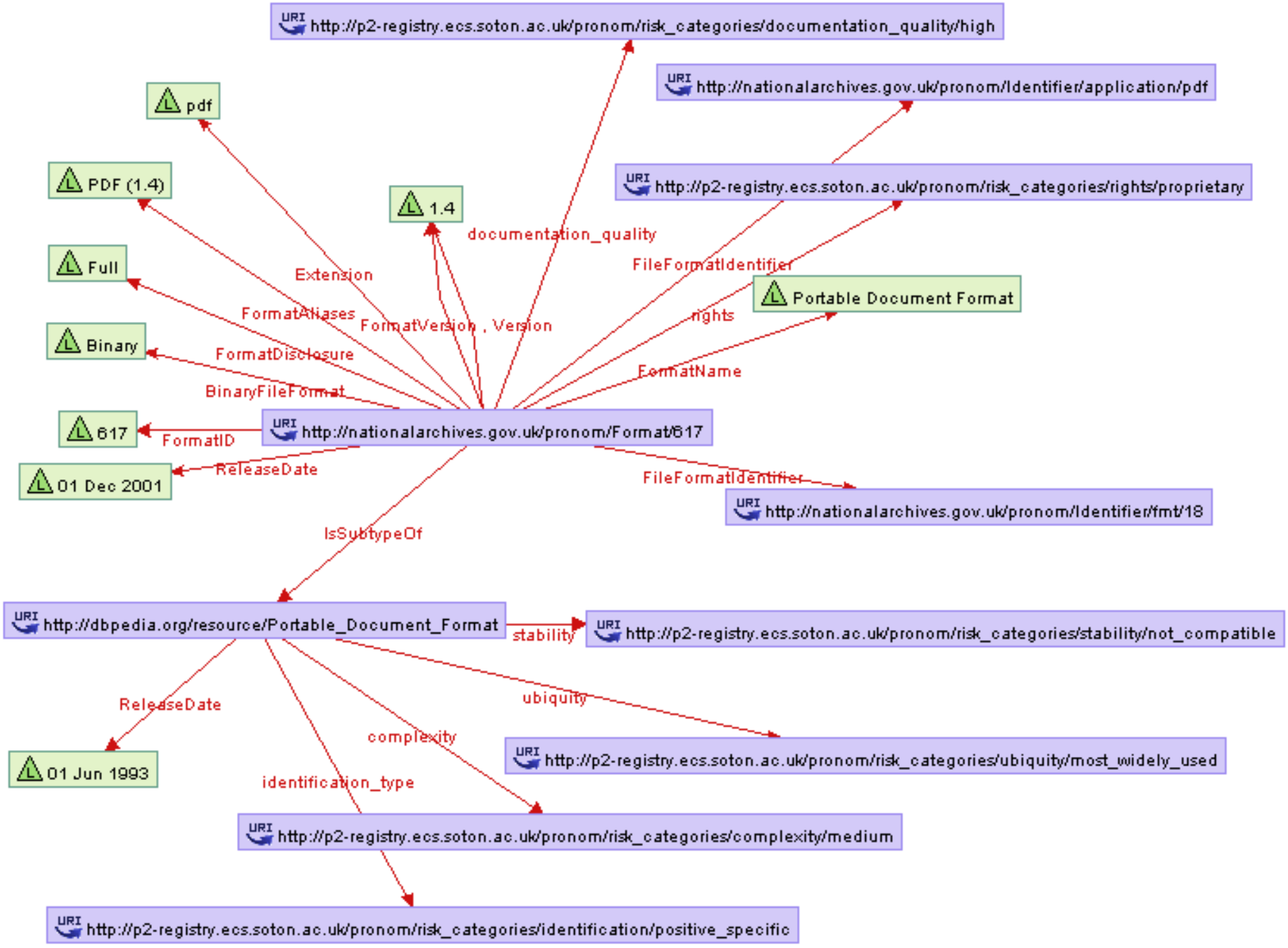
dcraw recoveres edges

dcraw

Apple

Decision criteria and measurements

- Extraction of structured data
  - FITS
  - Jhove
  - Experiment metadata
- Runtime measurements of actions
  - in a controlled environment
  - MiniMEE: Quality-aware migration
- Comparison of objects
  - Static comparison, interpretation comparison, Observation
  - For images: ImageMagick *compare* and others
- **Shared information sources**
  - **UDFR, PRONOM**
  - **P2: Semantic web format registry**

