Digital Preservation

Preservation Planning revisited … and Quality Assurance

Christoph Becker

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

Agenda

- Preservation Planning revisited
  - What is a preservation plan?
  - How to create a preservation plan
  - Evaluation and plan definition
  - PP Case Studies
- Quality Assurance
  - What to measure
  - How to measure it
- Presentation of the DP-UE tasks

Preservation Planning

- A number of solutions exist
- All have specific strengths and weaknesses
- Individual requirements, obligations and constraints in every institution
- Decision between tools is complex
- Documentation and accountability is essential in decision-making

- Preservation Planning assists in decision making
- Evaluating preservation strategies on representative samples according to specific requirements and criteria

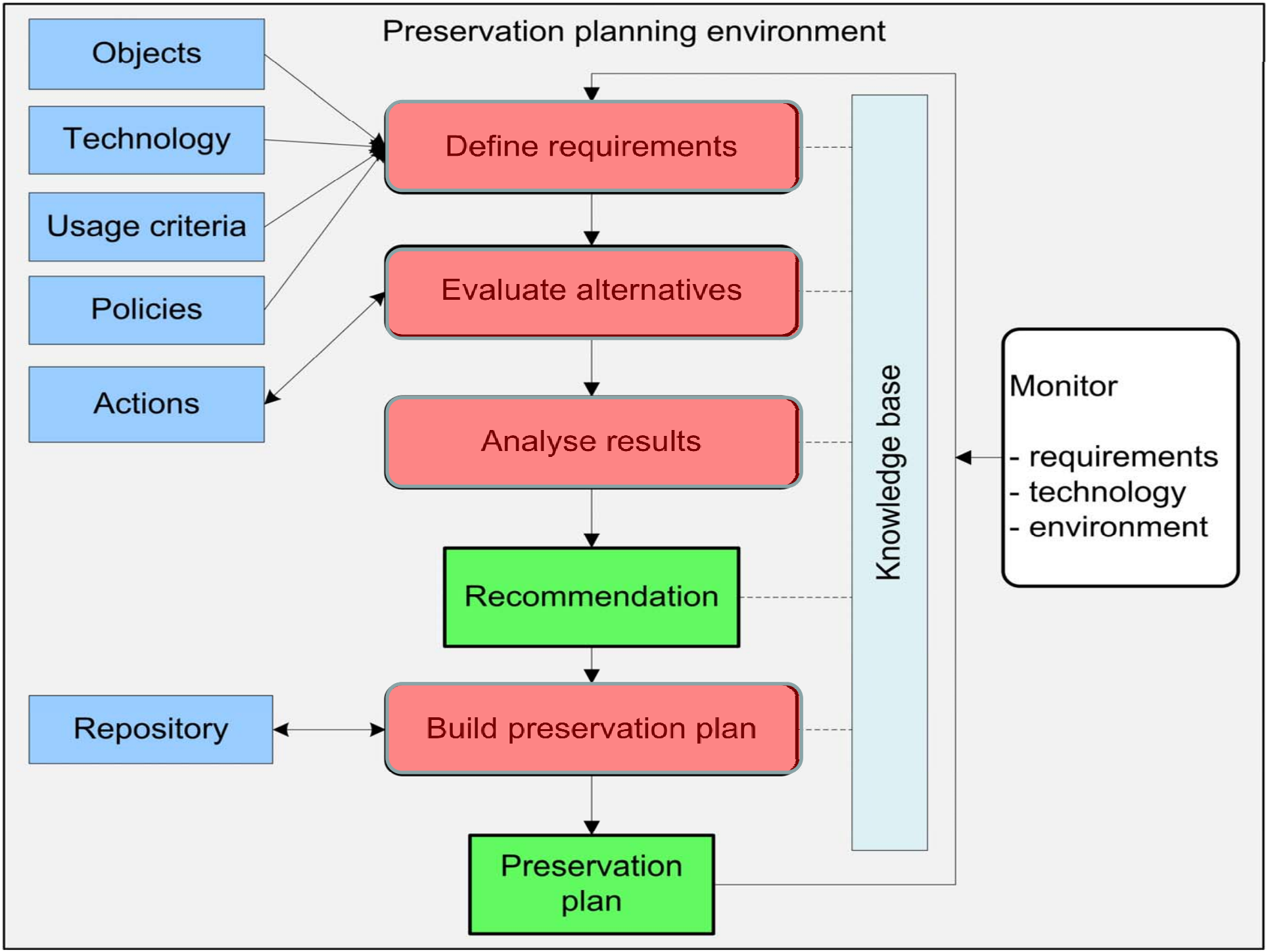
What is a preservation plan?

- ‘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).‘

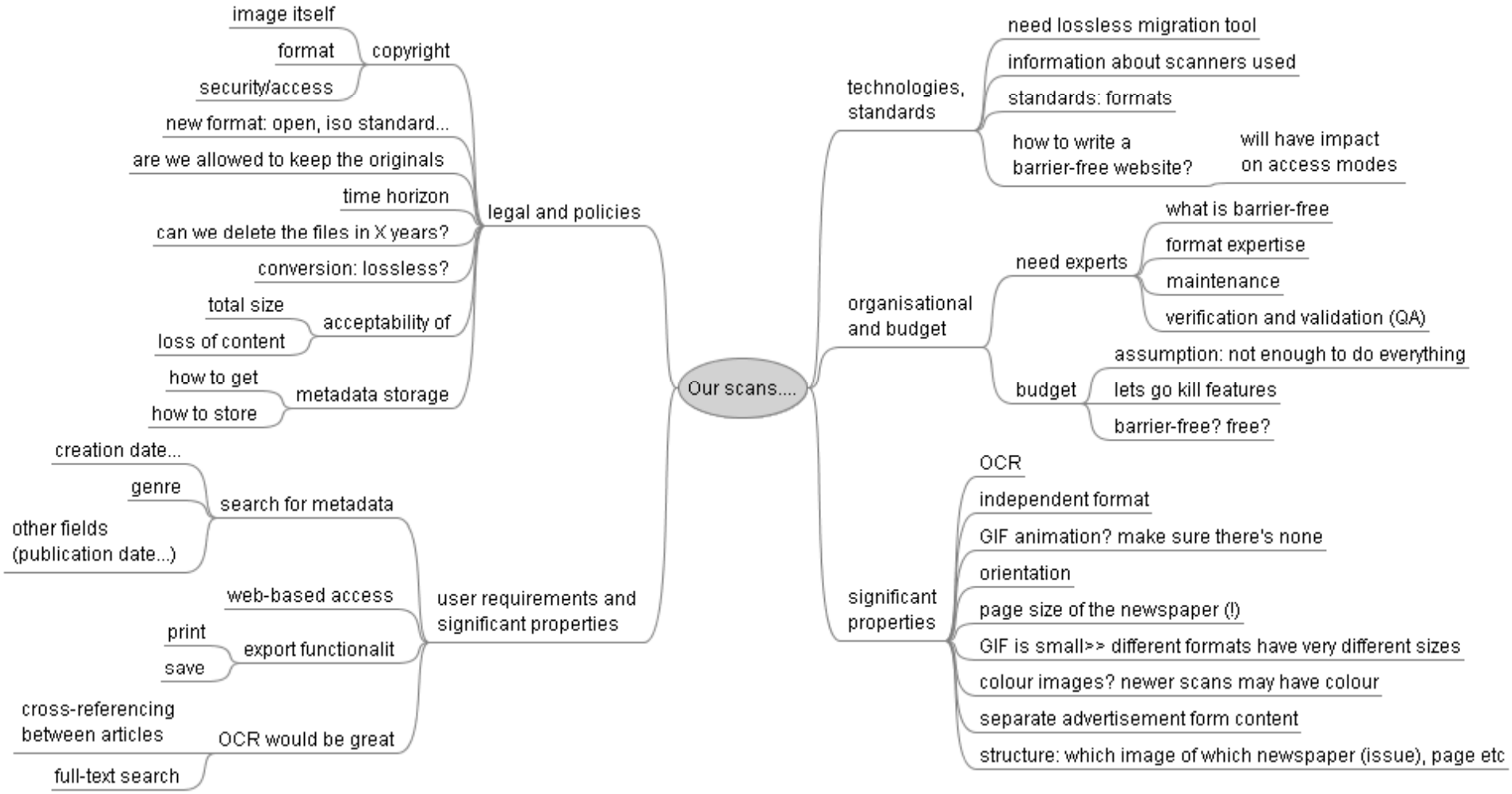
- The Preservation Plan takes into account the preservation **policies**, **legal obligations**, **organisational** and **technical constraints**, **user requirements** and **preservation goals**.
- It also describes the preservation **context**, the **evaluated alternative preservation strategies** and the **resulting decision** for one strategy, including the rationale of the decision.

What is *in* a preservation plan?

- Definition of scope
  - What to preserve
- Set of actions
  - How to preserve it
- Evaluation of actions, recommendation for one
  - How to do it and why do it this way
- Documentation of actions and reasons
  - Why did we decide that
- Conditions for QA and monitoring
  - What to look out for


Preservation planning environment
Objects
Technology
Usage criteria
Policies
Actions
Define requirements
Evaluate alternatives
Analyse results
Recommendation
Build preservation plan
Preservation plan
Repository
Knowledge base
Monitor
- requirements
- technology
- environment

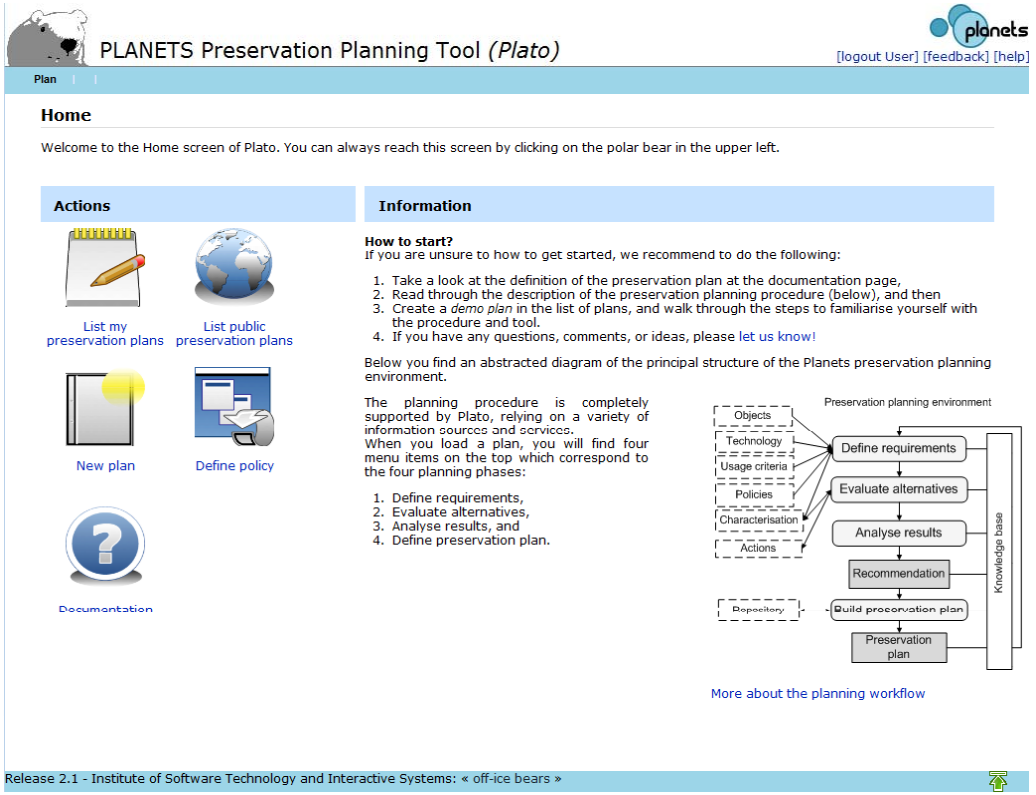
Aspects collected last week...



