

Preservation planning

April 4, 2011

Christoph Becker

Vienna University of Technology

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

- Context
 - Decision problems in digital preservation
- Overview
 - What is a preservation plan?
 - How to create a preservation plan
- Requirements
 - High-level requirements
 - Exercise

- Digital Preservation is
 - ... Information Management with a long-term view
 - ... Interoperability over time
 - ... Alignment between content, technology and consumers
 - ... Relevant in very diverse scenarios
 - From digitised books to eScience and engineering
 - From web archiving to business environments
- DP always needs to consider the contextual influences of organisations and systems



- The mission of digital preservation
 - Overcome obsolescence threats
 - ensure (current and) future access in a usable form for specific user communities
- The mission of preservation *planning*
 - Defining the right actions → Component selection
 - Questions:
 - How to select the right action in a given scenario?
 - How to ensure trust?
 - How to enable scalability?

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

- **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?

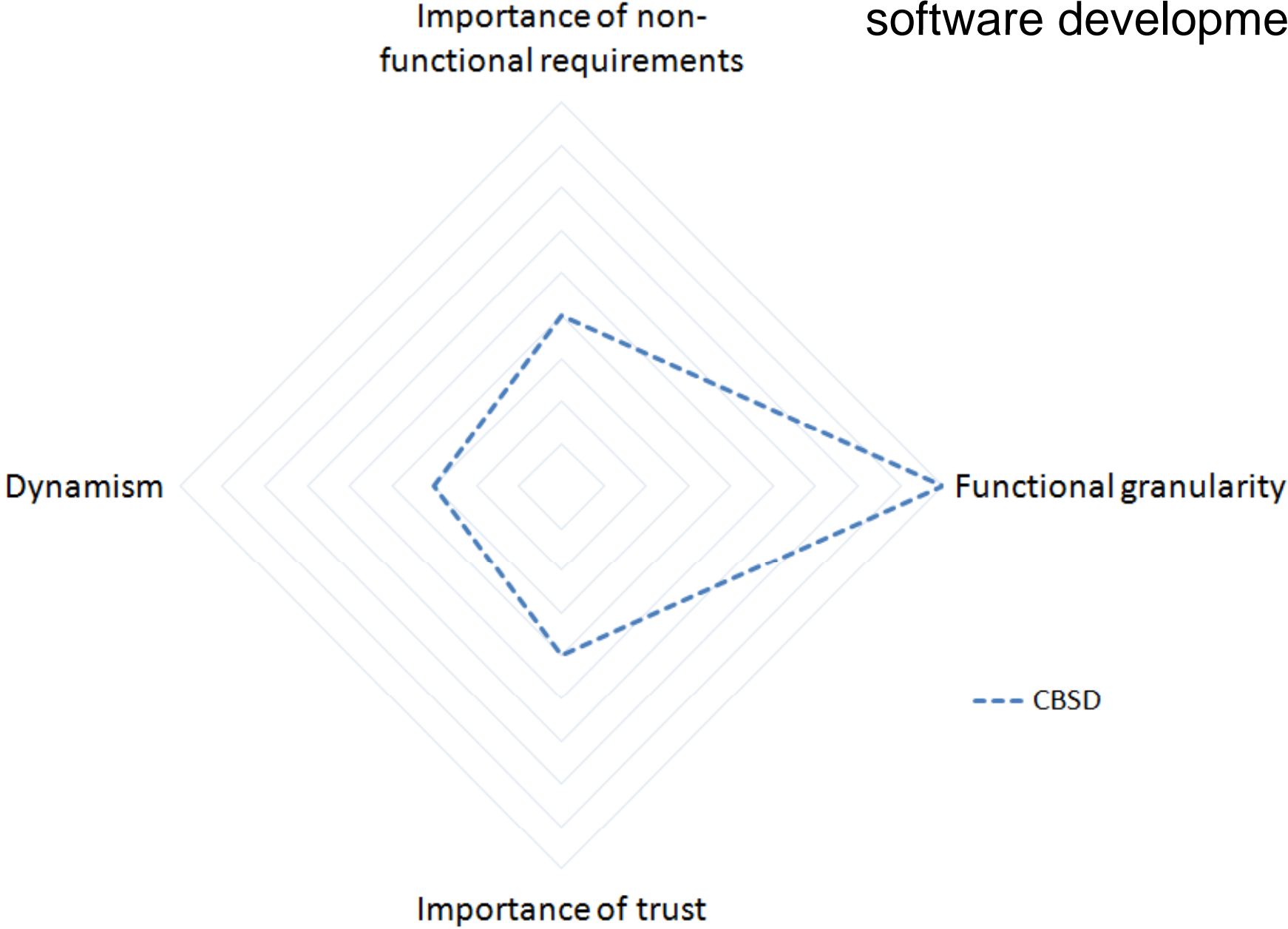
- **How can we ensure that decision processes scale up?**
 - How can we automate decision making?
 - How can we integrate continuous monitoring?
 - Which properties can be measured automatically, and how?
 - How can we create a controlled environment for observing the behaviour of components in a reproducible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

- Evaluating preservation actions
- Multi-objective decision making
- Component selection
- Trustworthy repositories
- Transparent documentation and evidence
- Automation and tool support
- Decision factors and measurements

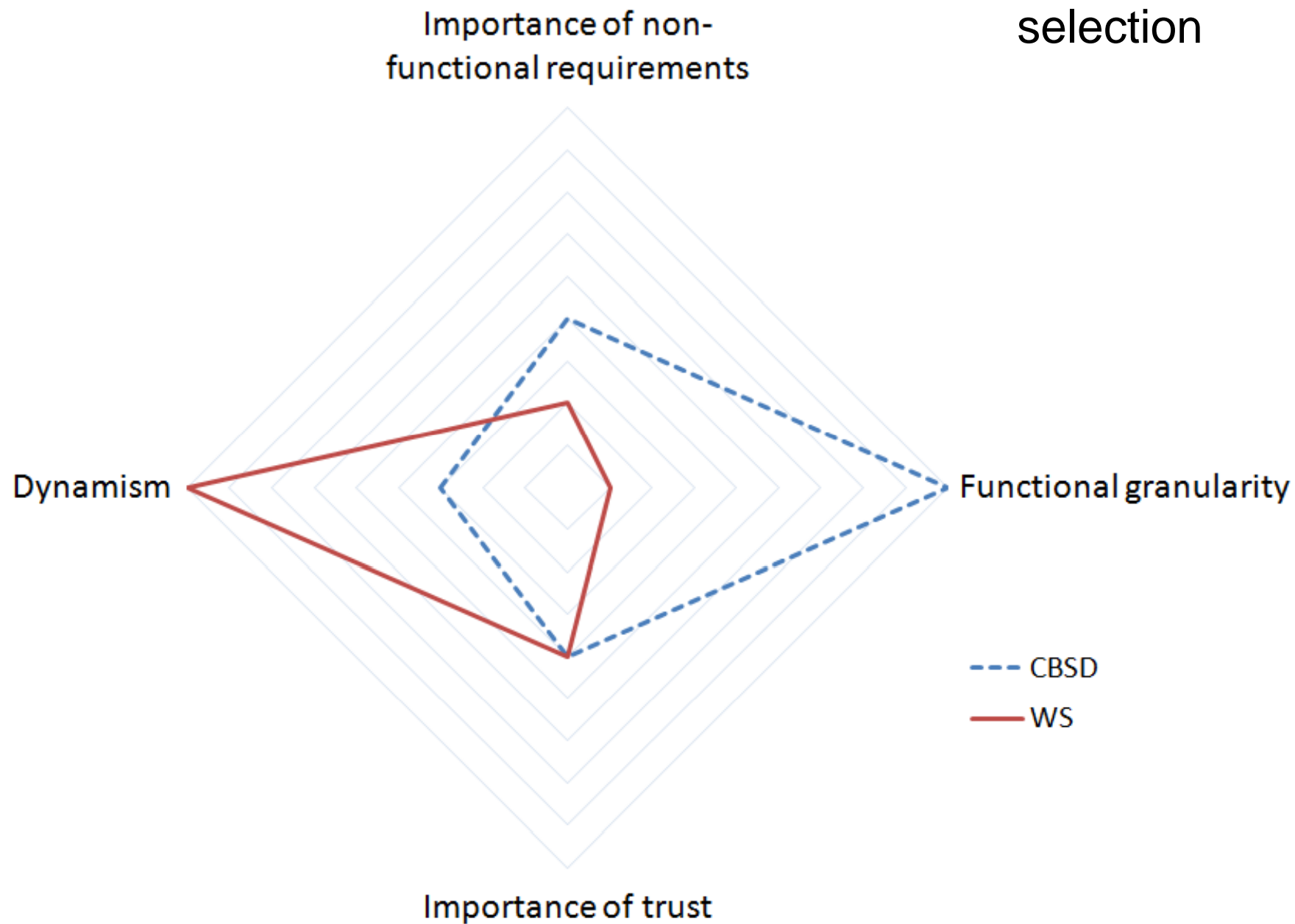
Evaluating preservation actions

- Several actions available (migration, emulation, ...)
- Challenges:
 - 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
- Comparable to
 - Component selection
 - Web service selection