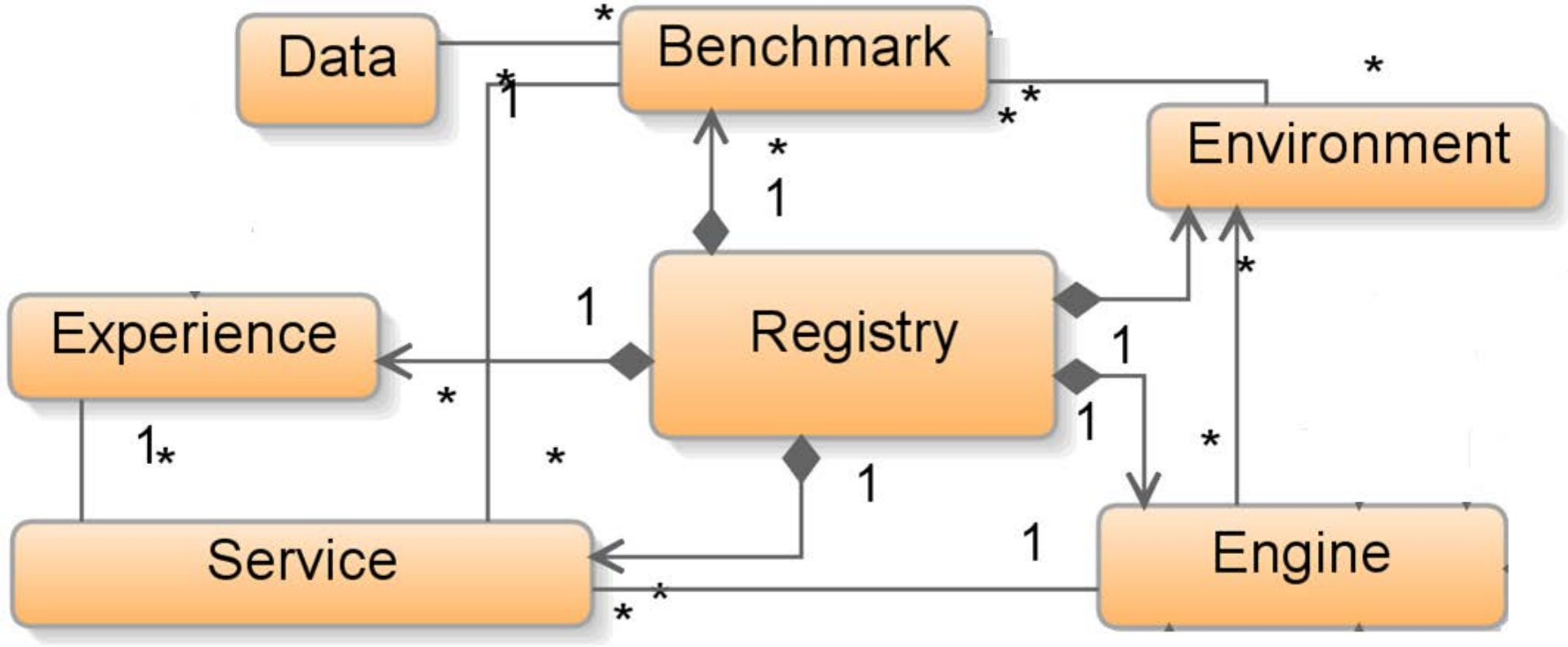

http://p2-registry.ecs.soton.ac.uk/pronom/risk_categories/documentation_quality/high
http://nationalarchives.gov.uk/pronom/Identifier/application/pdf
pdf
PDF (1.4)
1.4
http://p2-registry.ecs.soton.ac.uk/pronom/risk_categories/rights/proprietary
documentation_quality
Full
Extension
FileFormatIdentifier
Portable Document Format
FormatAliases
FormatVersion , Version
rights
Binary
FormatDisclosure
FormatName
BinaryFileFormat
617
FormatID
http://nationalarchives.gov.uk/pronom/Format/617
ReleaseDate
01 Dec 2001
FileFormatIdentifier
http://nationalarchives.gov.uk/pronom/Identifier/fmt/18
IsSubtypeOf
http://dbpedia.org/resource/Portable_Document_Format
stability
http://p2-registry.ecs.soton.ac.uk/pronom/risk_categories/stability/not_compatible
ReleaseDate
ubiquity
complexity
01 Jun 1993
http://p2-registry.ecs.soton.ac.uk/pronom/risk_categories/ubiquity/most_widely_used
identification_type
http://p2-registry.ecs.soton.ac.uk/pronom/risk_categories/complexity/medium
http://p2-registry.ecs.soton.ac.uk/pronom/risk_categories/identification/positive_specific

