



# Digital Preservation: How to Plan

Preservation Planning with Plato

Christoph Becker

Vienna University of Technology

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



Sofia, September 2009



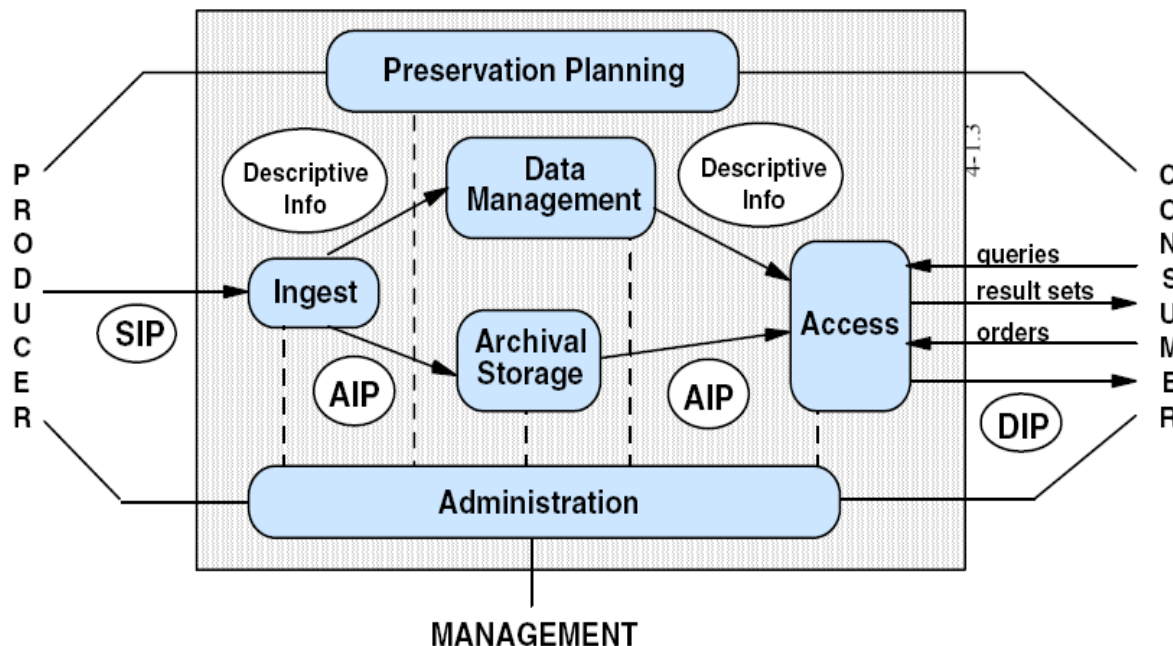
# Outline

- Why preservation planning?
  - Trusted digital repositories
  - Policies vs. plans
- Preservation Planning
  - What is a preservation plan?
  - How to create a preservation plan
  - The Planets Preservation Planning Workflow
  - Requirements definition
  - The planning tool Plato
- Part 2: Requirements discussion



# Trustworthiness in digital repositories

- Consumers need trust in digital repositories
- Producers need trust in digital repositories
- Repositories need trust in external providers
- Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)



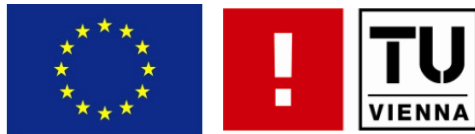
# TRAC and Preservation Planning: Example

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

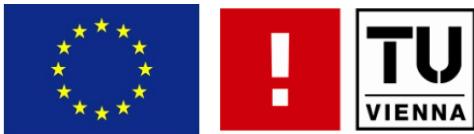
A3.6 Repository has a documented history of the changes to its operations, procedures, software, and hardware that, where appropriate, is linked to relevant preservation strategies and describes potential effects on preserving digital content.

