Digital Preservation Planning
July 29 2008, London, UK

The Planets Preservation Planning workflow and the planning tool Plato

organized in cooperation with DPC

Christoph Becker

Vienna University of Technology
http://www.ifs.tuwien.ac.at/~becker

Outline

- ❑ Preservation Planning
  - Evaluation of potential actions
- ❑ The Planets Preservation Planning Workflow
  - Underlying methodology
  - Workflow walkthrough
  - The planning tool Plato
- ❑ Requirements definition exercise
  - Group assignment
  - Schedule

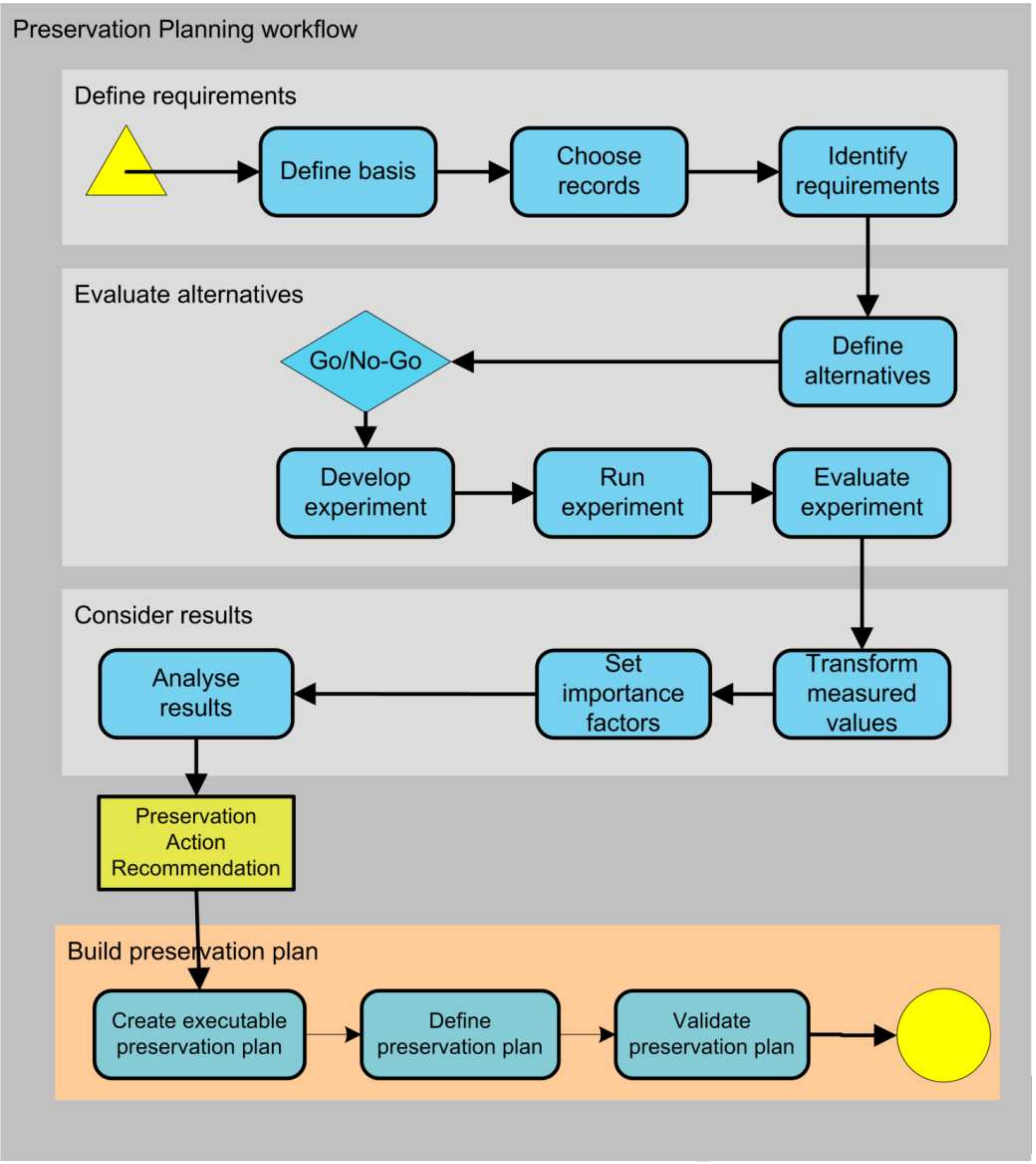
Evaluating preservation strategies

- ❑ Variety of solutions and tools exist
- ❑ Each strategy has unique strengths and weaknesses
- ❑ Requirements vary across settings
- ❑ Decision on which solution to adopt is complex
- ❑ Documentation and accountability is essential

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

Planets Preservation Planning Workflow

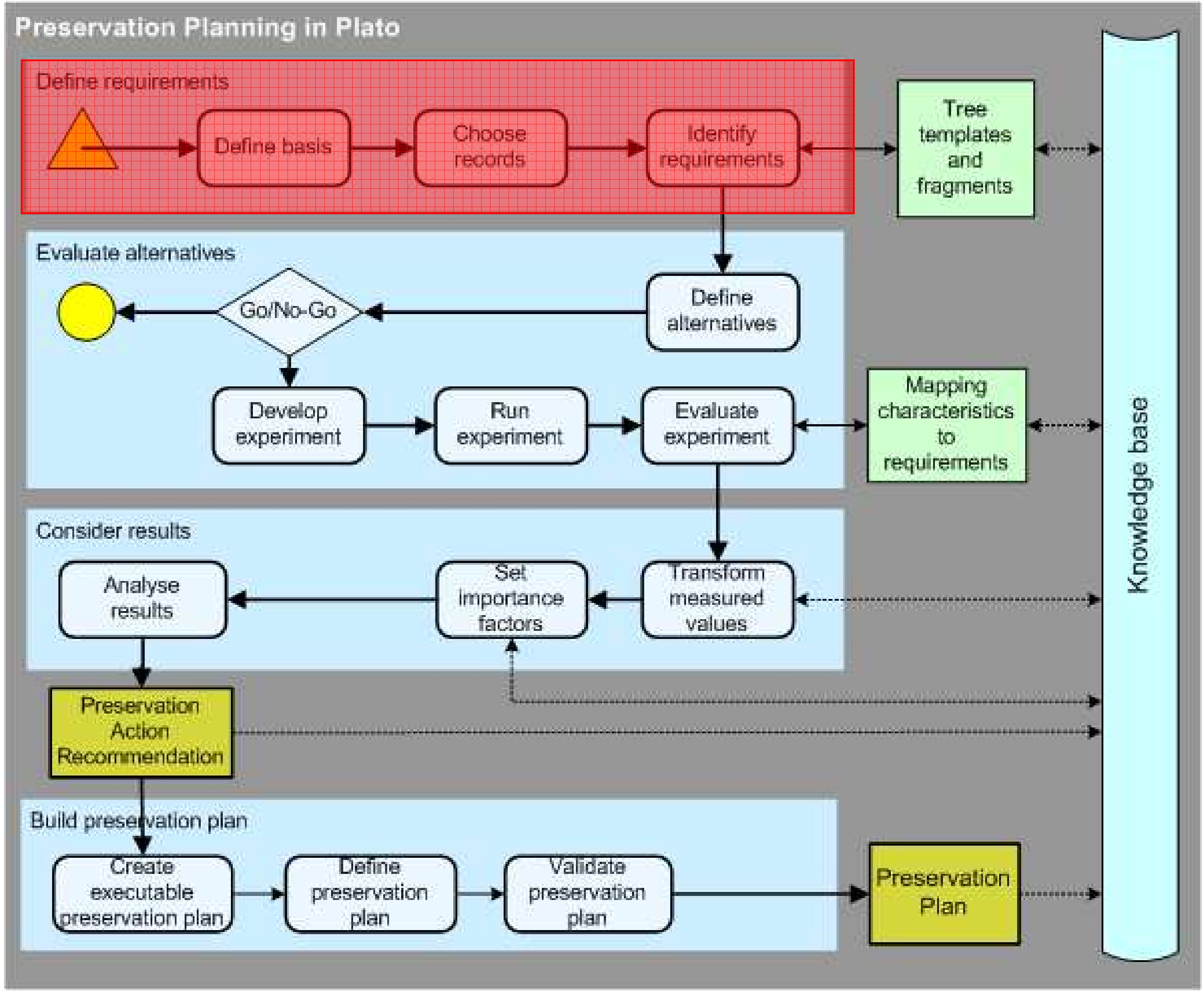
- ❑ Define requirements
- ❑ Evaluate potential actions
- ❑ Analyse results
- ❑ Build a preservation plan



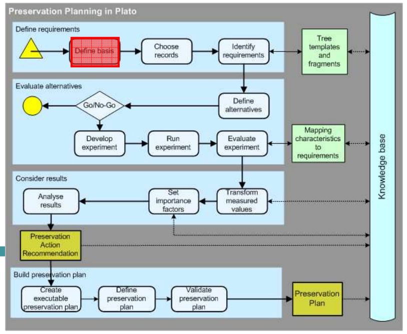
Preservation Planning in Plato

- Web based planning tool implementing the Planets preservation planning workflow
- Integration of registries and services for
  - File format identification
  - Preservation action
  - Characterisation and comparison
- Knowledge base
- A distributed architecture of preservation services

PP Workflow

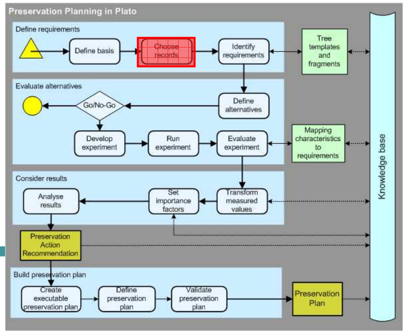

Preservation Planning in Plato
Define requirements
Define basis
Choose records
Identify requirements
Tree templates and fragments
Evaluate alternatives
Define alternatives
Go/No-Go
Develop experiment
Run experiment
Evaluate experiment
Mapping characteristics to requirements
Consider results
Transform measured values
Set importance factors
Analyse results
Preservation Action Recommendation
Build preservation plan
Create executable preservation plan
Define preservation plan
Validate preservation plan
Preservation Plan
Knowledge base

Define basis



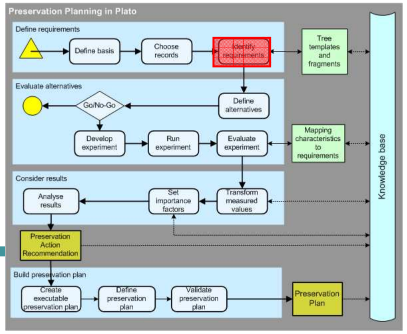
- What are the objects?
- What are the essential characteristics?
  - Content, context, structure, form and behaviour
- What are the requirements?
  - Authenticity, reliability, integrity, useability
  - Metadata (for different purposes)
- What preservation strategies will be applied and evaluated?