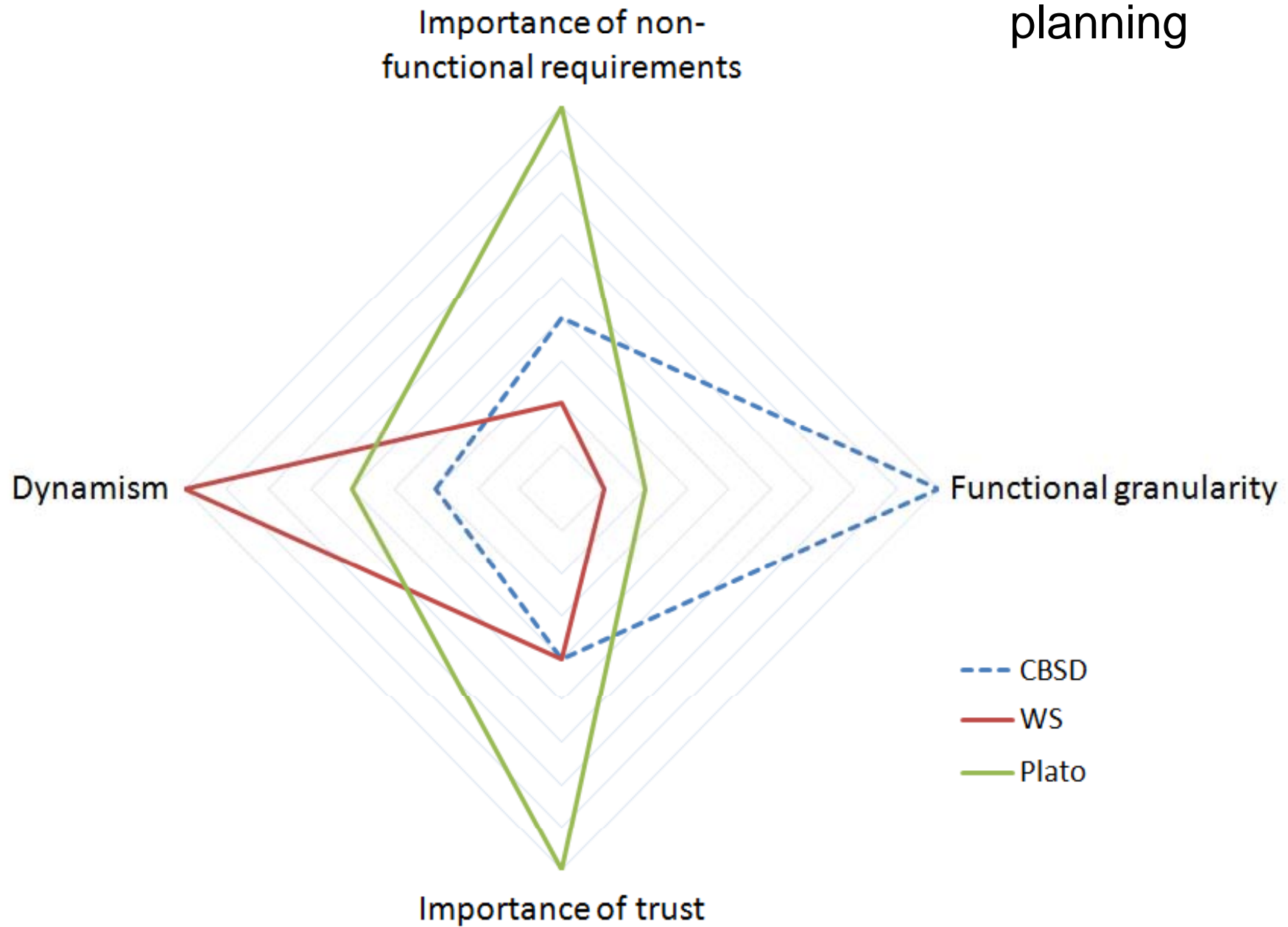
Component based software development



Web service selection



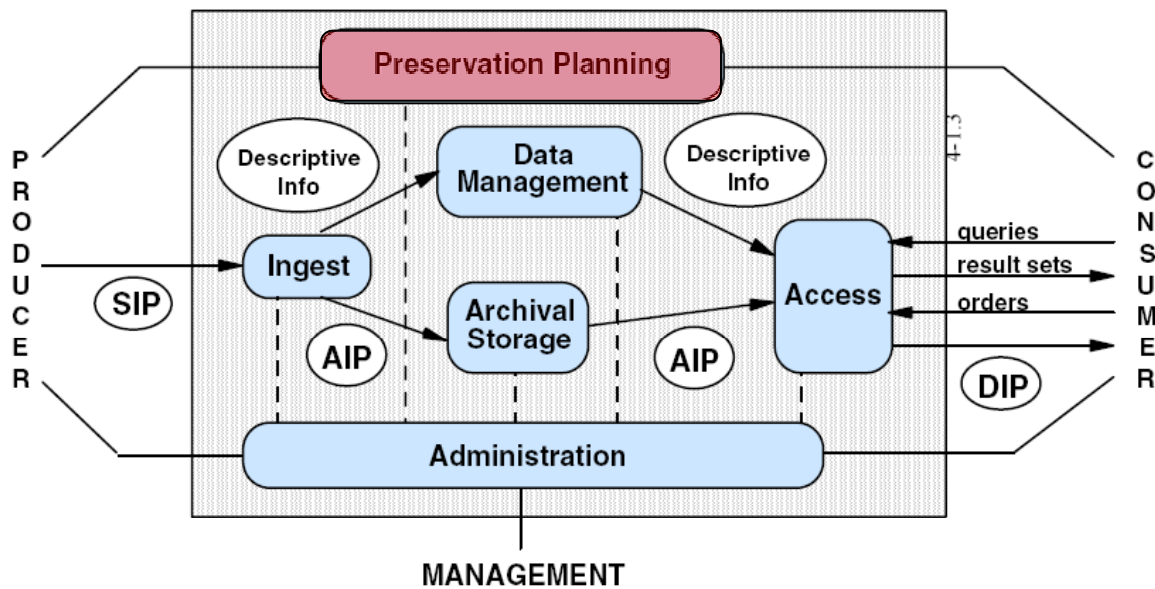
Preservation planning



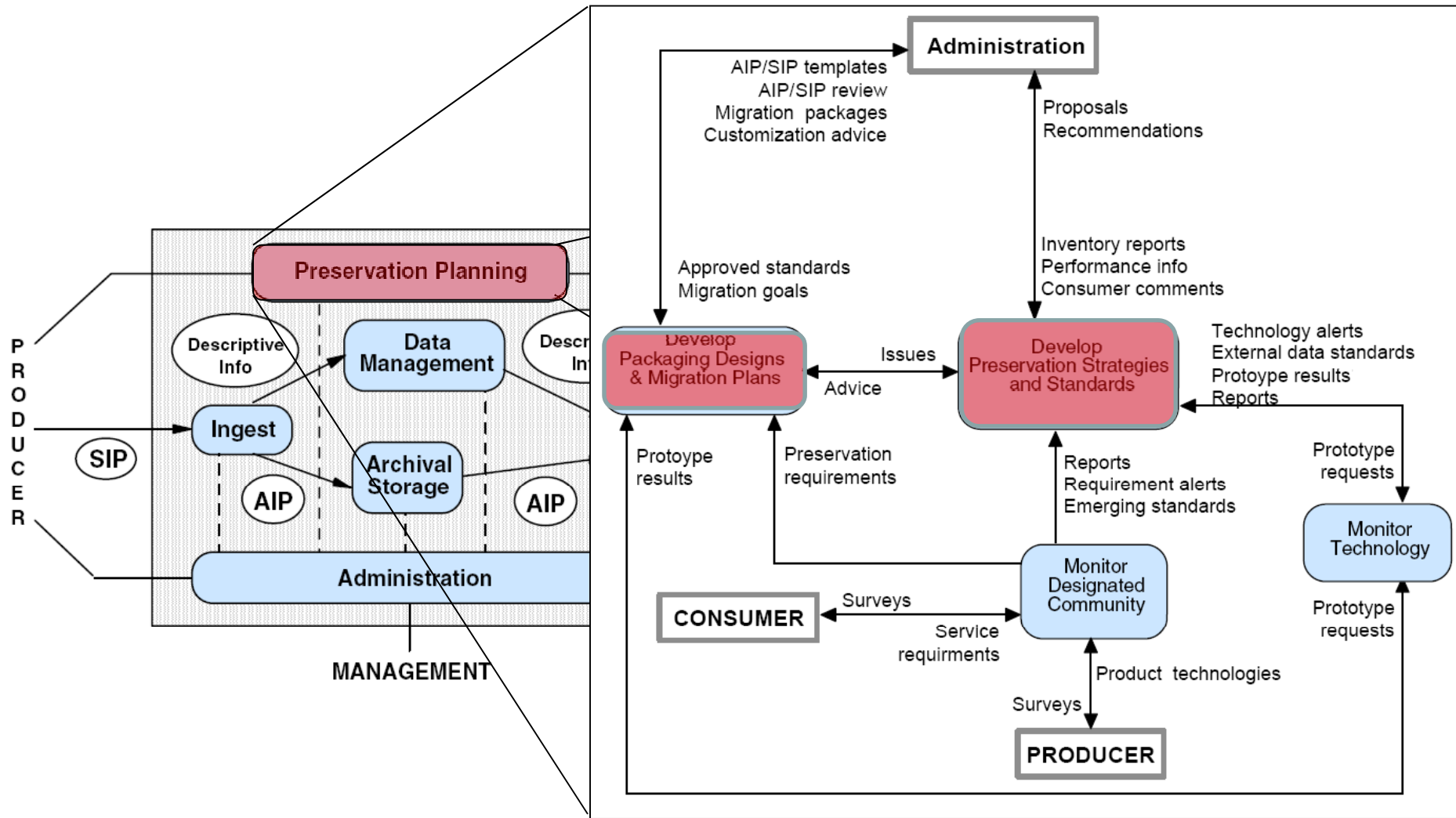
- Context: Trustworthy repositories
 - Open Archival Information Systems model (OAIS)
 - Trustworthy repositories criteria (TRAC, nestor)
 - Trust requires evidence
 - Evidence needs repeatable, objective facts

- Preservation planning approach
 - Evaluate potential actions objectively against scenario-specific requirements in a repeatable way
 - Sample-based experiments in controlled environment
 - Quantitative analysis of strengths and weaknesses
 - Evaluate suitability of each potential action

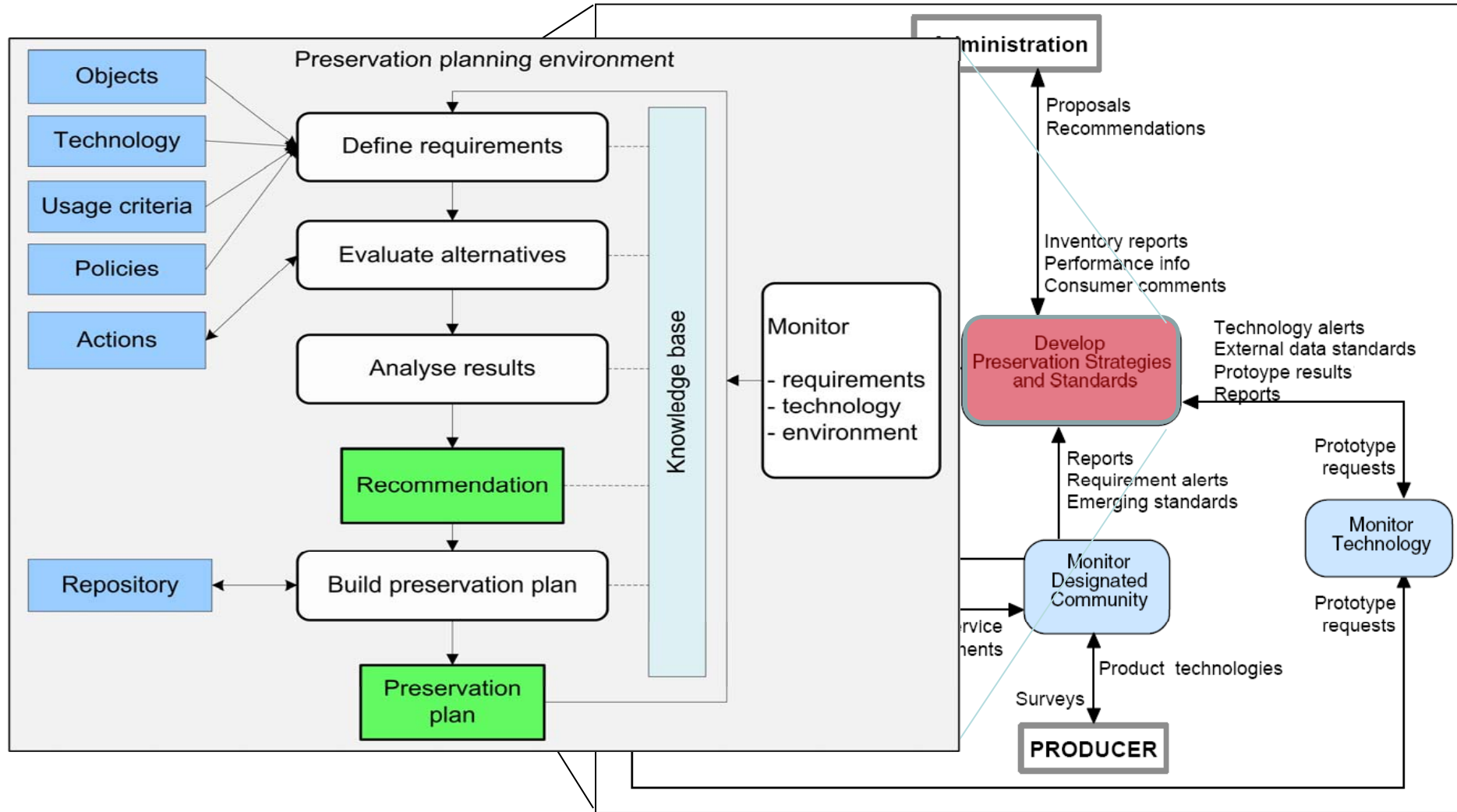
OAIS and Preservation Planning

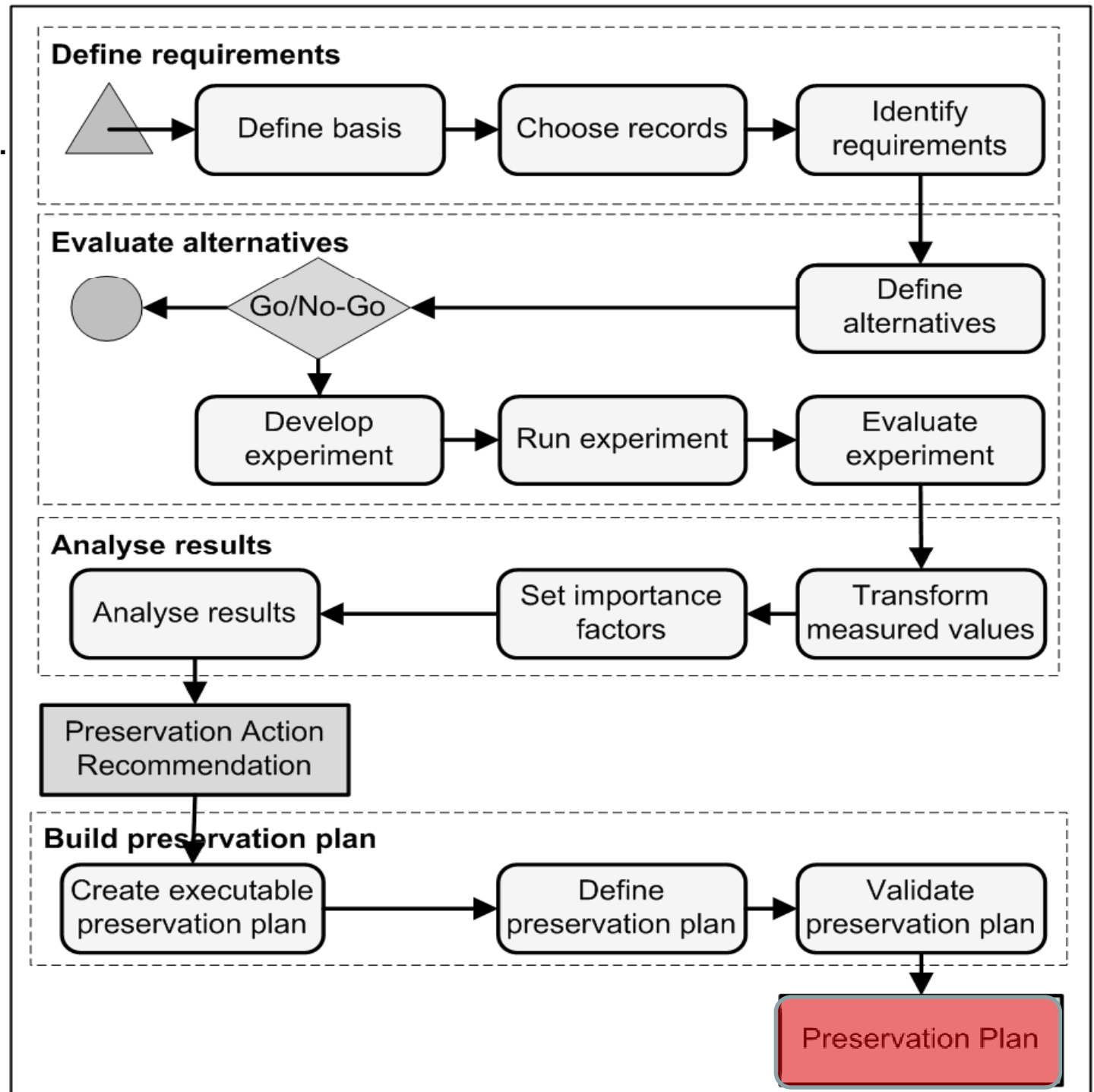


Preservation Planning and OAIS



Preservation Planning and OAIIS





What is 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

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.

Characteristics of a preservation plan

- Translation of a preservation policy
- Specification of how to treat a collection in a given institutional setting
- Monitored for
 - ✓ changes in technology
 - ✓ changes in organisational setting
 - ✓ changes in user requirements
 - ✓ changes in available tools
 - ✓ changes in preservation methods
- Species concrete action
 - ✓ The **preservation action plan** can be an executable workflow definition, detailing actions and required technical environment
 - ✓ The preservation plan provides the context/background of the preservation action plan

The content of a preservation plan

1. Identification
2. Status
 - ✓ What was the immediate reason for this plan?
 - ✓ Has it been approved and if so, when and by whom
 - ✓ How does it relate to other plans related to a specific type of objects?
3. Description of institutional setting
4. Description of the collection (digital objects)
5. Purpose and requirements
6. Evidence of decision for a specific preservation action
 - ✓ what is the foundation of the decision
 - ✓ description of evaluation of possible actions
7. Costs considerations
8. Trigger for re-evaluation
9. Roles and responsibilities
10. Preservation action plan
 - ✓ executable program

A 3.2 Repository has procedures and policies in place, and mechanisms for their review, update, and development as the repository grows and as technology and community practice evolve.

- Policies, plans, monitoring

A3.6 Repository has a documented history of the changes to its operations, procedures, software, and hardware that, where appropriate, is linked to relevant preservation strategies and describes potential effects on preserving digital content.

