Preservation planning 2

What to decide and how

March 31, 2014

Kresimir Duretec

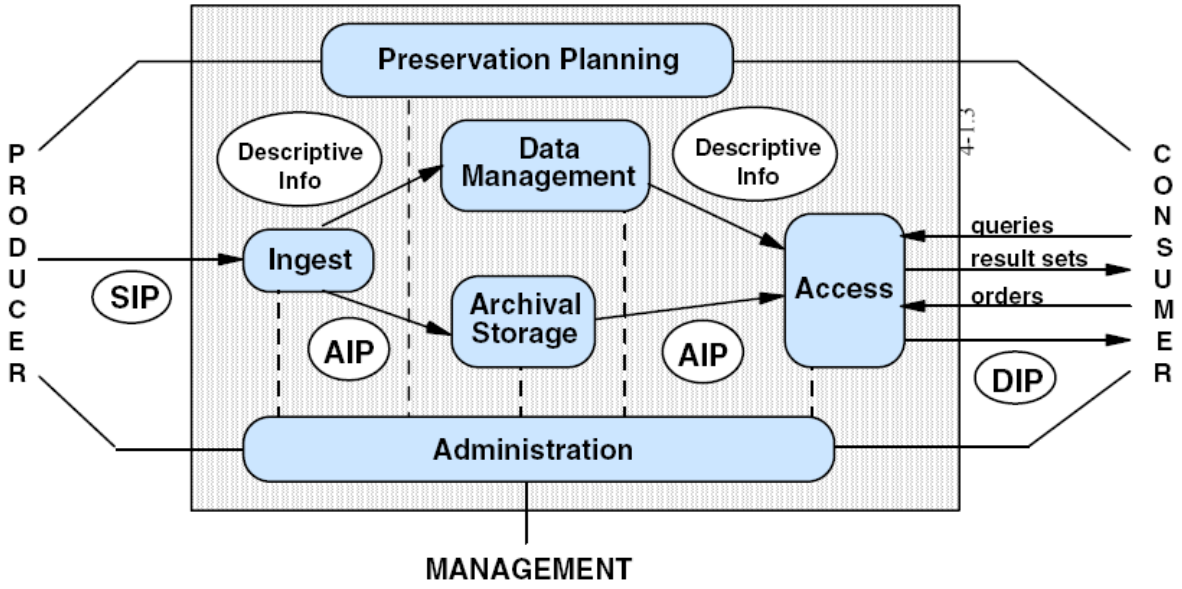
Preservation Operations

- Core operations for preservation
  - Analyse content
  - **Perform preservation actions**
  - Perform Quality Assurance
  - Manage metadata
  - Report

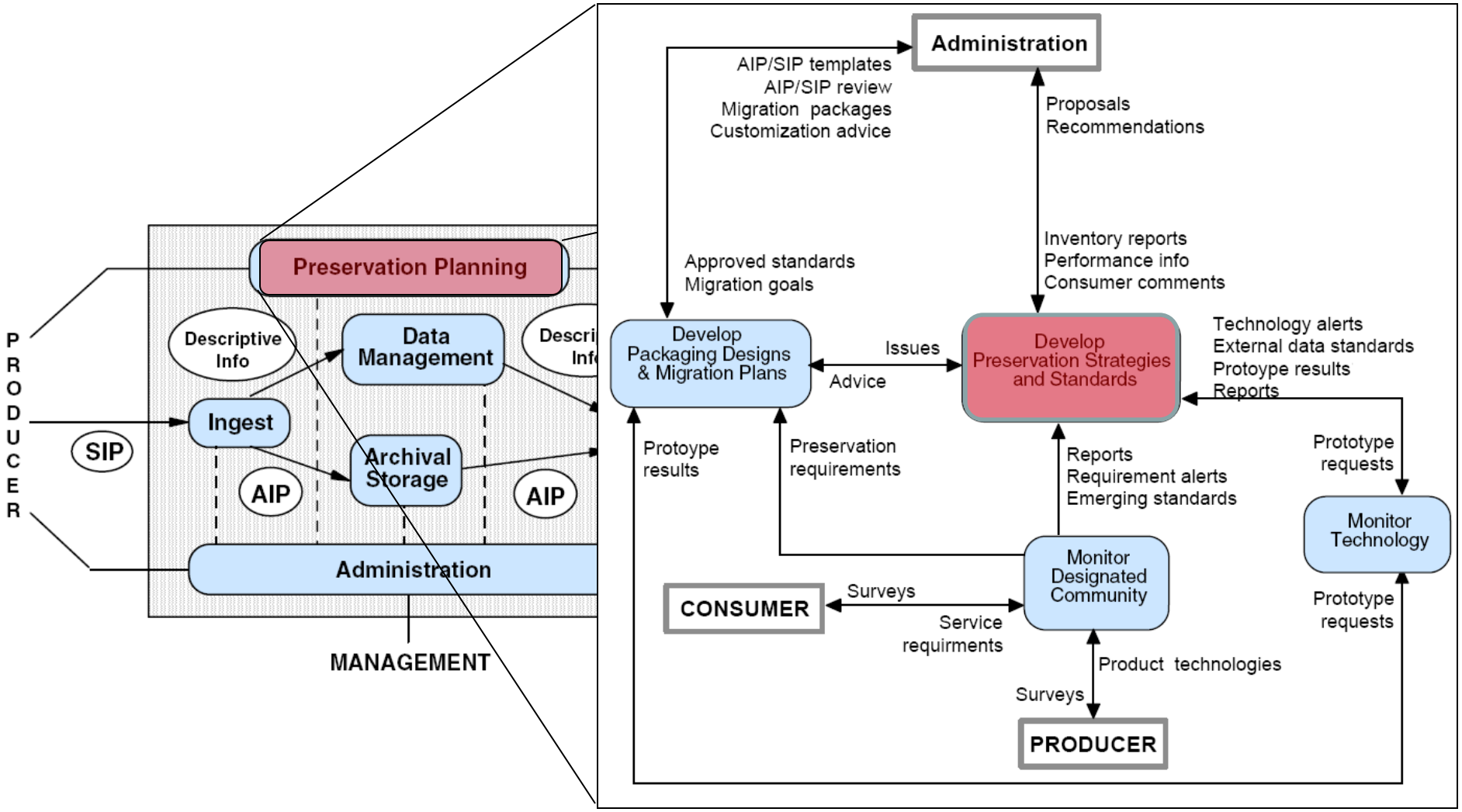
Perform preservation actions

- Several preservation strategies developed
  - For each strategy: several tools available
    - For each tool: several parameter settings available
- How do you know which one is most suitable?
- What are the needs of your users? Now? In the future?
- Which aspects of an object do you want to preserve?
- What are the requirements?
- How to prove in 10, 20, 50, 100 years, that the decision was correct / acceptable at the time it was made?