Choose objects/records



- Different object types
  - Text documents, audio, video, e-mail, multimedia, databases, data sets, ...
- Distinction between
  - Physical (technical) object = computer file, and
  - The intellectual object (e.g. what is shown on the screen)
- Choice of objects affects the evaluation

Identify requirements

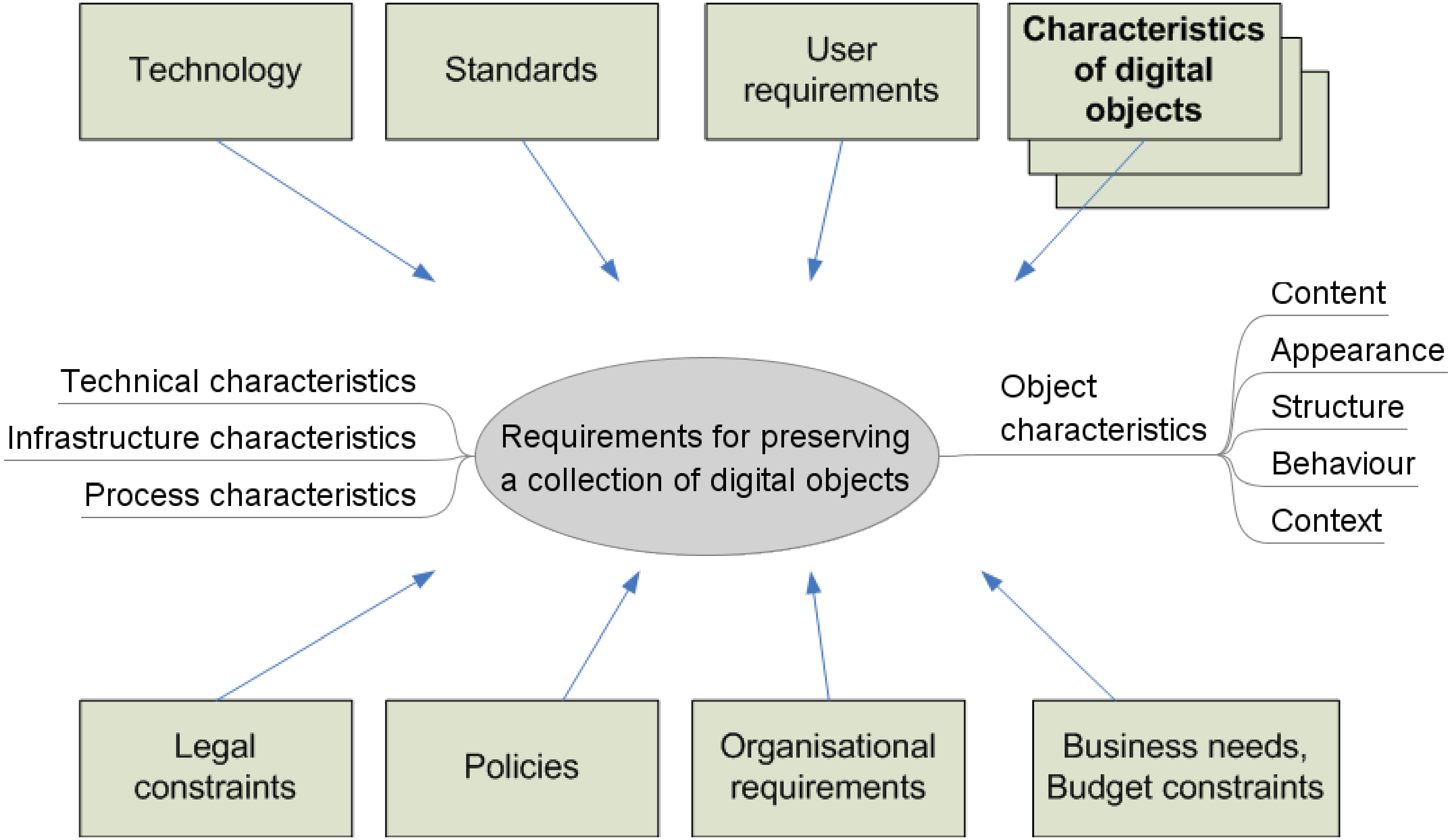


- Define all relevant goals and characteristics (high-level, detail) with respect to a given application domain
- Usually four major groups:
  - object characteristics (content, metadata ...)
  - record characteristics (context, relations, ...)
  - process characteristics (scalability, error detection,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

The Objective Tree

- Define all relevant goals and characteristics (high-level, detail) with respect to a given application domain
- Put the requirements in relation to each other → Tree structure
- Top-down or bottom-up
  - Start from high-level goals and break down to specific criteria
  - Collect criteria and organize in tree structure

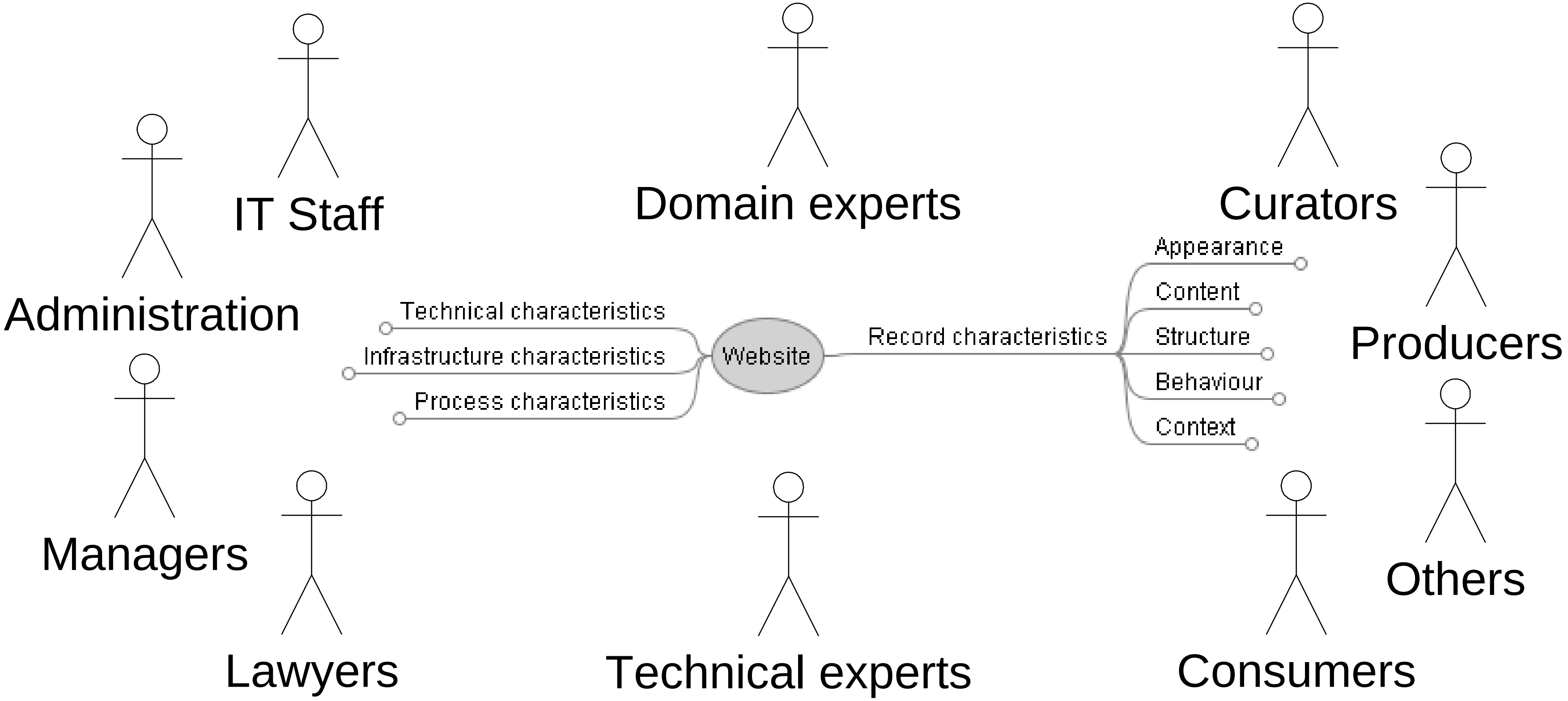
Requirements and Influence Factors



Technology
Standards
User requirements
Characteristics of digital objects
Technical characteristics
Infrastructure characteristics
Process characteristics
Requirements for preserving a collection of digital objects
Object characteristics
Content
Appearance
Structure
Behaviour
Context
Legal constraints
Policies
Organisational requirements
Business needs, Budget constraints