- Preservation plans need traceability

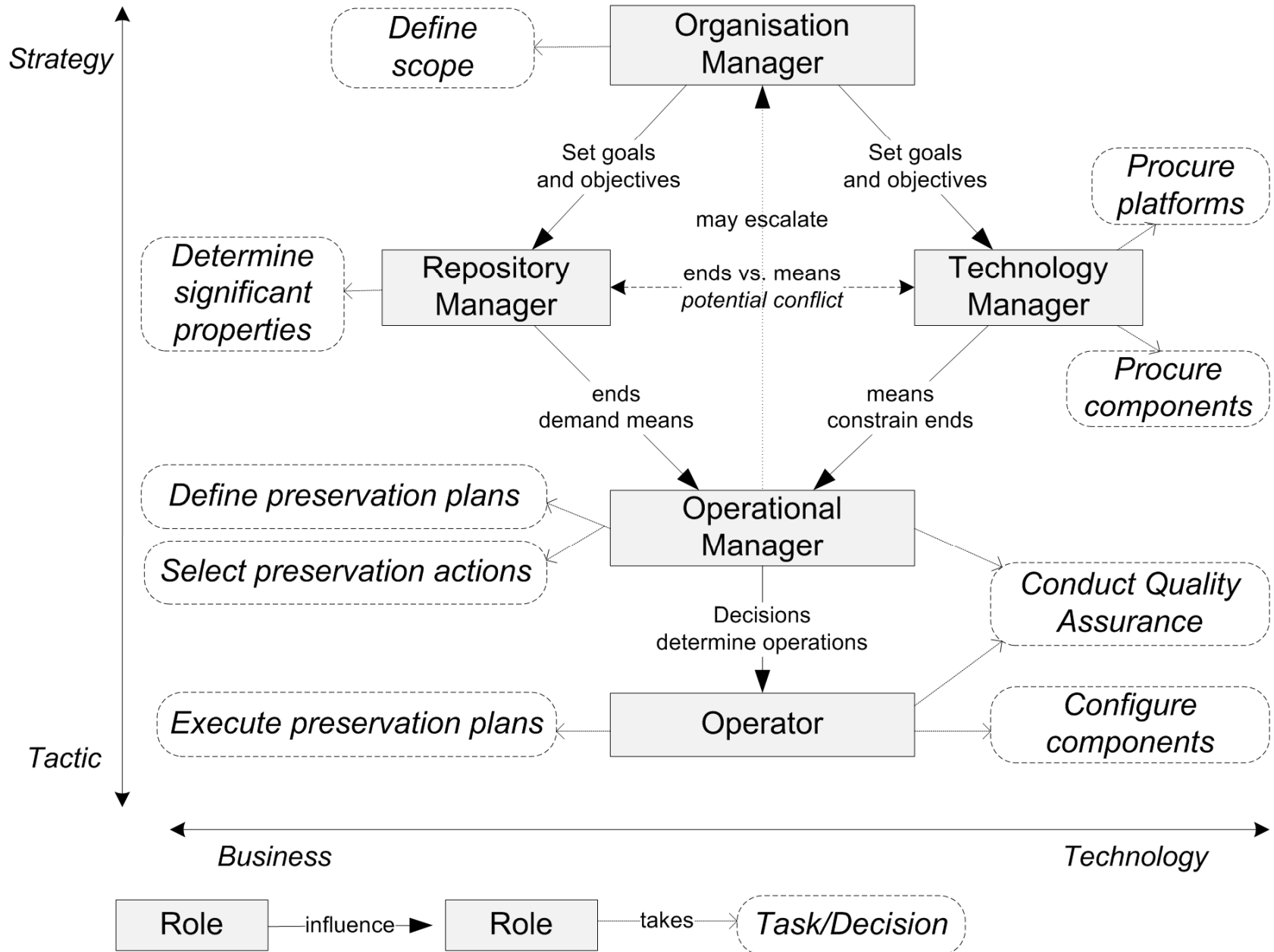
B3.1 Repository has documented preservation strategies.

- Preservation Plan

B3.3 Repository has mechanisms to change its preservation plans as a result of its monitoring activities.

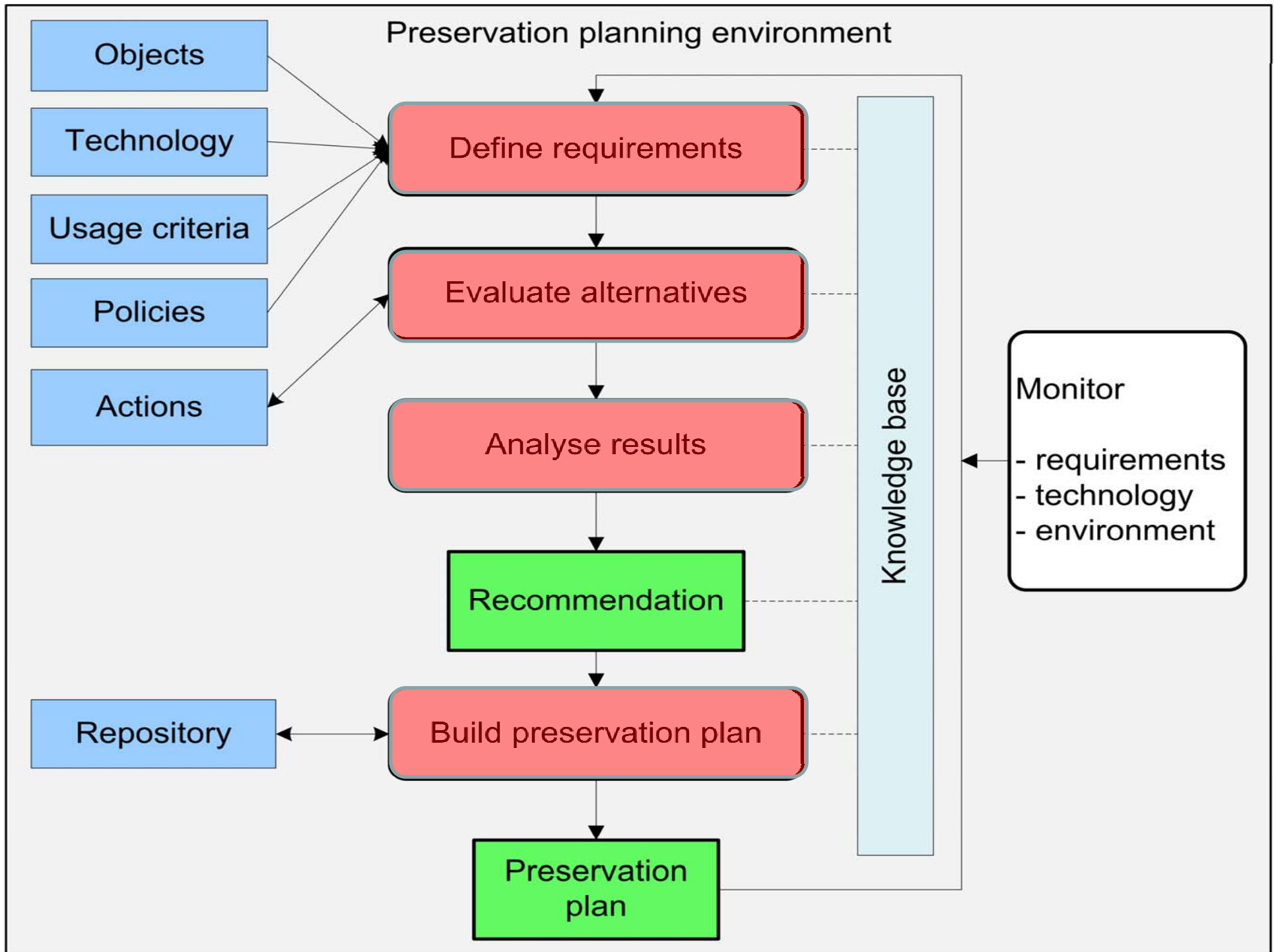
- Monitor environment
- Update preservation plans

A DP decision space

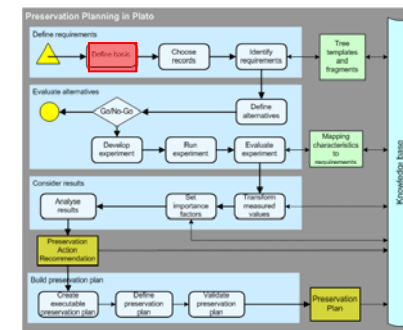


Agenda

- Preservation planning methodology
 - Walk-through
 - Objective Trees
 - Case Studies
 - Tool support
- Exercise
- Summary and Outlook



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

- Mandate/ vision / mission statements
- Policy documents (if they exist)
- Project plans
- Guidelines
- Procedures/ rules

- Example policy statements of institutions with a digital preservation programme
 - UK Data Archive
 - National Archives of Australia
 - ISO/TR 18492:2005
Long-term preservation of electronic document-based information

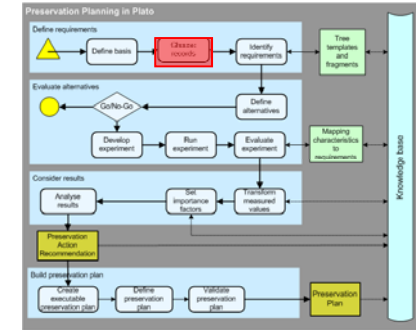
- < UK Data Archive Preservation Policy
- <http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0308.pdf>
- p. 11: “The UKDA has chosen to implement a preservation strategy based upon open and available file formats, data migration and media refreshment.”
- What does this choice mean in practice? Two examples:
 - Emulation is –apparently– not a preservation strategy that will be chosen; all obsolete files will be migrated.
 - Migration to open file formats will be preferred.

- < An Approach to the Preservation of Digital Records
- http://www.naa.gov.au/images/an-approach-green-paper_tcm2-888.pdf
- p. 14: “The digital preservation program must be able to preserve any digital record that is brought into National Archives’ custody regardless of the application or system it is from or data format it is stored in.”
- What does this choice mean in practice? One example:
 - all records that are accepted, should be preserved, regardless file format, medium, application, etc.
 - transform to open standard + keep ‘original’ format

- International standard: Long-term preservation of electronic *document-based information*
- p. 12: Migration to standard formats
Storage repositories should consider **migrating** electronic document-based information from the wide variety of formats used by creators or recipients to a smaller number of “standardized” formats upon their transfer to the custody of the repository.
“Standardized” formats could be a consensus on formats that are widely used and are likely to cover a majority of a particular class of electronic document-based information. Proprietary file formats should be avoided. Among the technology neutral formats that merit consideration are PDF/A-1, XML, TIFF and JPEG.

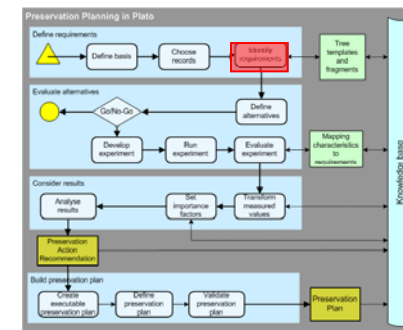
- Framework (requirements) are identified at a high level
- Some requirements can already be explicit. Examples:
 - choice for one strategy
(e.g. migration to open document format)
 - choice that some types of records/documents can be denied
because e.g. an *exotic* file format is used

Choose objects/records



- Representative for the objects in the collection
- Right choice of samples is essential
- 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

Identify requirements

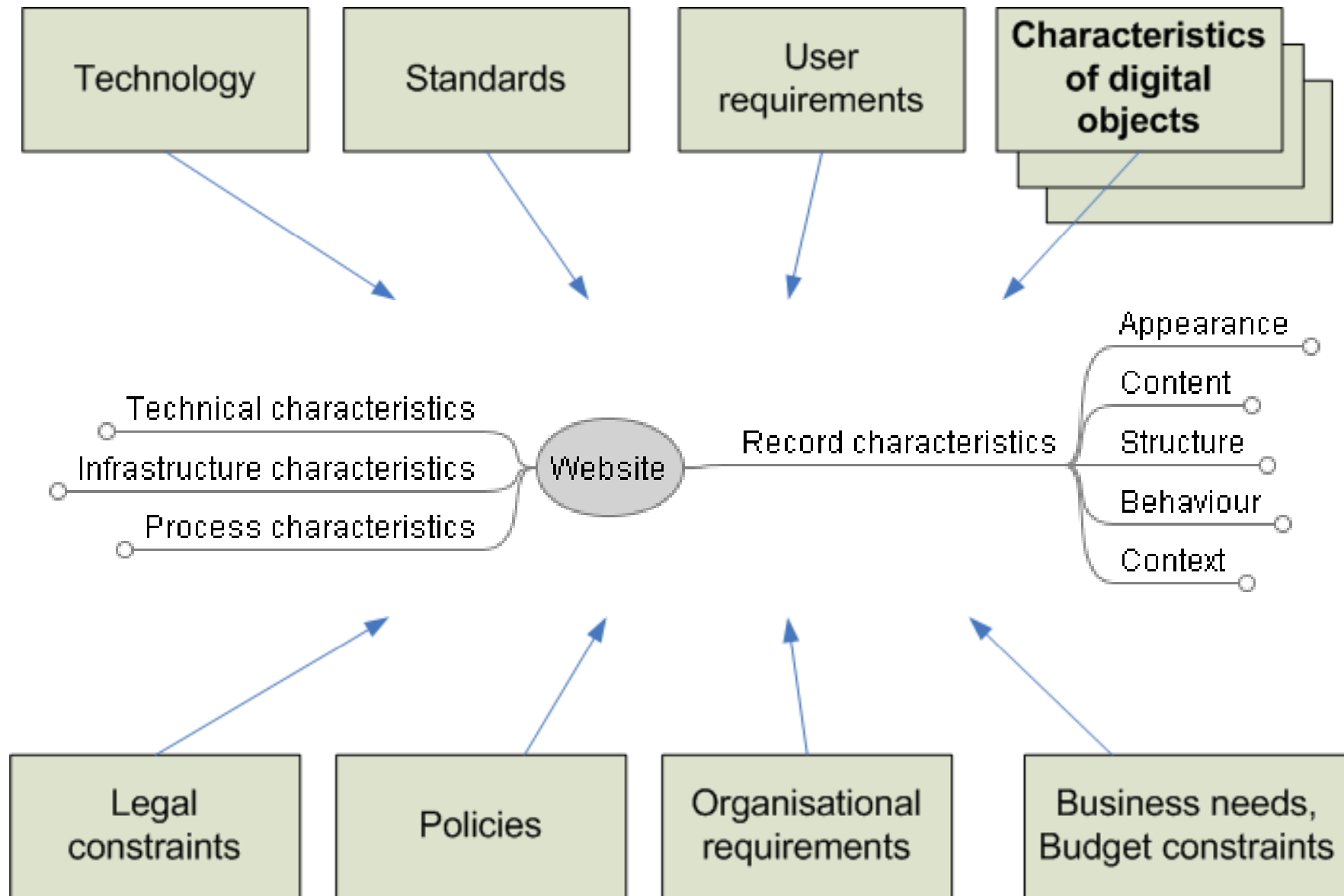


- Define all relevant goals and characteristics (high-level, detail) with respect to a given application domain
- Usually four major groups:
 - object characteristics (content, metadata ...)
 - record characteristics (context, relations, ...)
 - process characteristics (scalability, error detection, ...)
 - costs (set-up, per object, HW/SW, personnel, ...)
- Put the objects in relation to each other (hierarchical)