The planning tool PLATO

- Plato is a web application based on J2EE technologies
  - (Jboss Seam, Facelets, Richfaces, AJAX, JPA...)
- It supports the complete planning workflow
  - Characterisation of sample objects
  - Requirements definition, mindmap integration, knowledge base
  - Action discovery and invocation
  - Automated experiments
  - Visual analysis of results
  - Plan specification
  - Traceable documentation

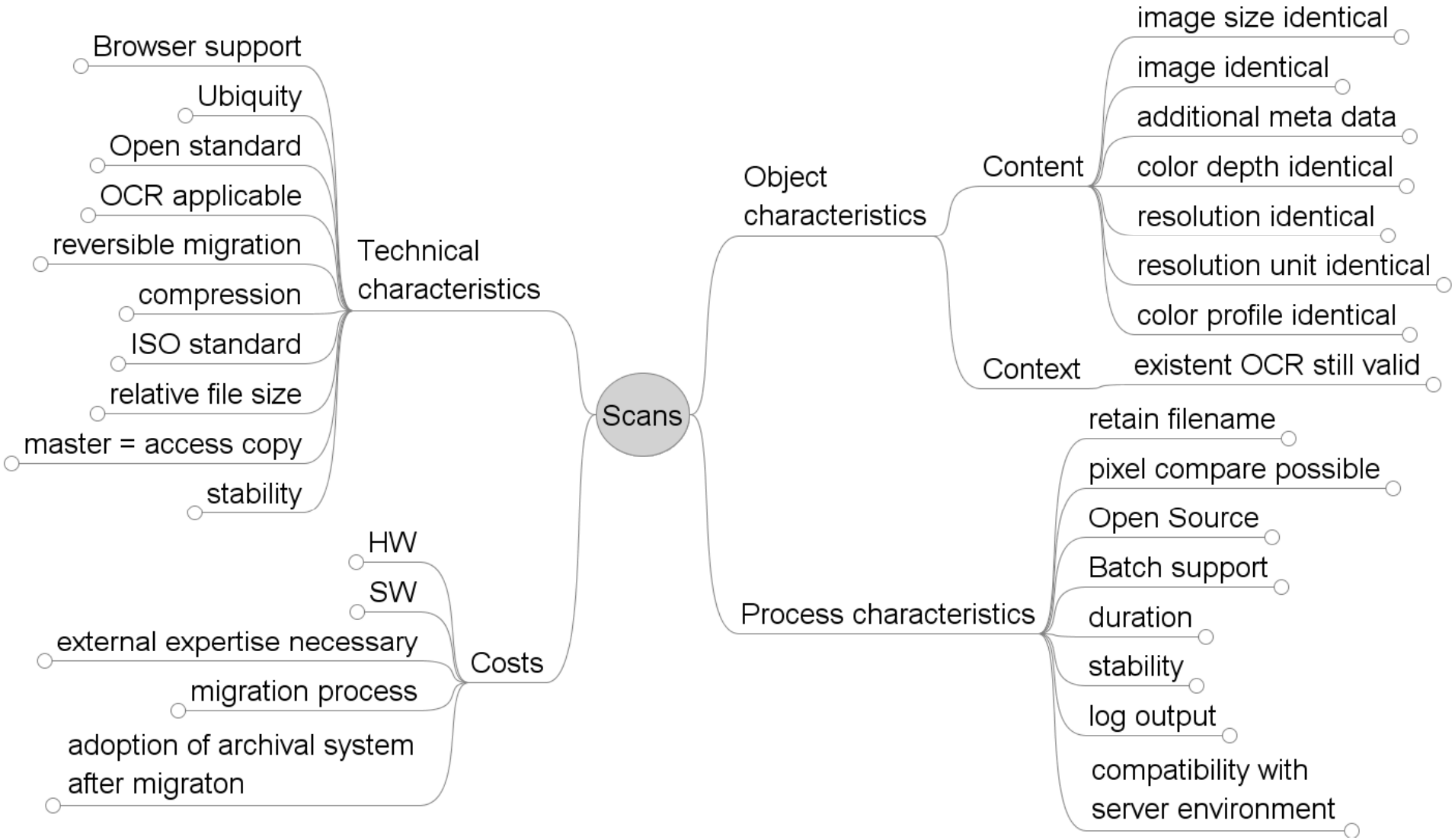
...



A real case in Plato

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

Scanned books requirements



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

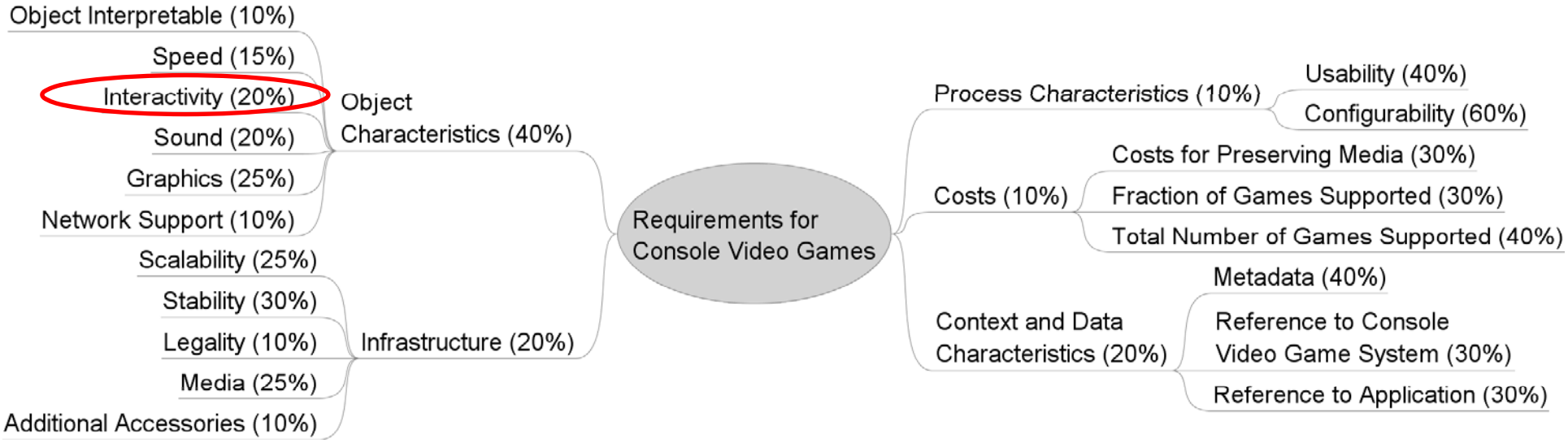
Node		Results
Scans	Keep status quo:	4.70
	ImageMagick - TIFF to JP2:	4.09
	Kakadu - TIFF to JP2:	4.06
	GeoJasper - TIFF to JP2:	4.03
Object characteristics	Keep status quo:	1.25
	ImageMagick - TIFF to JP2:	1.04
	Kakadu - TIFF to JP2:	1.04
	GeoJasper - TIFF to JP2:	1.04
Content	Keep status quo:	2.50
	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

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

Console video games study

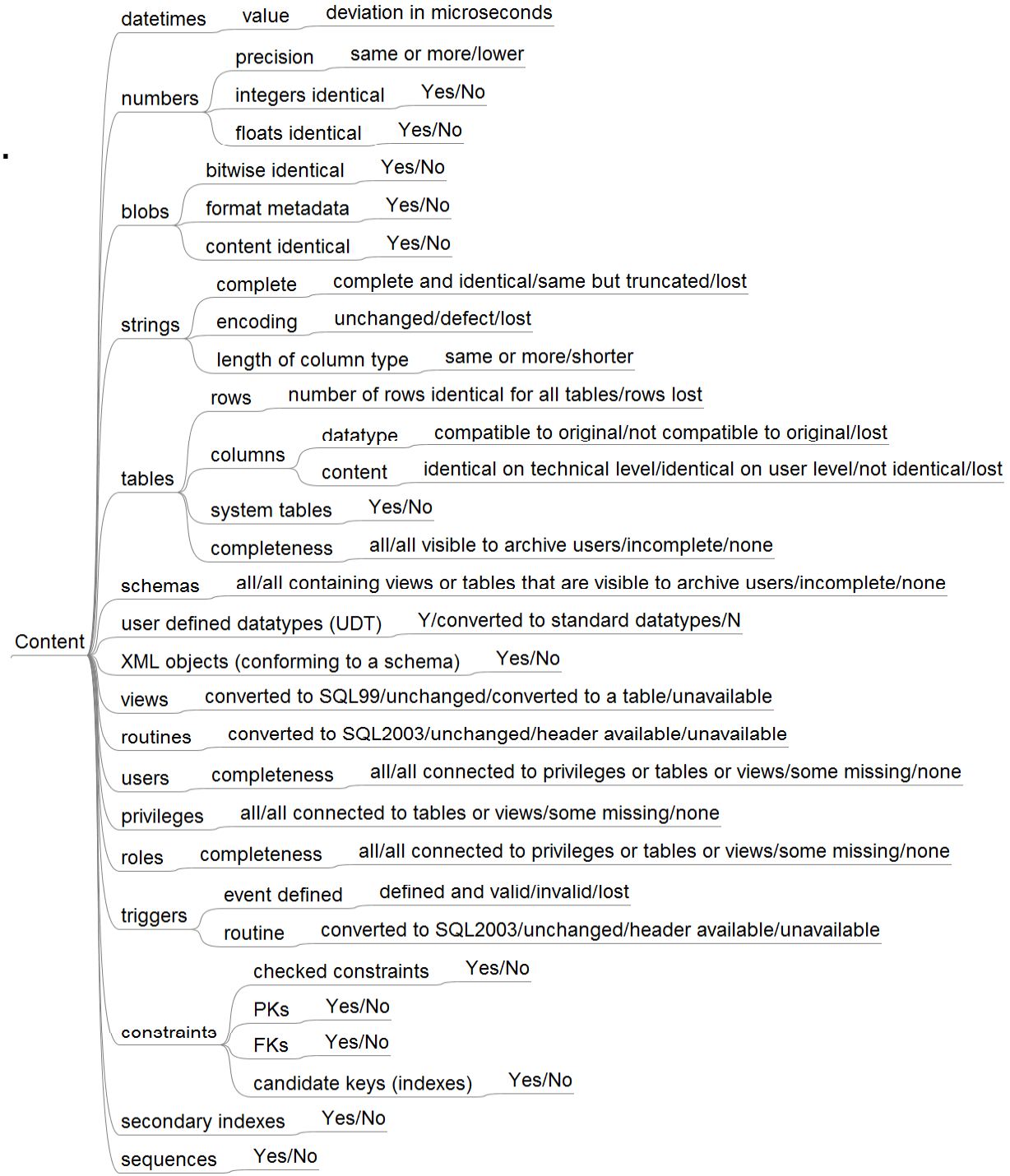


Alternative	Sample object	WSSample	WMSample	WSTotal	WMTotal
ZSNES 1.51	Super Mario World	3,45	2,75	3,28	2,68
	Super Scope 6	3,30	2,70		
	Starfox	3,38	2,78		
SNES9X 1.51	Super Mario World	3,43	2,82	3,31	2,70
	Super Scope 6	3,28	2,68		
	Starfox	3,38	2,78		
MESS 0.119	Super Mario World	3,56	2,88	2,68	0,00
	Super Scope 6	3,47	2,79		
	Starfox	2,47	0,00		
VLC 0.8.6c/MP4	Super Mario World	4,65	0,00	4,65	0,00

Table 5.1: Evaluation results for preserving games for the Nintendo SNES

Database study

Content branch



Results: Weighted multiplication

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

Node	Results	
Requirements	Archive to XML:	3.88
	Keep original DB:	3.72
	CSV export:	0.00
Object characteristics	Archive to XML:	2.14
	Keep original DB:	2.32
	CSV export:	0.00
Content	Archive to XML:	2.38
	Keep original DB:	2.51
	CSV export:	0.00
appearance	Archive to XML:	1.14
	Keep original DB:	1.16
	CSV export:	1.06
context	Archive to XML:	1.23
	Keep original DB:	1.17
	CSV export:	0.88
behaviour	Archive to XML:	1.00
	Keep original DB:	1.12
	CSV export:	1.00
Format characteristics	Archive to XML:	1.36
	Keep original DB:	1.20
	CSV export:	1.13
Tool characteristics	Archive to XML:	1.34

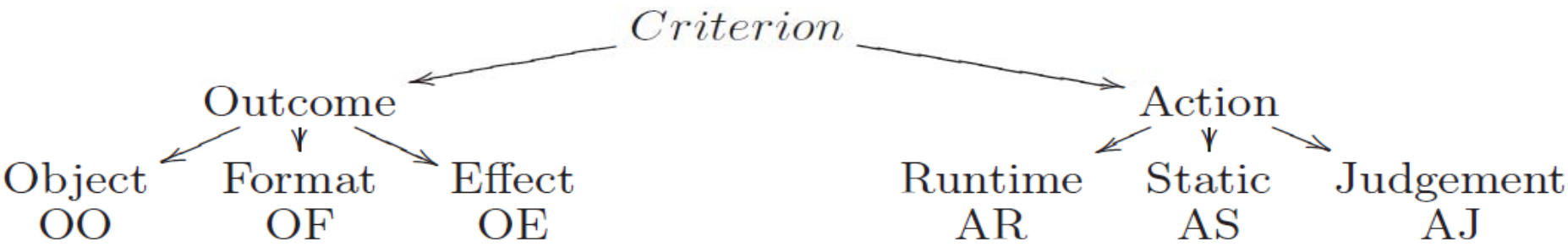
Results: Weighted sum

Result-Tree with all Alternatives, Aggregation method: Weighted sum.