Stakeholders

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



IT Staff

Domain experts

Curators

Administration

Producers

Technical characteristics

Infrastructure characteristics

Process characteristics

Website

Record characteristics

Appearance

Content

Structure

Behaviour

Context

Managers

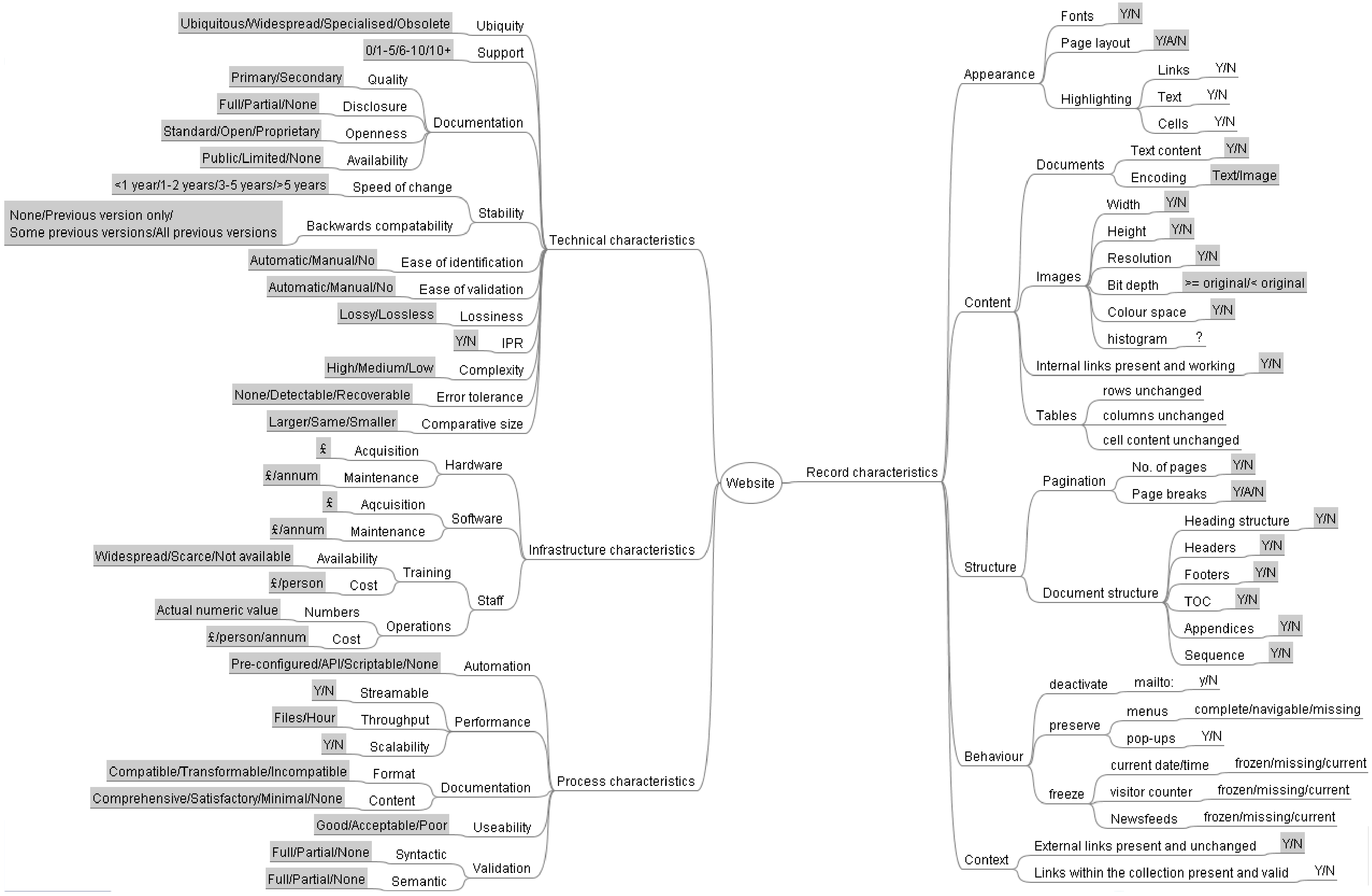
Others

Lawyers

Technical experts

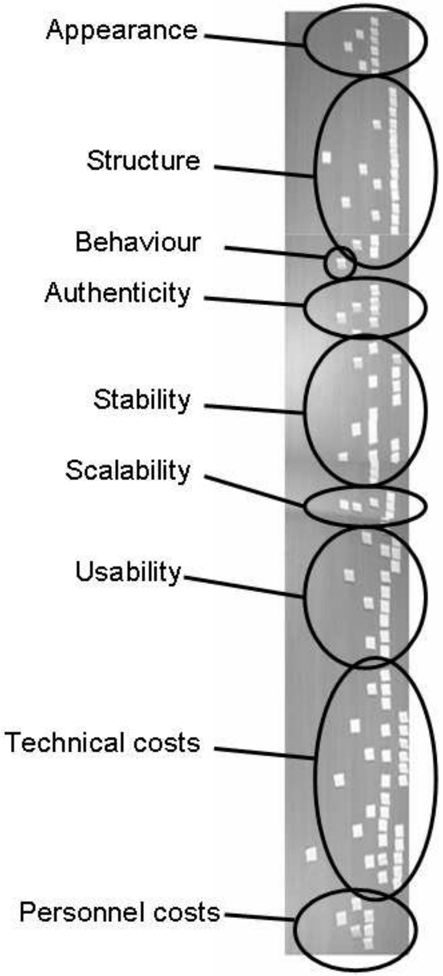
Consumers

An Objective Tree

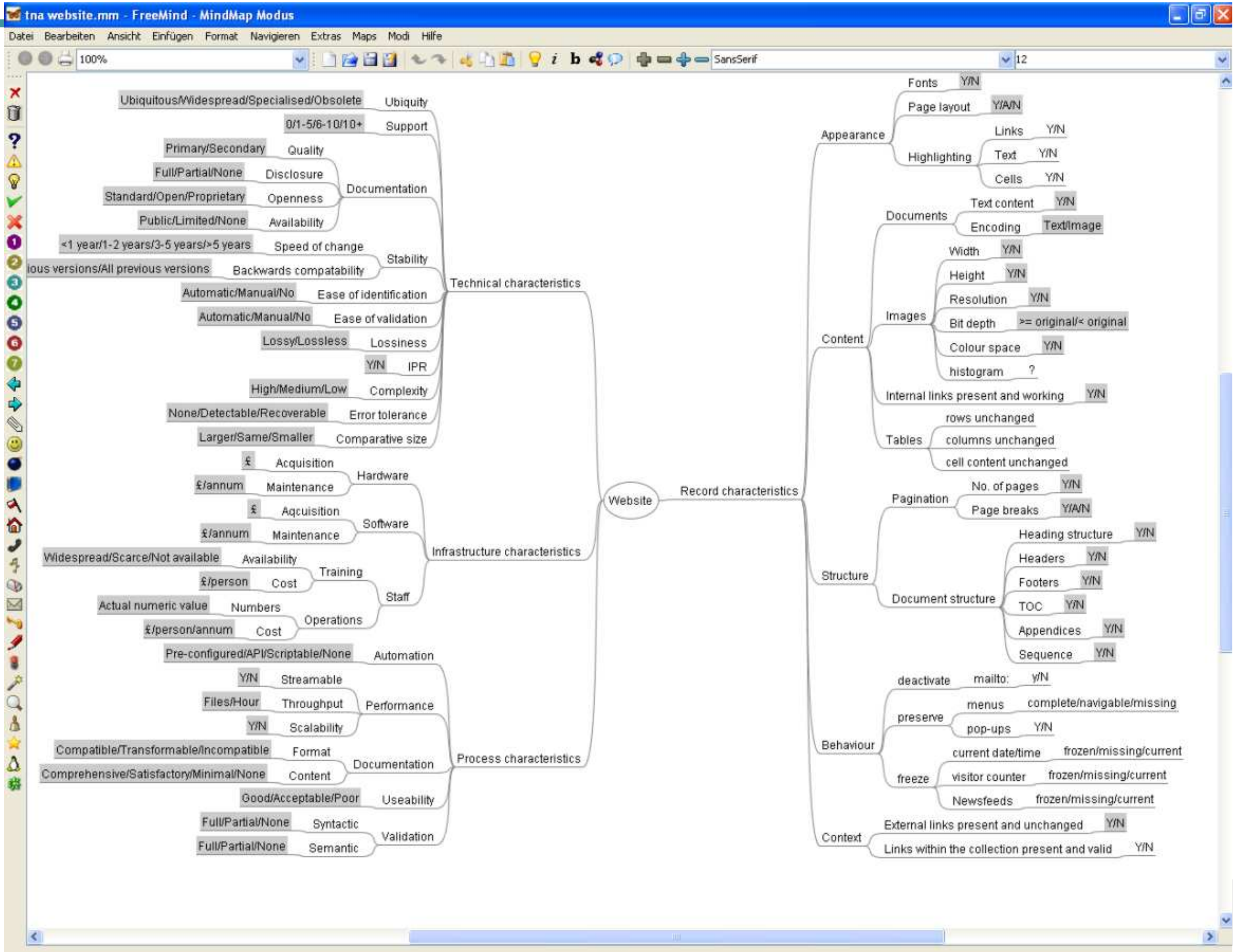


- Website
  - Technical characteristics
    - Ubiquity: Ubiquitous/Widespread/Specialised/Obsolete
    - Support: 0/1-5/6-10/10+
    - Documentation
      - Quality: Primary/Secondary
      - Disclosure: Full/Partial/None
      - Openness: Standard/Open/Proprietary
      - Availability: Public/Limited/None
    - Stability
      - Speed of change: <1 year/1-2 years/3-5 years/>5 years
      - Backwards compatability: None/Previous version only/ Some previous versions/All previous versions
    - Ease of identification: Automatic/Manual/No
    - Ease of validation: Automatic/Manual/No
    - Lossiness: Lossy/Lossless
    - IPR: Y/N
    - Complexity: High/Medium/Low
    - Error tolerance: None/Detectable/Recoverable
    - Comparative size: Larger/Same/Smaller
  - Infrastructure characteristics
    - Hardware
      - Acquisition: £
      - Maintenance: £/annum
    - Software
      - Aqcuisition: £
      - Maintenance: £/annum
    - Staff
      - Training
        - Availability: Widespread/Scarce/Not available
        - Cost: £/person
      - Operations
        - Numbers: Actual numeric value
        - Cost: £/person/annum
  - Process characteristics
    - Automation: Pre-configured/API/Scriptable/None
    - Performance
      - Streamable: Y/N
      - Throughput: Files/Hour
      - Scalability: Y/N
    - Documentation
      - Format: Compatible/Transformable/Incompatible
      - Content: Comprehensive/Satisfactory/Minimal/None
    - Useability: Good/Acceptable/Poor
    - Validation
      - Syntactic: Full/Partial/None
      - Semantic: Full/Partial/None
  - Record characteristics
    - Appearance
      - Fonts: Y/N
      - Page layout: Y/A/N
      - Highlighting
        - Links: Y/N
        - Text: Y/N
        - Cells: Y/N
    - Content
      - Documents
        - Text content: Y/N
        - Encoding: Text/Image
      - Images
        - Width: Y/N
        - Height: Y/N
        - Resolution: Y/N
        - Bit depth: >= original/< original
        - Colour space: Y/N
        - histogram: ?
      - Internal links present and working: Y/N
      - Tables
        - rows unchanged
        - columns unchanged
        - cell content unchanged
    - Structure
      - Pagination
        - No. of pages: Y/N
        - Page breaks: Y/A/N
      - Document structure
        - Heading structure: Y/N
        - Headers: Y/N
        - Footers: Y/N
        - TOC: Y/N
        - Appendices: Y/N
        - Sequence: Y/N
    - Behaviour
      - deactivate
        - mailto: y/N
      - preserve
        - menus: complete/navigable/missing
        - pop-ups: Y/N
      - freeze
        - current date/time: frozen/missing/current
        - visitor counter: frozen/missing/current
        - Newsfeeds: frozen/missing/current
    - Context
      - External links present and unchanged: Y/N
      - Links within the collection present and valid: Y/N

Analog…

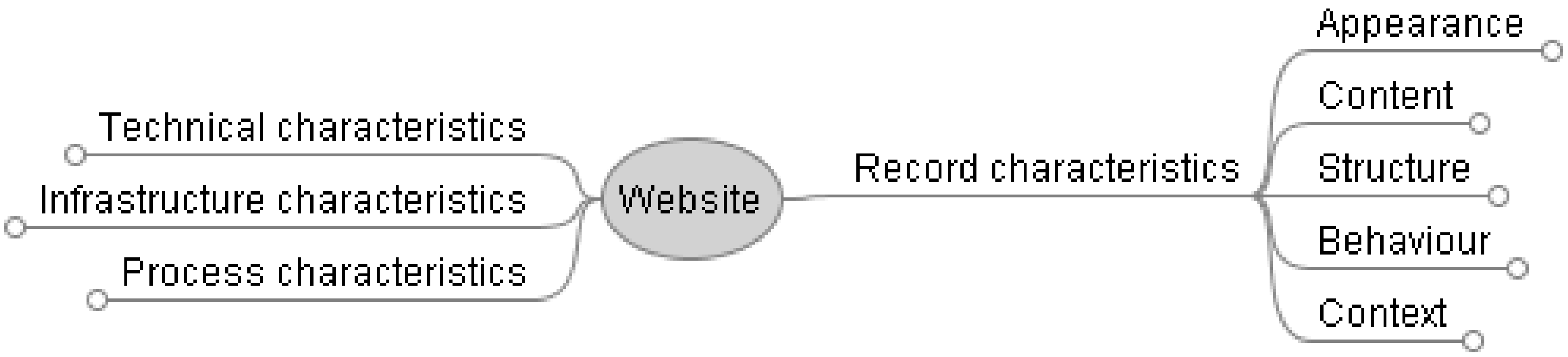

Appearance
Structure
Behaviour
Authenticity
Stability
Scalability
Usability
Technical costs
Personnel costs