Measuring runtime behaviour of actions

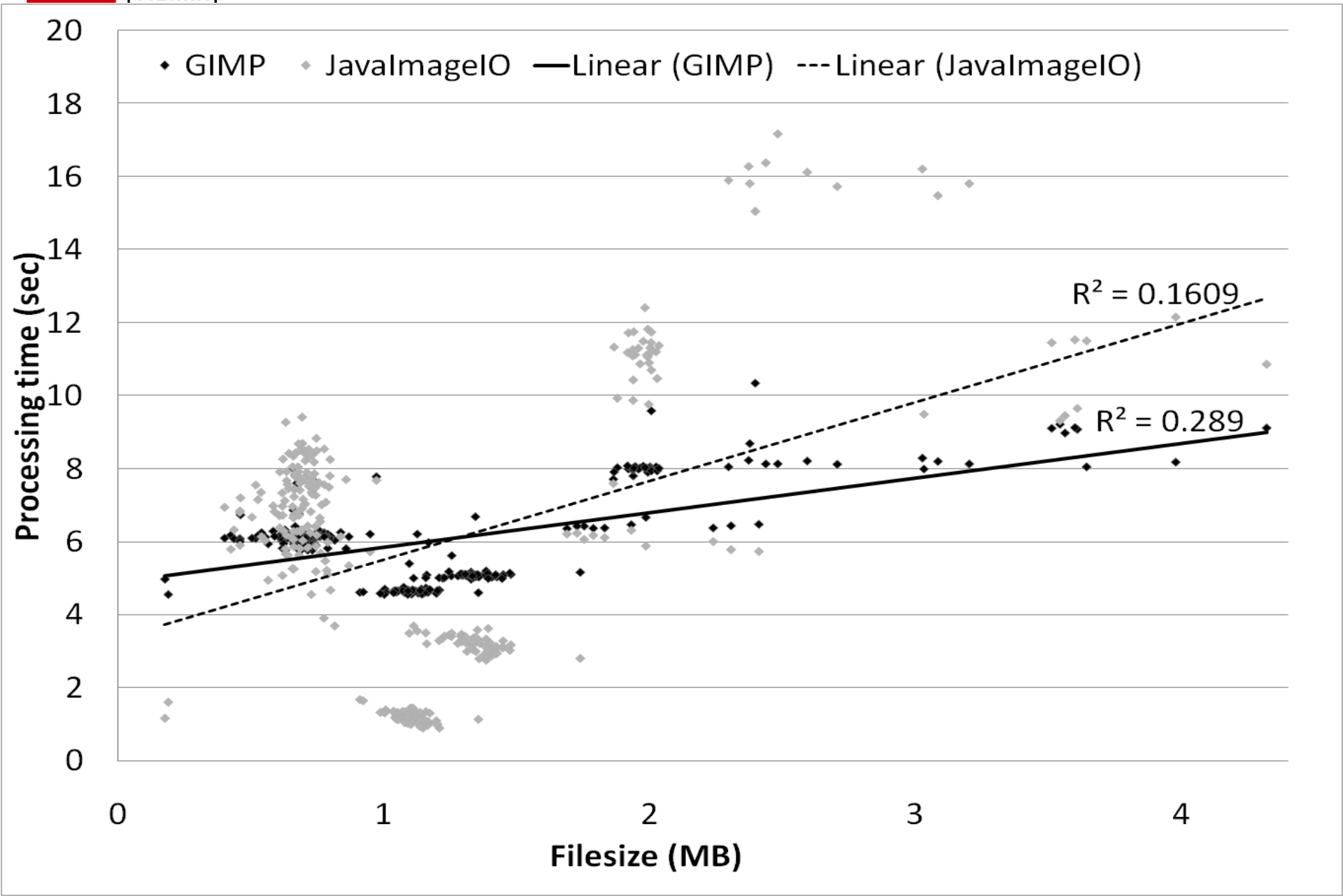
- Integration architecture
  - Dynamic discovery and invocation of preservation actions
  - Web services for loose coupling and flexible integration
  - Little or no information about process on provider side
- MiniMEE: Quality-aware migration
  - Non-intrusive provider-side instrumentation
  - Migration engine monitors components at runtime
  - Transparent invocation in controlled environment
  - Performance information delivered to the requester
- Framework design
- Measurement tools and observations

Core elements of the framework

- Engines make components quality-aware
- Environments have associated benchmark scores
- Registry accumulates experience