This tree contains only strategies that do not have knock-out evaluation criteria; see above

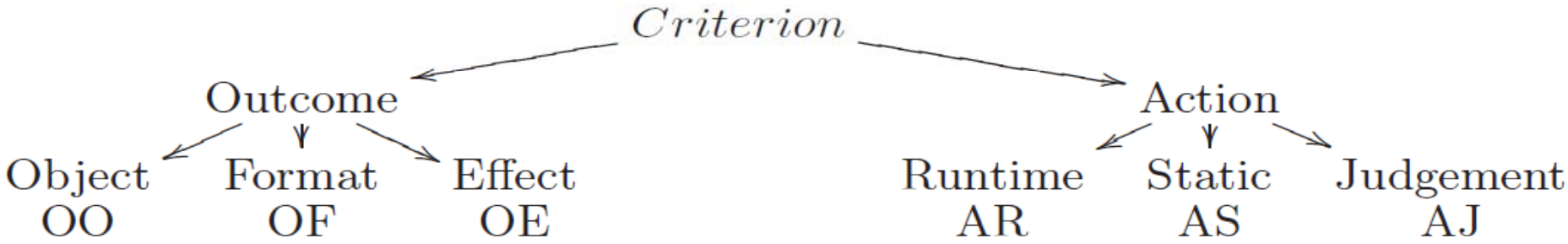
Node	Results	
Requirements	Archive to XML:	4.25
	Keep original DB:	4.07
Object characteristics	Archive to XML:	2.40
	Keep original DB:	2.60
Content	Archive to XML:	2.63
	Keep original DB:	2.76
appearance	Archive to XML:	0.41
	Keep original DB:	0.45
context	Archive to XML:	0.69
	Keep original DB:	0.57
behaviour	Archive to XML:	0.07
	Keep original DB:	0.35
Format characteristics	Archive to XML:	0.95
	Keep original DB:	0.57

Decision criteria



- Each criterion concerns either the action or its outcome
- Outcome
  - Object (Authenticity, Fixity, editability)
  - Format (Licensing, Standardisation, Complexity…)
  - Effect (Filesize, Costs)

Decision criteria



- Each criterion concerns either the action or its outcome
- Action
  - Runtime properties (performance, stability, logging…)
  - Static (price, license…)
  - Judgement (configuration interface usability…)

How to measure?

Category	Abbr.	Example	Data collection and measurements
Outcome object	OO	*Image pixelwise identical* (RMSE)	Measurements of input and output, measurements taken in controlled experimentation
Outcome format	OF	*Format is ISO standardised* (boolean)	Measurements of output, trusted external data sources
Outcome effect	OE	*Annual bitstream preservation costs* (€)	Measurements of output, trusted external data sources, models, partly manual calculation and validation, sharing
Action runtime	AR	*Throughput* (MB per ms)	Measurements taken in controlled experimentation
Action static	AS	*License costs per CPU* (€)	Trusted external data sources, manual evaluation and validation, sharing
Action judgement	AJ	*Configuration interface usability* (excellent, sufficient, poor)	Manual judgement, sharing

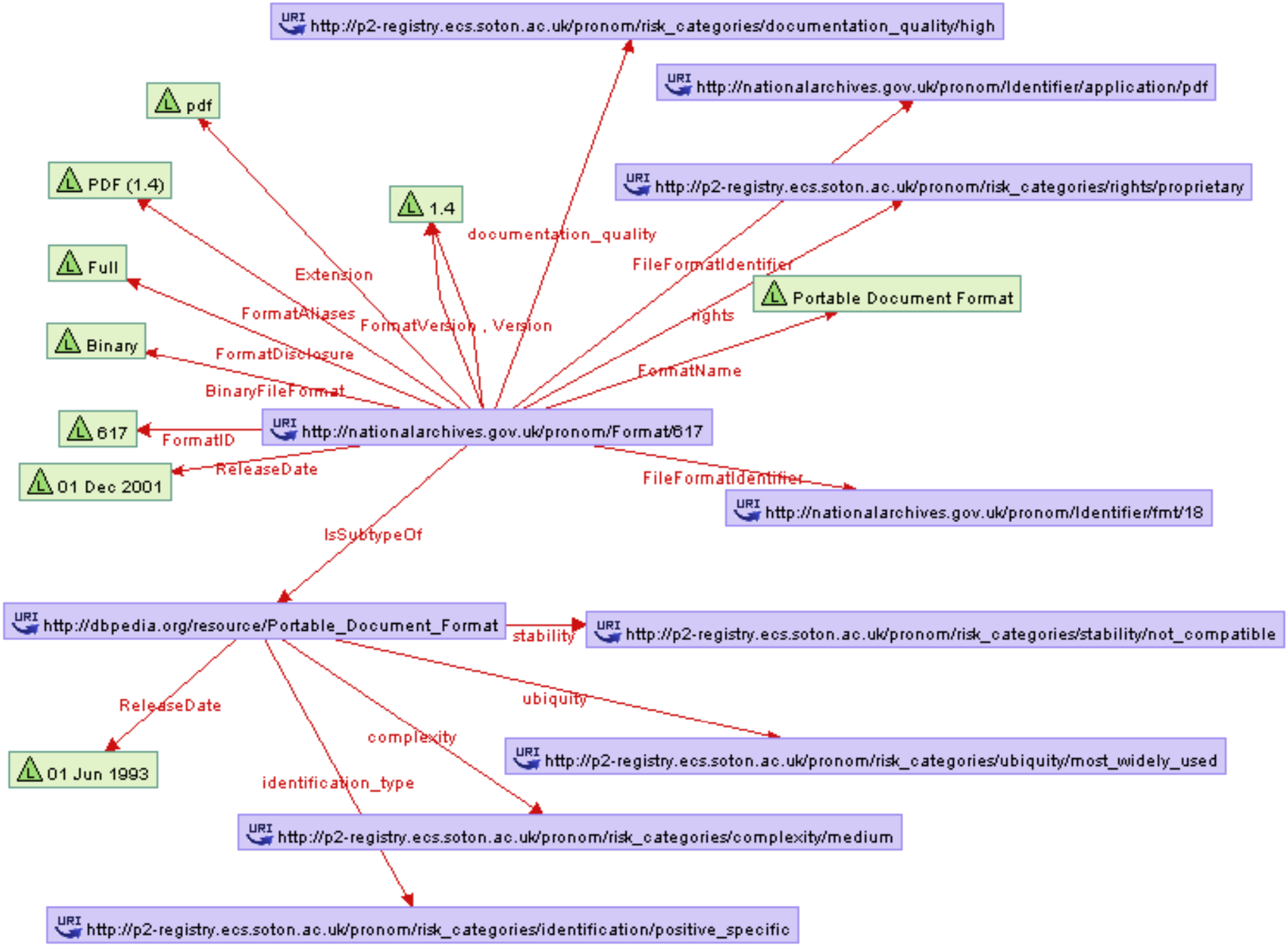
Some file format requirements

- Specifications available
  - Is an XML specification enough?
  - Syntacs **and semantics** needed
- Standardized (ISO, ANSI, ITEF, ...)
- Accepted and widely used
- Not covered by patent
- Free of compression
- Free of any cryptographical techniques

- Flexible and extensible?
- Anything else?

Data sources

- PRONOM
  - Sparse data
- www.digitalpreservation.gov/formats
  - Incomplete
- Wikipedia
  - reliable?
- The web
  - unstructured
- P2: Combination of PRONOM with dbpedia
  - Linked Data
  - ~45.000 statements
  - Still far from complete


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
FileFormatIdentifier
Full
Extension
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
01 Jun 1993
complexity
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

How to measure?

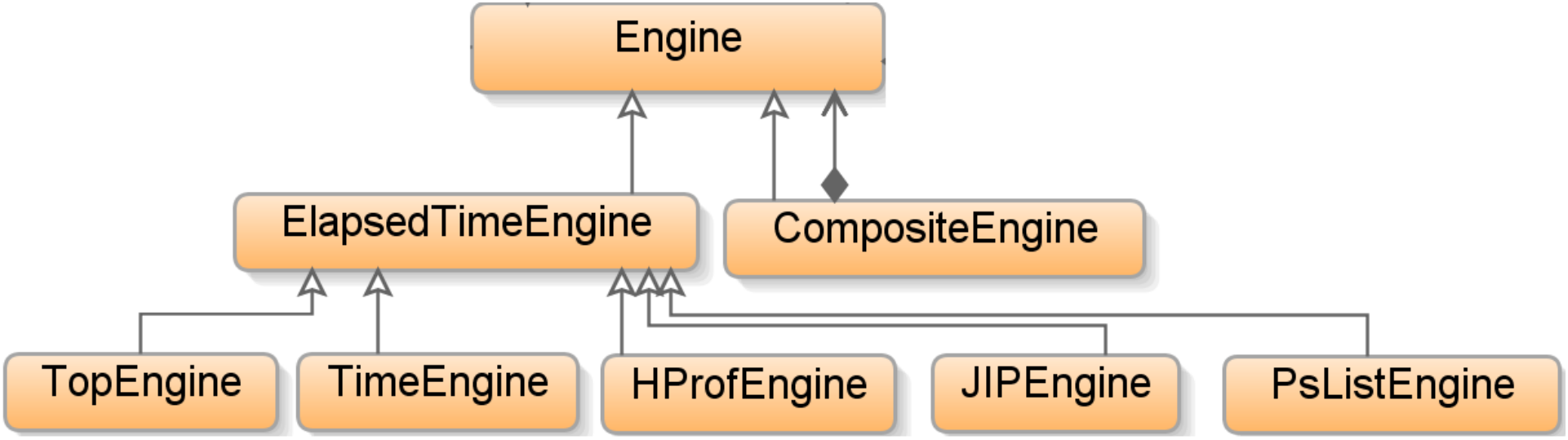
Category	Abbr.	Example	Data collection and measurements
Outcome object	OO	*Image pixelwise identical* (RMSE)	Measurements of input and output, measurements taken in controlled experimentation
Outcome format	OF	*Format is ISO standardised* (boolean)	Measurements of output, trusted external data sources
Outcome effect	OE	*Annual bitstream preservation costs* (€)	Measurements of output, trusted external data sources, models, partly manual calculation and validation, sharing
Action runtime	AR	*Throughput* (MB per ms)	Measurements taken in controlled experimentation
Action static	AS	*License costs per CPU* (€)	Trusted external data sources, manual evaluation and validation, sharing
Action judgement	AJ	*Configuration interface usability* (excellent, sufficient, poor)	Manual judgement, sharing

WS QoS measurement techniques

- Components generally behind web service...
- Intermediaries
  - Traffic routed through them
- Probing
  - Independent party invokes services and collects QoS attributes
- Sniffing
  - Monitor traffic on client side
- Provider-side instrumentation
  - Invasive vs. Non-invasive
  - Access to code?
- Non-invasive provider-side service instrumentation