… or born-digital



Case Study: Web archiving

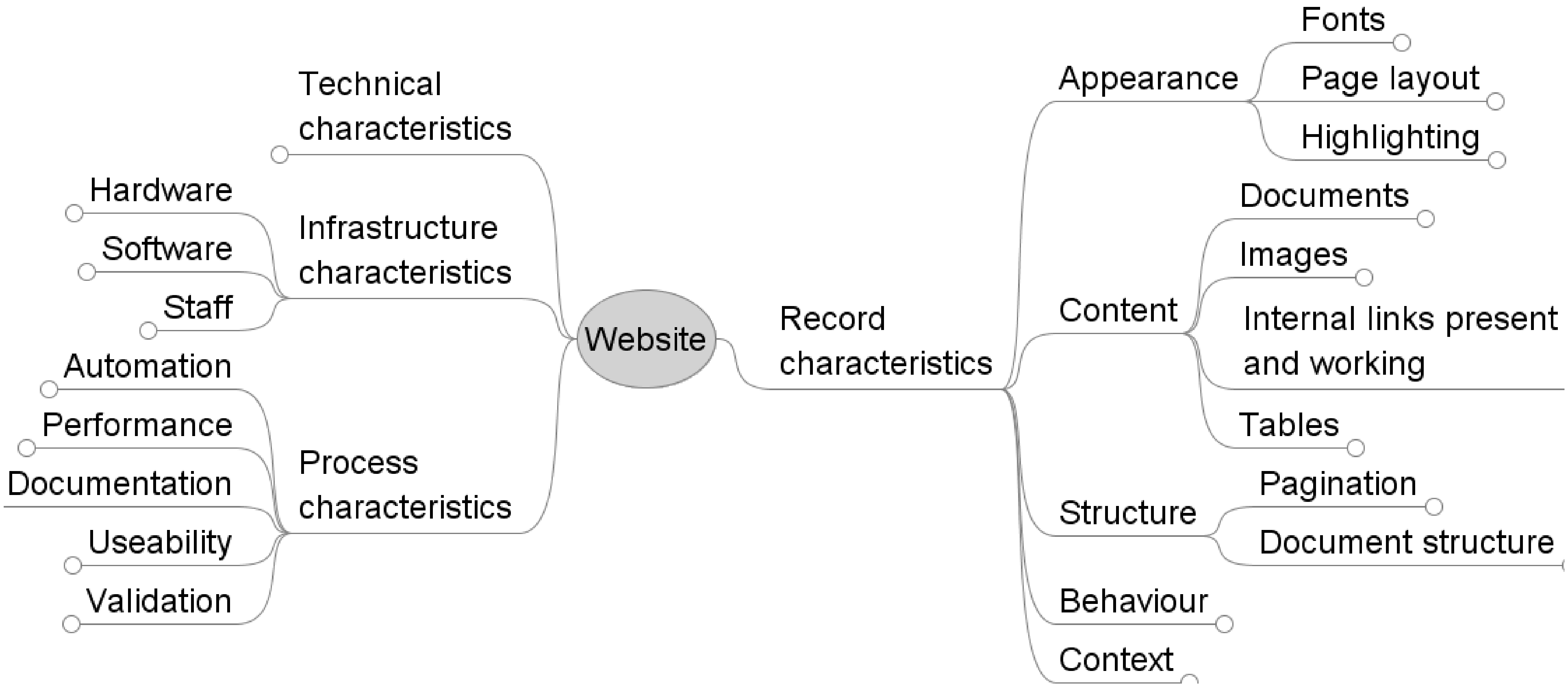
- Static web pages from the public domain
- Includes documents in formats such as doc, pdf
- Images
- No interactive content shall be preserved