Comparing tool performance



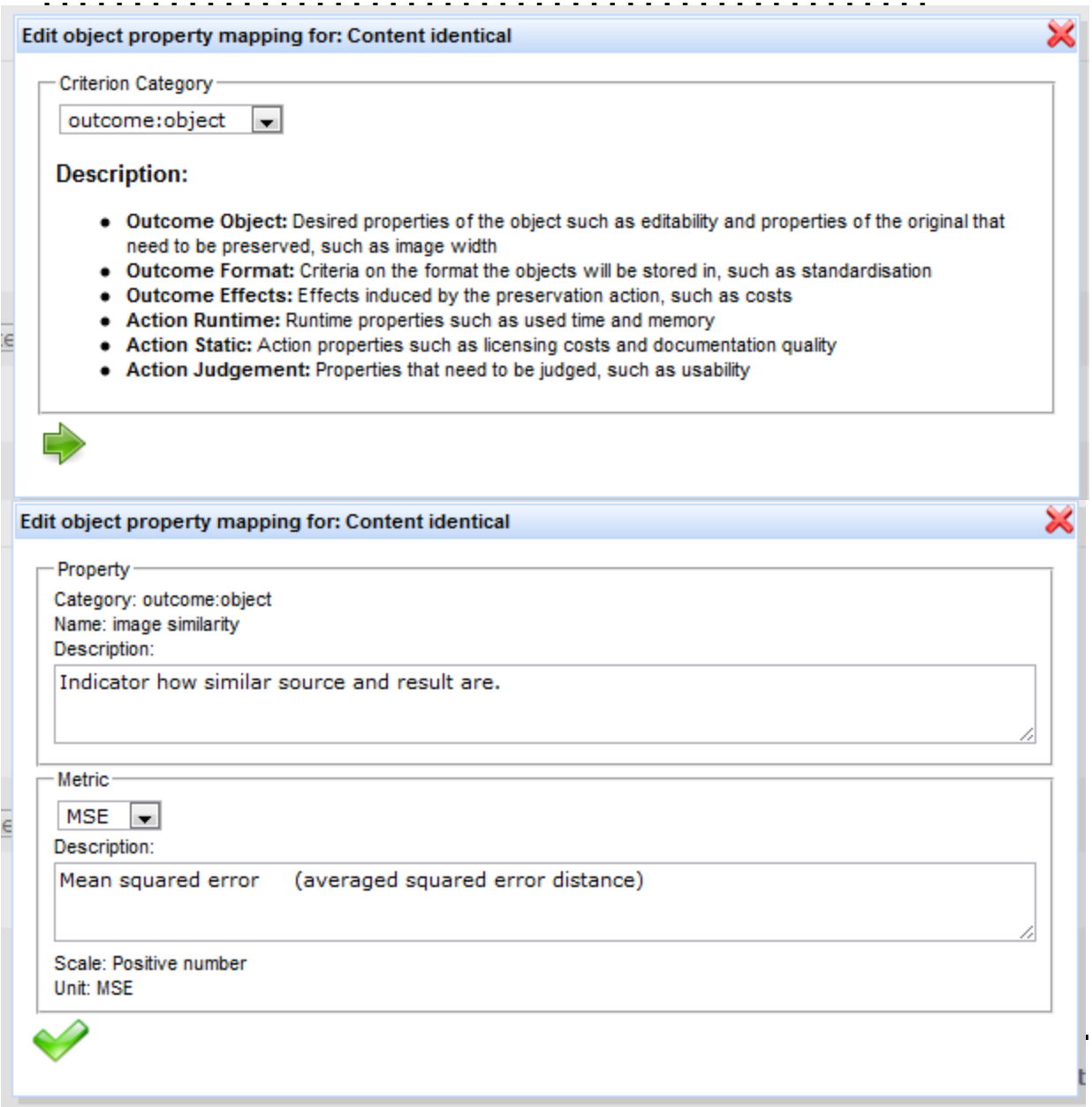
Identified, unambigous criteria

- Planning framework supports decision maker
  - ... But did previously not provide support for sharing experience and draw conclusions
- Assessment needs to be specific to an organisation
  - ... But the objective facts that are the basis for assessment should be collected independently
  - ... And we want to be able to share them
- Plato 3: Decision criteria can be linked to identified properties
  - These properties are objective and have clear semantics
  - Assessment (utility function) is specified separately
- Plato contains growing knowledge base of properties
  - These can be assigned to decision criteria in the tree
  - .. Which enables experience sharing