The Objective Tree

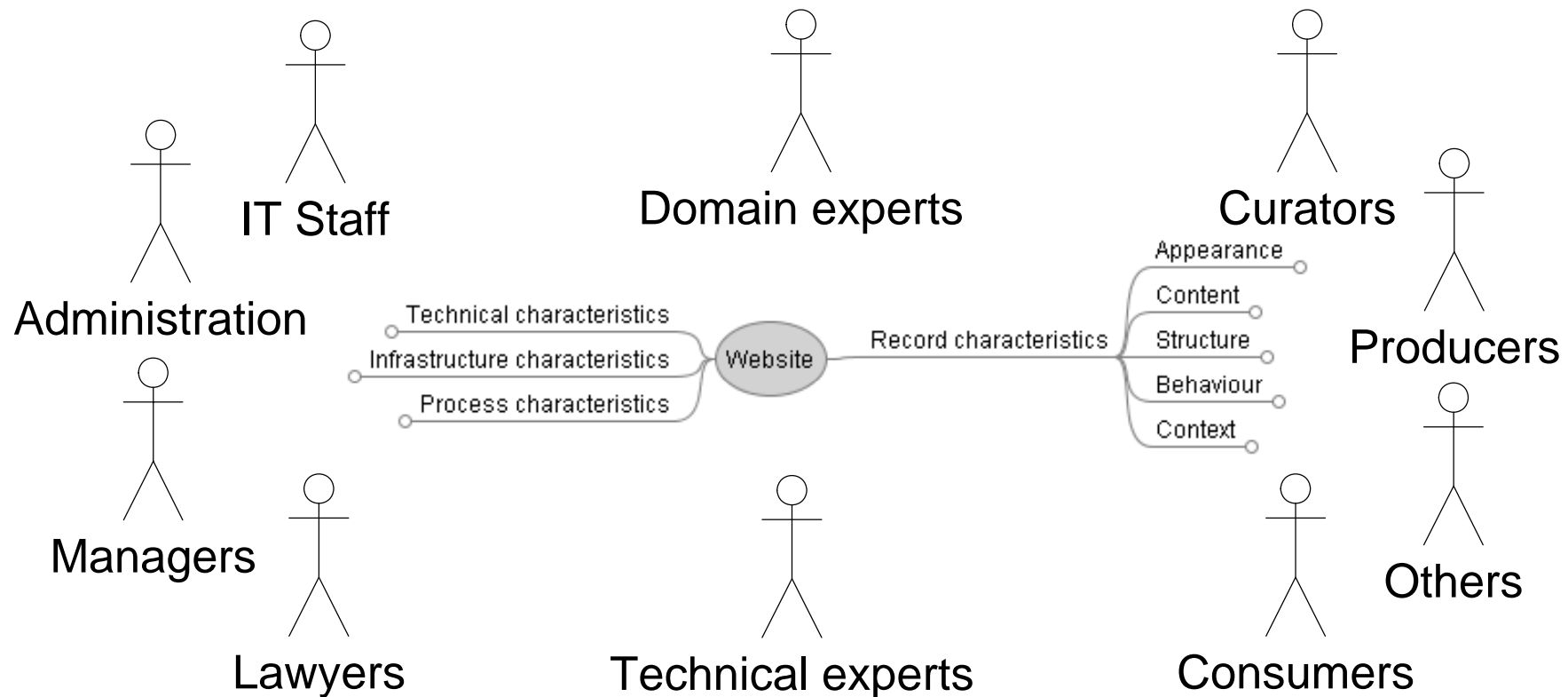
- Tree structure describing requirements and goals
 - A weighted hierarchy of objectives leading into measurable criteria
 - A utility function for each criterion specifies the organisation's assessment for the range of possible values
- Created top-down or bottom-up
 - Start from high-level goals and break down to specific criteria
 - Collect criteria and organize in tree structure

Influence Factors



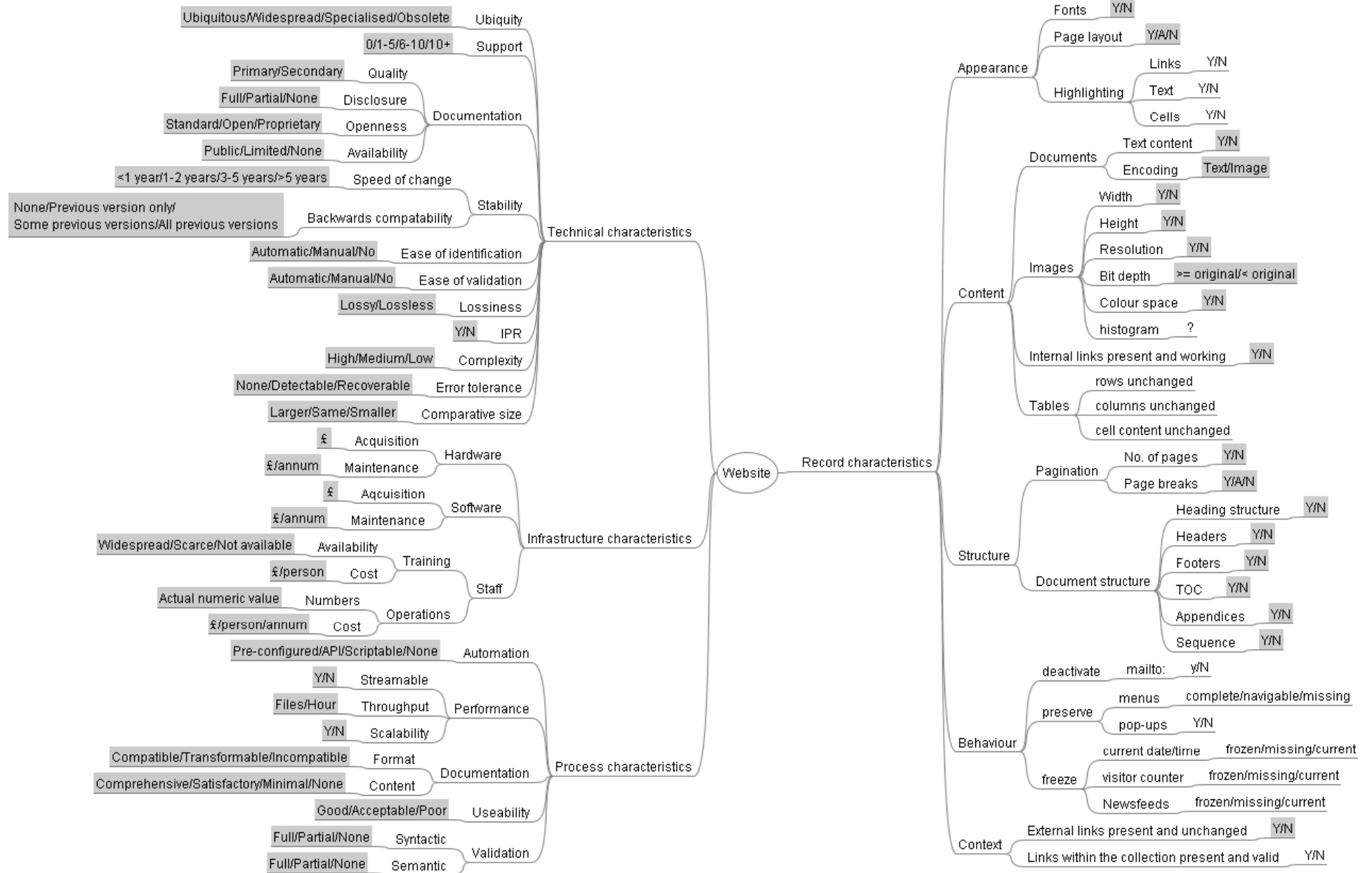
Stakeholders

- Input needed from a wide range of persons, depending on the institutional context and the collection

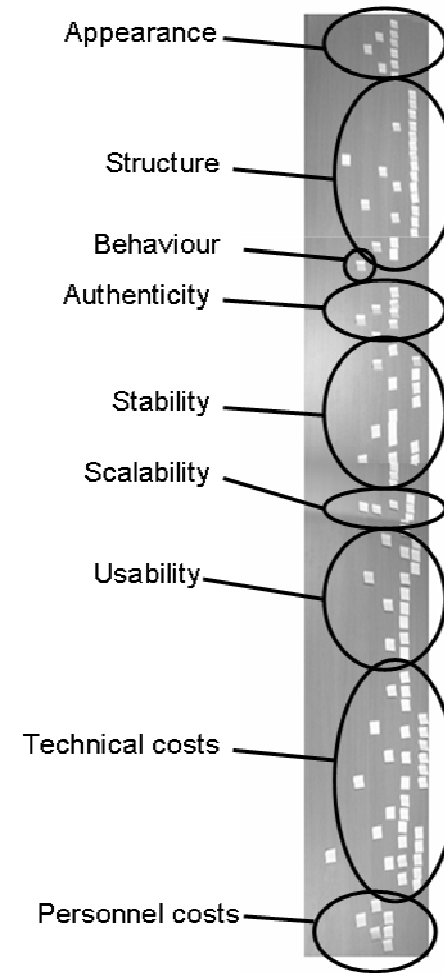




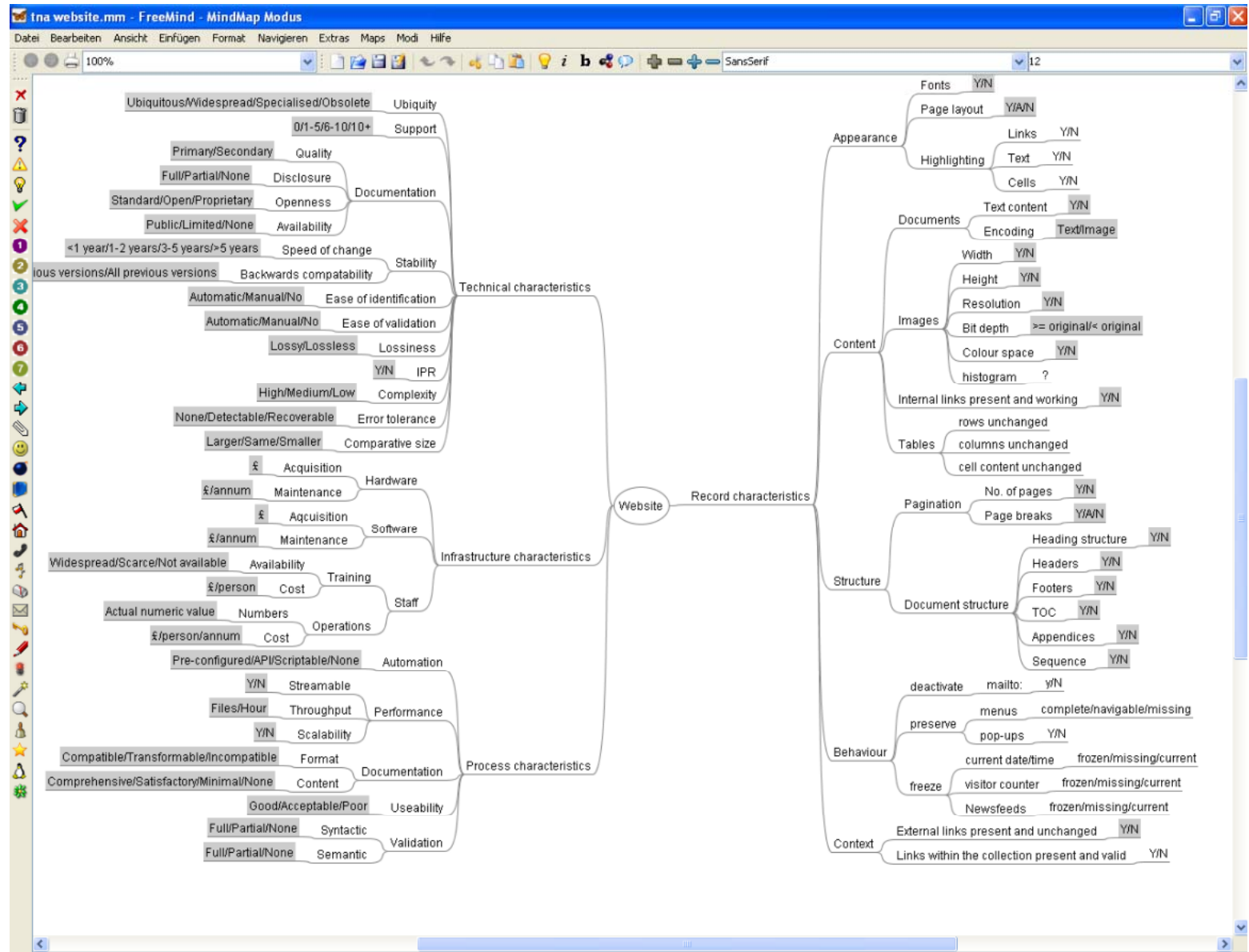
The Objective Tree



Analog...



... or born-digital



- Goal of digital preservation is to serve (future) users in providing usable and authentic information
- What are needs/requirements of users?
 - easy access
 - knowledge about origin of documents/ to be able to interpret them
 - to use them to their own convenience
- Example requirements
 - some users prefer that all information is presented in a uniform way
 - some users prefer that they can search full-text in documents (consequence: don't migrate texts to image files)
 - ...