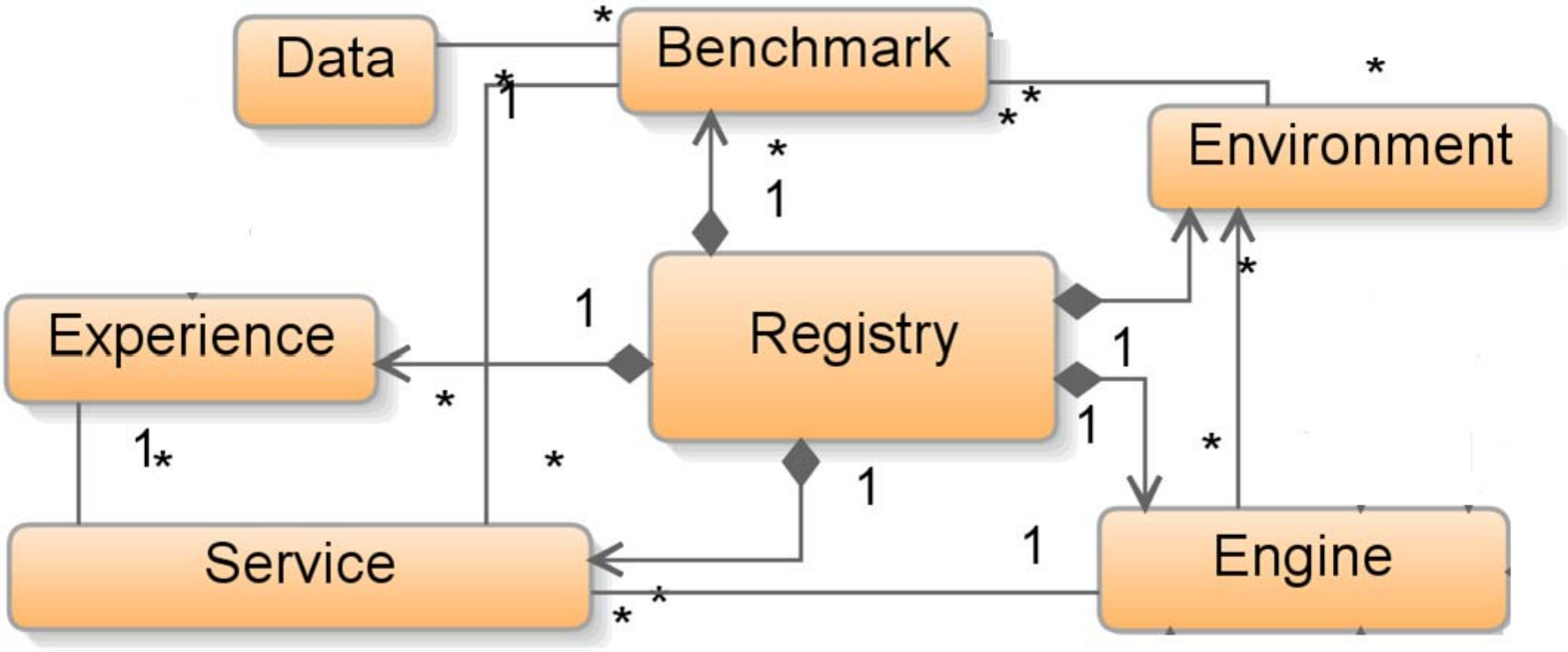
Measurement techniques

- Measuring CPU time and memory usage:
  - Elapsed time
  - Linux, Unix: TOP, time
  - Windows: PsList
  - Java: JIP, HPROF

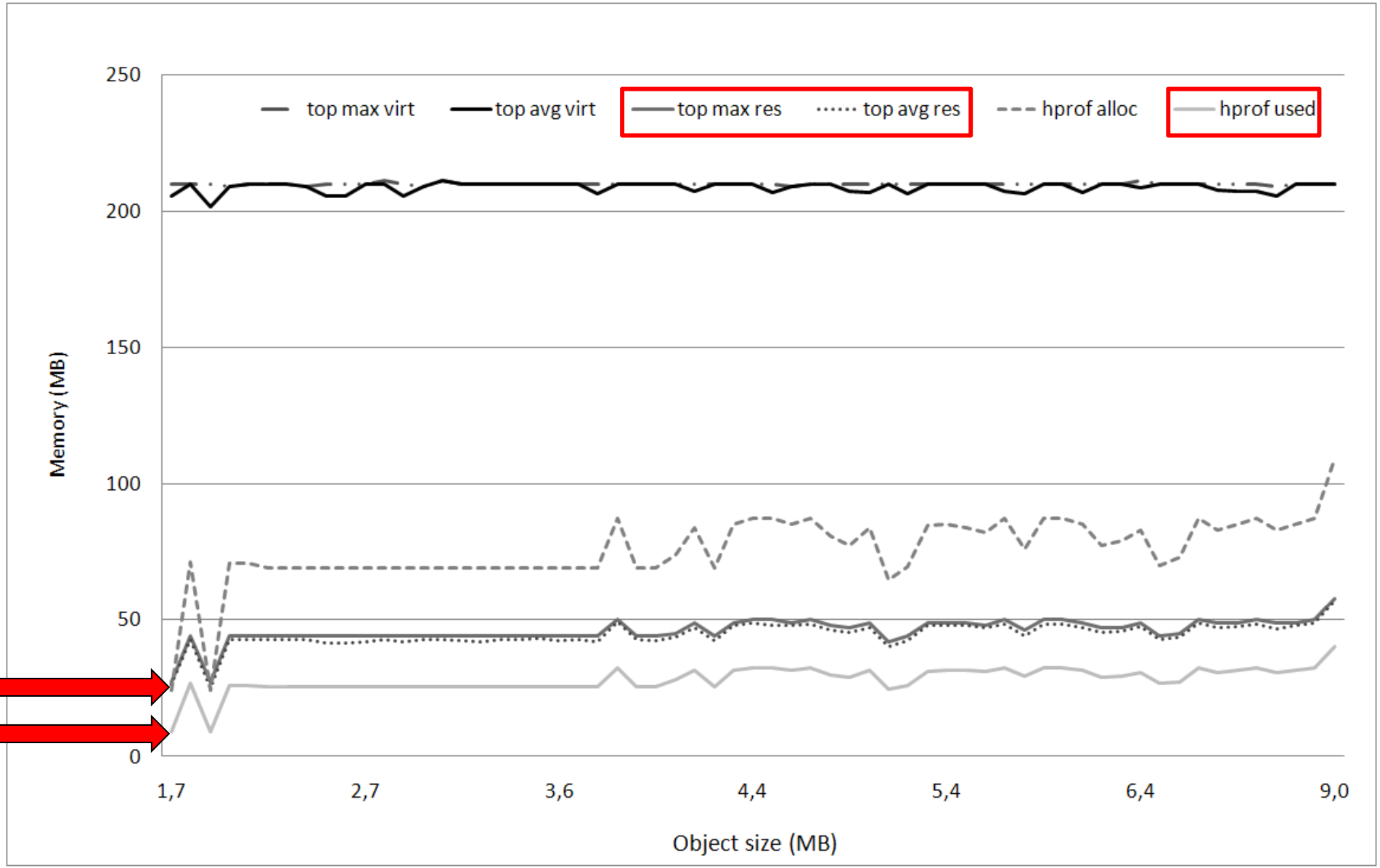


Design for a controlled evaluation environment

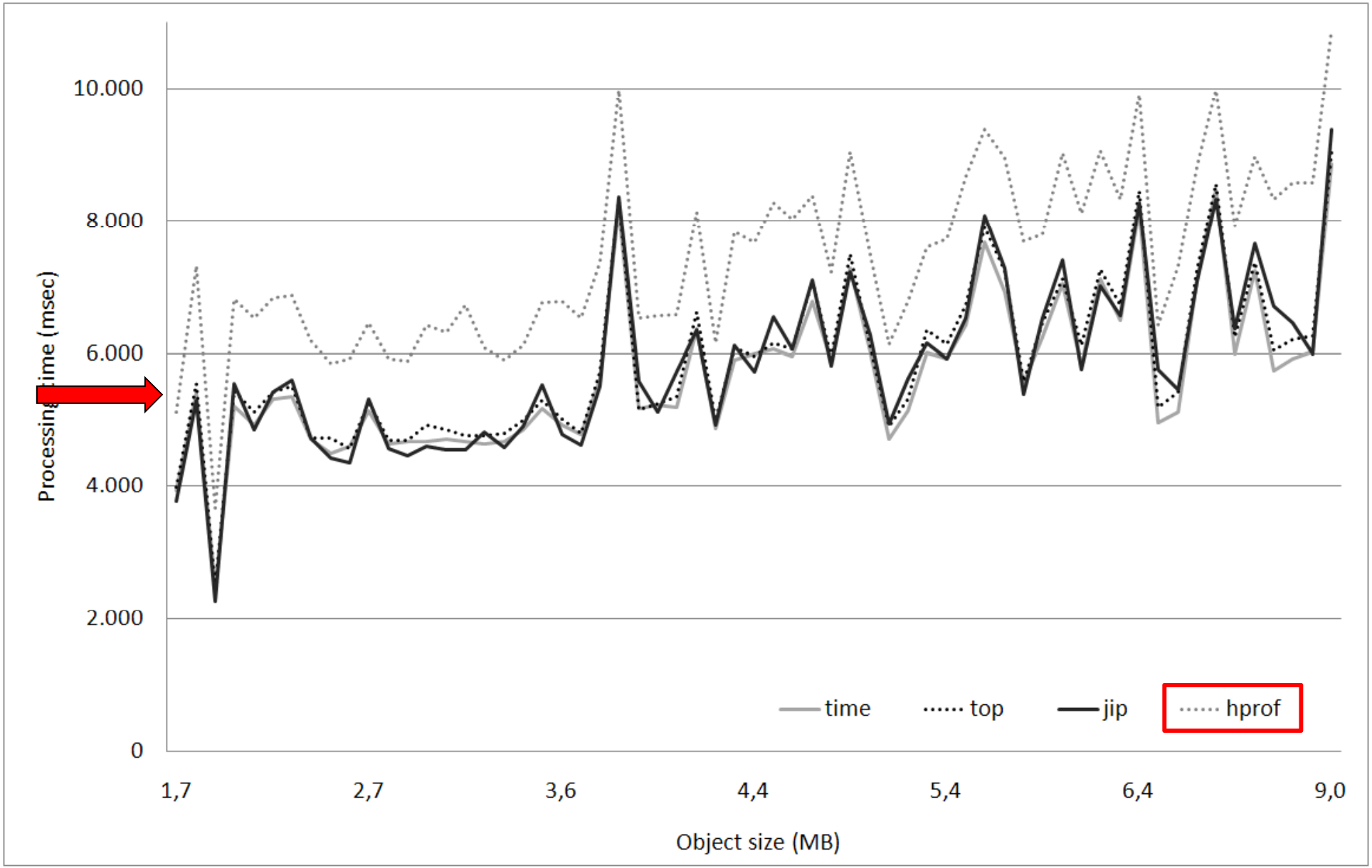
- Non-invasive provider-side service instrumentation
- Engines make components quality-aware
- Environments have associated benchmark scores
- Registry accumulates experience