Extensible measurement framework

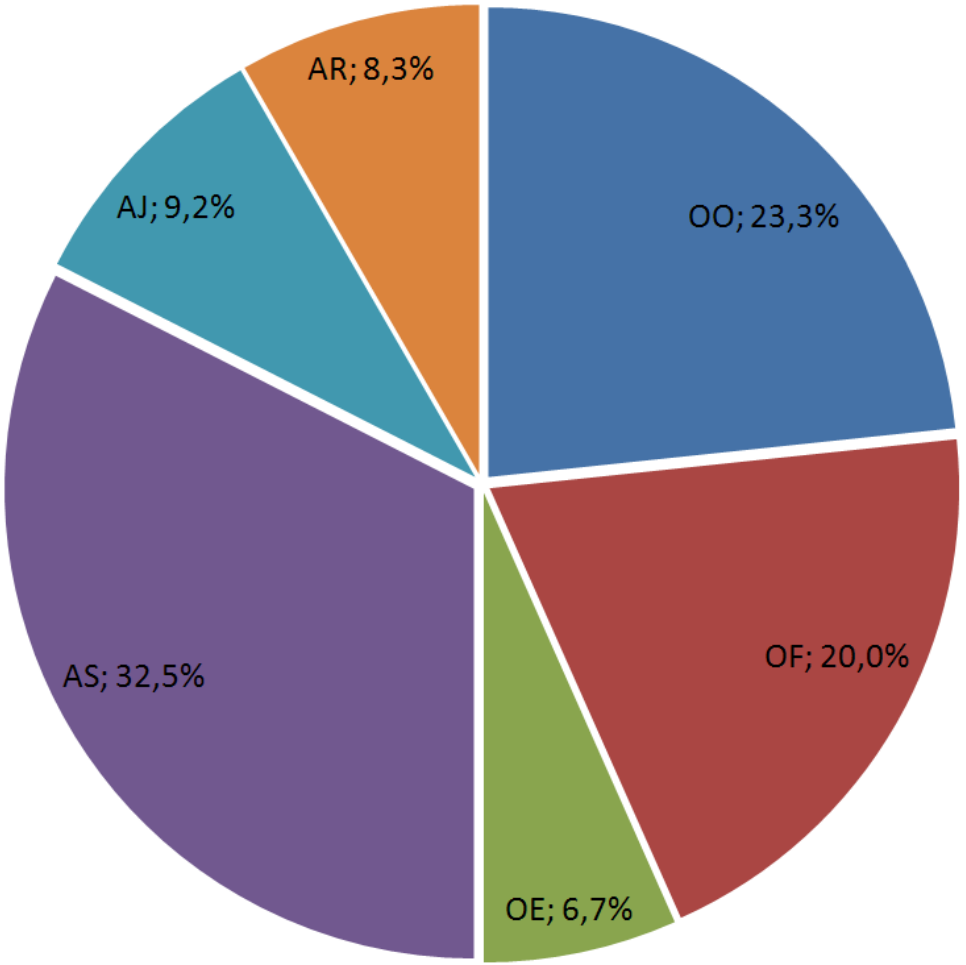
- Plato knowledge base contains *Evaluators*
- Evaluators declare properties they can measure
  - *ExperimentEvaluator*
  - *FitsEvaluator*
  - *ObjectEvaluator*
  - *ImageComparisonEvaluator*
  - *MiniREEFEvaluator*
  - *XCLEvaluator*
- Integrated in Plato
- Assignment of measurable properties/metrics to criteria
  - *action://runtime/performance/memory/peak*
  - *outcome://object/image/similarity#equal*
  - *outcome://object/format/conforms*
  - *outcome://format/numberOfTools/Open*
  - *outcome://object/image/metadata/creationDate#equal*
  - *.....*