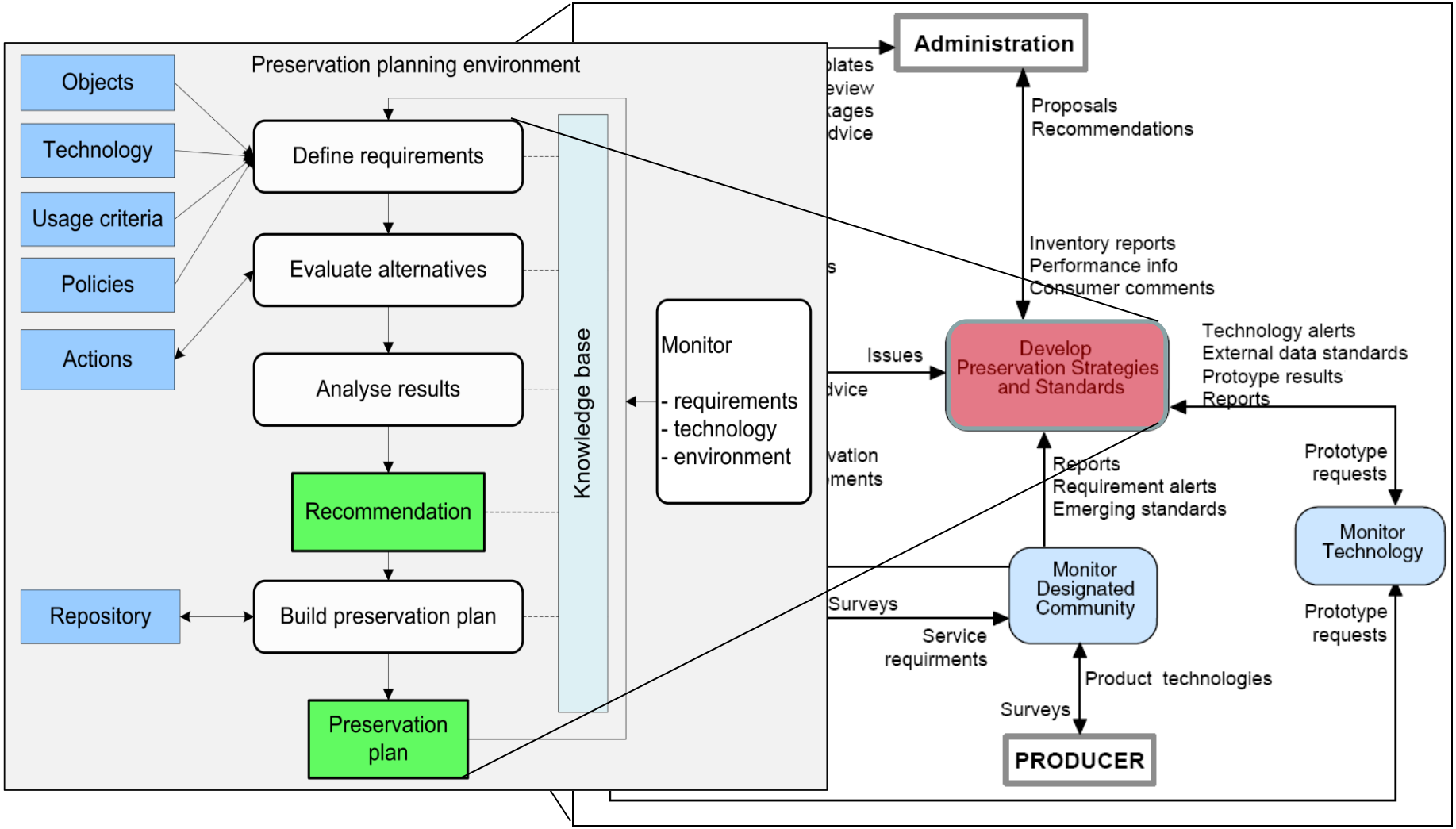
Preservation Planning



Preservation Planning



Preservation Planning



Preservation Planning: Question 1

- **How can we select the optimal preservation action for a given setting?**
  - What are the drivers and constraints on the decision space?
  - What are the goals and objectives?
  - What are the factors influencing the decision makers' preferences?
  - How can we model multiple competing objectives and requirements?
  - How should we evaluate software components?

Preservation Planning: Question 2

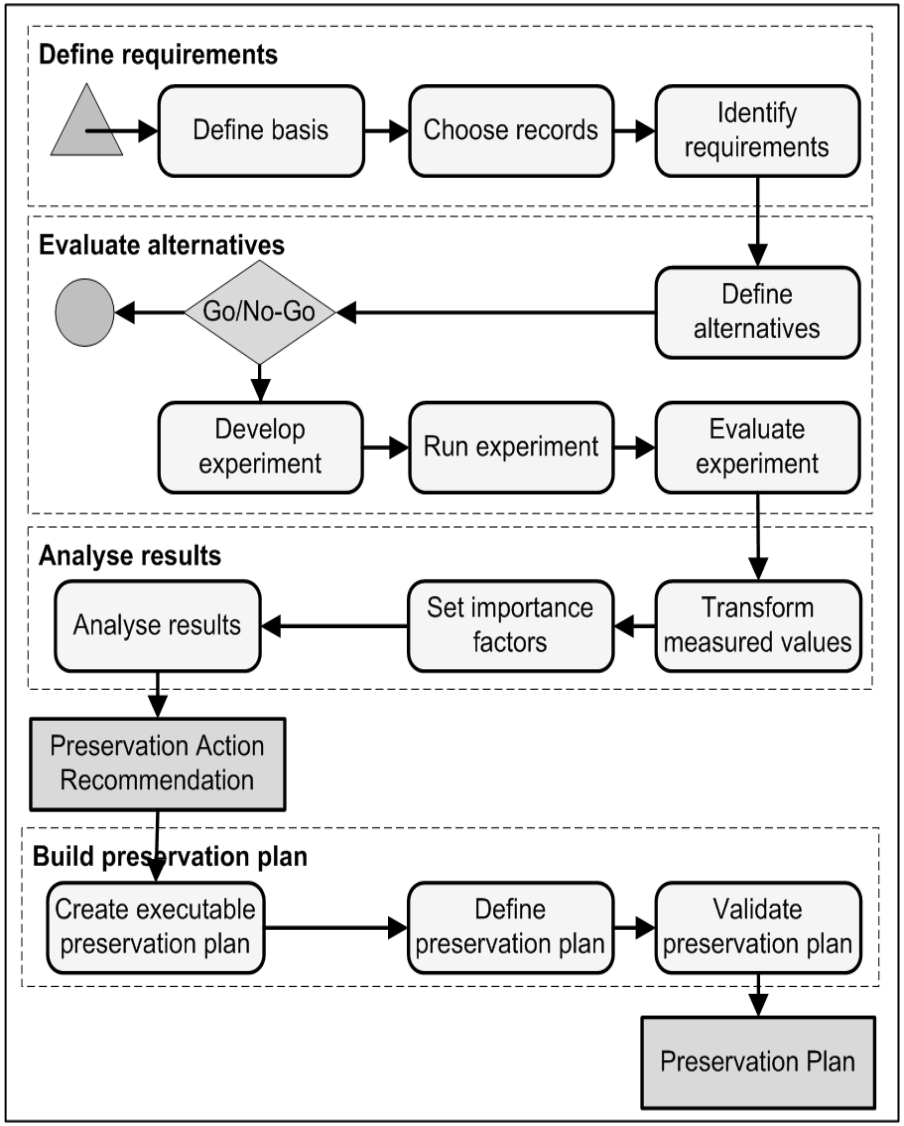
- **How can we ensure trustworthy preservation planning?**
  - What are the requirements on trust that need to be addressed?
  - What decision steps and evidence need to be documented?
  - What are the aspects that a plan needs to address, and what are the elements needed to cover them?
  - How can we ensure reliable evaluation procedures and repeatable evidence?

Preservation Planning: Question 3

- **How can we ensure that decision processes scale up?**
  - How can we automate decision making?
  - How can we integrate continuous monitoring?
  - Which properties can be measured automatically, and how?
  - How can we create a controlled environment for observing the behaviour of components in a reproducible way?