Object characteristics

- Content
- Structure
- Appearance
- Behaviour
- Context

A bit more detail…

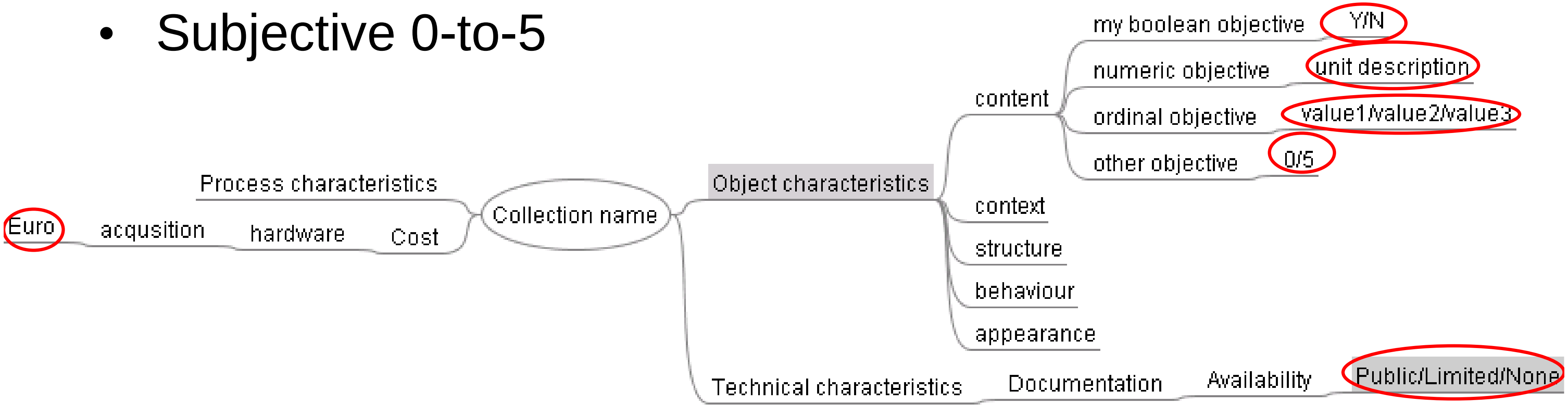


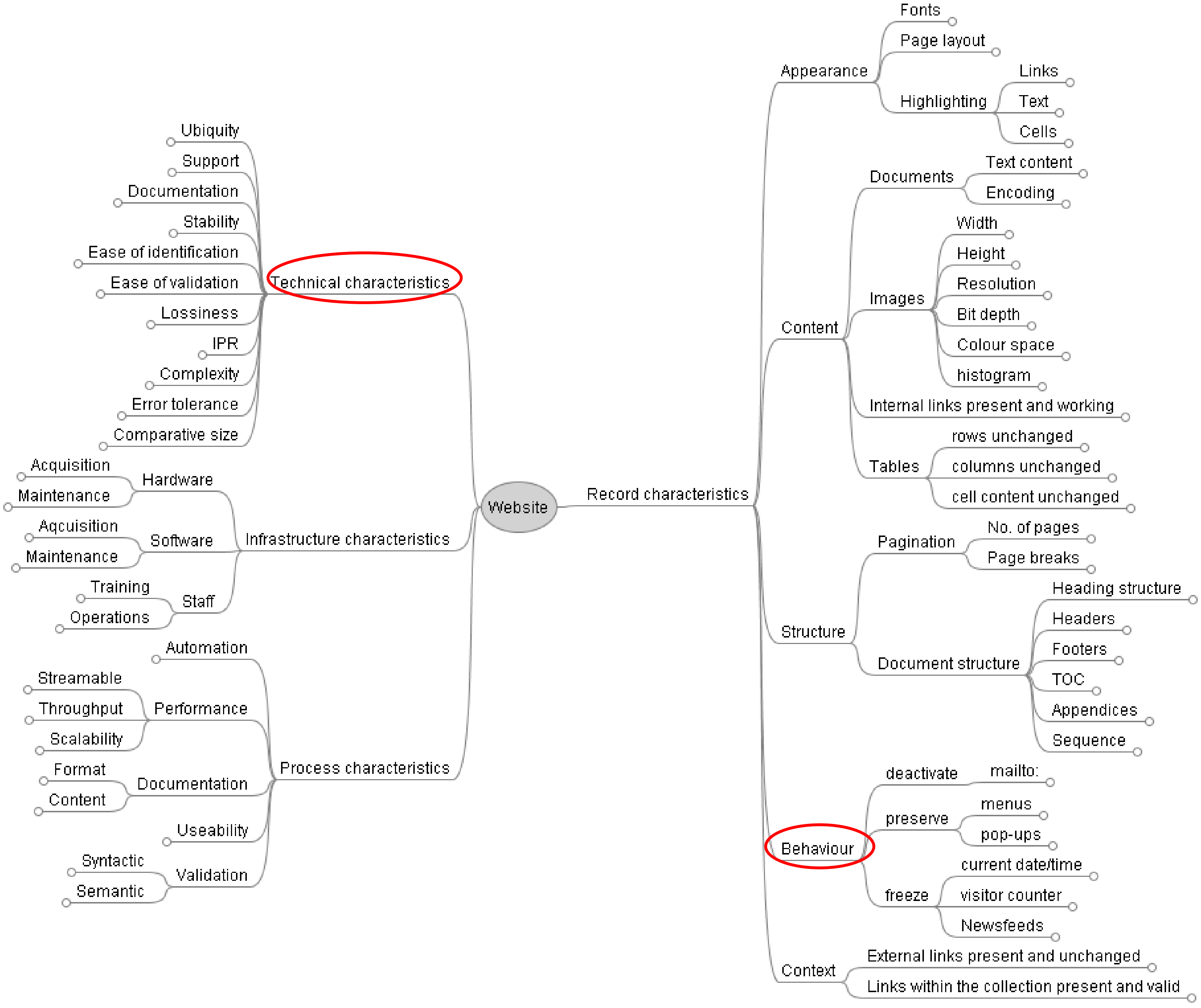
Assign Measurable Units

- ❑ 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

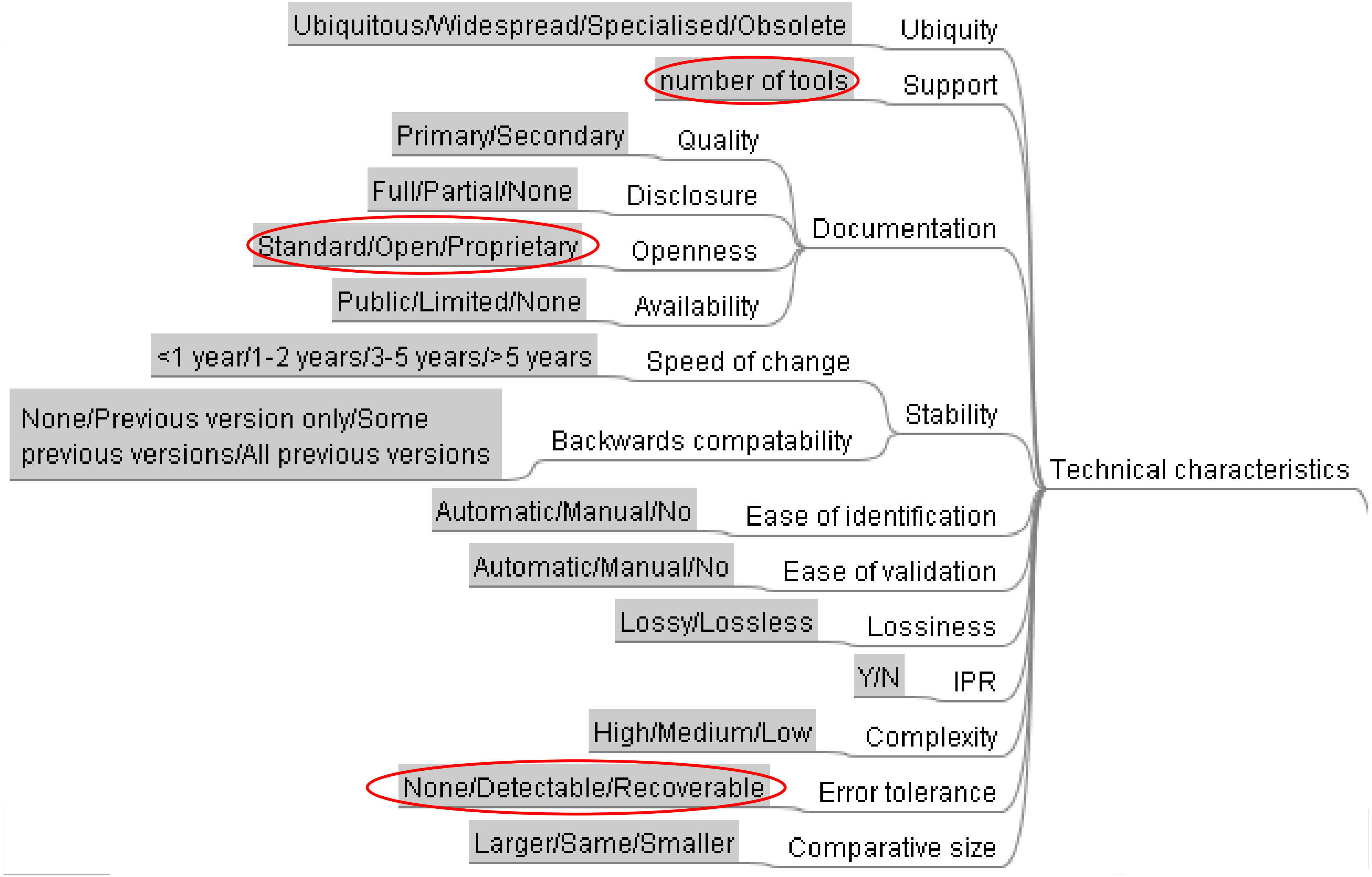
- Numeric
- Yes/No (Y/N)
- Yes/Acceptable/No (Y/A/N)
- Ordinal: define the possible values
- Subjective 0-to-5





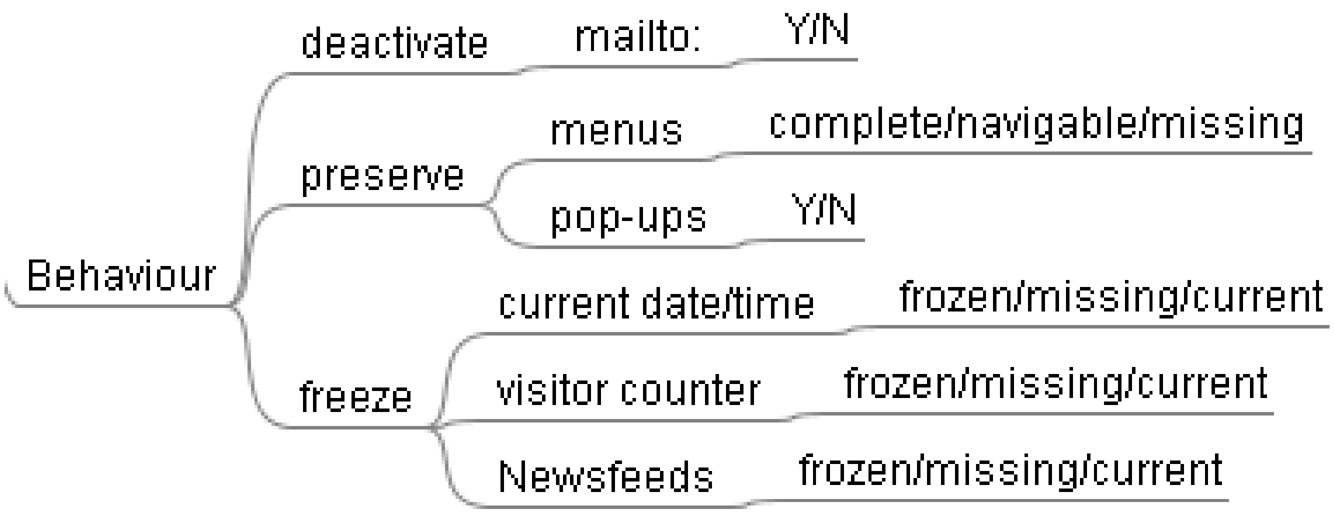
- Website
  - Technical characteristics
    - Ubiquity
    - Support
    - Documentation
    - Stability
    - Ease of identification
    - Ease of validation
    - Lossiness
    - IPR
    - Complexity
    - Error tolerance
    - Comparative size
  - Infrastructure characteristics
    - Hardware
      - Acquisition
      - Maintenance
    - Software
      - Aqcuisition
      - Maintenance
    - Staff
      - Training
      - Operations
  - Process characteristics
    - Automation
    - Performance
      - Streamable
      - Throughput
      - Scalability
    - Documentation
      - Format
      - Content
    - Useability
    - Validation
      - Syntactic
      - Semantic
  - Record characteristics
    - Appearance
      - Fonts
      - Page layout
      - Highlighting
        - Links
        - Text
        - Cells
    - Content
      - Documents
        - Text content
        - Encoding
      - Images
        - Width
        - Height
        - Resolution
        - Bit depth
        - Colour space
        - histogram
      - Internal links present and working
      - Tables
        - rows unchanged
        - columns unchanged
        - cell content unchanged
    - Structure
      - Pagination
        - No. of pages
        - Page breaks
      - Document structure
        - Heading structure
        - Headers
        - Footers
        - TOC
        - Appendices
        - Sequence
    - Behaviour
      - deactivate
        - mailto:
      - preserve
        - menus
        - pop-ups
      - freeze
        - current date/time
        - visitor counter
        - Newsfeeds
    - Context
      - External links present and unchanged
      - Links within the collection present and valid

File format characteristics



- Technical characteristics
  - Ubiquity
    - Ubiquitous/Widespread/Specialised/Obsolete
  - Support
    - number of tools
  - Documentation
    - Quality
      - Primary/Secondary
    - Disclosure
      - Full/Partial/None
    - Openness
      - Standard/Open/Proprietary
    - Availability
      - Public/Limited/None
  - Stability
    - Speed of change
      - <1 year/1-2 years/3-5 years/>5 years
    - Backwards compatability
      - None/Previous version only/Some previous versions/All previous versions
  - Ease of identification
    - Automatic/Manual/No
  - Ease of validation
    - Automatic/Manual/No
  - Lossiness
    - Lossy/Lossless
  - IPR
    - Y/N
  - Complexity
    - High/Medium/Low
  - Error tolerance
    - None/Detectable/Recoverable
  - Comparative size
    - Larger/Same/Smaller

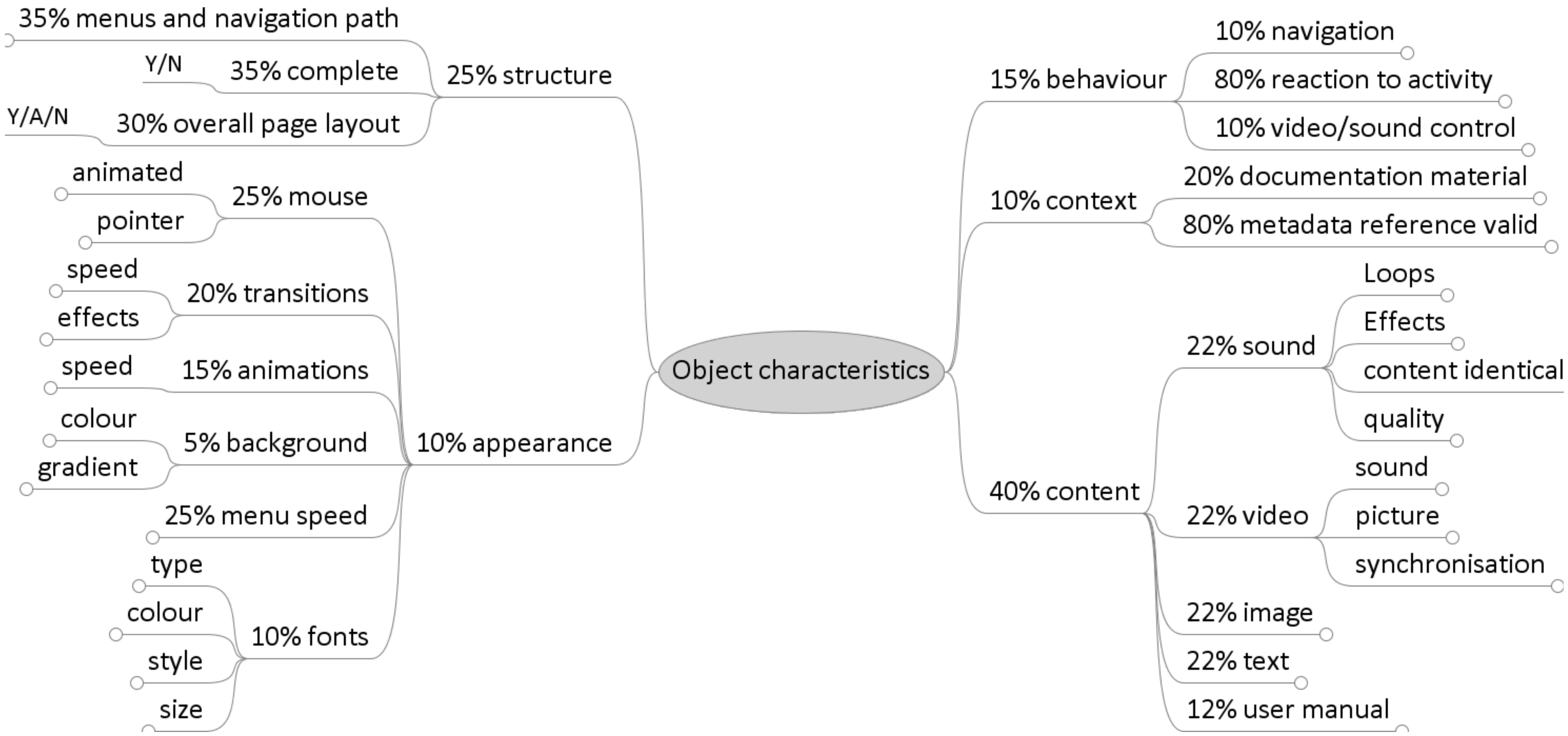
Behaviour



- Behaviour
  - deactivate
    - mailto:
      - Y/N
  - preserve
    - menus
      - complete/navigable/missing
    - pop-ups
      - Y/N
  - freeze
    - current date/time
      - frozen/missing/current
    - visitor counter
      - frozen/missing/current
    - Newsfeeds
      - frozen/missing/current