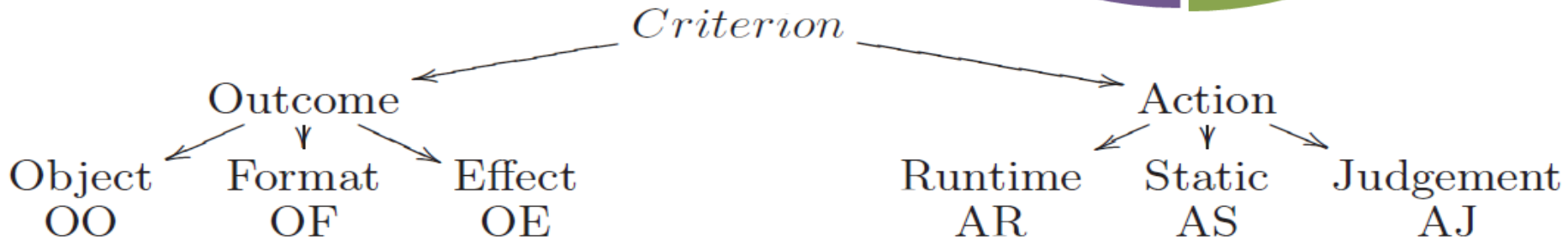
Assignment of measurements in Plato



Case studies

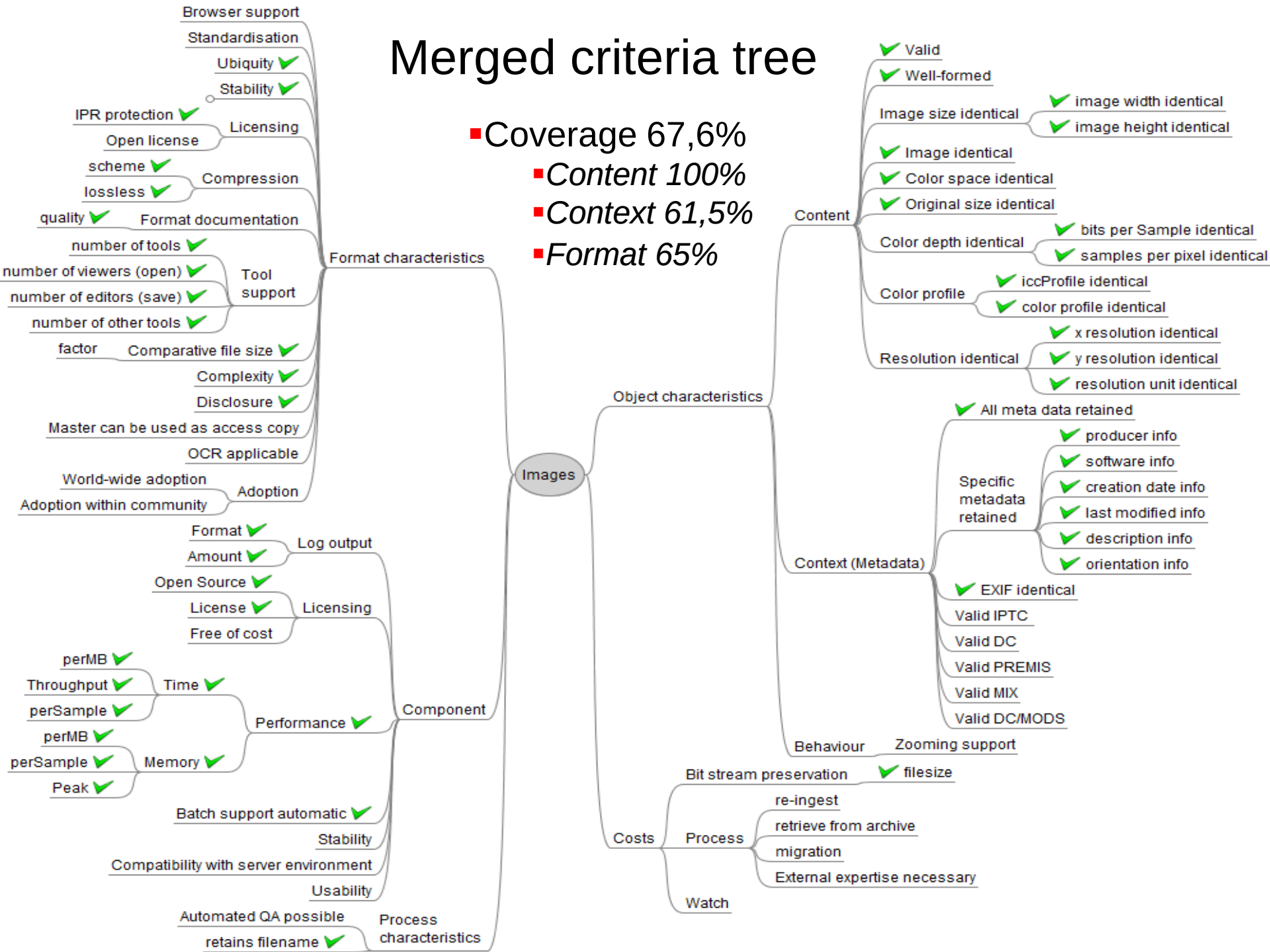
- Distribution in four case studies on scanned images