- What needs to be preserved?
 - Authenticity
 - Reliability
 - Integrity
 - Usability
 - Accuracy

 - Content
 - Context
 - Structure
 - Appearance
 - Behaviour

Assign Measurable Units

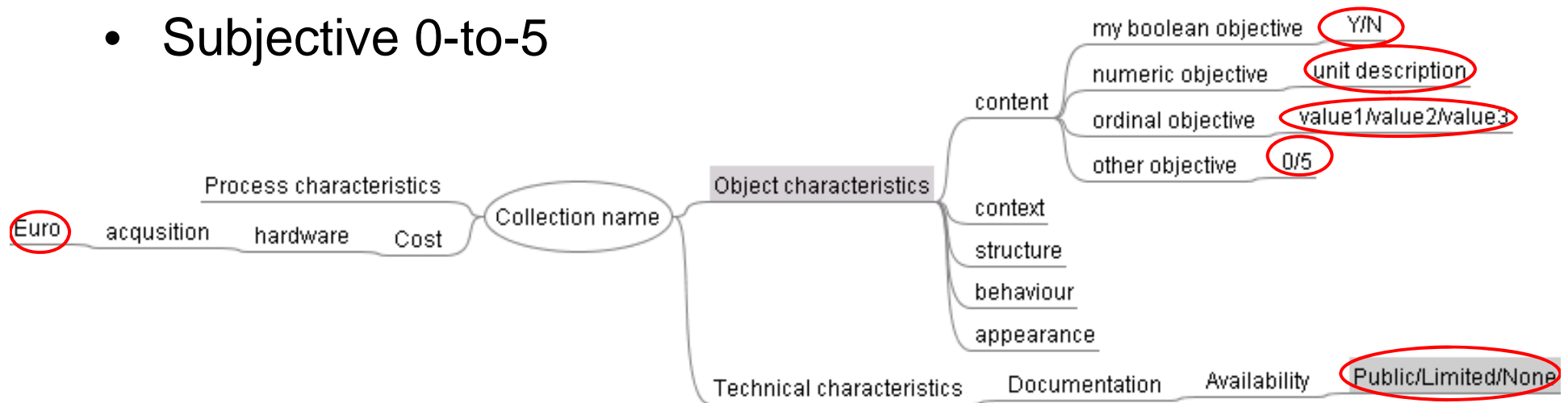
- Leaf criteria should be objectively measurable
 - Seconds per object
 - Euro per object
 - Bits of colour depth

- Subjective scales where necessary
 - Adoption of file format
 - Amount of (expected) support

- Quantitative results

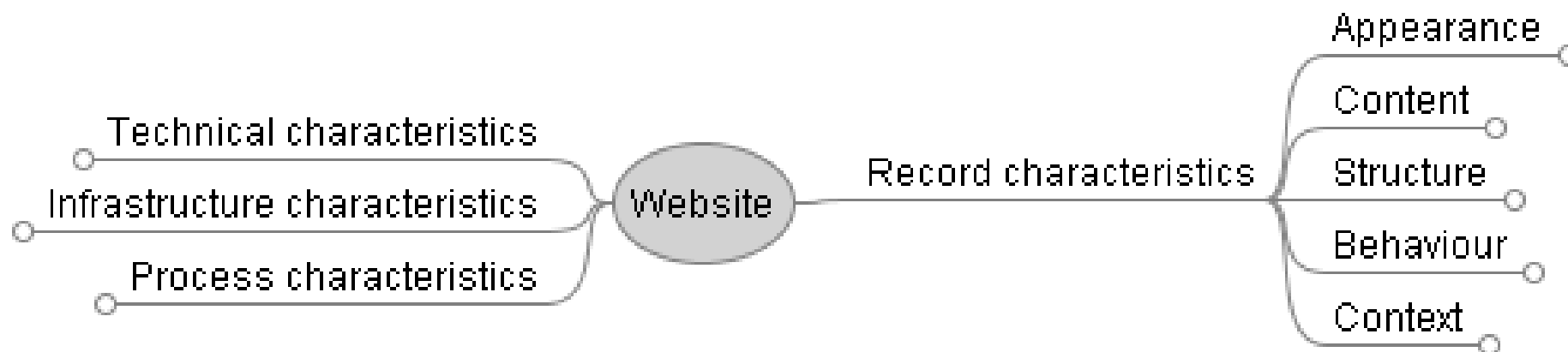
Types of scales

- Numeric
- Yes/No (Y/N)
- Yes/Acceptable/No (Y/A/N)
- Ordinal: define the possible values
- Subjective 0-to-5

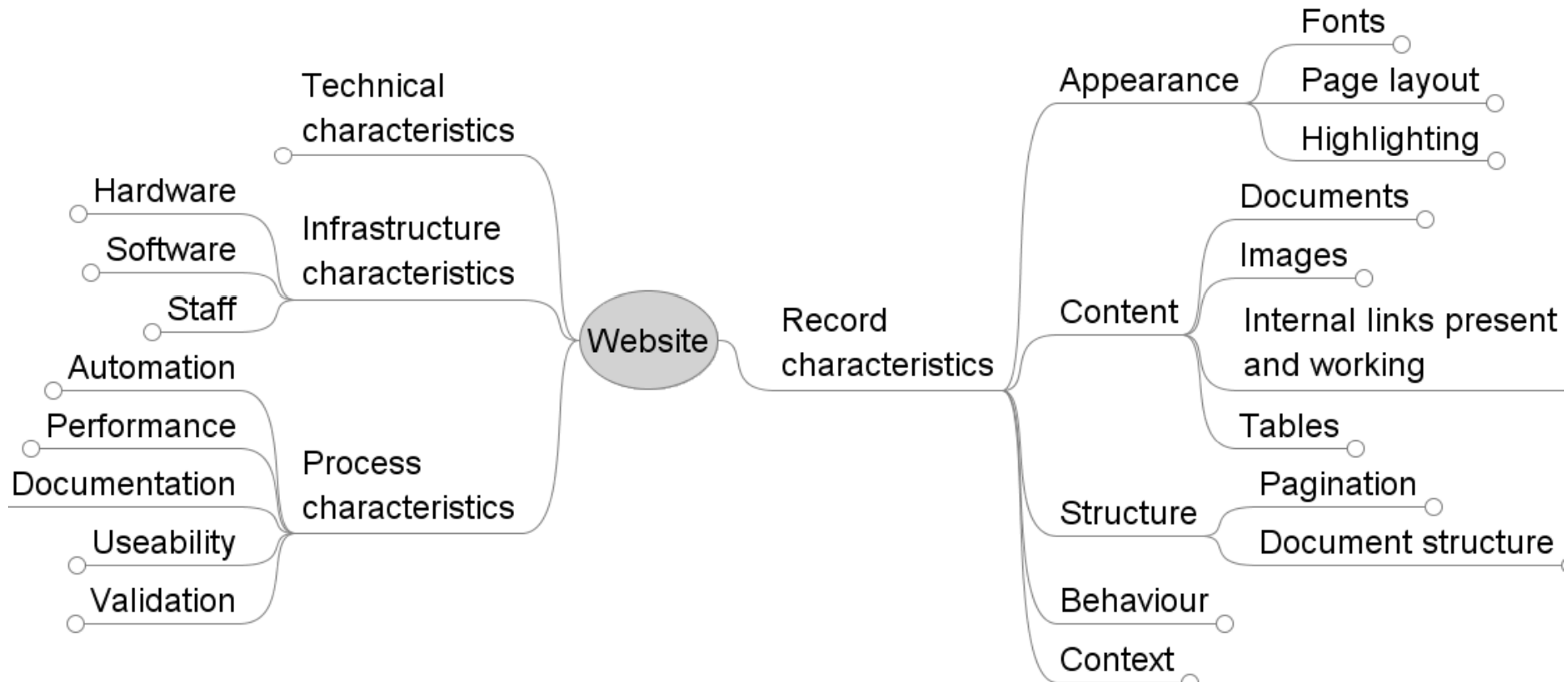


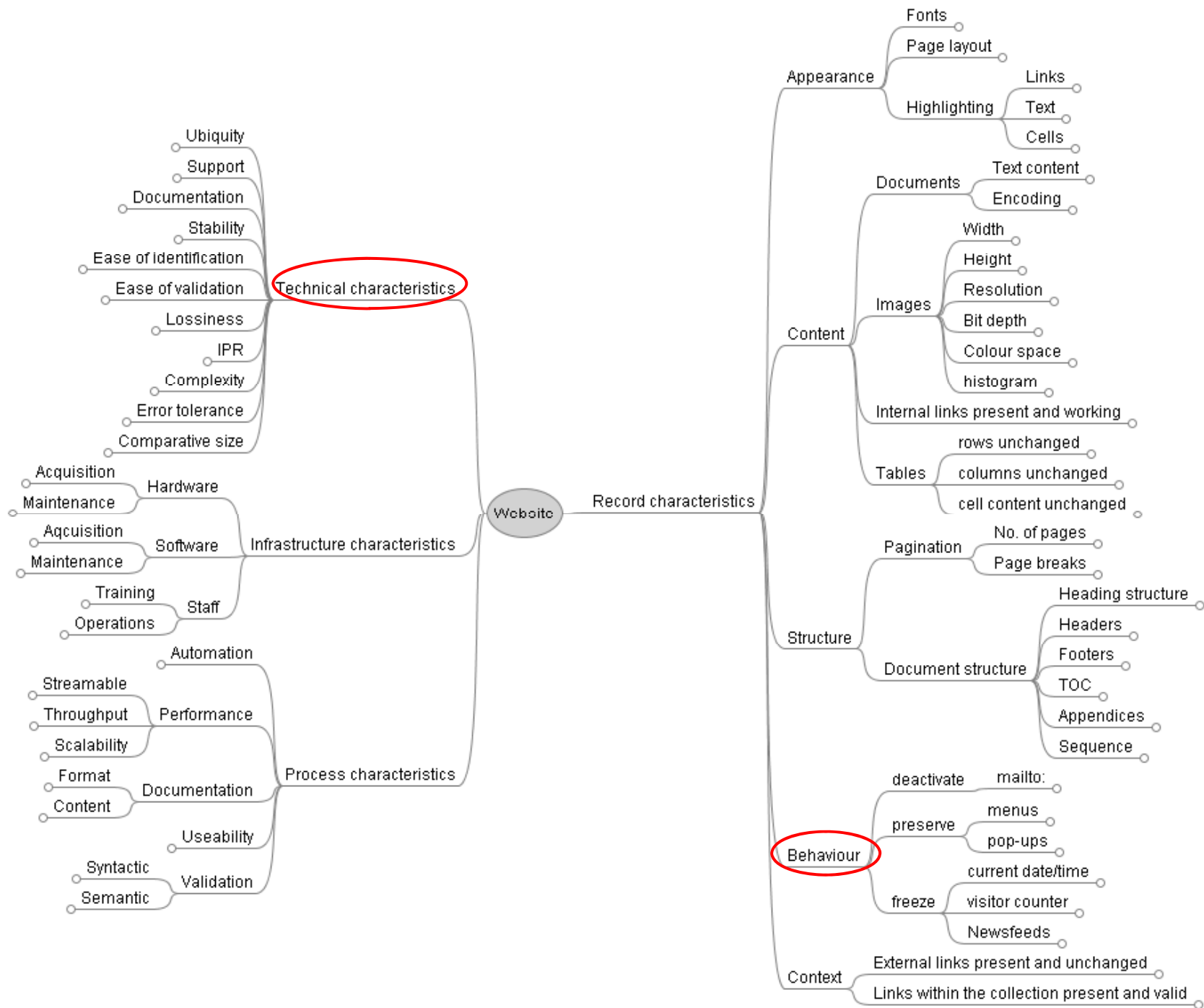
Case Study: Web archiving

- Static web pages from the public domain
- Includes documents in formats such as doc, pdf
- Images
- No interactive content shall be preserved