Planning workflow

- 4 major steps
  - Define requirements
  - Evaluate alternatives
  - Analyse results
  - Build preservation plan



Define basis

- What are the objects?
- What are the fundamental requirements?
  - Authenticity, reliability, integrity, usability
  - Metadata (for different purposes)
- What are the applying policies, legal constraints, regulations…
  - User groups, target community
  - Institutional settings

Define sample objects

- Representative for the objects in the collection
- They should cover all essential features and characteristics of the collection in question
- As few as possible, as many as needed
- Often between 3-10

- ... c3po

Requirements definition

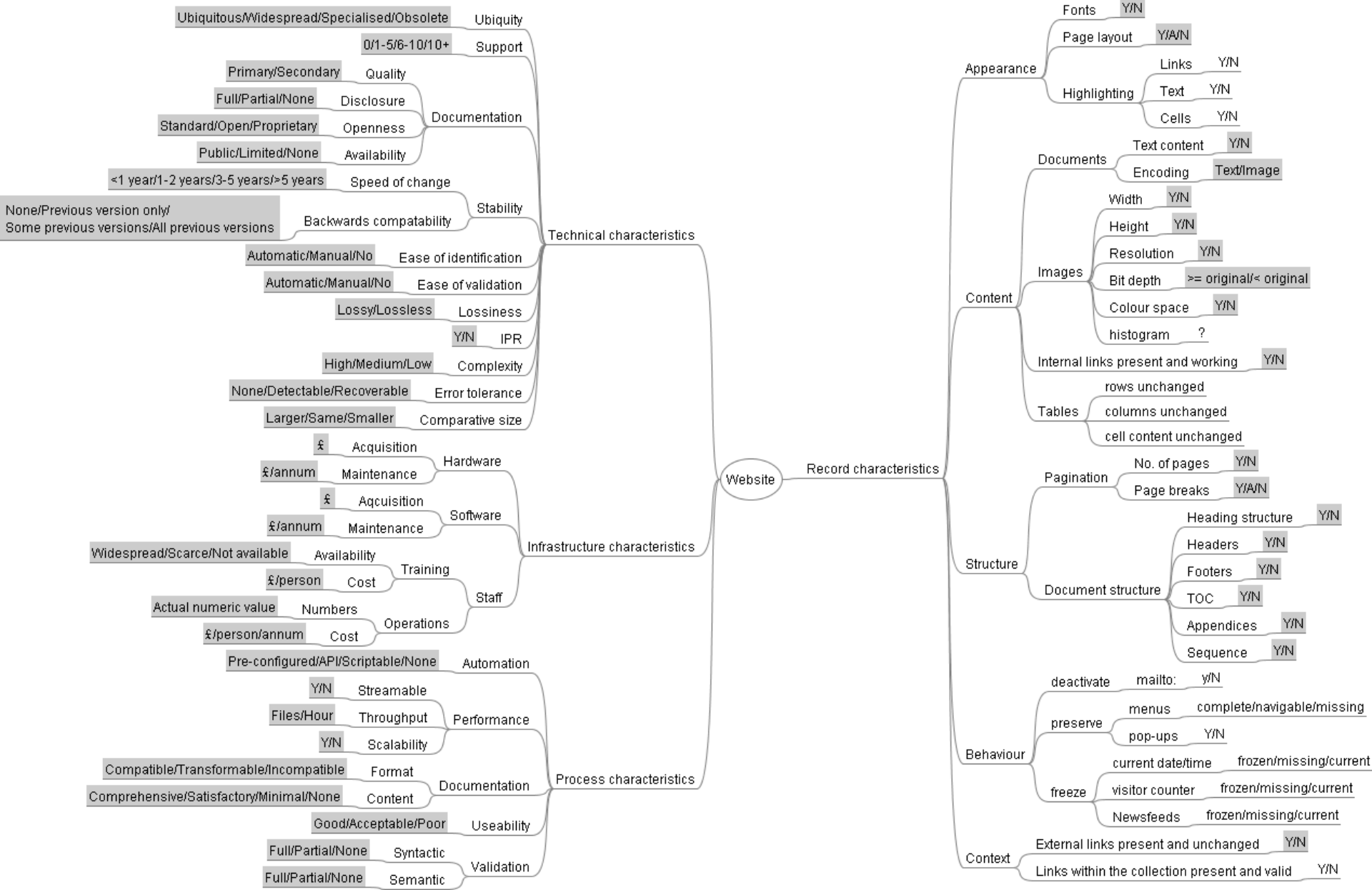
- What are our goals and objectives?
- How do we measure achievement of our goals?
- Which drivers have an influence on which objectives?
- Define complementary criteria for all objectives
  - Trade-offs between objectives might eventually be necessary
    - Usability vs. authenticity
    - Structure vs. independency
    - Access vs. costs
    - …
- How can we ensure criteria are free of ambiguity?

The Objective Tree

- Tree structure describing requirements and goals
  - A weighted hierarchy of objectives leading into measurable criteria
  - A utility function for each criterion specifies the organisation's assessment for the range of possible values

- Created top-down or bottom-up
  - Start from high-level goals and break down to specific criteria
  - Collect criteria and organize in tree structure

An Objective Tree



Decision criteria: What to measure?

- 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…)

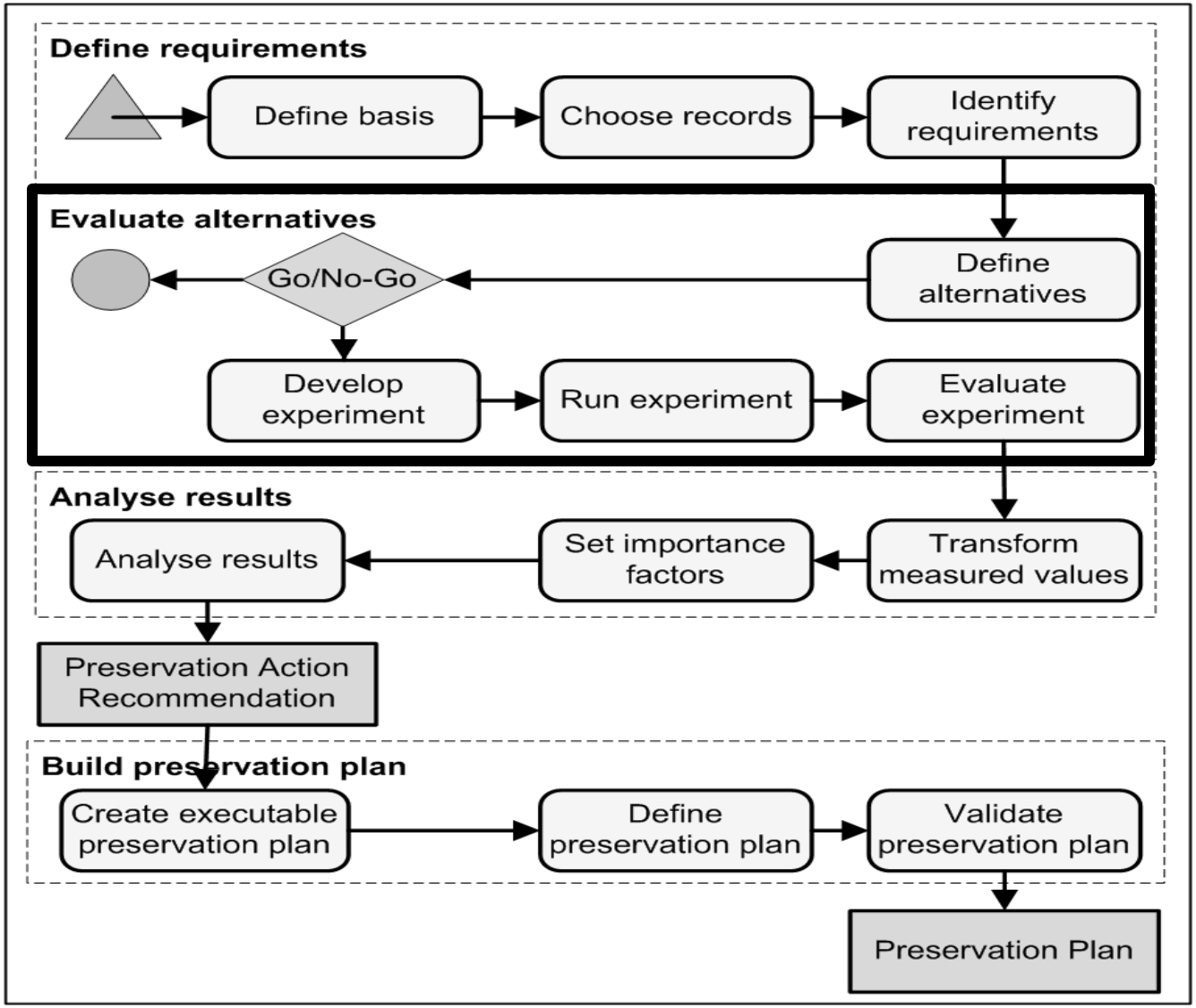
Decision criteria: What to measure?