Merged criteria tree

- Coverage 67,6%
  - *Content 100%*
  - *Context 61,5%*
  - *Format 65%*

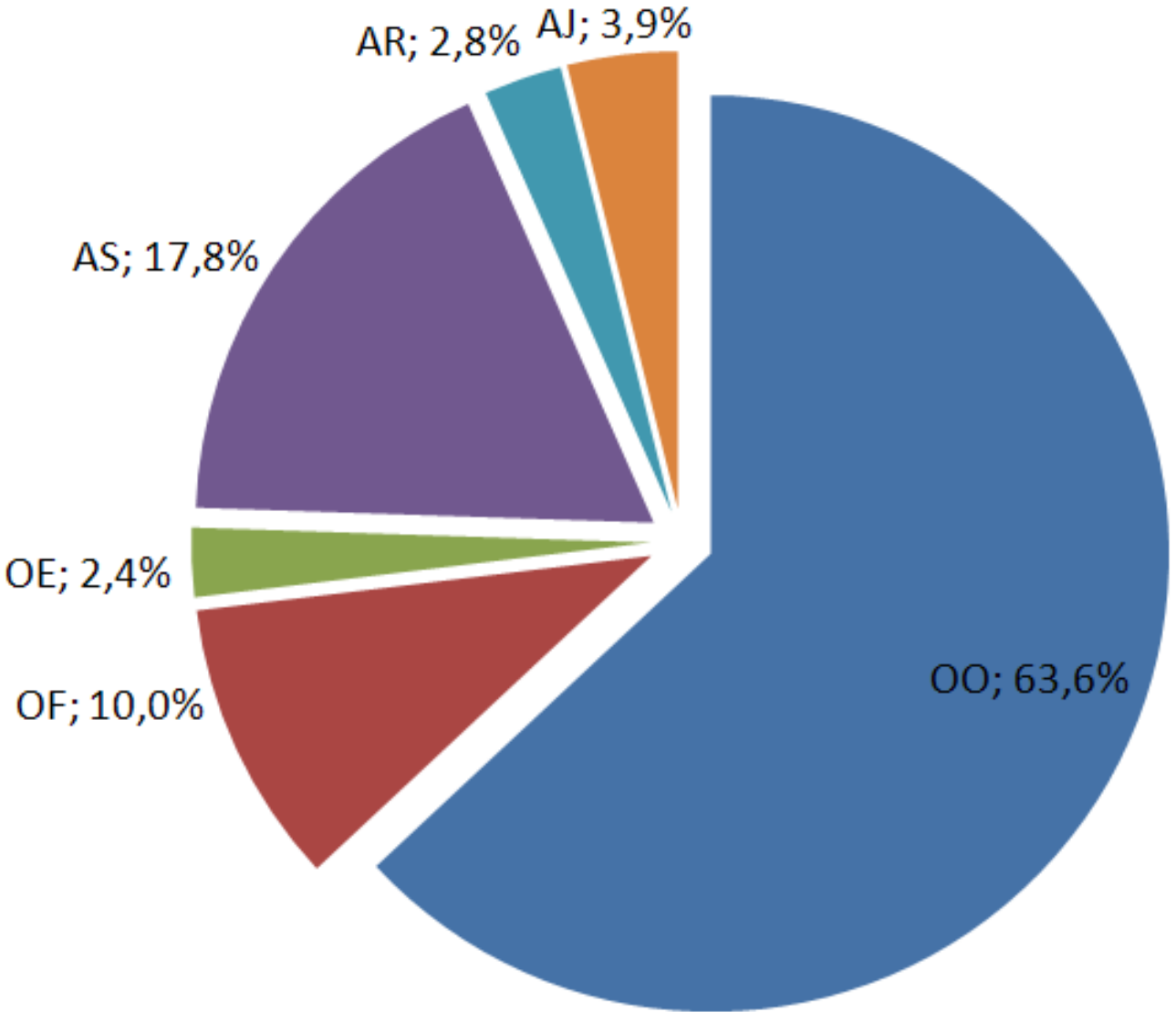


Where are we now?

- 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
  - Measurement of some properties is very difficult
  - Measurement of emulation is virtually non-existent
  - Many complex properties cannot be measured yet
  - … but a lot of people are working on that ☺

Case studies

- Distribution in thirteen cases on various types of content



Example criteria we can't measure (yet)

- Images
  - Define a metric that quantifies the distance of two images represented in different colour spaces
- Databases:
  - Measure the accuracy of the content of user defined datatypes that have been normalised into (represented as) standard SQL data types
- Games:
  - Measure controller response delay (in an emulation environment)
  - Verify the synchronicity of video and audio
- Documents:
  - Measure the identity of footnote numbering between a .doc and the converted .pdf/a
- ……

Sneak preview: A knowledge browser

- The number of properties is growing…
- For a given property, we want to know
  - How often was it used, and where? Is it relevant for our specific scenario?
  - How often did it lead to a rejection of candidate due to an unacceptable evaluation result?
  - What is the impact of this property? What if the evaluation changes?
- New analysis module in the planning tool

Sneak preview: A knowledge browser

Plato 3

The PLANETS Preservation Planning Tool

Plan

Knowledge browser

General Statistic	Plans	Templates
relevant trees	11	0
overall leaves	254	0
mapped leaves	11	0

Property statistics	
Available Properties	313
Properties used at least once	7
Criteria used at least once	7

Criterion selection

Category	*Property*	*Metric*
(all)	image: sampling frequency unit	(none)
outcome:object	image: similarity	equal
outcome:format	image: text quality	
outcome:effect	image: width	
action	image: x sampling frequency	
	image: y sampling frequency	
	relative filesize	