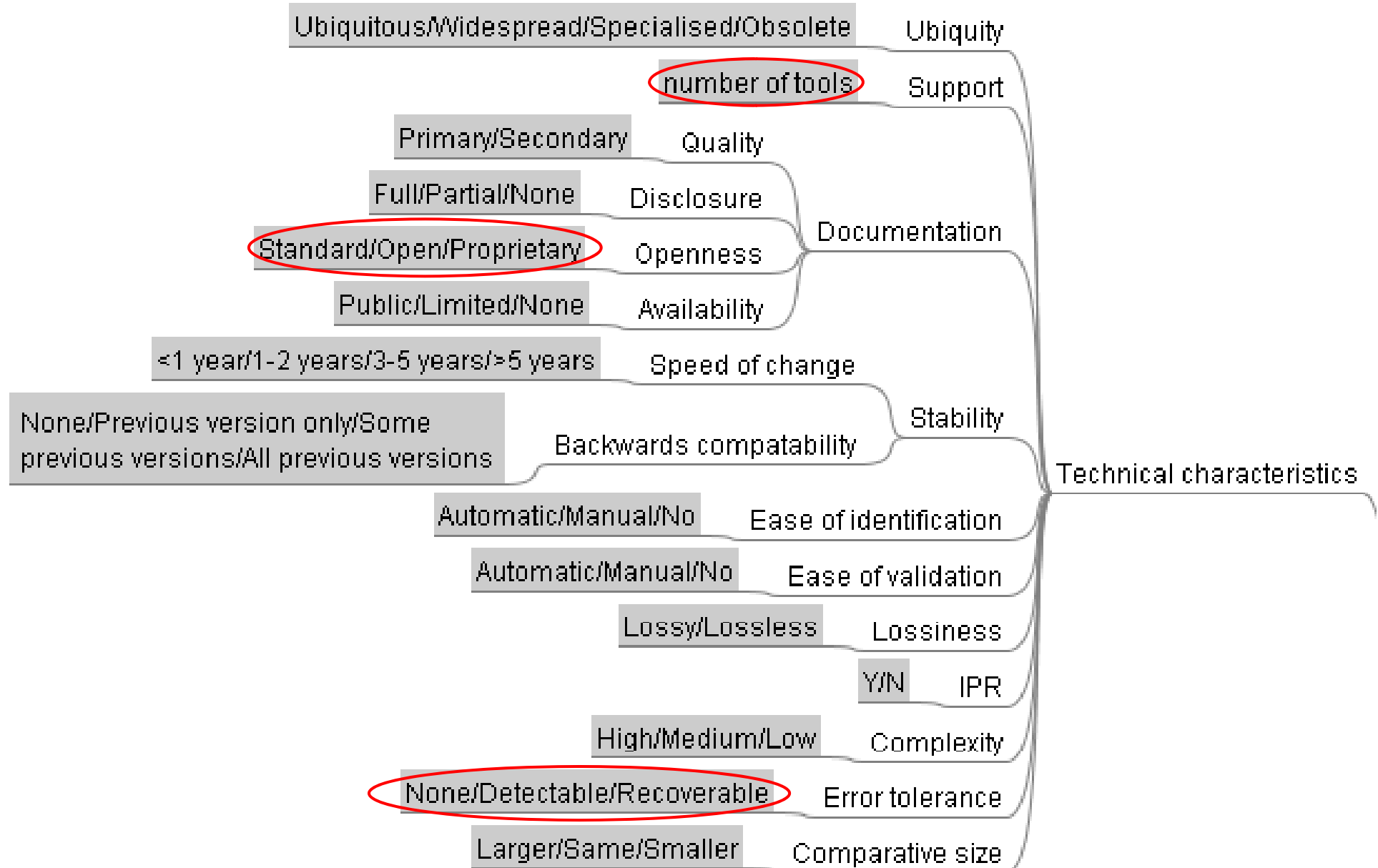


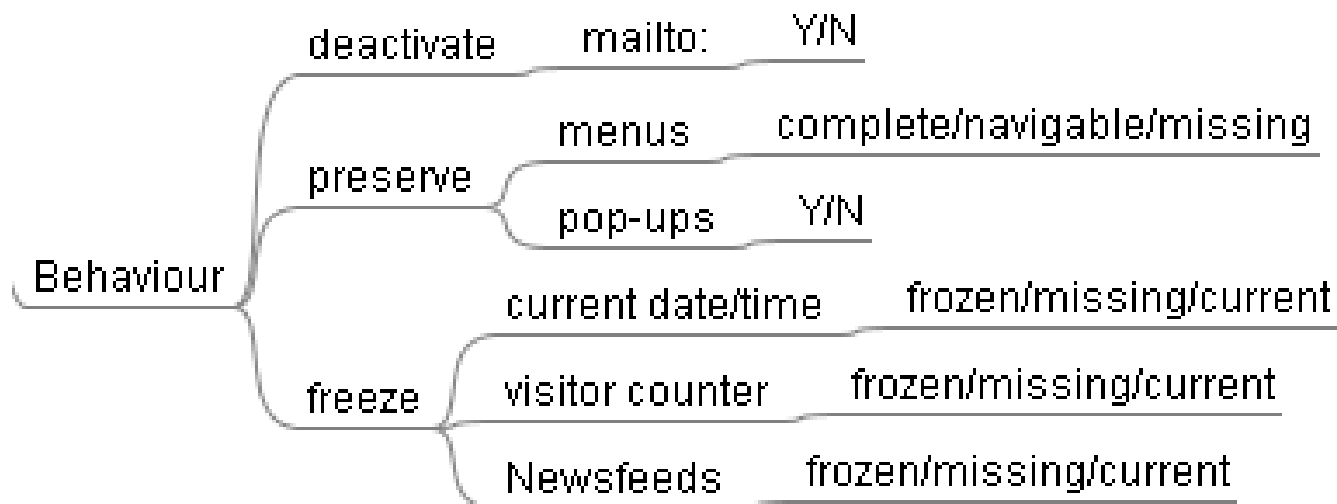
A bit more detail...



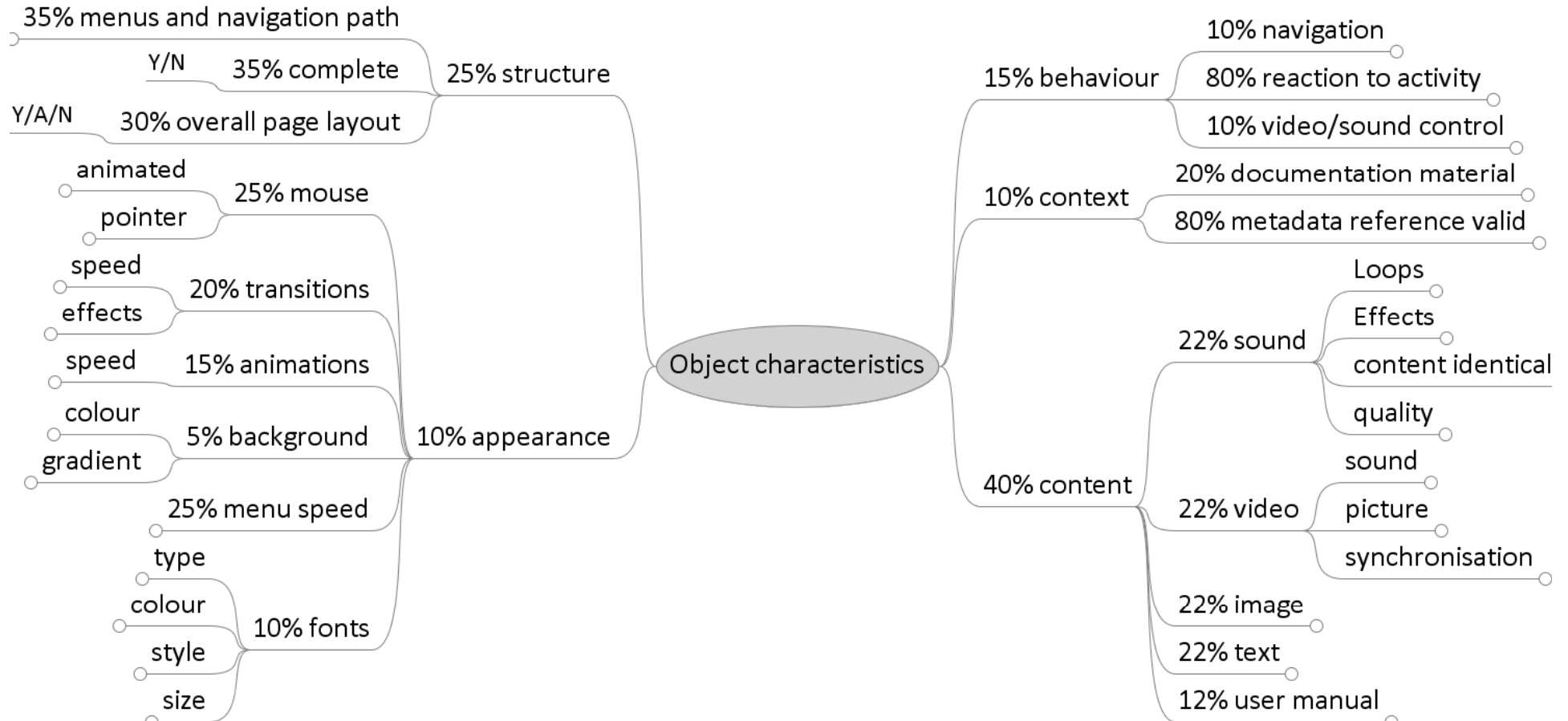


File format characteristics





- Visitor counter and similar things can be
 - Frozen at the point of harvesting
 - Left out
 - Still counting while being accessed in the archive (Is this desirable?)

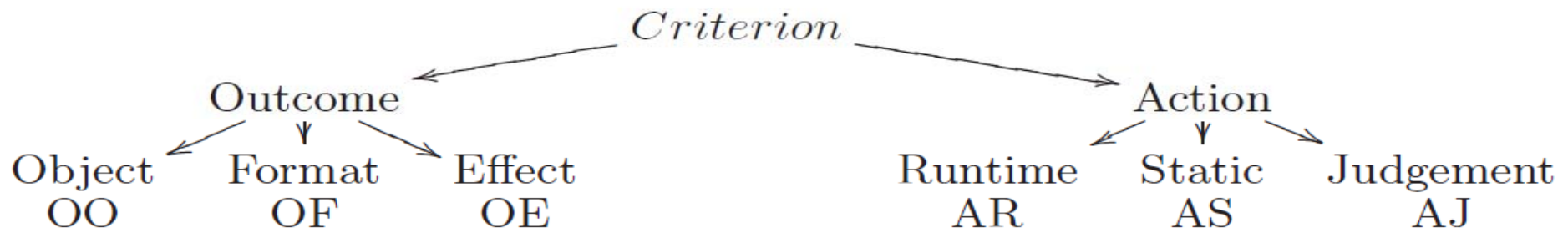


Behaviour

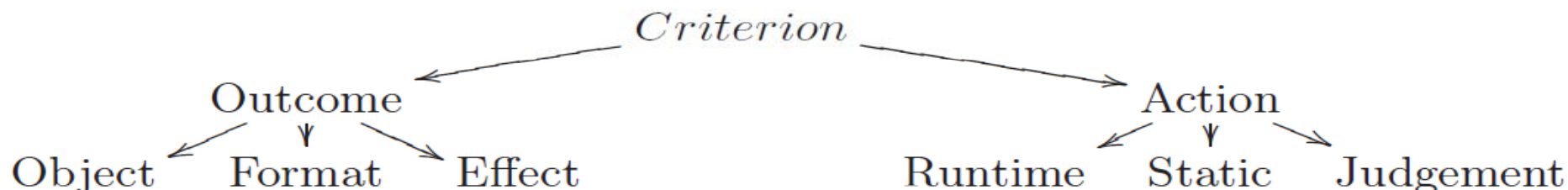
- Interactive presentations exhibit two facets
 - Graph-like navigation structure
 - Navigation along the paths

Node	Scale	Restriction
Object characteristics		
behaviour		
navigation	Ordinal	interactive and integrated/navigatable/none
reaction to activity		
mouse		
position	Boolean	
clicks	Boolean	
keyboard	Boolean	
video/sound control		
structure		
menus and navigation path	Ordinal	complete and free/partial (linear)/none
complete	Boolean	
overall page layout	Ordinal	Y/A/N

- **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...)



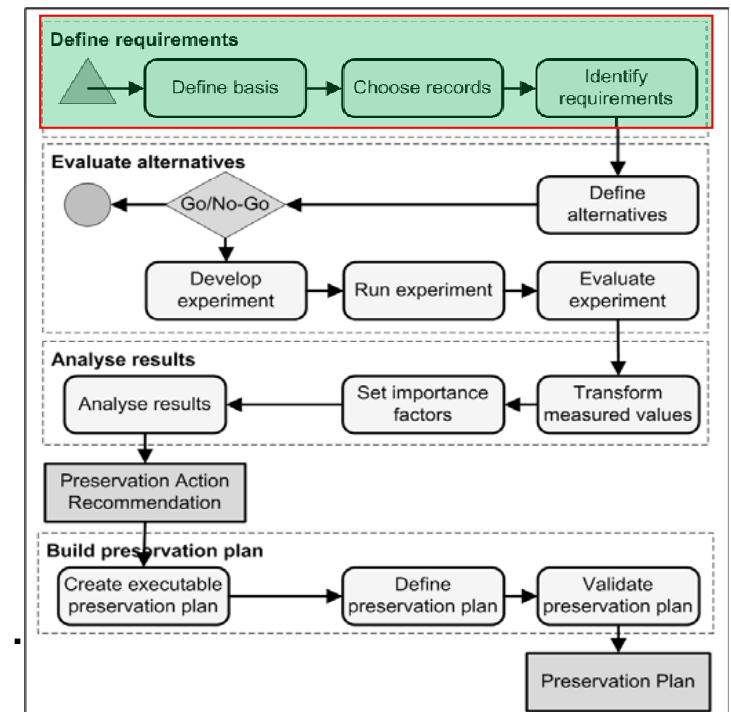
Criteria and measurements



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	

Results of Phase 1

- Defined and documented the context of a preservation problem
 - Which types of objects
 - Which environment
 - What are the obligations and constraints
- Defined and documented representative samples for performing experiments
- Defined and documented goals and requirements



- ... 10 minutes break ...
- An arctic mountain adventure
 - Requirements
 - Exercise

Practice time!

A digital preservation scenario

Context: National library

- We are a national library
- Legal mandate: Make publicly available the cultural heritage of our times to the people, free of charge, barrier-free, now **and in the future**
- Legal mandate and budgeting might not fit together perfectly
- Find optimal solution given the constraints
- Be able to prove that we did everything to our best knowledge and using state-of-the-art technology

Scanned newspapers archive

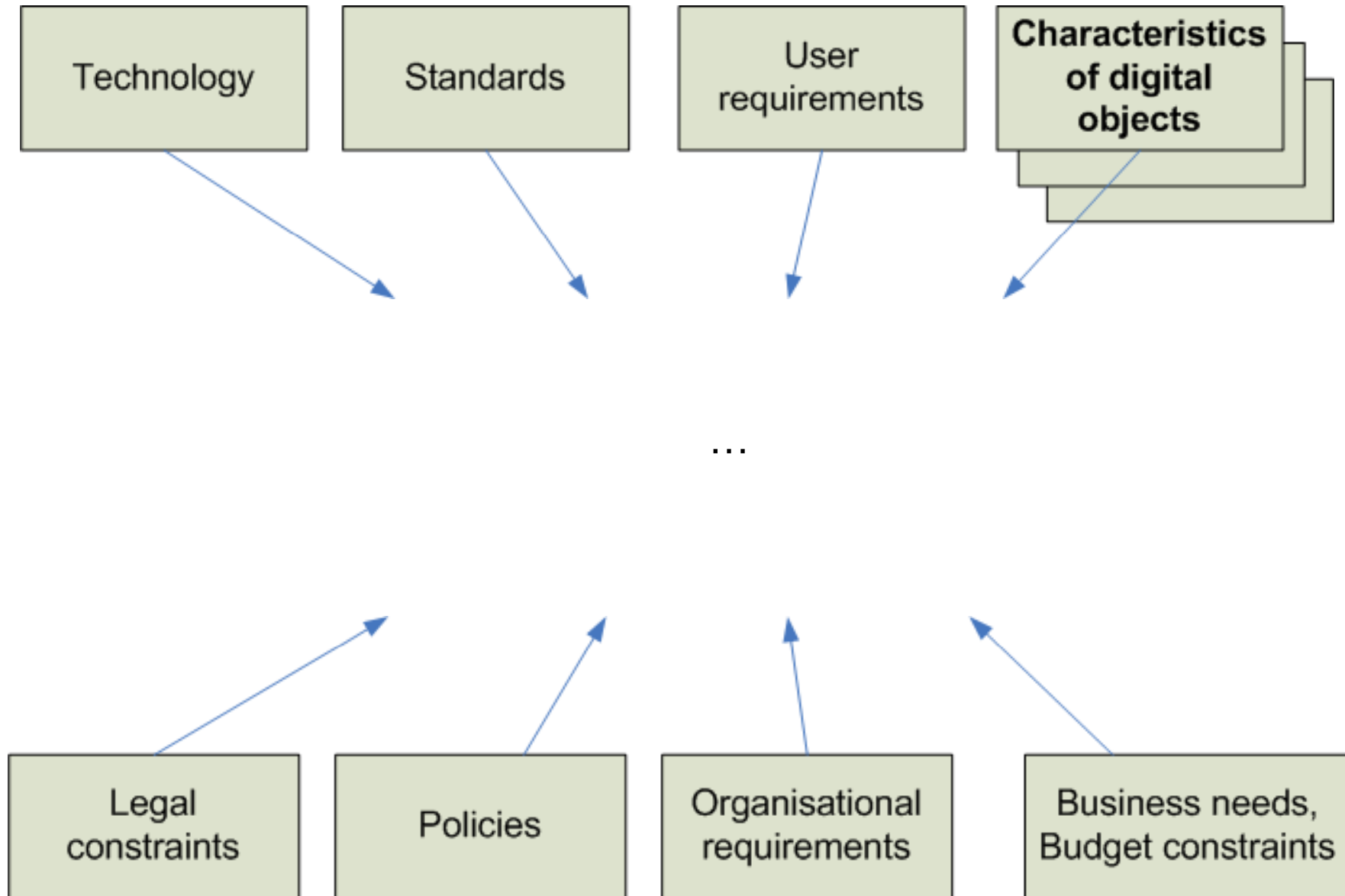
- One of the (many) collections we have is a set of newspaper scans
- ~8.000.000 GIF images, mostly black-and-white or grayscale
- Different size, resolution, age, created with different tools,...
- Should be preserved in an openly specified format, cheap but safe, and easily accessible to the public (online)