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
Action judgement	Technical interoperability, configuration flexibility	Manual judgement, sharing	

Results of Phase 1

- Defined and documented the context of a preservation problem
  - Which types of objects
  - Which environment
  - Purpose and target consumers
  - Obligations and constraints
- Defined and documented representative samples for performing experiments
- Defined and documented goals and objectives
  - From goals and requirements to measurable criteria

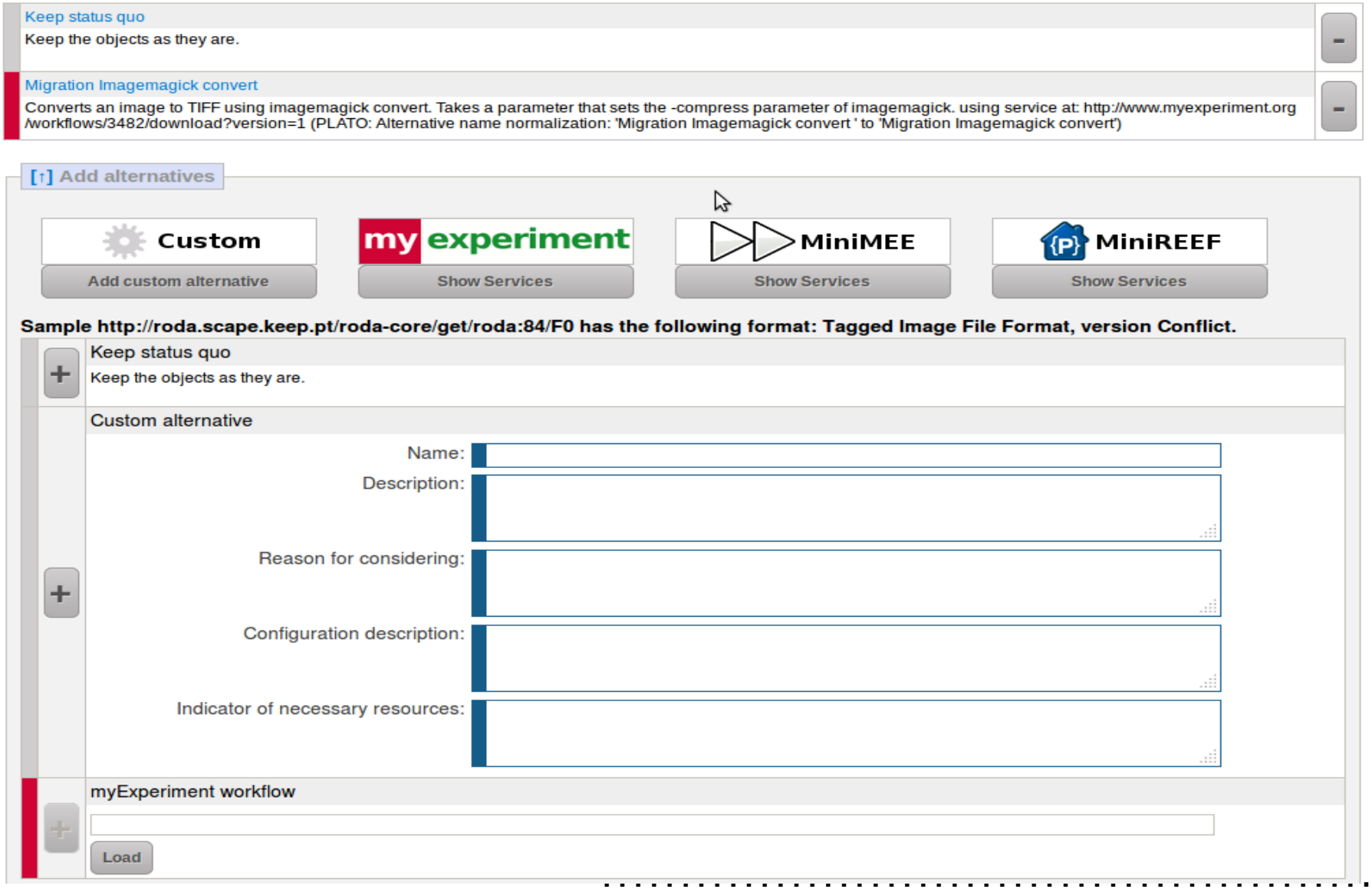
Evaluate alternatives



Define alternatives

- Given the type of objects and requirements, what strategies would be best suitable/are possible?
  - Migration
  - Emulation
  - Both
  - Other?
- For each alternative precise definition of
  - Which tool (OS, version,...)
  - Which functions of the tool in which order
  - Which parameters

Component discovery in Plato 4



Define alternatives

- Detailed design and overview of the resources for each alternative
  - human resources (qualification, roles, responsibility, …)
  - technical requirements (hardware and software components)
  - time (time to set-up, run experiment,...)
  - cost (costs of the experiments,...)