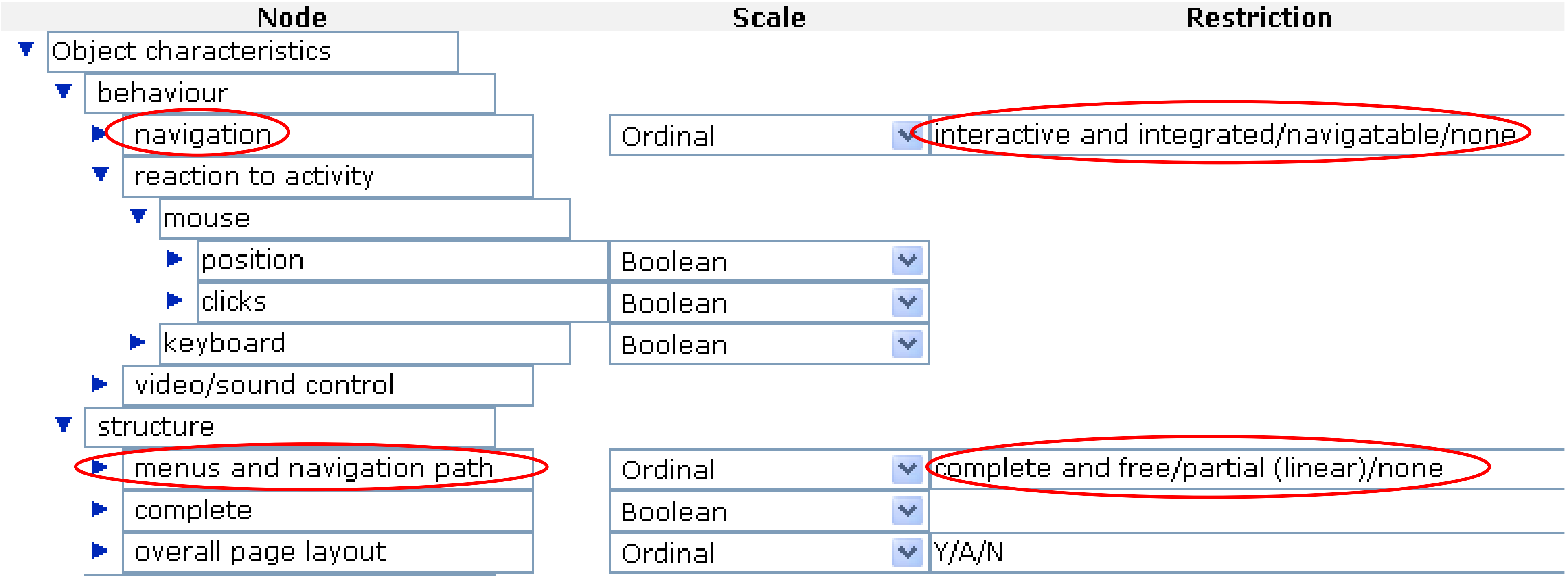
- Visitor counter and similar things can be
  - Frozen at the point of harvesting
  - Left out
  - Still counting while being accessed in the archive (Is this desirable?)

Interactive multimedia



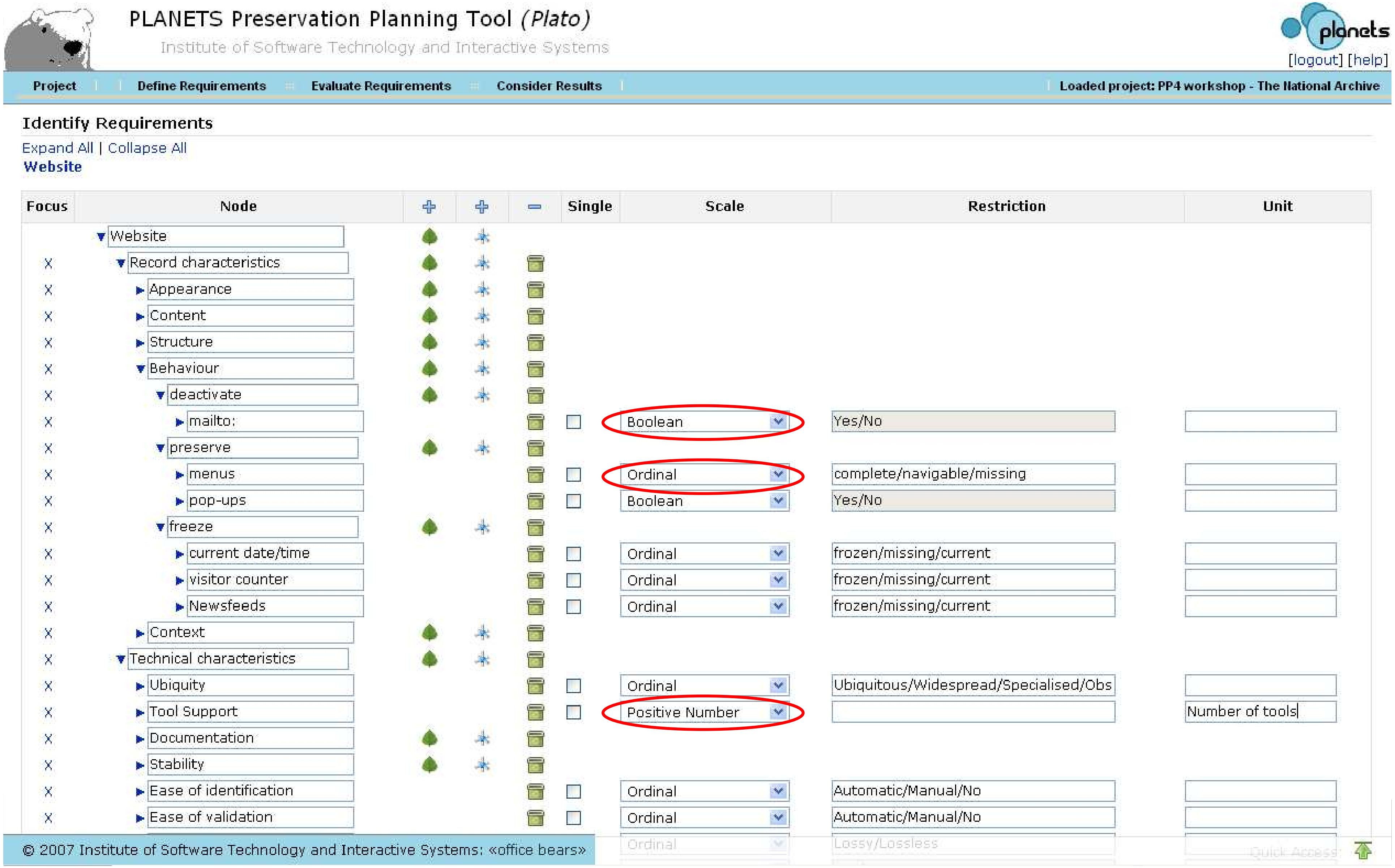
Behaviour

- 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

Objective Tree



PLANETS Preservation Planning Tool *(Plato)*

Institute of Software Technology and Interactive Systems

planets
[logout] [help]
Project | Define Requirements | Evaluate Requirements | Consider Results
Loaded project: PP4 workshop - The National Archive