Create a preservation plan

- Define the scope, the scenario, the constraints
- Describe the content we have to care for
- Specify our requirements
- Shortlist of potential alternative strategies
- Evaluate them
- Select the best and implement it
- Monitor it closely to detect deviations

- Today... part 1
 - Define the scope, the scenario, the constraints
 - Describe the content we have to care for
 - Specify high-level requirements
- Next week... part 2
 - Revisit requirements
 - Evaluate alternatives and analyse results
 - Define the preservation plan

Requirements definition



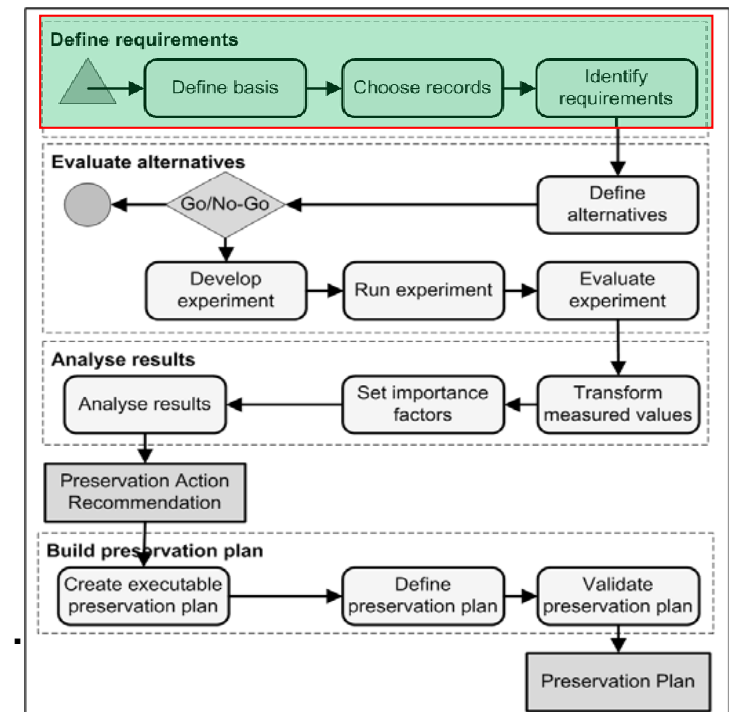


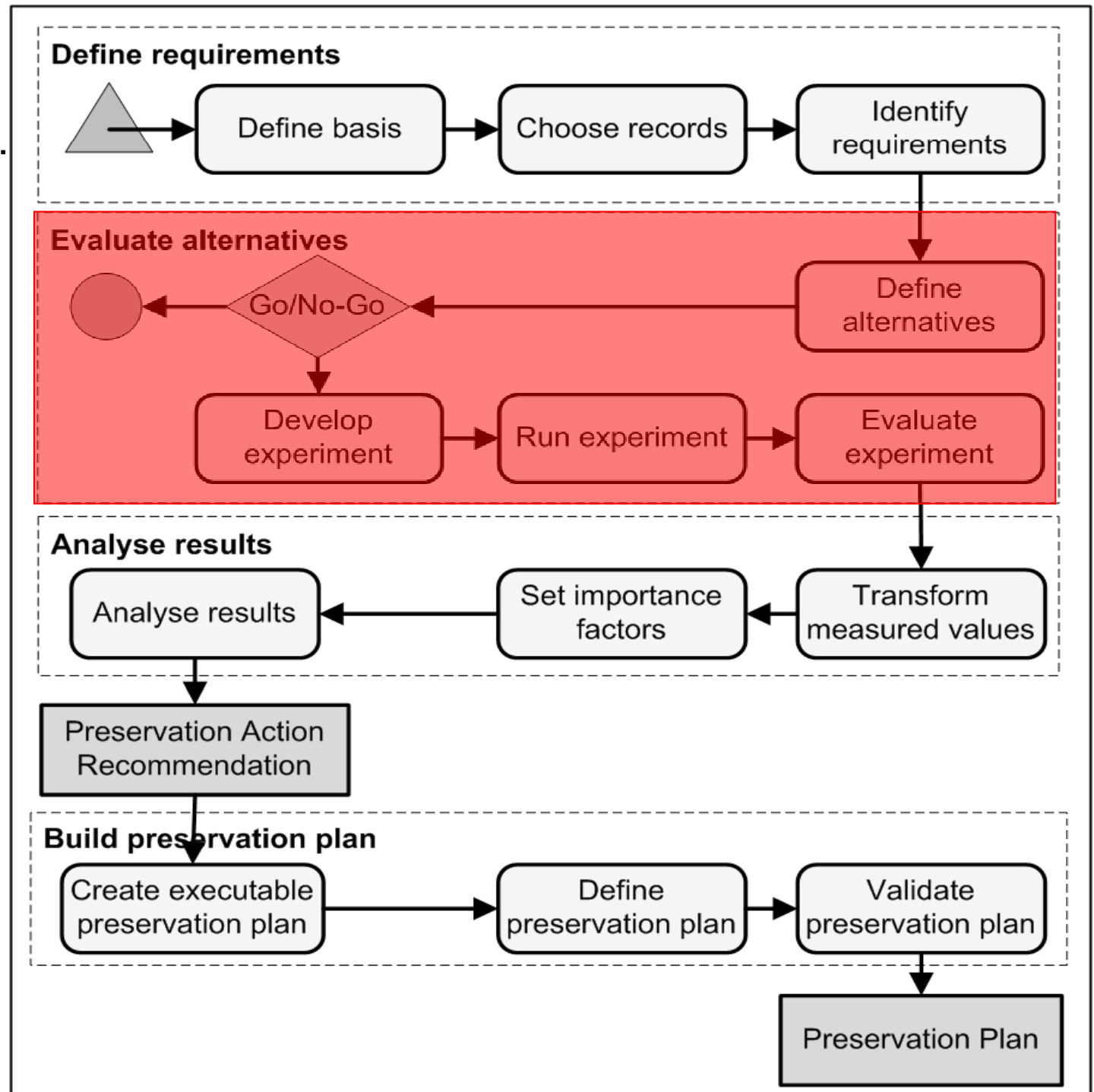
Requirements definition

...

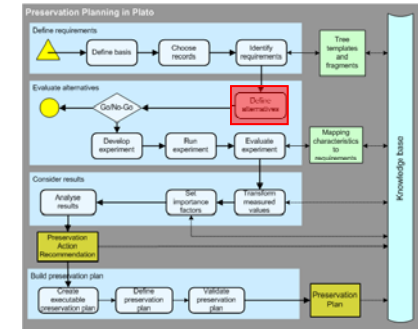
Results of Phase 1

- Defined and documented the context of a preservation problem
 - Which types of objects
 - Which environment
 - What are the obligations and constraints
- Defined and documented representative samples for performing experiments
- Defined and documented goals and requirements





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



Define the alternatives of the Project

ID	Name	Description	Remove
196616	TIFF (tool A)	Convert to TIFF using the well-tested and expensive tool 'A'	Remove
196613	TIFF (tool B)	Convert to TIFF/4 using this new tool named 'B'	Remove
196614	GIF (tool C)	Convert to GIF using the well-tested tool 'C'	Remove
196615	PNG (tool D)	Convert to PNG using the well-tested tool 'D'	Remove

[Add new Alternative](#)

[Save](#)

[Discard changes](#)

[Proceed](#)

Create alternatives from applicable services

Sample record #1 has format JPEG File Interchange Format, 1.01.

You can look up services that are able to handle this object type in the following registries:

Planets Preservation Action Tool registry



[Show Preservation Services](#)

Planets Service Registry



[Show Preservation Services](#)

CRiB Service Registry

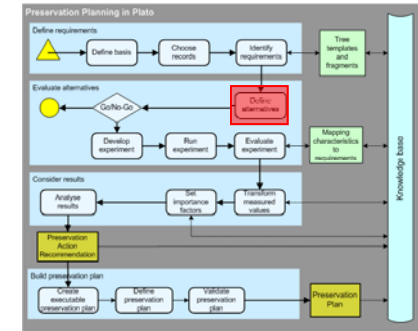


[Show Preservation Services](#)

	Preservation Action	Target Format	Info
<input type="checkbox"/>	JPG > BMP	Windows Bitmap, version 3.0	JPG>BMP
<input checked="" type="checkbox"/>	JPG > TIF	Tagged Image File Format, version 3	JPG>BMP>TIF
<input type="checkbox"/>	JPG > TIF #2	Tagged Image File Format, version 3	JPG>TIF
<input checked="" type="checkbox"/>	JPG > TIF_2	Tagged Image File Format, version 3	JPG>TIF_2
<input type="checkbox"/>	JPG > PNG	Portable Network Graphics, version 1.0	JPG>PNG
<input type="checkbox"/>	JPG > JP2	JPEG 2000	JPG>JP2

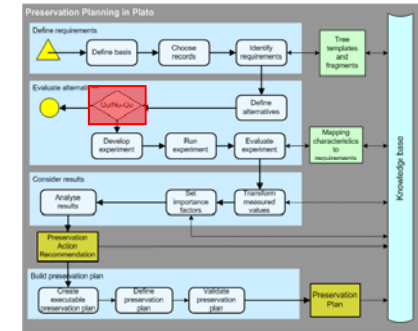
[Create alternatives for selected services](#)

Specify resources

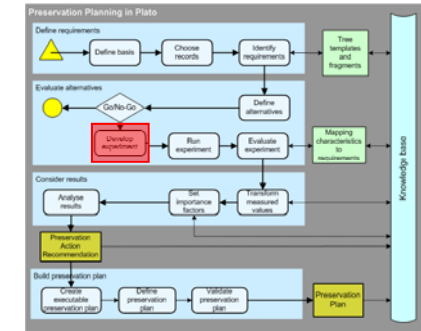


- Detailed design and overview of the resources for each alternative
 - human resources (qualification, roles, responsibility, ...)
 - technical requirements (hardware and software components)
 - time (time to set-up, run experiment,...)
 - cost (costs of the experiments,...)

Go/No-Go

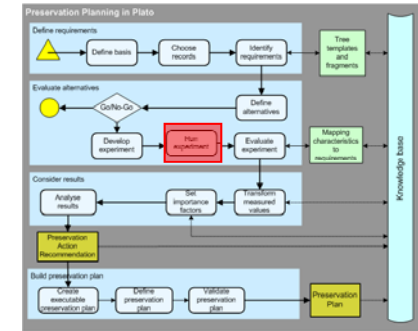


- Deliberate step for taking a decision whether it will be useful and cost-effective to continue the procedure, given
 - The resources to be spent (people, money)
 - The availability of tools and solutions,
 - The expected result(s).
- Review of the experiment/ evaluation process design so far
 - Is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



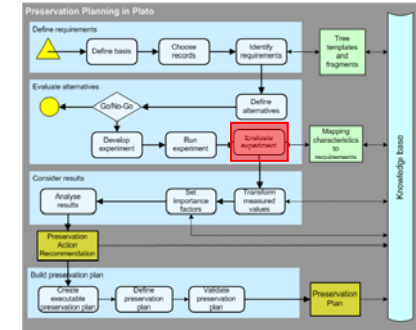
- Formulate for each evaluation or experiment or preservation process detailed
 - Development plan
 - steps to build and test software components
 - procedures and preparation
 - parameter settings for integrating preservation services
 - Test plan (mechanisms how to)
 - Evaluation/experiment plan (workflow/sequence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