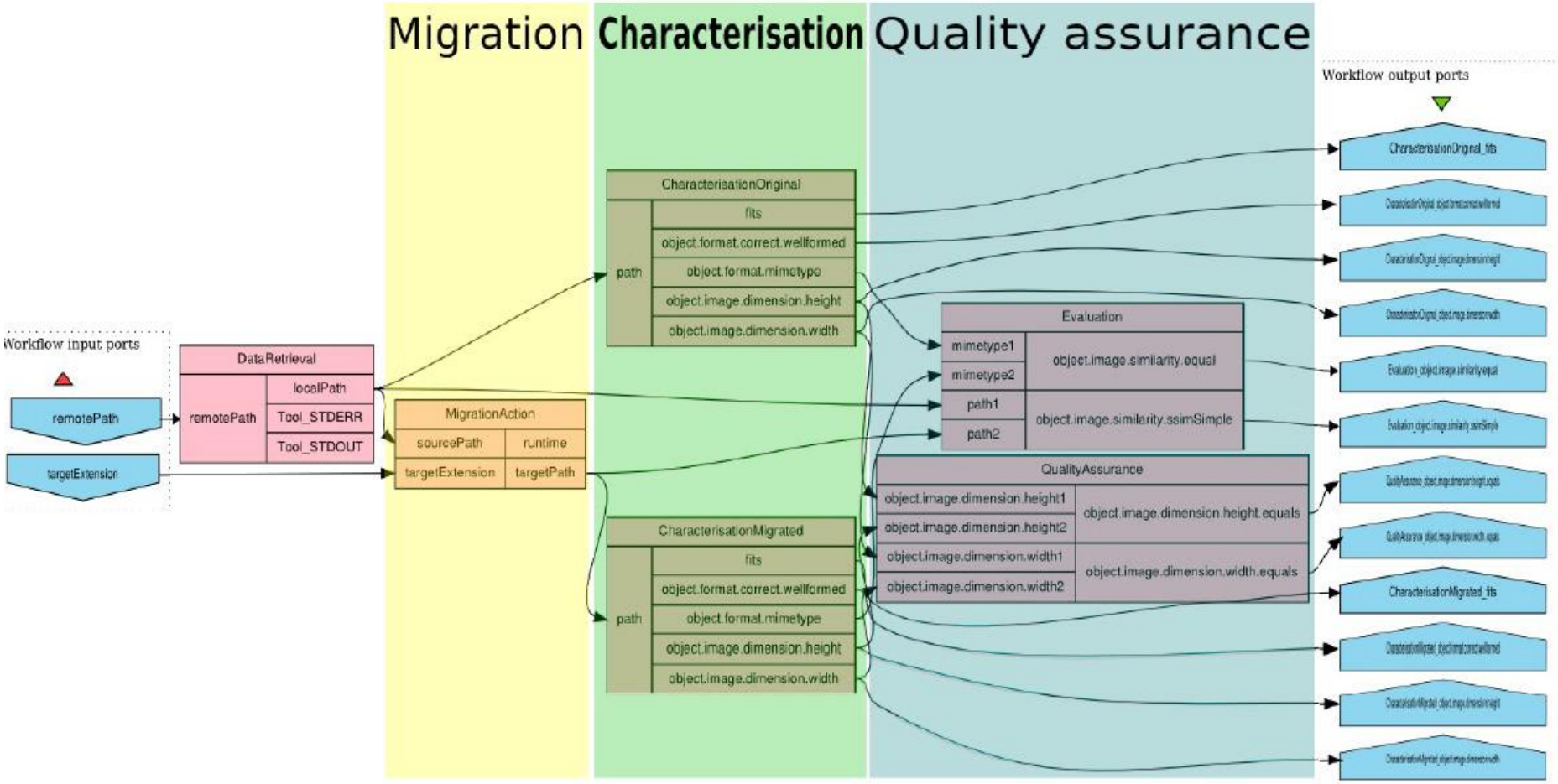
Go/No-Go

- Deliberate step for taking a decision whether it will be useful and cost-effective to continue the procedure, given
  - The resources to be spent (people, money)
  - The availability of tools and solutions,
  - The expected result(s).
- Review of the experiment/ evaluation process design so far
  - Is the design complete, correct and optimal?
- Need to document the decision
- If insufficient: can it be redressed or not?

Develop experiment

- Formulate for each evaluation or experiment or preservation process detailed
  - Development plan
    - steps to build and test software components
    - procedures and preparation
    - parameter settings for integrating preservation services
  - Test plan (mechanisms how to)
  - Evaluation/experiment plan (workflow/sequence of activities)

A planning experiment workflow in Taverna



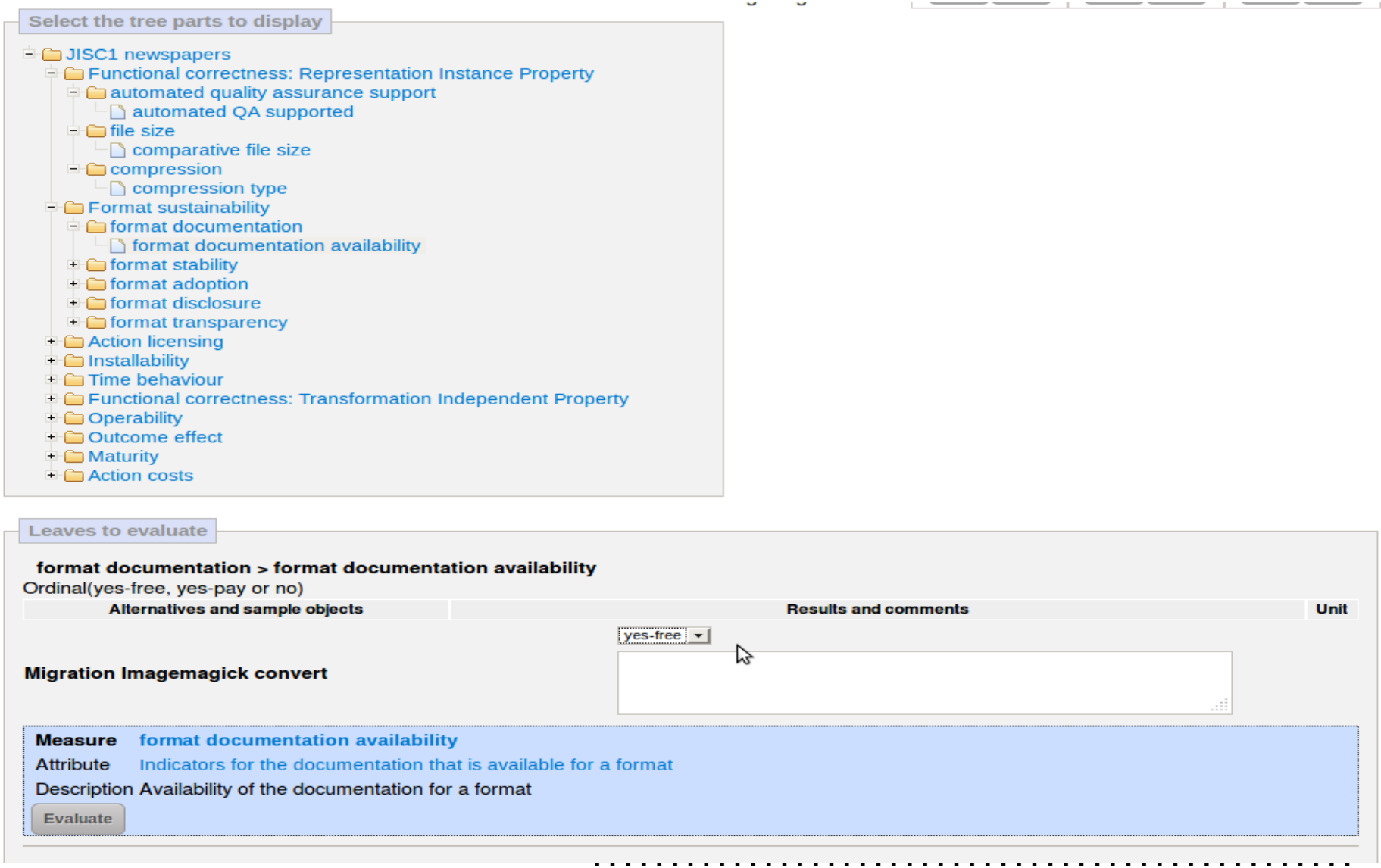
Run experiment

- Before conducting an evaluation or running an experiment, the experiment process as designed has to be tested
  - It may lead to re-design or even termination of the evaluation/ experiment process
- The results will be evaluated in the next stage
- The whole process needs to be documented

Evaluate experiment

- Evaluate the outcome of each alternative for each leaf of the objective tree
  - The evaluation will identify
    - Need for repeating the process
    - Unexpected (or undesired) results
- Includes both technical and intellectual aspects