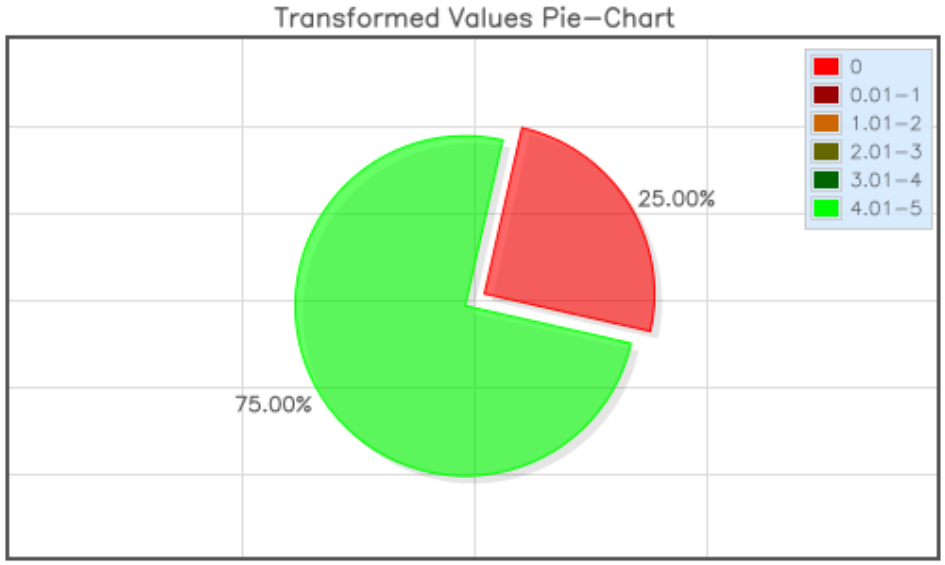
Properties in Category:239

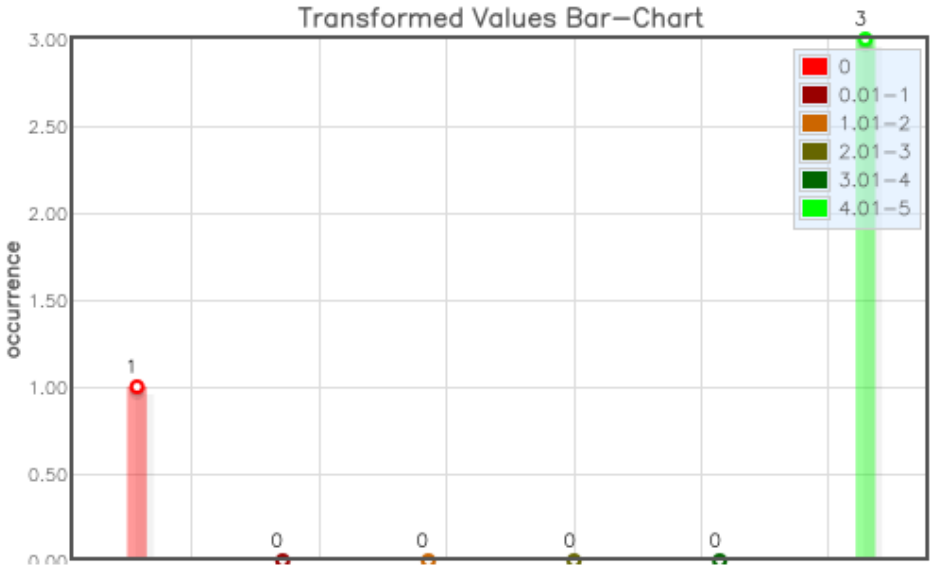
☐ display only used properties

Sneak preview: A knowledge browser

Criterion Characteristic	Value
property description	Width of an image, measured in number of pixels on the x-axis.
metric	Comparison of two values for equality
scale	Boolean

Criterion Statistic	Plans	Templates
leaves in category	5	0
leaves using property	2	0
criterion mapped	2	0
AVG weight	0,43	0
AVG total weight	0,3	0
potential knock-outs	2	
actual knock-outs	1	
measurements obtained	4	
Yes : 3		
No : 1		





Sneak preview: A knowledge browser

Importance factors	
AVG potential output range (occurrence based)	2,505
MAX potential output range	2,75
MIN potential output range	2,26
AVG actual output range (occurrence based)	1,13
MAX actual output range	2,26
MIN actual output range	0

Important Properties Table				
Category	Property	Metric	avg potential output range	avg actual output range
OUTCOME_OBJECT	image: height	equal	2.505	2.505
OUTCOME_OBJECT	image: width	equal	2.505	1.13
OUTCOME_OBJECT	format: valid and well-formed		1.98	0.0
OUTCOME_FORMAT	standardization		0.522	0.372
OUTCOME_FORMAT	ubiquity		0.4485	0.1995
OUTCOME_FORMAT	compression		0.144	0.144
OUTCOME_FORMAT	image: zoom possible		0.018	0.0