Run experiment

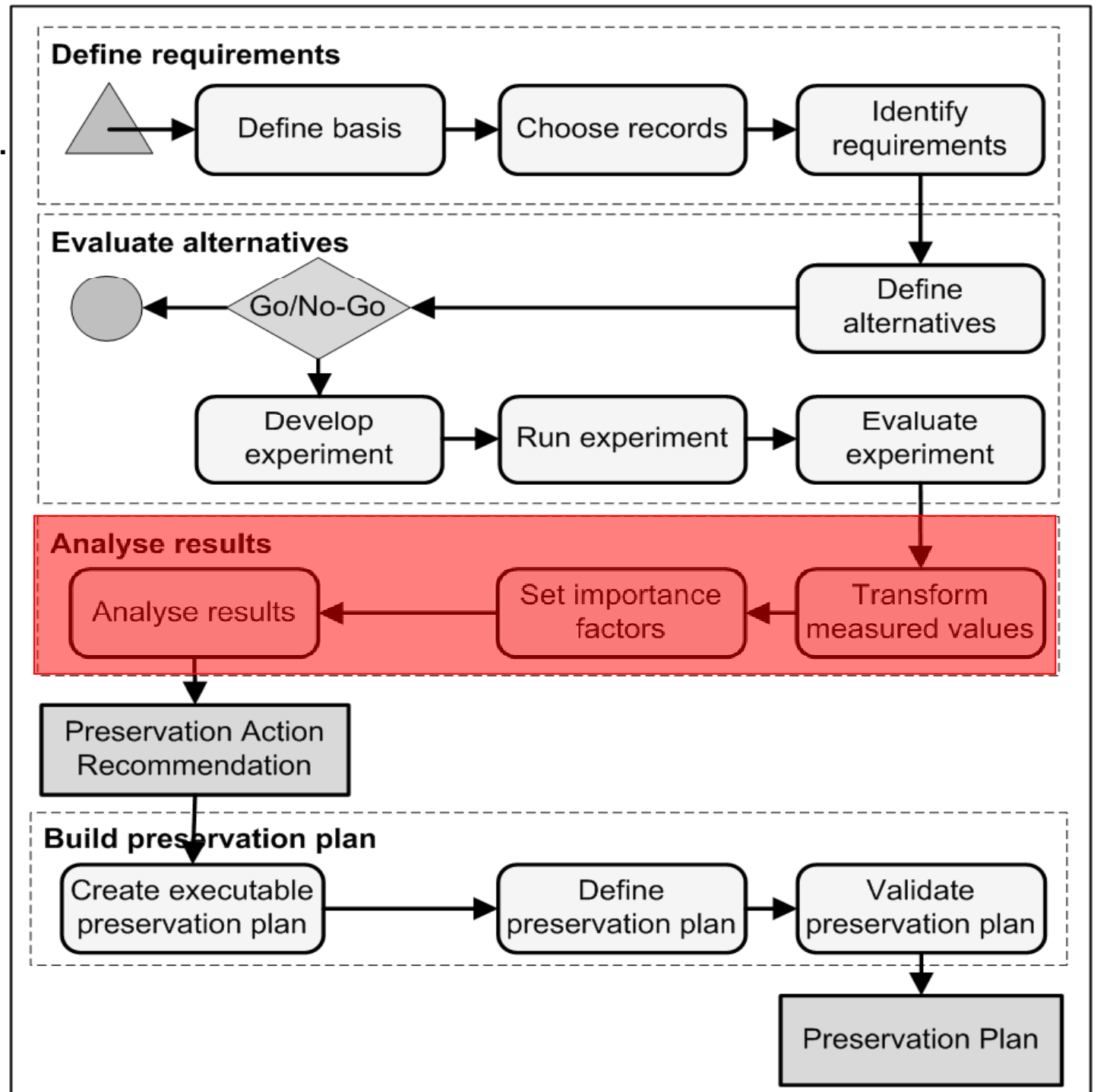


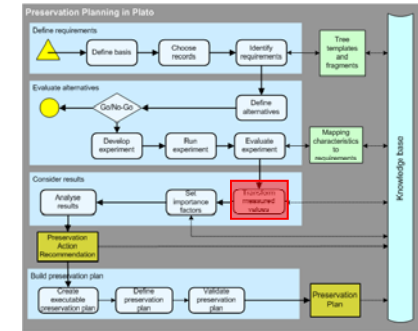
- Before conducting an evaluation or running an experiment, the experiment process as designed has to be tested
 - It may lead to re-design or even termination of the evaluation/ experiment process
- The results will be evaluated in the next stage
- The whole process needs to be documented

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

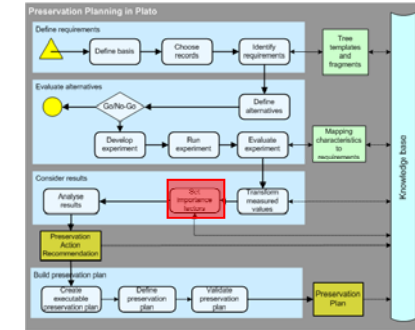




- Measures come in seconds, euro, bits, goodness values,...
- Need to make them comparable
- Transform measured values to uniform scale
- Transformation tables for each leaf criterion
- Linear transformation, logarithmic, special scale
- Scale 1-5 plus "not-acceptable"

Set importance factors

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



PLANETS Preservation Planning Tool - Mozilla Firefox

http://localhost:8080/plato/workflow/importancefactors.seam

PLANETS Preservation Planning Tool (*Plato*)

Project | Define Requirements | Evaluate Requirements | Consider Results | Project 'Minimalist'

Set Importance Factors

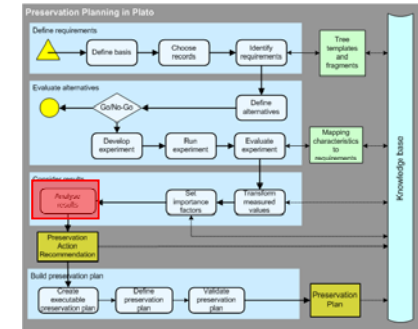
Balance weights automatically

[Expand All](#) | [Collapse All](#)

Object characteristics

Focus	Name	Weight	Lock	Total weight
	▼ Object characteristics	0	<input type="checkbox"/>	1
X	▶ behaviour	0	<input checked="" type="checkbox"/>	0.15
X	▶ structure	0	<input checked="" type="checkbox"/>	0.25
X	▶ context	0	<input type="checkbox"/>	0.1
X	▶ appearance	0	<input type="checkbox"/>	0.1
X	▶ content	0	<input checked="" type="checkbox"/>	0.4

Save Proceed



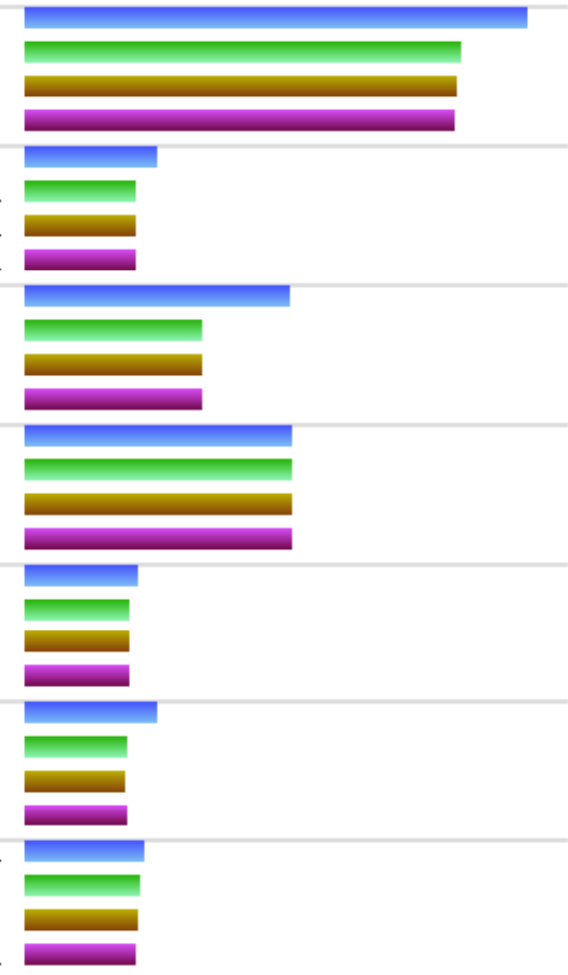
- Aggregate Values

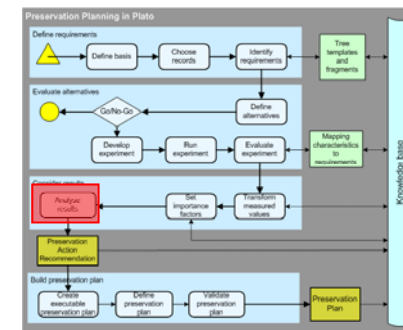
- Multiply the transformed measured values in the leaf nodes with the leaf weights
- Sum up the transformed weighted values over all branches of the tree
- Creates performance values for each alternative on each of the sub-criteria identified

Results: Weighted sum

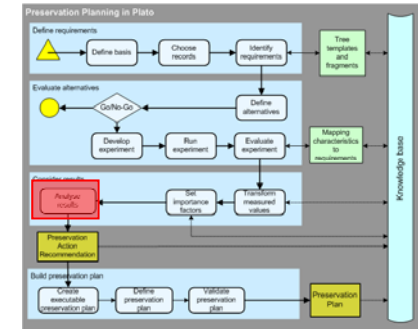
Result-Tree with all Alternatives, Aggregation method: Weighted sum.

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

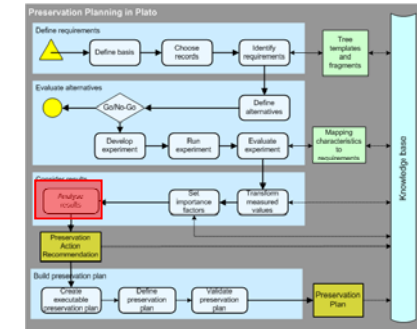
Node	Results
<ul style="list-style-type: none"> Scans <ul style="list-style-type: none"> Keep status quo: 4.70 ImageMagick - TIFF to JP2: 4.09 Kakadu - TIFF to JP2: 4.06 GeoJasper - TIFF to JP2: 4.03 Object characteristics <ul style="list-style-type: none"> Keep status quo: 1.25 ImageMagick - TIFF to JP2: 1.04 Kakadu - TIFF to JP2: 1.04 GeoJasper - TIFF to JP2: 1.04 Content <ul style="list-style-type: none"> Keep status quo: 2.50 ImageMagick - TIFF to JP2: 1.68 Kakadu - TIFF to JP2: 1.68 GeoJasper - TIFF to JP2: 1.68 Context <ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Keep status quo: 1.06 ImageMagick - TIFF to JP2: 0.98 Kakadu - TIFF to JP2: 0.98 GeoJasper - TIFF to JP2: 0.98 Costs <ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> Keep status quo: 1.14 ImageMagick - TIFF to JP2: 1.10 Kakadu - TIFF to JP2: 1.08 GeoJasper - TIFF to JP2: 1.04 	



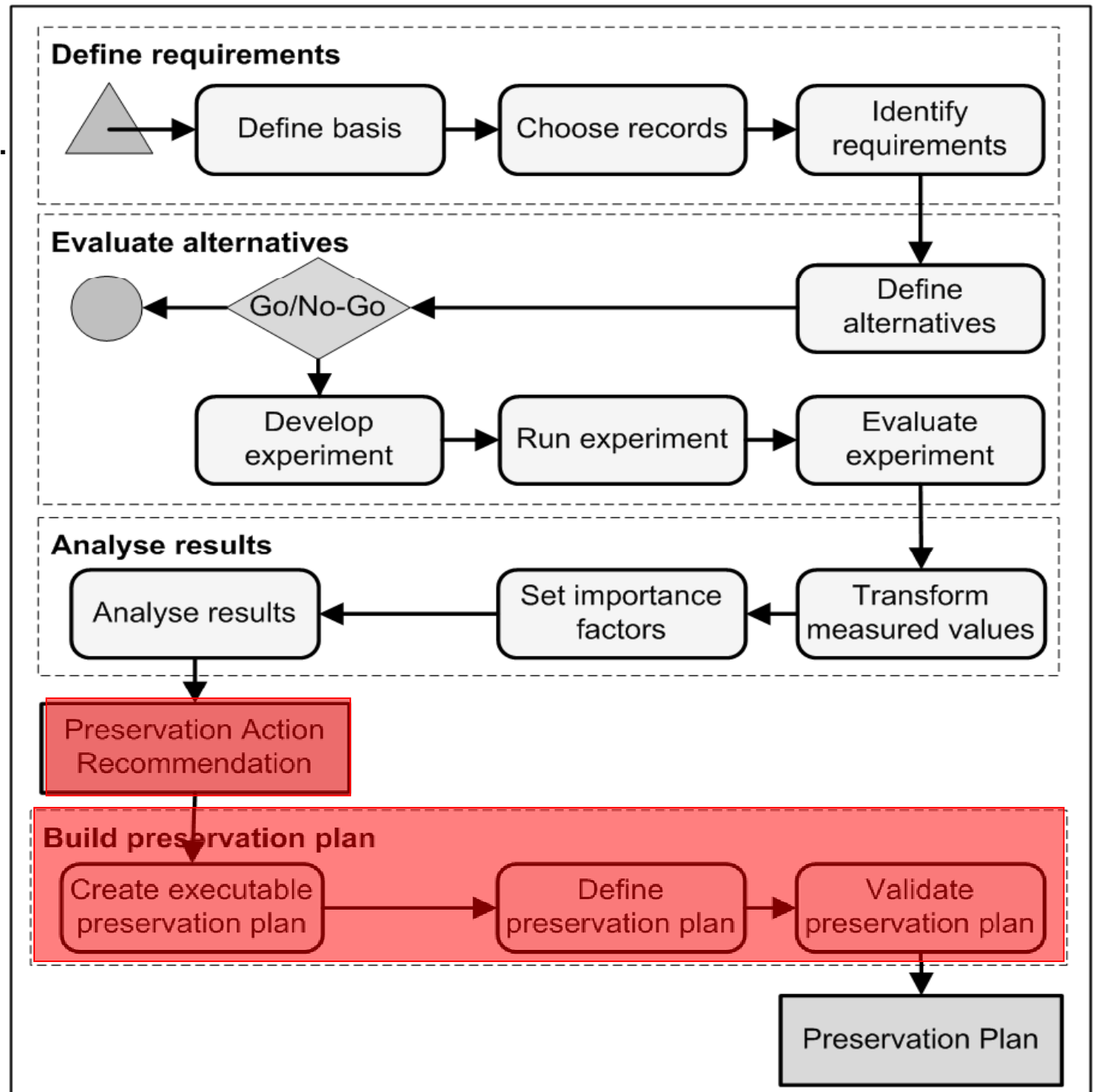
- Single performance value for each alternative to rank the alternatives
- Single performance values for each alternative for each sub-set of criteria to identify the best combination of alternatives
- Sensitivity Analysis: Analysis of the influence of small changes in the weight on the final value
- Basis for making Informed, well-documented, repeatable, accountable decisions

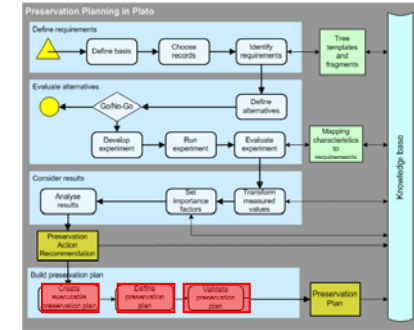


- Rank alternatives according to overall utility value at root
- Performance of each alternative
 - overall
 - for each sub-criterion (branch)
- Allows performance measurement of combinations of strategies
- Final sensitivity analysis against minor fluctuations in
 - measured values
 - importance factors



- 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





- 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 according to OAIS
 - Sign and deploy

Conclusions

- A simple, methodologically sound model to specify and document requirements
- Repeatable and documented evaluation for informed and accountable decisions
- Set of templates to assist institutions
- Generic workflow that can easily 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>

Questions?

... And a little bit of **READING** for next week:

- “From TIFF to JPEG 2000?”

<http://www.dlib.org/dlib/november09/kulovits/11kulovits.html>

- **Next week: Presentation of the exercise topics!**
- (Digital Preservation UE, 2.0)