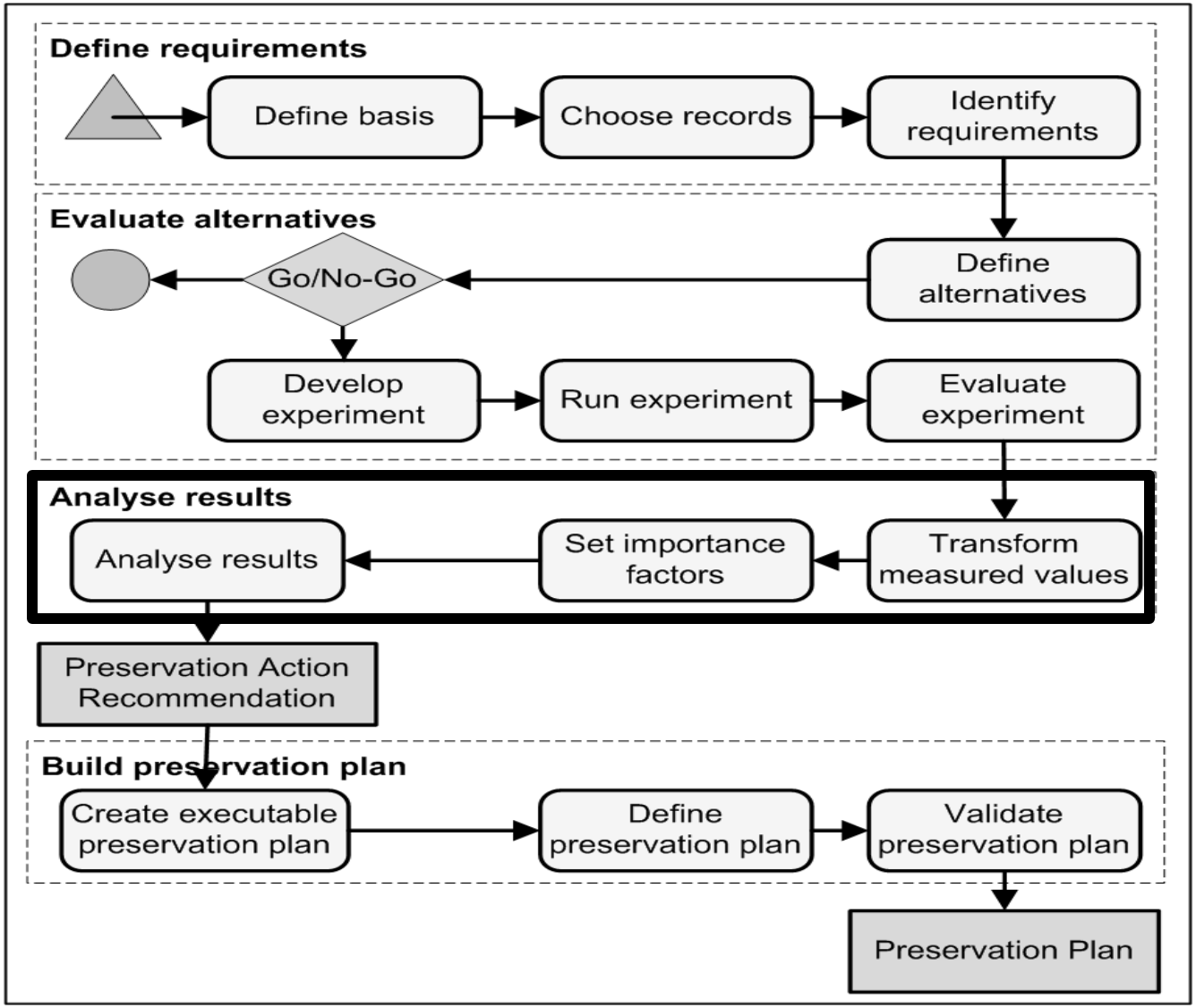
Evaluate experiment in Plato 4



Results of Phase 2

- Possible alternatives defined
- Decision is taken on which experiments to evaluate
- Experiments are developed and executed
- Results are collected and stored in an objective tree

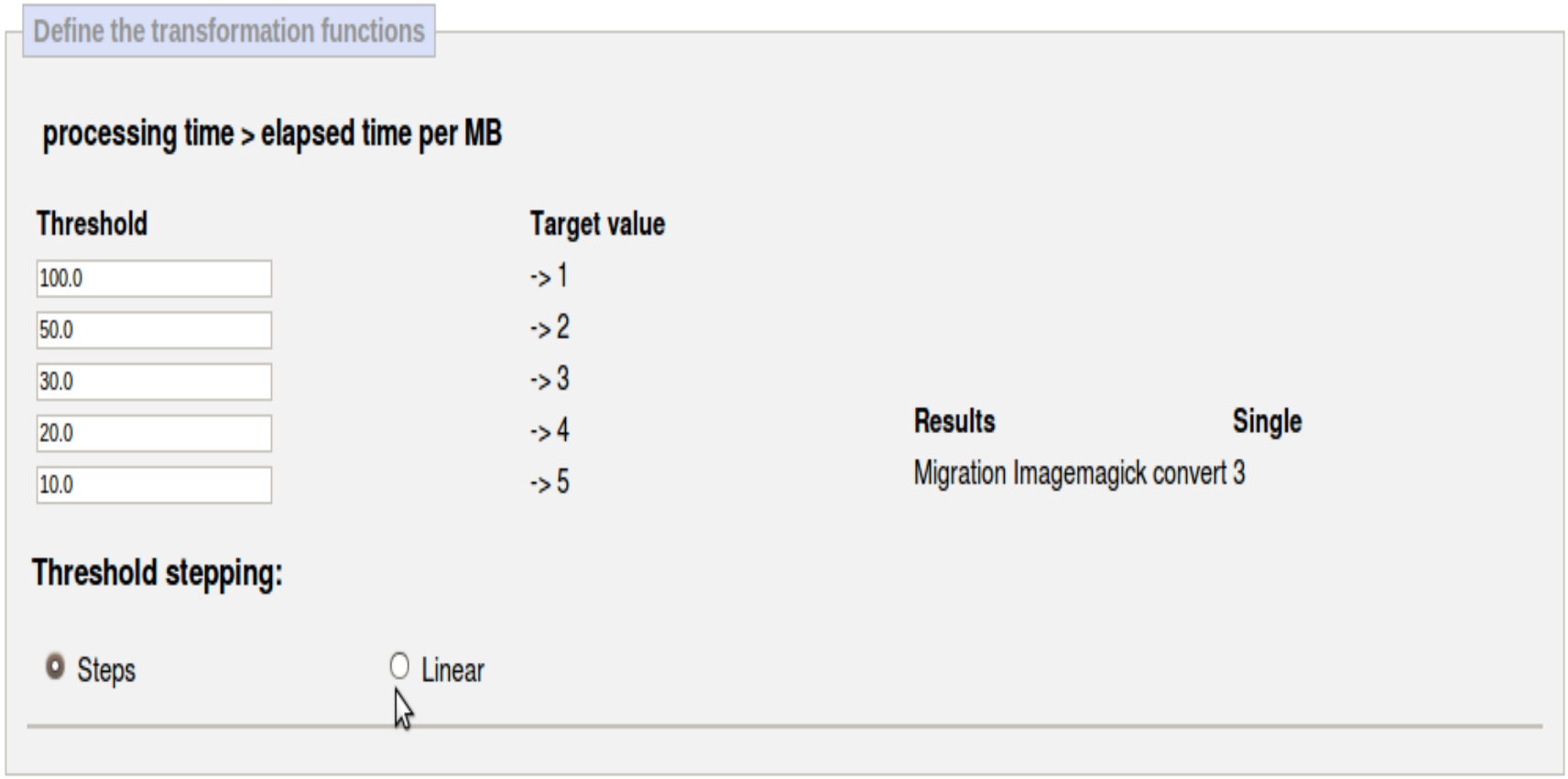
Analyse results



Transform measured values

- Measures come in seconds, euro, bits,…
- Need to make them comparable
- Transform measured values to uniform scale
- Transformation tables for each leaf criterion
- Linear transformation, logarithmic, special scale
- Scale 0-5 (0 is "not-acceptable“)