Identify Requirements

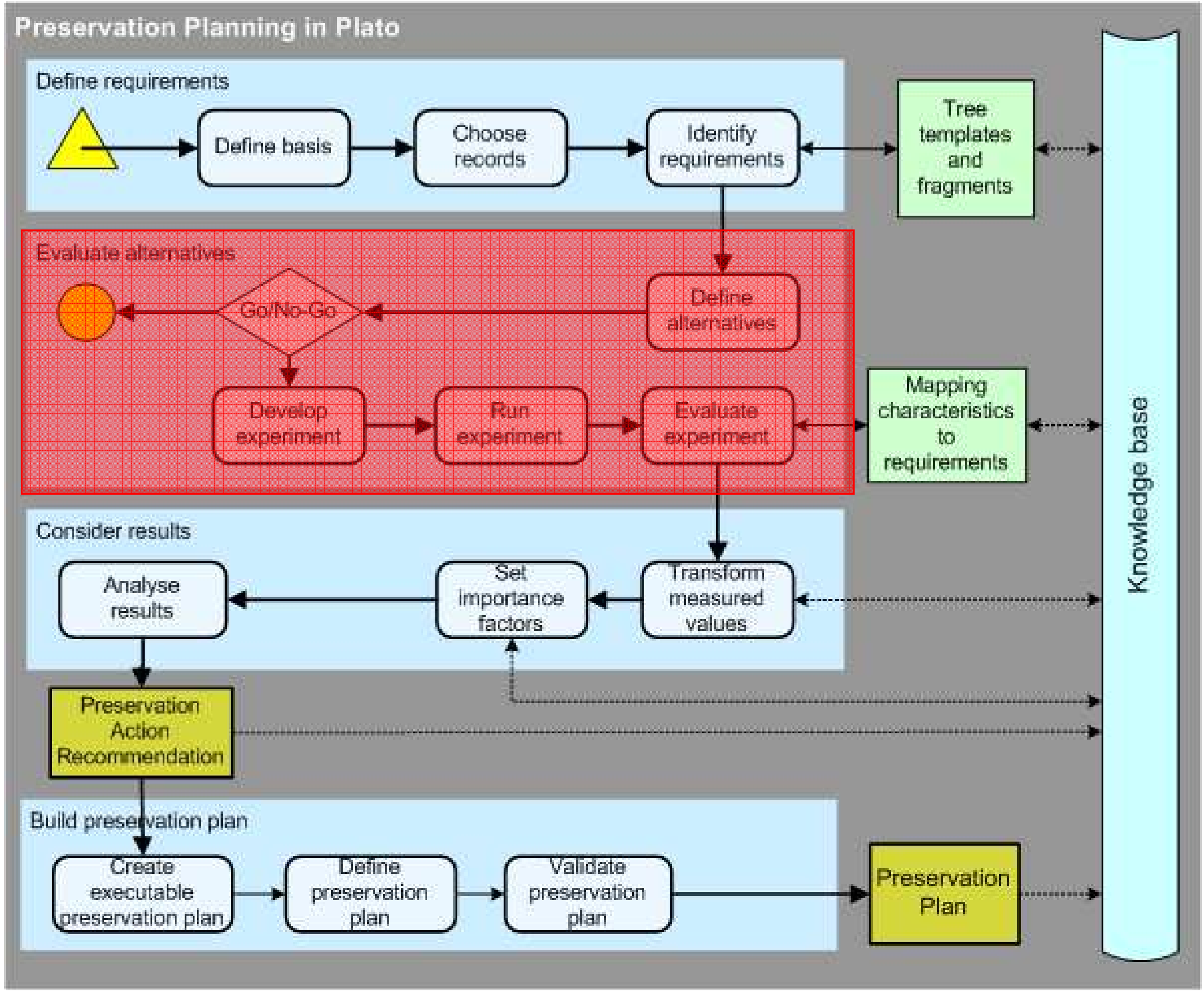
Expand All | Collapse All

Website

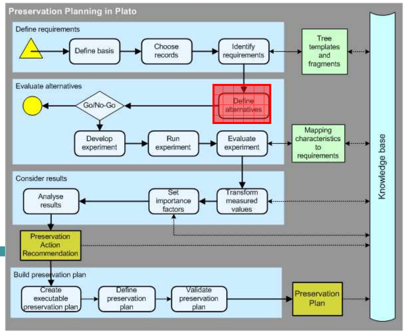
Focus	Node	+	+	−	Single	Scale	Restriction	Unit
	▼ Website	●	✳					
X	▼ Record characteristics	●	✳	🗑				
X	▶ Appearance	●	✳	🗑				
X	▶ Content	●	✳	🗑				
X	▶ Structure	●	✳	🗑				
X	▼ Behaviour	●	✳	🗑				
X	▼ deactivate	●	✳	🗑				
X	▶ mailto:			🗑	□	Boolean	Yes/No	
X	▼ preserve	●	✳	🗑				
X	▶ menus			🗑	□	Ordinal	complete/navigable/missing	
X	▶ pop-ups			🗑	□	Boolean	Yes/No	
X	▼ freeze	●	✳	🗑				
X	▶ current date/time			🗑	□	Ordinal	frozen/missing/current	
X	▶ visitor counter			🗑	□	Ordinal	frozen/missing/current	
X	▶ Newsfeeds			🗑	□	Ordinal	frozen/missing/current	
X	▶ Context	●	✳	🗑				
X	▼ Technical characteristics	●	✳	🗑				
X	▶ Ubiquity			🗑	□	Ordinal	Ubiquitous/Widespread/Specialised/Obs	
X	▶ Tool Support			🗑	□	Positive Number		Number of tools
X	▶ Documentation	●	✳	🗑				
X	▶ Stability	●	✳	🗑				
X	▶ Ease of identification			🗑	□	Ordinal	Automatic/Manual/No	
X	▶ Ease of validation			🗑	□	Ordinal	Automatic/Manual/No	
						Ordinal	Lossy/Lossless	

© 2007 Institute of Software Technology and Interactive Systems: «office bears»

PP Workflow

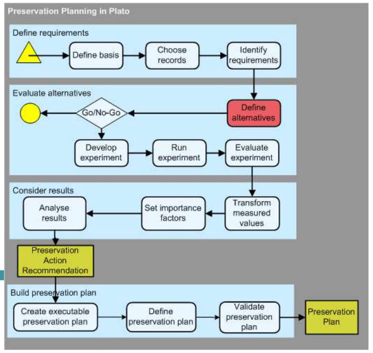

Preservation Planning in Plato
Define requirements
Define basis
Choose records
Identify requirements
Tree templates and fragments
Evaluate alternatives
Define alternatives
Go/No-Go
Develop experiment
Run experiment
Evaluate experiment
Mapping characteristics to requirements
Consider results
Transform measured values
Set importance factors
Analyse results
Preservation Action Recommendation
Build preservation plan
Create executable preservation plan
Define preservation plan
Validate preservation plan
Preservation Plan
Knowledge base

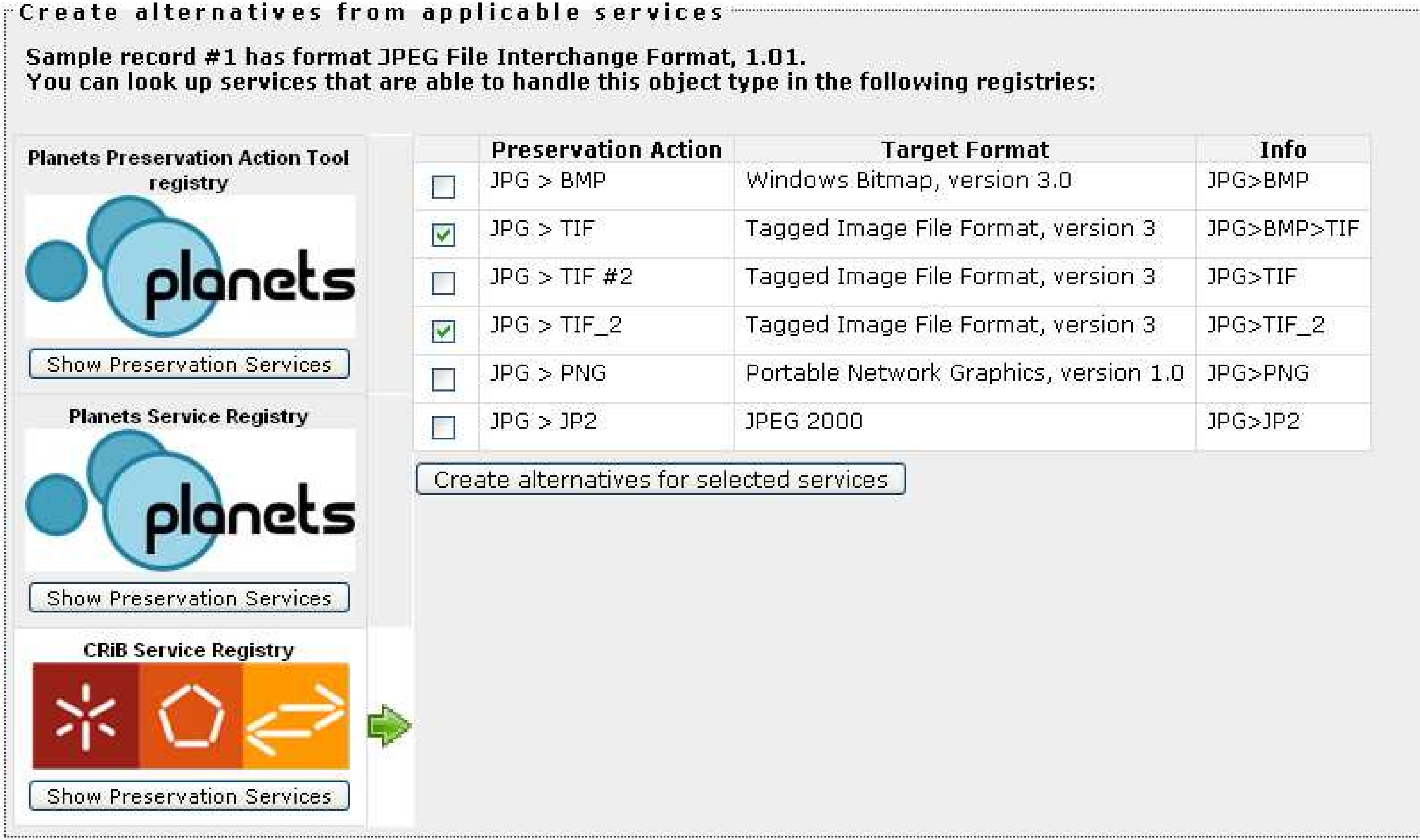
Define alternatives



- 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

Discovering possible actions





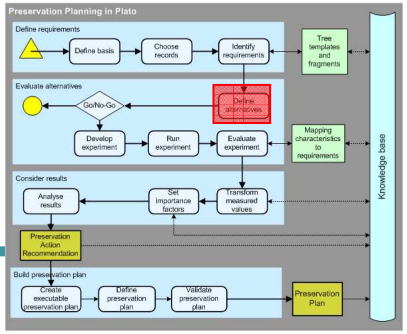
Create alternatives from applicable services

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
☐	JPG > BMP	Windows Bitmap, version 3.0	JPG>BMP
☑	JPG > TIF	Tagged Image File Format, version 3	JPG>BMP>TIF
☐	JPG > TIF #2	Tagged Image File Format, version 3	JPG>TIF
☑	JPG > TIF_2	Tagged Image File Format, version 3	JPG>TIF_2
☐	JPG > PNG	Portable Network Graphics, version 1.0	JPG>PNG
☐	JPG > JP2	JPEG 2000	JPG>JP2

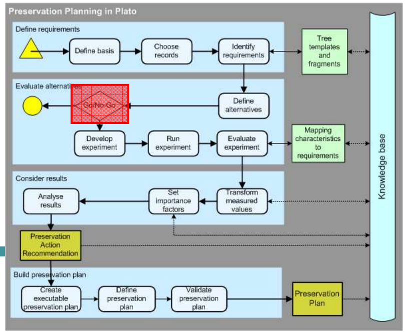
Specify resources



❑ Detailed design and overview of the resources for each alternative

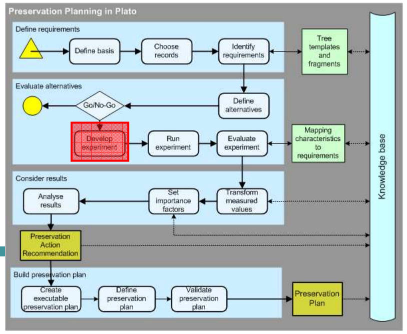
- human resources (qualification, roles, responsibility, …)
- technical requirements (hardware and software components)
- time (time to set-up, run experiment,...)
- cost (costs of the experiments,...)

Go/No-Go



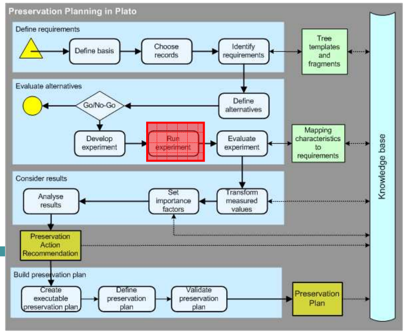
- Deliberate step for taking a decision whether it will be useful and cost-effective to continue the procedure, given
  - The resources to be spent (people, money)
  - The availability of tools and solutions,
  - The expected result(s).
- Review of the experiment/ evaluation process design so far
  -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

Develop experiment



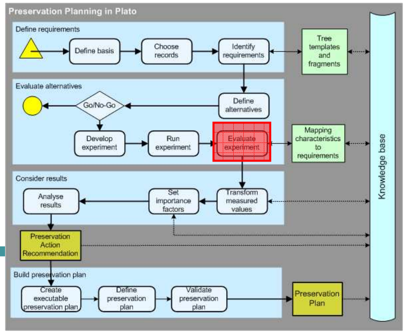
- Formulate for each evaluation or experiment or preservation process detailed
  - Development plan
    - steps to build and test software components
    - procedures and preparation
    - parameter settings for integrating preservation services
  - Test plan (mechanisms how to)
  - Evaluation/experiment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