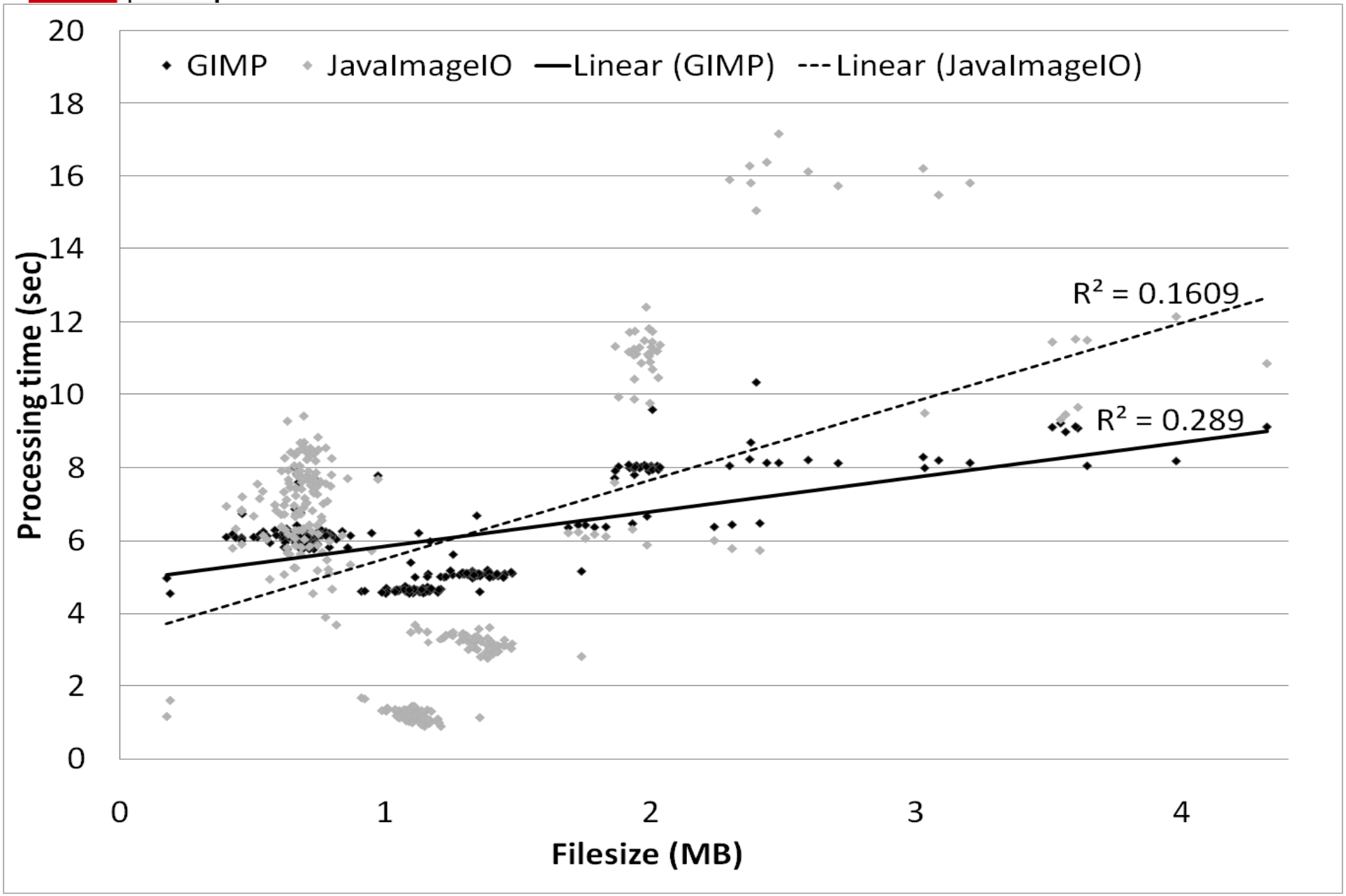
Profiling memory usage of Java tools



Profiling timing of Java tools

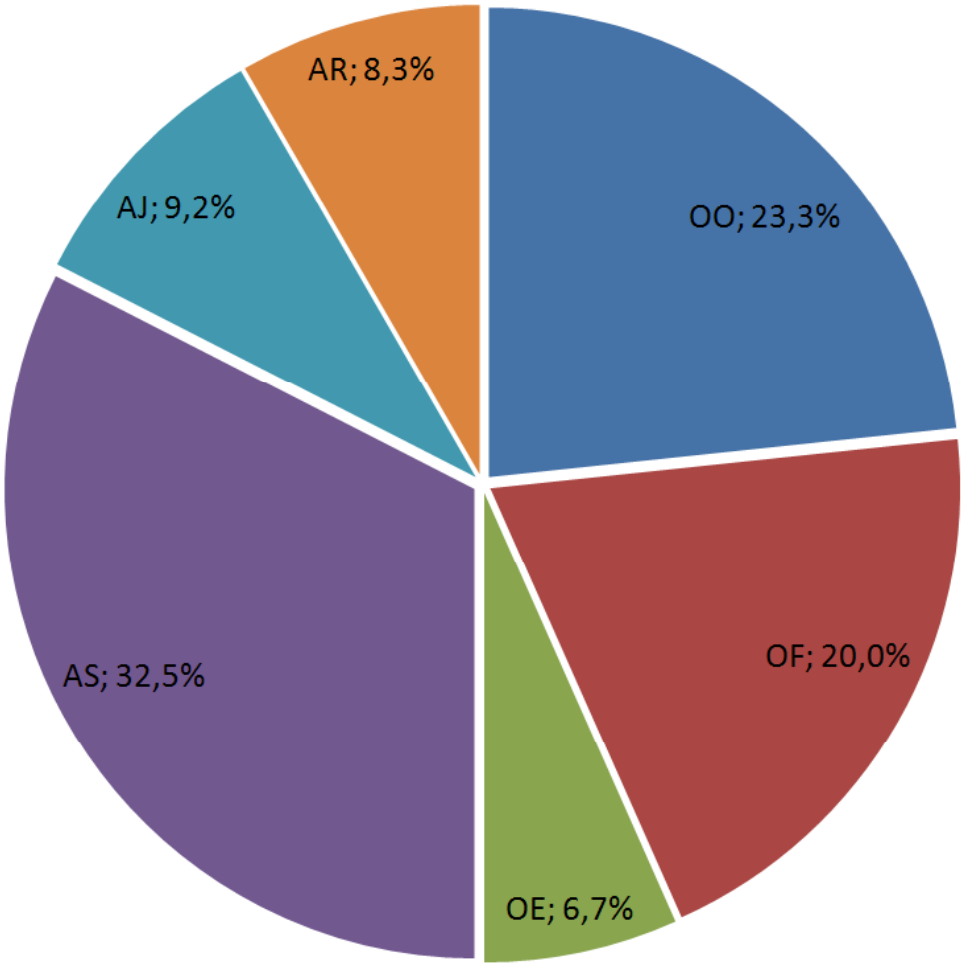


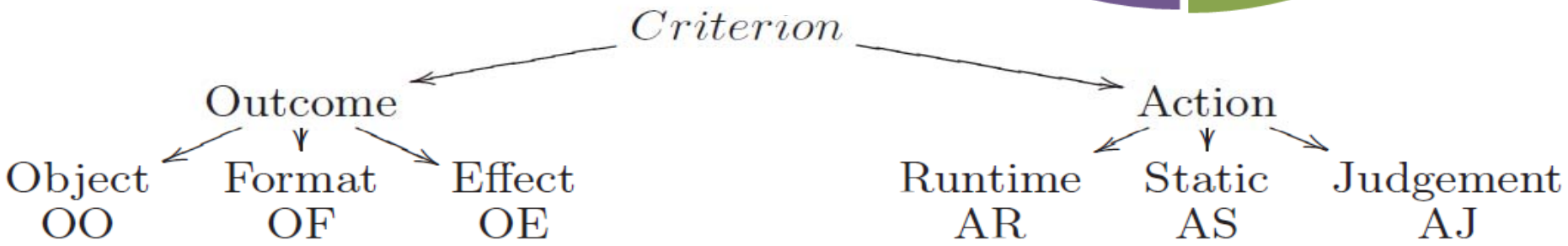
Comparing tool performance



Decision criteria

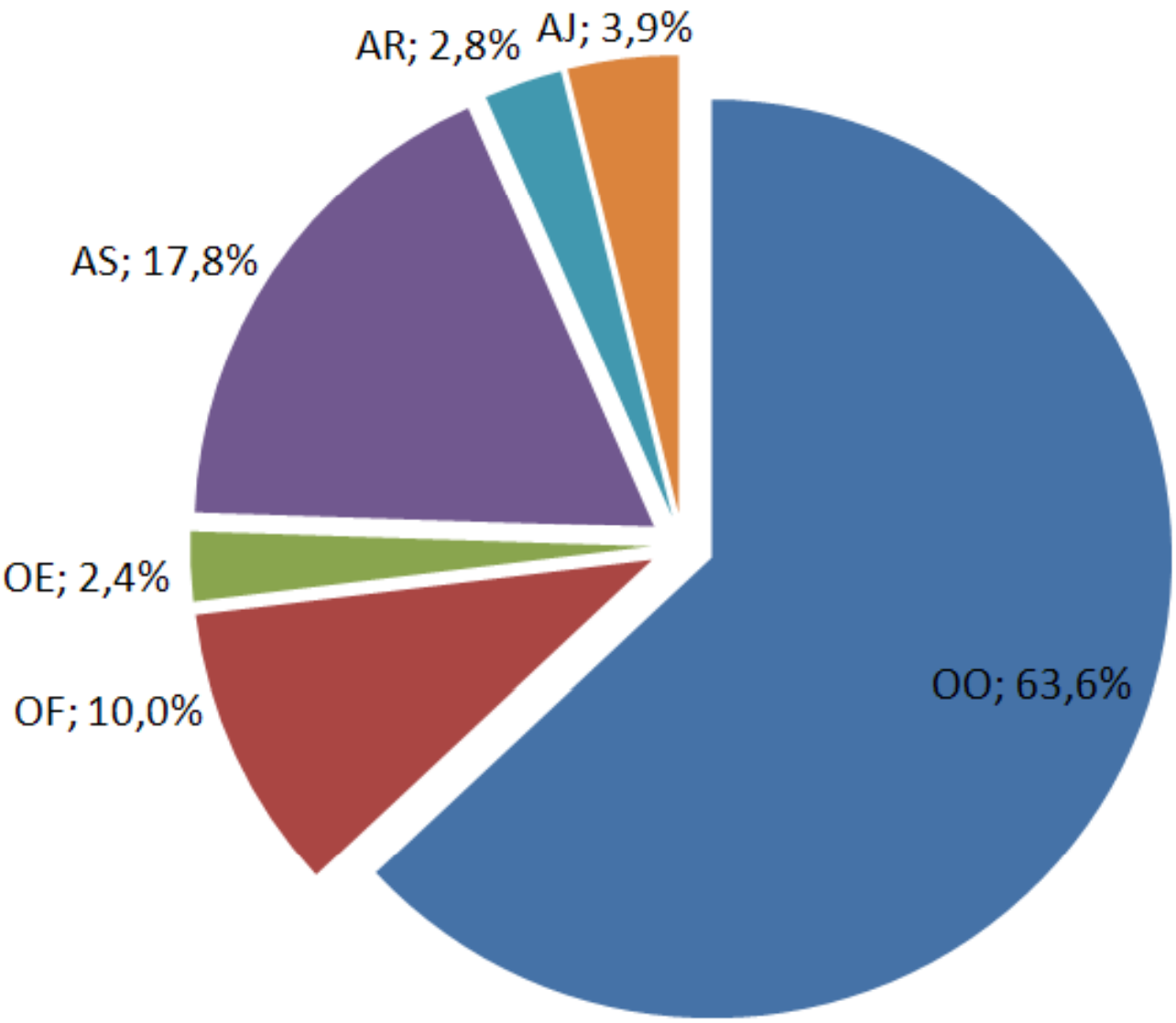
- Distribution in four case studies on scanned images





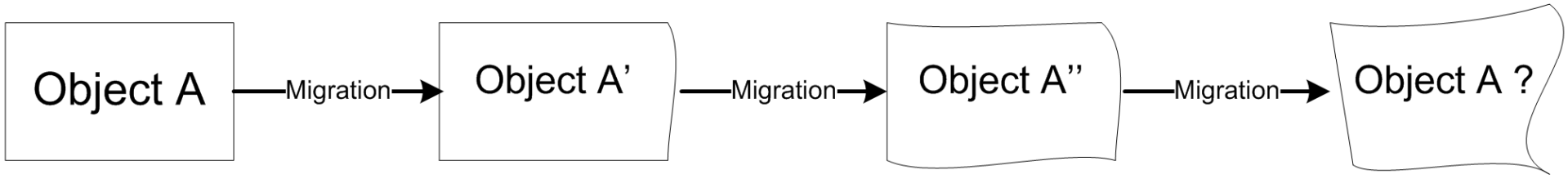
Decision criteria

- Distribution in thirteen cases on various types of content



Core requirement: Keep object intact

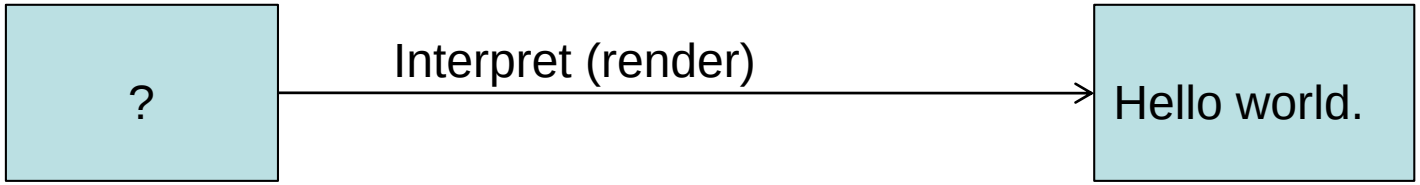
- Essential object characteristics
  - Content
  - Appearance
  - Structure
  - Behaviour
  - Context