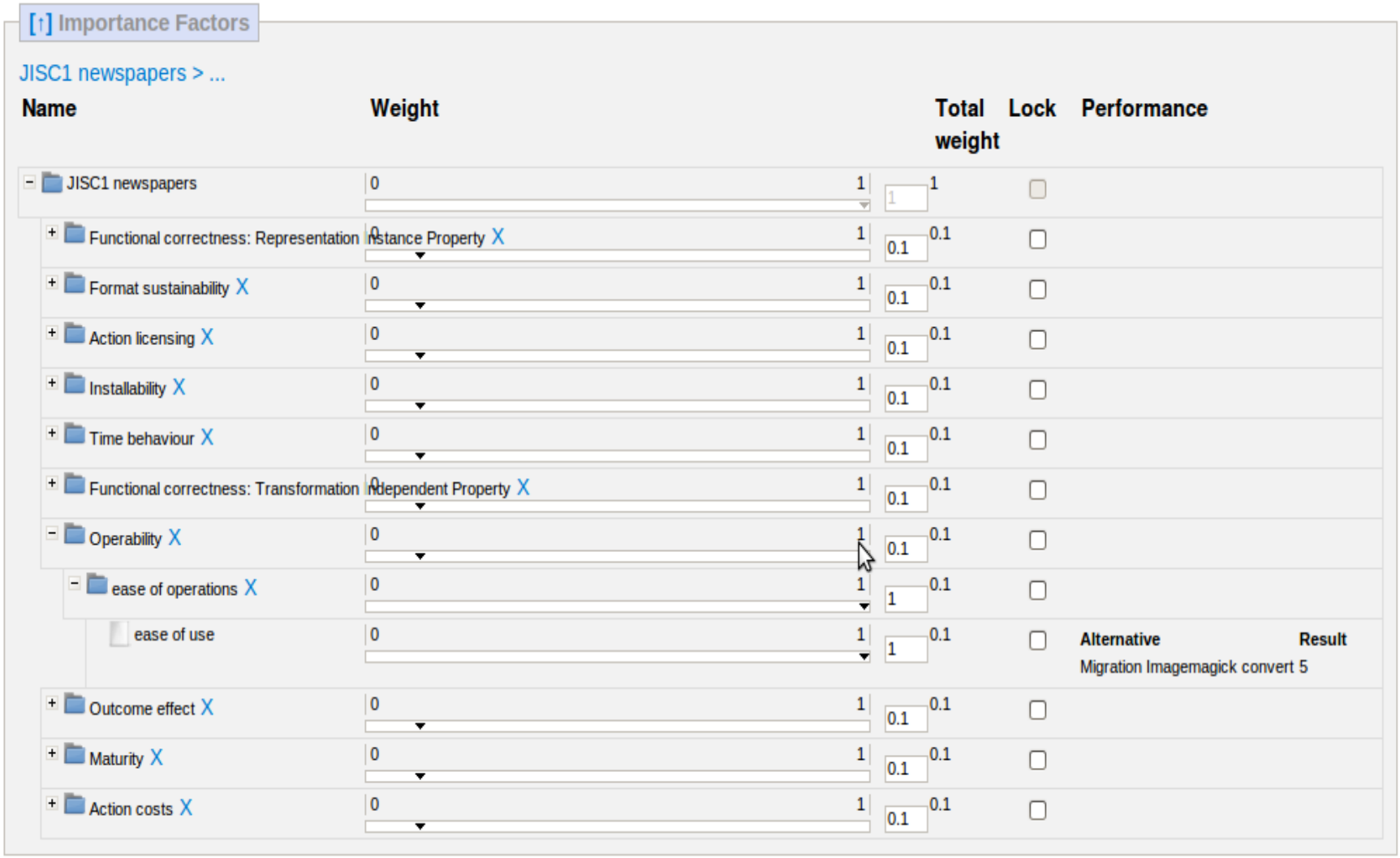
Transform measured values in Plato 4



Set importance factors

- Definition which criteria are more important
- Depends on individual preferences and requirements
- Influence on the final ranking
- Aggregation of weights

Set importance factors in Plato 4



Analyse results

- Aggregate Values
  - Creates performance values for each alternative on each of the sub-criteria identified
- Weighted Sum
  - Multiply the transformed measured values in the leaf nodes with the leaf weights
  - Sum up the transformed weighted values over all branches of the tree
- Weighted Multiplication
  - Take the transformed measured values in the leaf nodes to the power of the leaf weights
  - Multiply up the transformed weighted values over all branches of the tree

Analyse results in Plato 4

Results: Weighted multiplication

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

Node	Results	
Newspaper objectives	GIF > TIF	2.97
	GIF > Text #3	3.59
Object characteristics	GIF > TIF	1.16
	GIF > Text #3	1.52
Appearance	GIF > TIF	1.52
	GIF > Text #3	3.29
Image quality	GIF > TIF	1.00
	GIF > Text #3	1.50
Text quality	GIF > TIF	1.00
	GIF > Text #3	1.50
Image width match	GIF > TIF	1.00
	GIF > Text #3	1.21
Image height match	GIF > TIF	1.00
	GIF > Text #3	1.21
Image colour space match	GIF > TIF	1.52
	GIF > Text #3	1.00
Technical characteristics	GIF > TIF	1.25
	GIF > Text #3	1.16
Process Characteristics	GIF > TIF	1.37
	GIF > Text #3	1.37
Costs	GIF > TIF	1.50
	GIF > Text #3	1.50

Analyse results

- Rank alternatives according to overall utility value at root
- Performance of each alternative
  - overall
  - for each sub-criterion (branch)
- Allows performance measurement of combinations of strategies
- Final sensitivity analysis against minor fluctuations in
  - measured values
  - importance factors