Run experiment



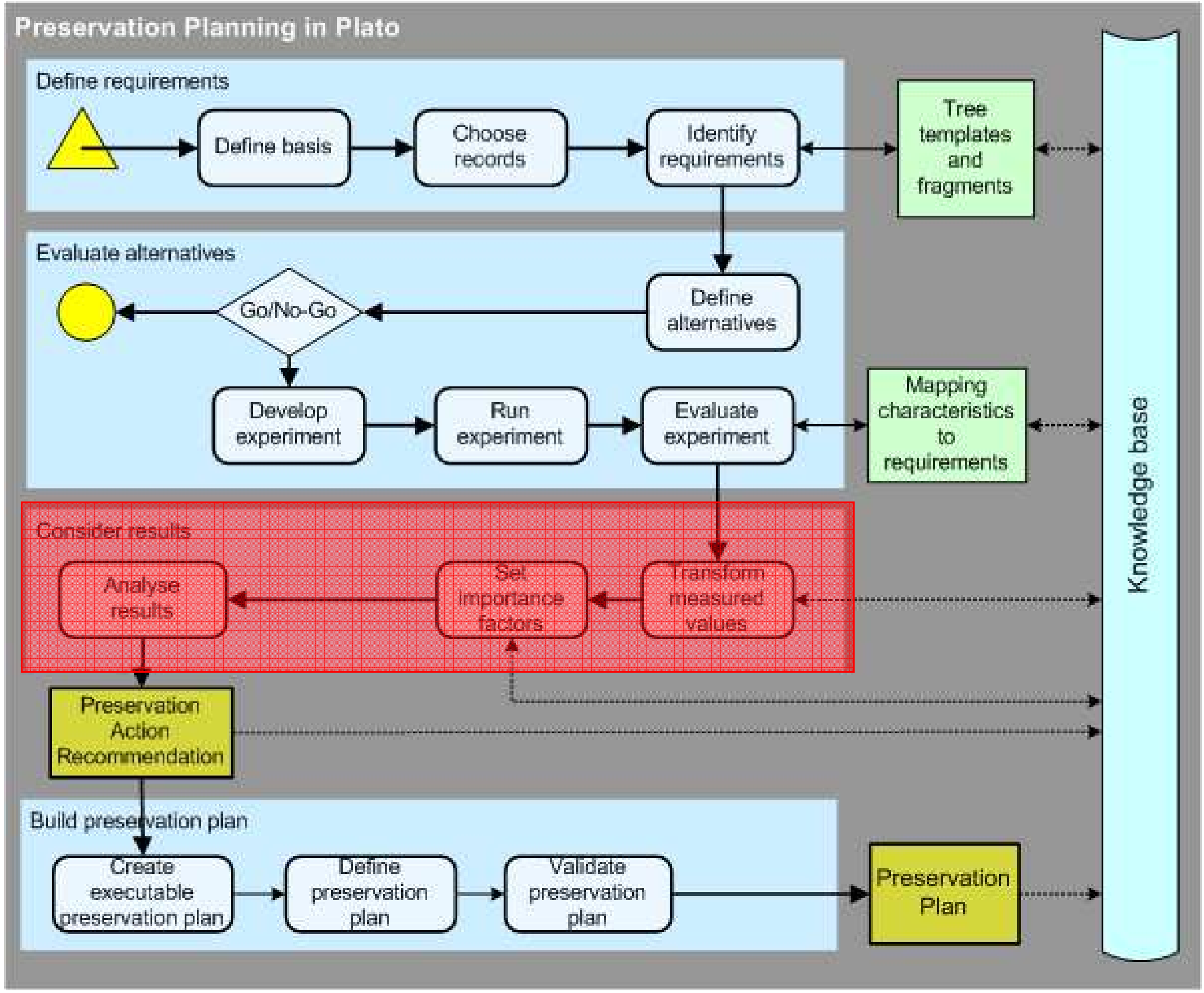
- Before conducting an evaluation or running an experiment, the experiment process as designed has to be tested
  - It may lead to re-design or even termination of the evaluation/ experiment process
- The results will be evaluated in the next stage
- The whole process needs to be documented

Evaluate experiment

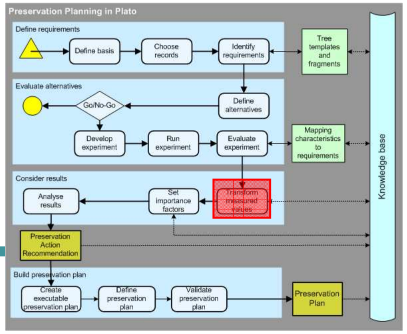


- Evaluate the outcome of each alternative for each leaf of the objective tree
- The evaluation will identify
  - Need for repeating the process
  - Unexpected (or undesired) results
- Includes both technical and intellectual aspects
- Evaluation may include comparing the results of more than one experiment/evaluation.

PP Workflow

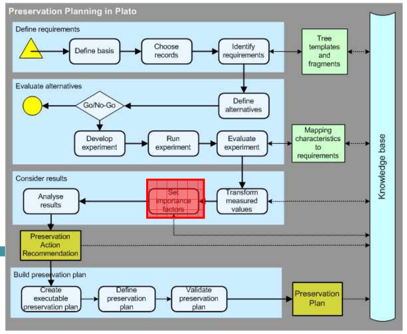

Preservation Planning in Plato
Define requirements
Define basis
Choose records
Identify requirements
Tree templates and fragments
Evaluate alternatives
Define alternatives
Go/No-Go
Develop experiment
Run experiment
Evaluate experiment
Mapping characteristics to requirements
Consider results
Analyse results
Set importance factors
Transform measured values
Preservation Action Recommendation
Build preservation plan
Create executable preservation plan
Define preservation plan
Validate preservation plan
Preservation Plan
Knowledge base

Transform measured values



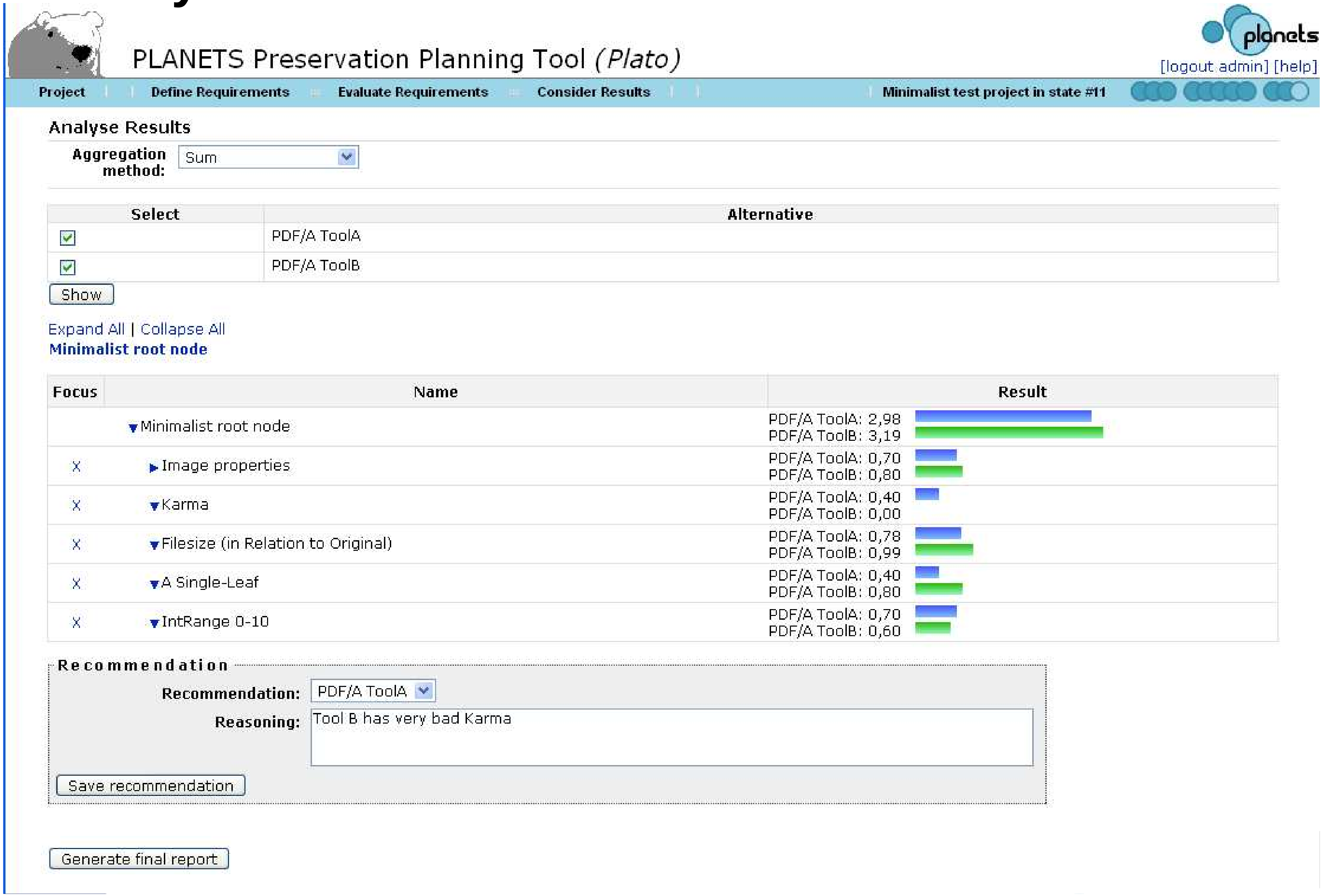
- ❑ Measures come in seconds, euro, bits, goodness values,…
- ❑ Need to make them comparable
- ❑ Transform measured values to uniform scale
- ❑ Transformation tables for each leaf criterion
- ❑ Linear transformation, logarithmic, special scale
- ❑ Scale 1-5 plus "not-acceptable"

Set importance factors



- Definition which criteria are more important
- Depends on individual preferences and requirements
- Adaptation for each implementation
- High influence on the final ranking
- Aggregation of weights

Analyse results



PLANETS Preservation Planning Tool *(Plato)*
planets
[logout admin] [help]
Project | Define Requirements | Evaluate Requirements | Consider Results
Minimalist test project in state #11

Analyse Results

Aggregation method: Sum

Select	Alternative
☑	PDF/A ToolA
☑	PDF/A ToolB

Show
Expand All | Collapse All

Minimalist root node

Focus	Name	Result
	▼Minimalist root node	PDF/A ToolA: 2,98 PDF/A ToolB: 3,19
X	▶Image properties	PDF/A ToolA: 0,70 PDF/A ToolB: 0,80
X	▼Karma	PDF/A ToolA: 0,40 PDF/A ToolB: 0,00
X	▼Filesize (in Relation to Original)	PDF/A ToolA: 0,78 PDF/A ToolB: 0,99
X	▼A Single-Leaf	PDF/A ToolA: 0,40 PDF/A ToolB: 0,80
X	▼IntRange 0-10	PDF/A ToolA: 0,70 PDF/A ToolB: 0,60

Recommendation

Recommendation: PDF/A ToolA

Reasoning: Tool B has very bad Karma

Questions?

becker@ifs.tuwien.ac.at

www.ifs.tuwien.ac.at/dp/plato
www.planets-project.eu

Outline

- ❑ Preservation Planning
  - ▪ Evaluation of potential actions
- ❑ The Planets Preservation Planning Workflow
  - ▪ Underlying methodology
  - ▪ Workflow walkthrough
  - ▪ The planning tool Plato
- ❑ Requirements definition exercise
  - ▪ Group assignment
  - ▪ Schedule