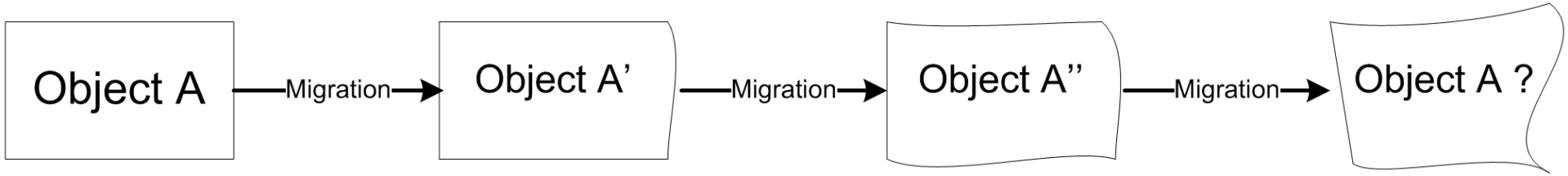


Validating a migrated image

- Yes, it's in JPEG 2000 format

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





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)*.

RAW

Adobe

dcraw recoveres edges

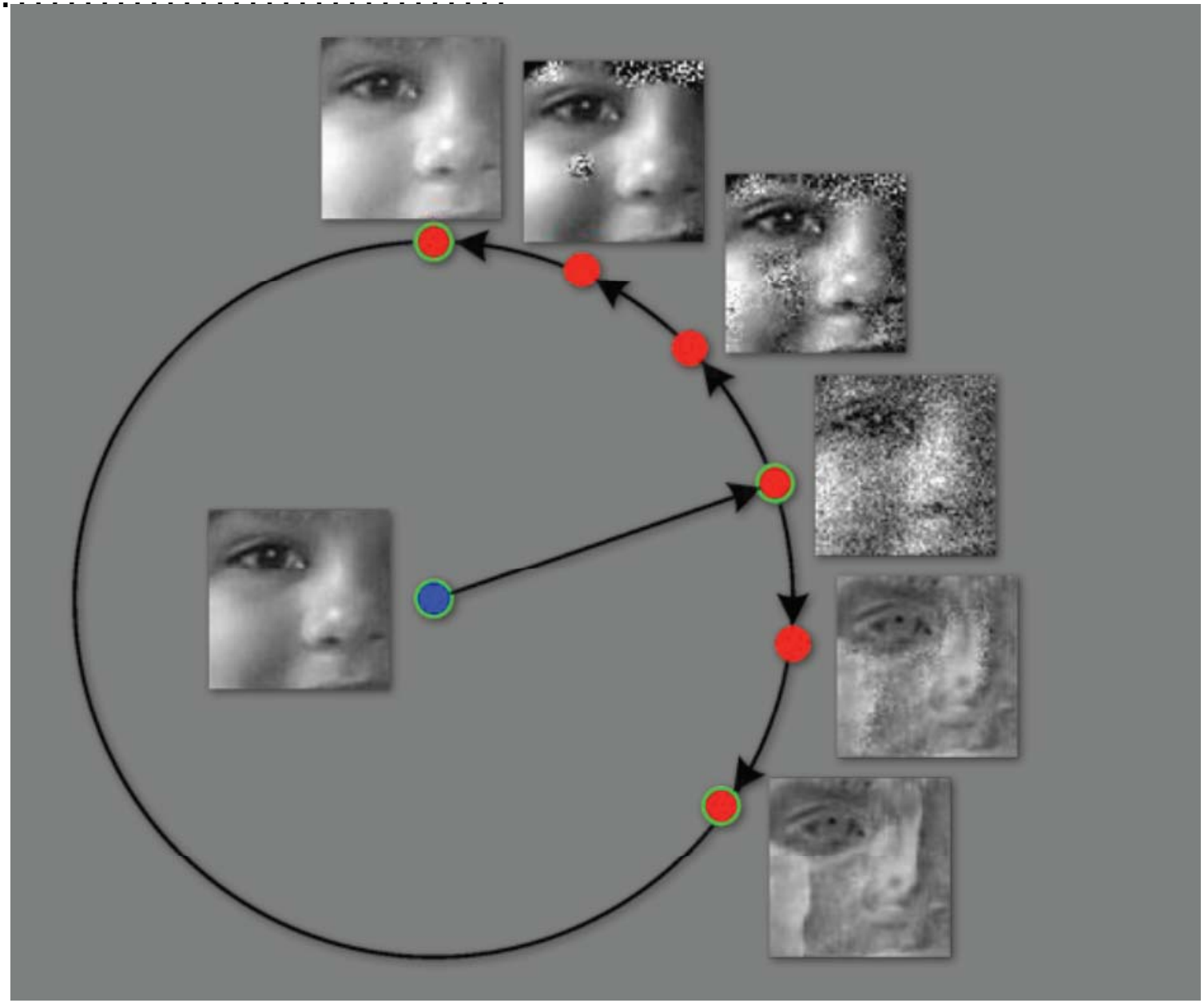
dcraw

Apple

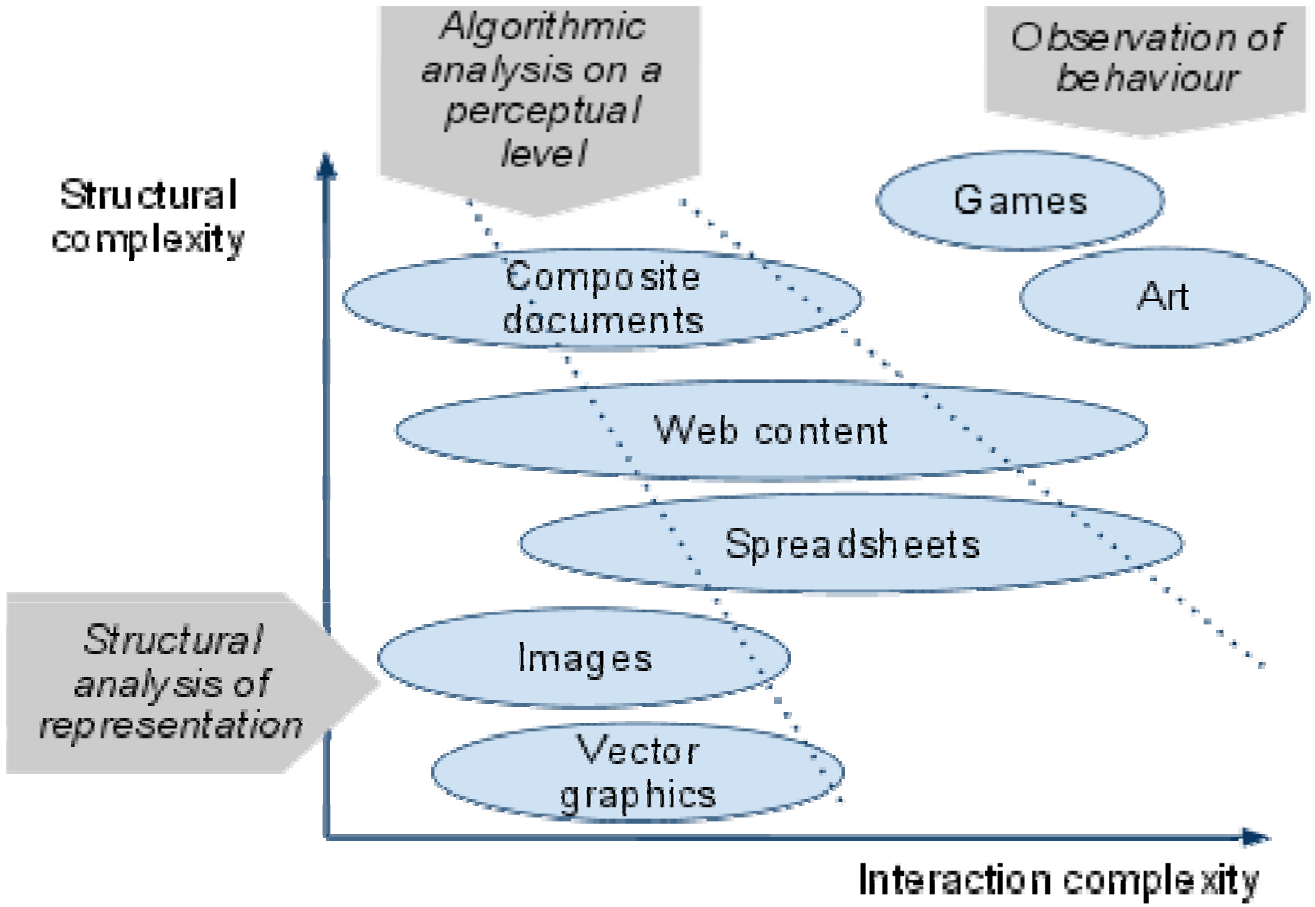
Distance metrics: How meaningful?