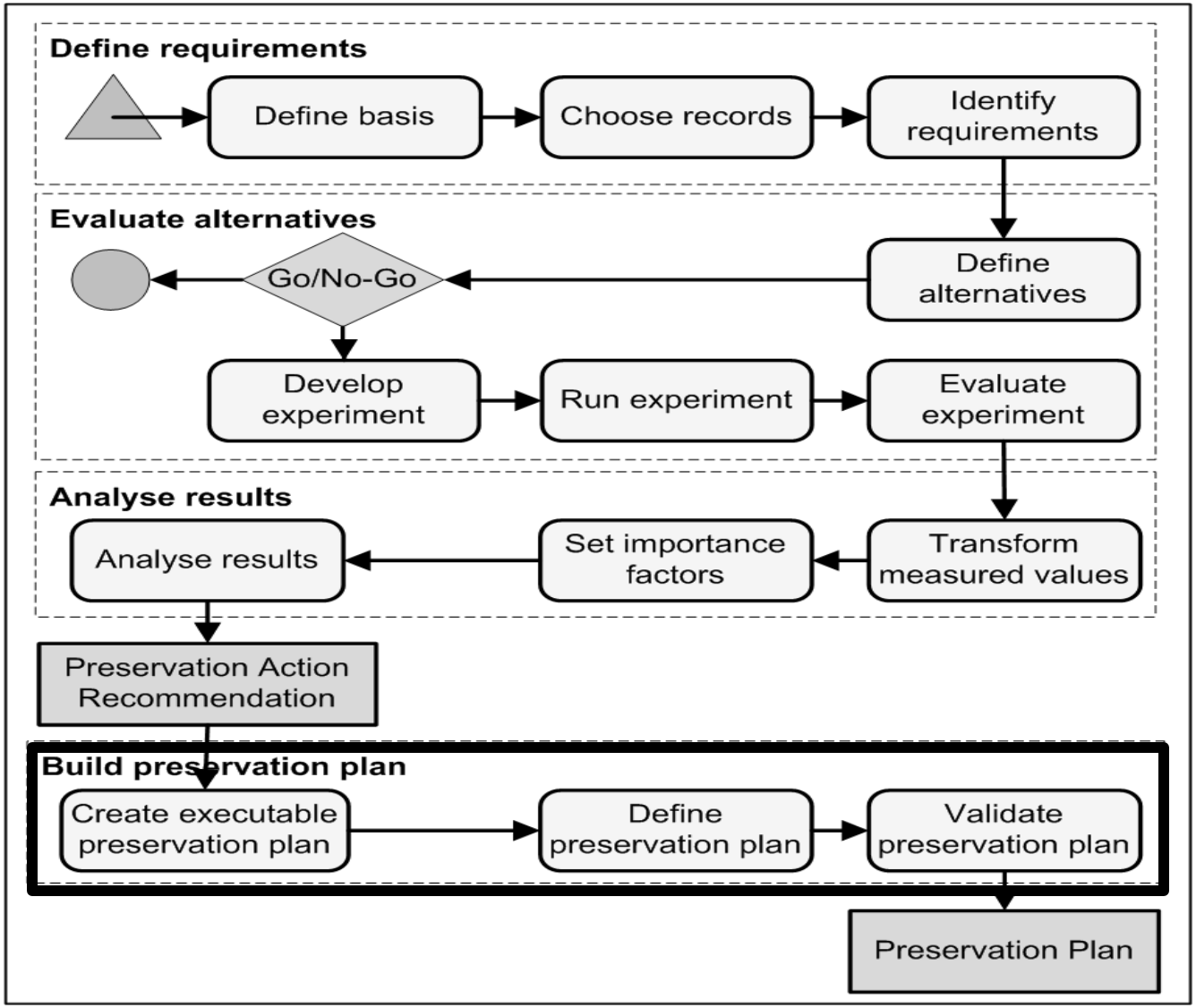
Analyse results

- The review of the results may help to refine
  - The evaluation process/procedure
  - The preservation planning environment itself
  - The evaluation metrics
  - Understanding of the essential characteristics of the objects,
  - and identify further evaluations, experiments
- The review should take into account all previous work done in the preservation planning environment
- The review should look at both the technical and intellectual aspects of digital objects

Result of Phase 3

- Experiment results are transformed to a scale (0 -5)
- Importance factors are adjusted so some objectives can have more influence on the final outcome
- Results are analysed
- Final preservation action recommendation is given

Build preservation plan



Build preservation plan

- Create executable elements of preservation plan
  - Sequence of preservation actions to call, parameters, …
  - Automatic steps + manual interventions where required
  - Automatic verification of results during deployment
- Define preservation plan
  - Create PP based on evidence produced during the PP process
  - Verify completeness of PP
- Seek approval and validation of PP
  - Management activity
  - Sign and deploy

Conclusions

- From strategy and policies to operations
- A simple, methodologically sound model to specify and document requirements
- Repeatable and documented evaluation for informed and accountable decisions
- Generic workflow that can be integrated in different institutional settings
- **Plato: Tool support to perform solid, well-documented analysis**
- Provides basic preservation plan

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

Challenges – Scalability

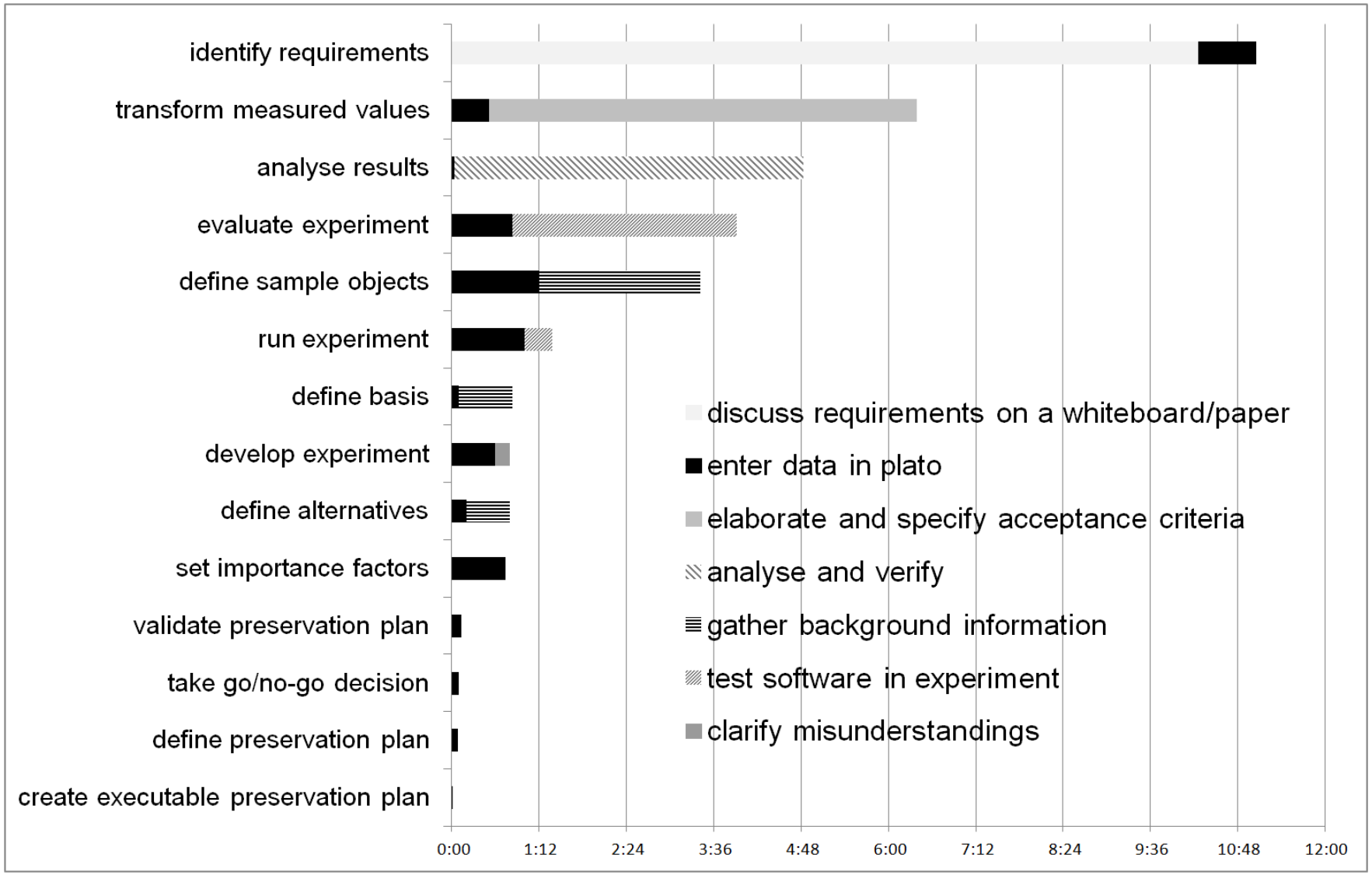
- **How can we ensure that decision processes scale up?**
  - How can we automate decision making?
  - How can we integrate continuous monitoring?
  - Which properties can be measured automatically, and how?
  - How can we create a controlled environment for observing the behaviour of components in a reproducible way?

Challenges – Scalability

- Plato provides sound and trustworthy planning process but it is effort intensive

- A case study at The State and University Library Denmark (SB)
  - Collection of broadcast recordings
  - Most of the collection is in WAV 22.05 kHz format
  - Around 150 000 files (21.5 TB) is in MP3 file format
  - Effort required to develop a plan ~ **35 PH** (person hours)

SB case study (efforts)



Improving the scalability

- Automated content analysis
  - eliminates manual sample description and selection
  - C3PO tool
- Formalized policies (more about policies in the next lectures)
  - enable tool automation
  - facilitate requirements reuse

Improving the scalability

- Manual experiment conduction and collecting the results is time consuming and error prone
- There is a need for a standardized experimentation environment
  - Plato 4 relies on Taverna workflow environment
- Provides different components (migration , QA) as Taverna workflows
  - reusability
  - leafs from the objective tree can be automatically measured

Task 2 Preservation planning

- single or group work
  - in case of a group work max 2 persons
- you will create a preservation plan for your own collection
  - images, music, videos, documents, ...
- following the preservation planning workflow you will
  - pick sample objects
  - define requirements (objective tree !)
  - define alternatives which you want to evaluate
  - conduct experiments
  - analyse results

- Details will be publised on TUWEL on 1.4