AE
PAE
RMSE
...
SSIM

Anything but “0” is a problematic result

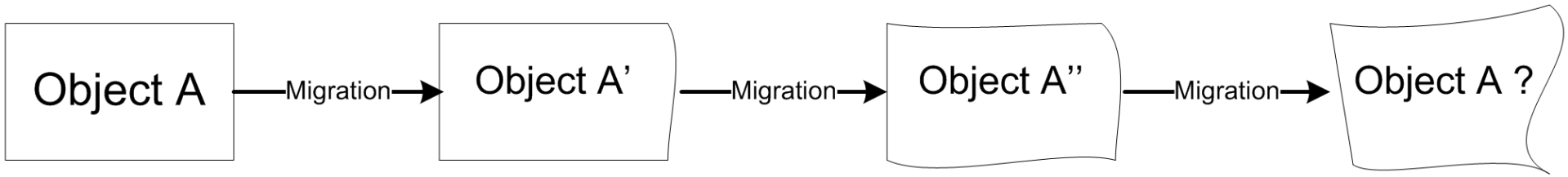


Approaches to analysing content

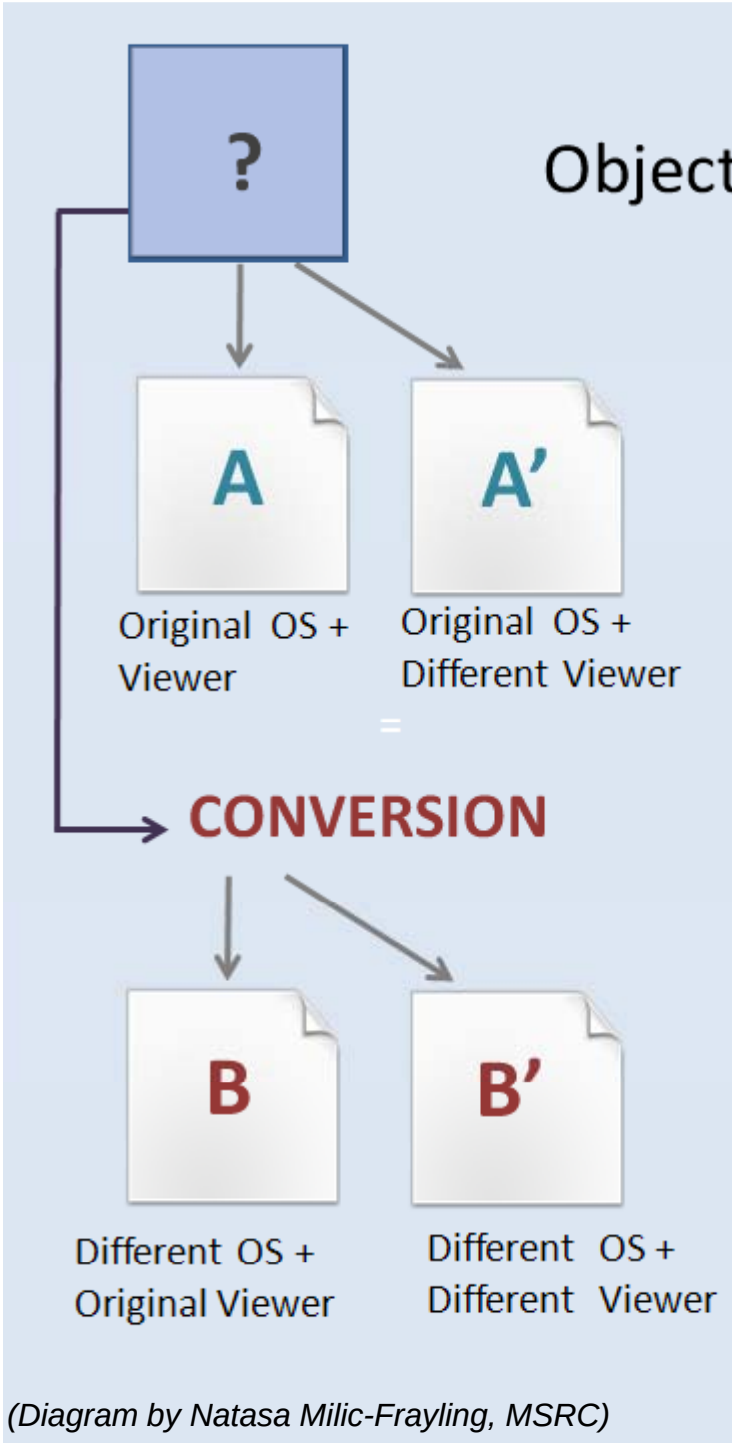


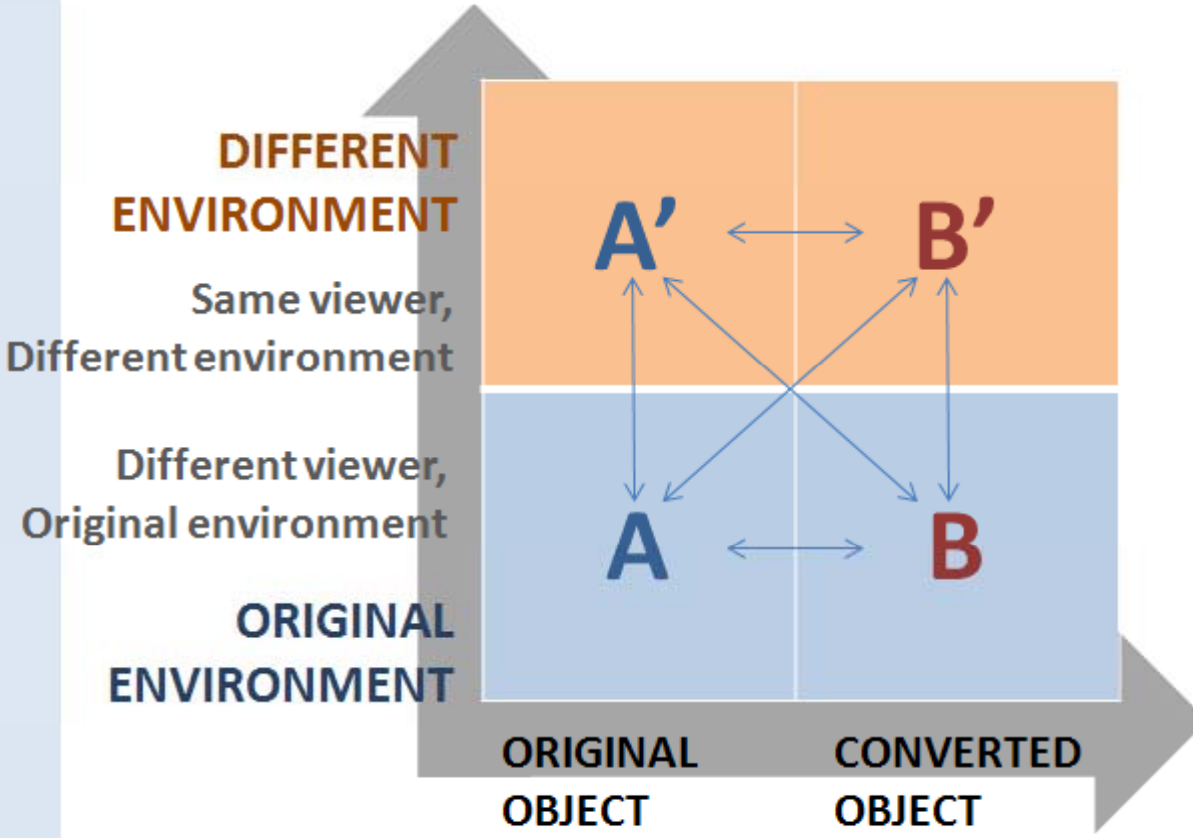
Exercise

- Download set of documents from:
  - www.ifs.tuwien.ac.at/~becker/teaching/dp/ss11/qa-exercise.zip
  - In the zip file you find:
    - Folder `collection-selection`
    - Requirements tree `electronic-documents.mm (+.png)`
- Collection-selection contains original and migrated files
  - E.g. FinalE5.doc, FinalE5.pdf
- Take requirements tree `electronic-documents` and evaluate object characteristics
  - 20 minutes evaluation
  - Note observations
  - Discussion



Objects, environments and dependencies





...networks of objects

(Diagram by Natasa Milic-Frayling, MSRC)

How to measure?

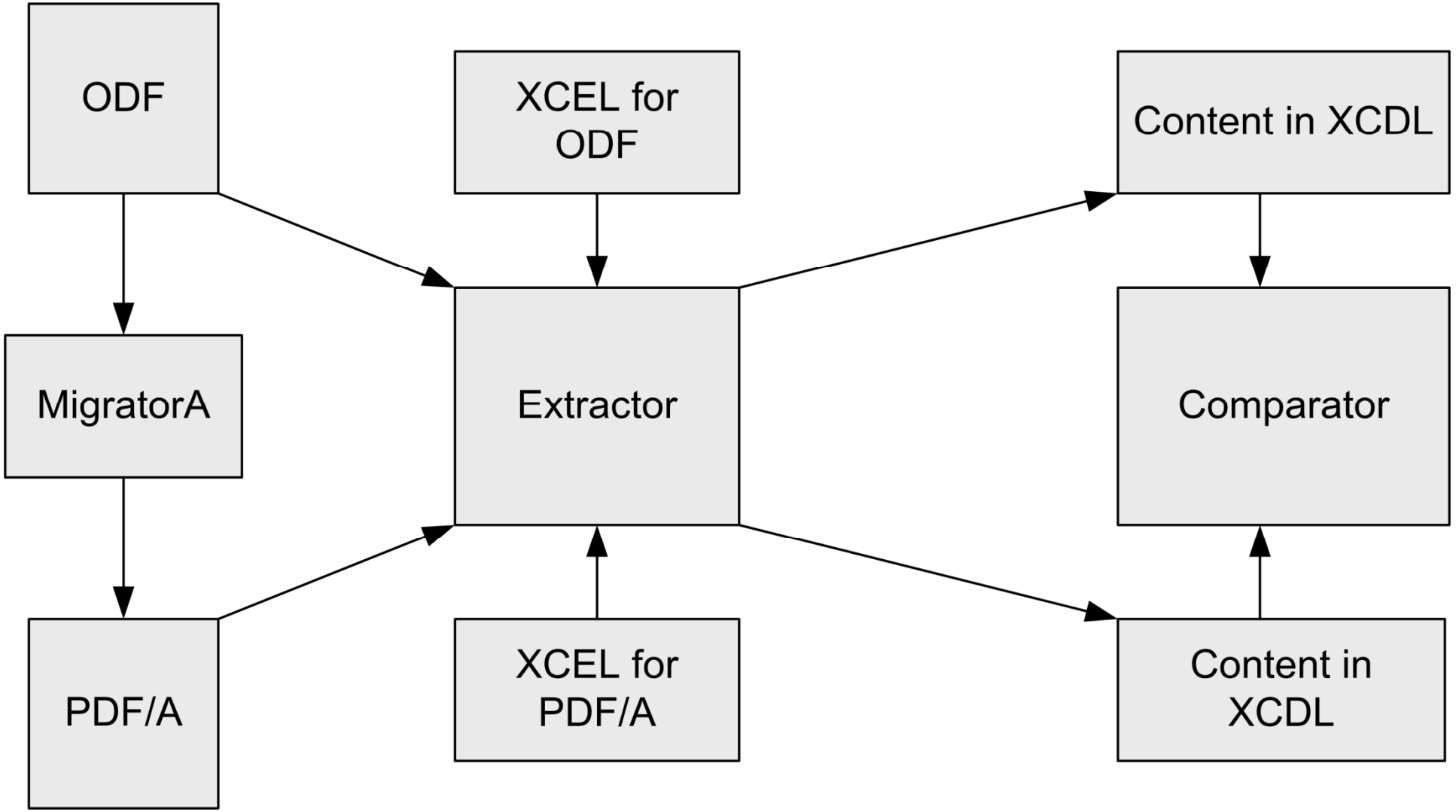
Category	Abbr.	Example	Data collection and measurements
Outcome object	OO	*Image pixelwise identical* (RMSE)	Measurements of input and output, measurements taken in controlled experimentation
Outcome format	OF	*Format is ISO standardised* (boolean)	Measurements of output, trusted external data sources
Outcome effect	OE	*Annual bitstream preservation costs* (€)	Measurements of output, trusted external data sources, models, partly manual calculation and validation, sharing
Action runtime	AR	*Throughput* (MB per ms)	Measurements taken in controlled experimentation
Action static	AS	*License costs per CPU* (€)	Trusted external data sources, manual evaluation and validation, sharing
Action judgement	AJ	*Configuration interface usability* (excellent, sufficient, poor)	Manual judgement, sharing

In-depth characterisation approaches

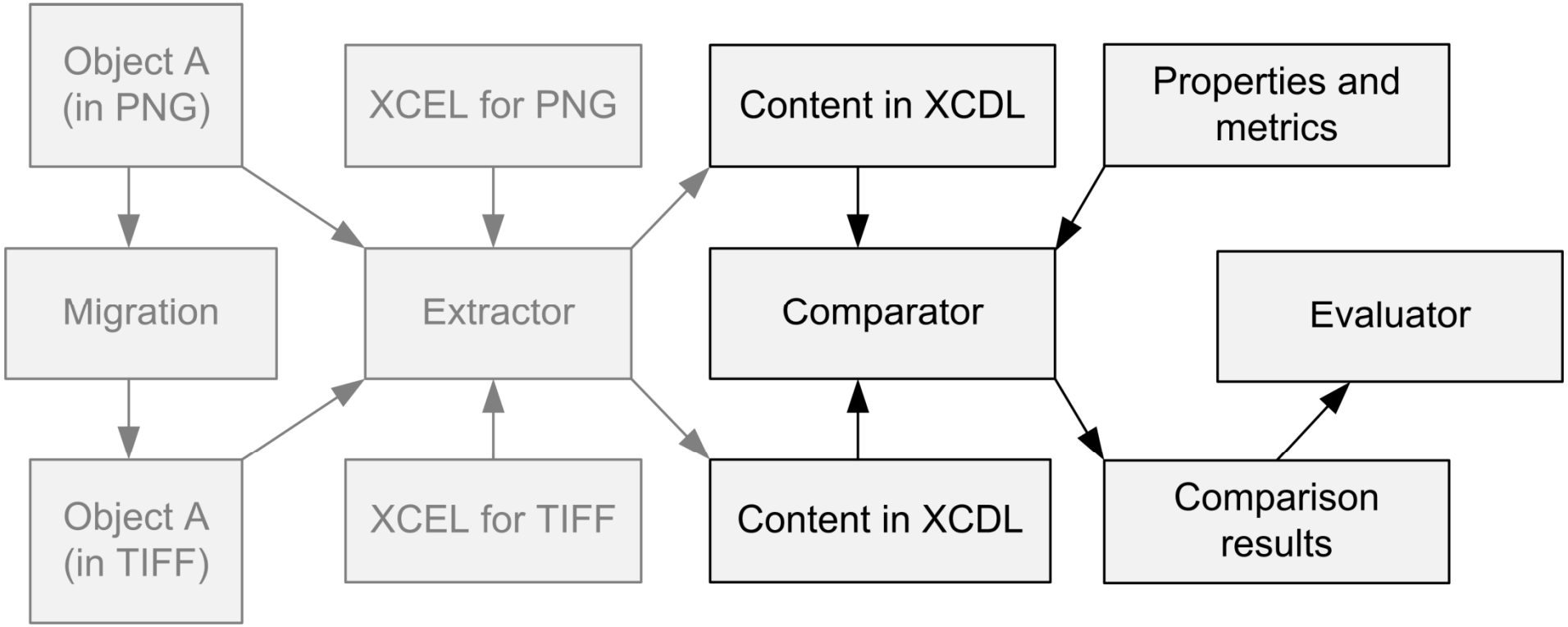
- XCL…
- eXtensible Characterisation Languages
  - XCDL, the description language
  - XCEL, the extraction language

- Bitstream Segment Graphs (BSG)
- New approach based on reasoning and rules

XCL: Structural analysis



Automating the evaluation



Bitstream Segment Graphs

- Use a graph to describe the structure of a file
- Define sets of rules used by a reasoner to create such a graph
- General rule base, format-specific rules
- Reasoner calculates BSG and “coverage” of a file
- BSG editor allows construction and exploration of the map

Bitstream Segment Graphs

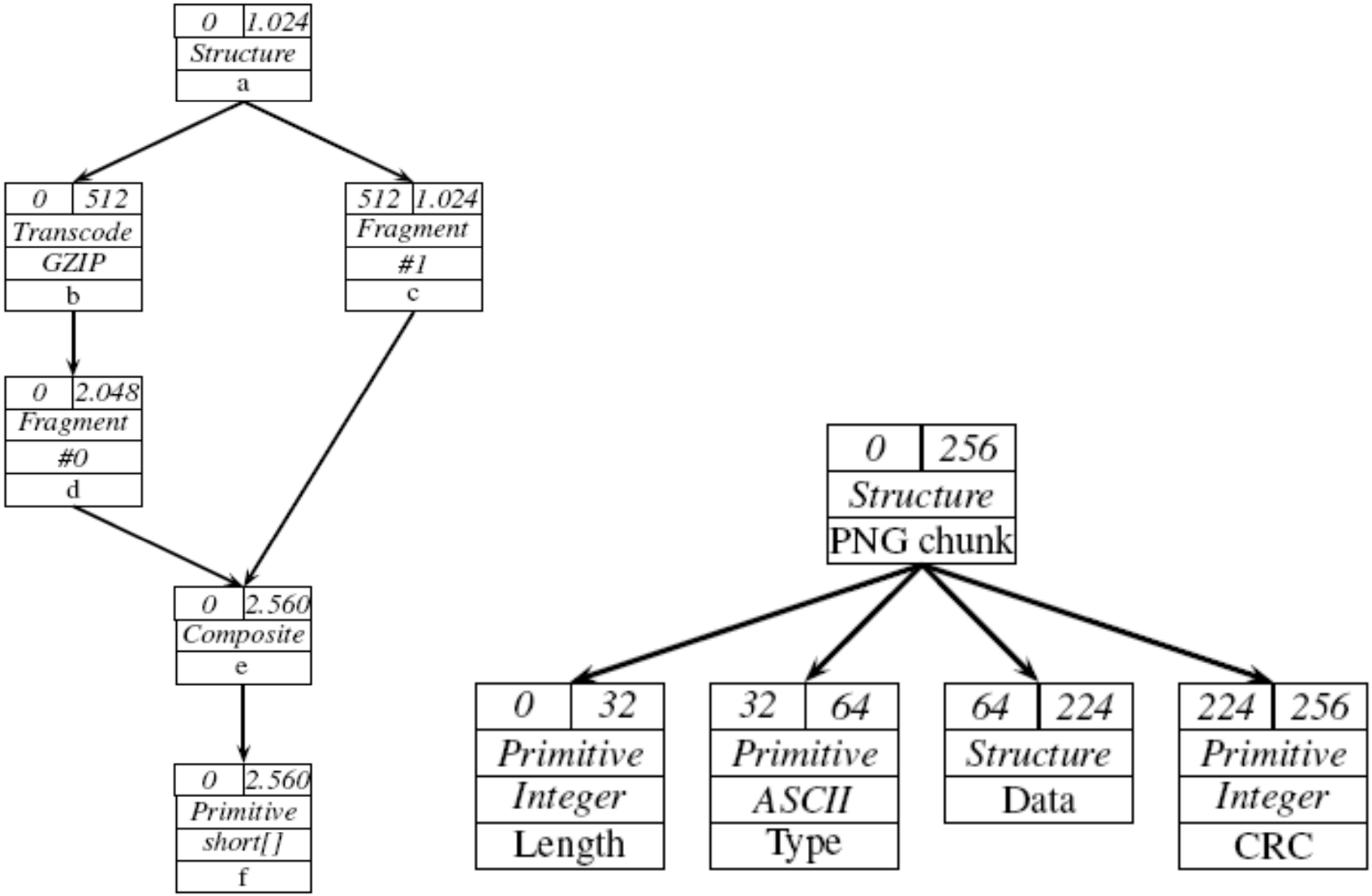


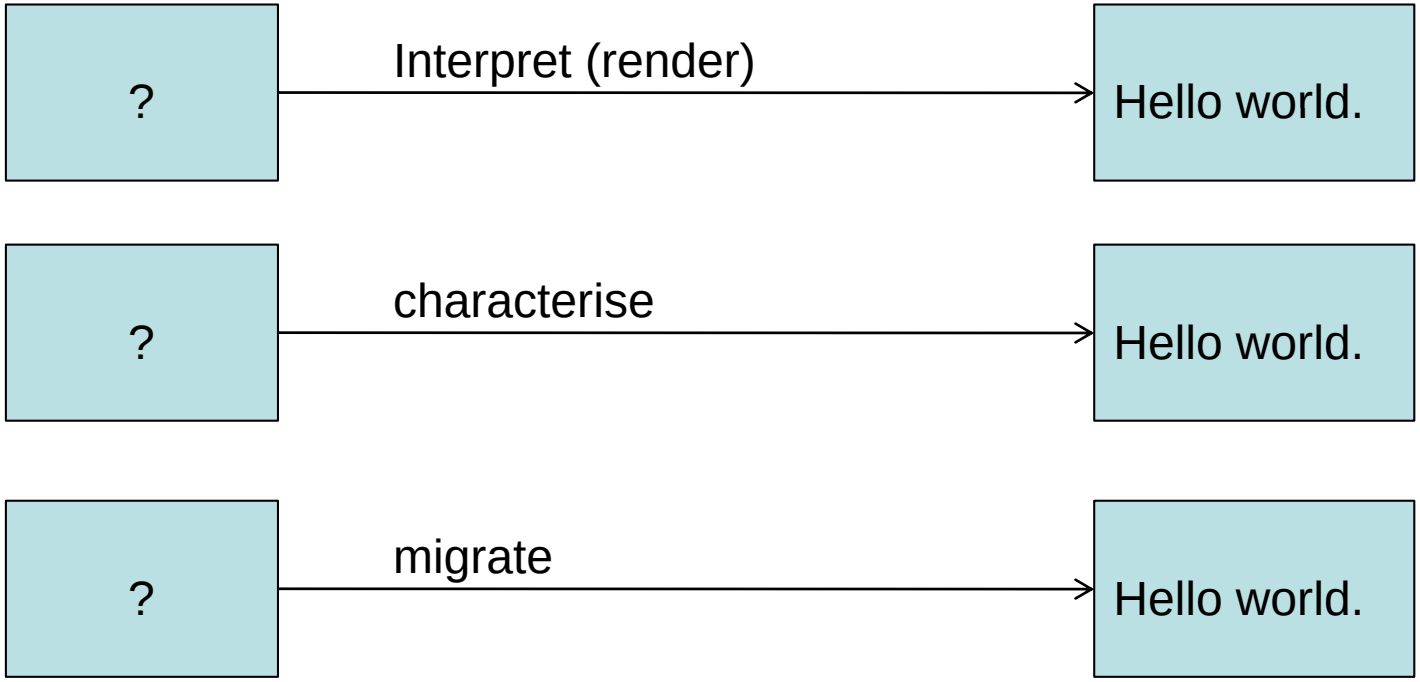
Figure 3: BSG instance for a PNG chunk.

Bitstream Segment Graphs

- DEMO

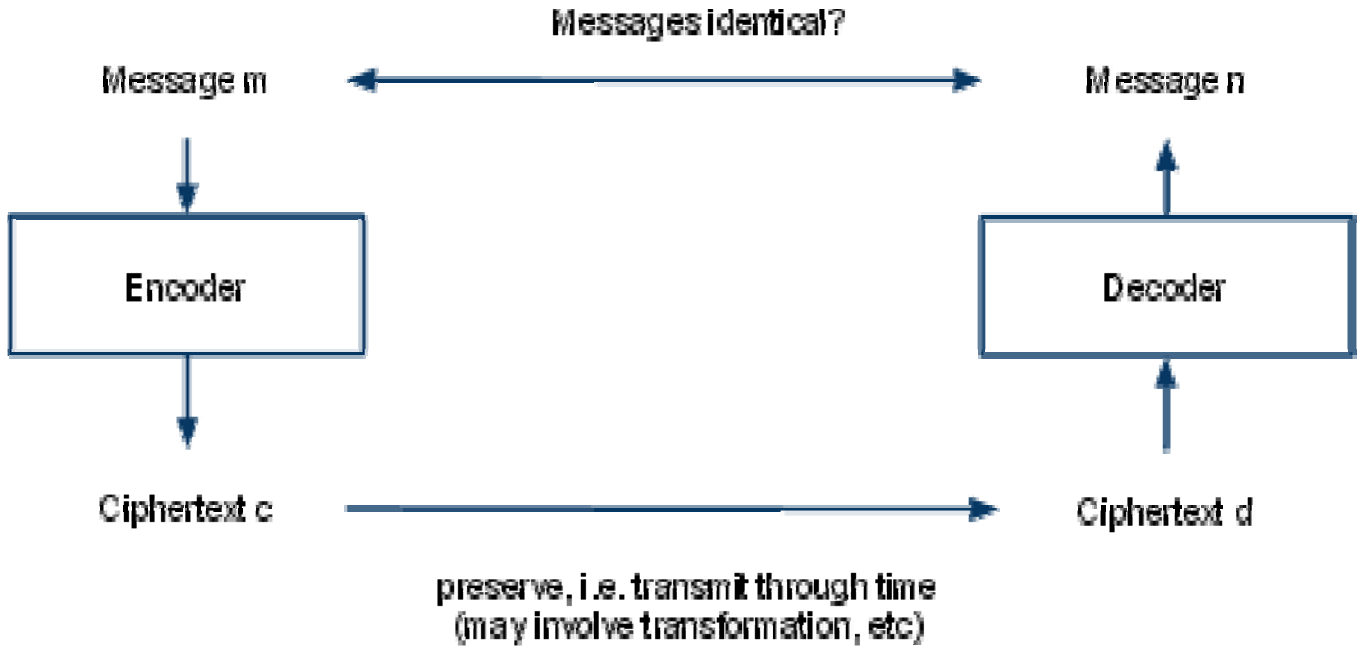
- http://wiki.dataformats.net/apeiron/launcher.jnlp

Interpretation levels



- Every characterisation is an interpretation
- Every interpretation is a transformation
- There is no ground truth (normally)…

... DP 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 receiver may not be the one who was targeted

Questions?

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

>>> DP UE Tasks are presented now