- Preservation plans need traceability

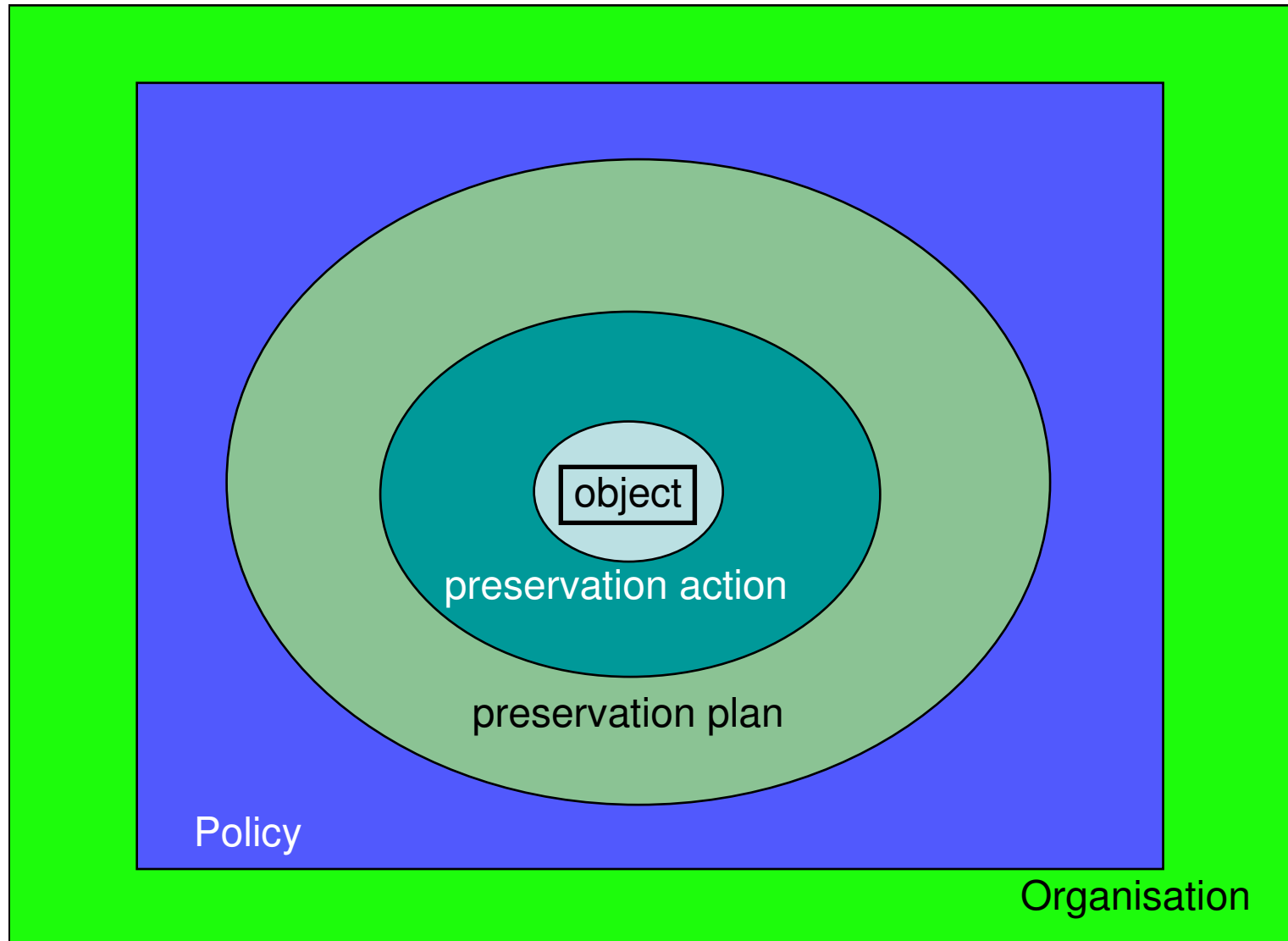


# Definition of a Preservation Plan

- ‘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 goal. It also describes the preservation context, the evaluated alternative preservation strategies and the resulting decision for one strategy, including the rationale of the decision.



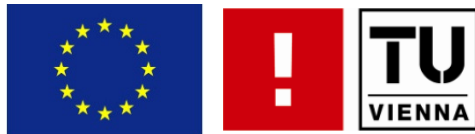
# Objects in context

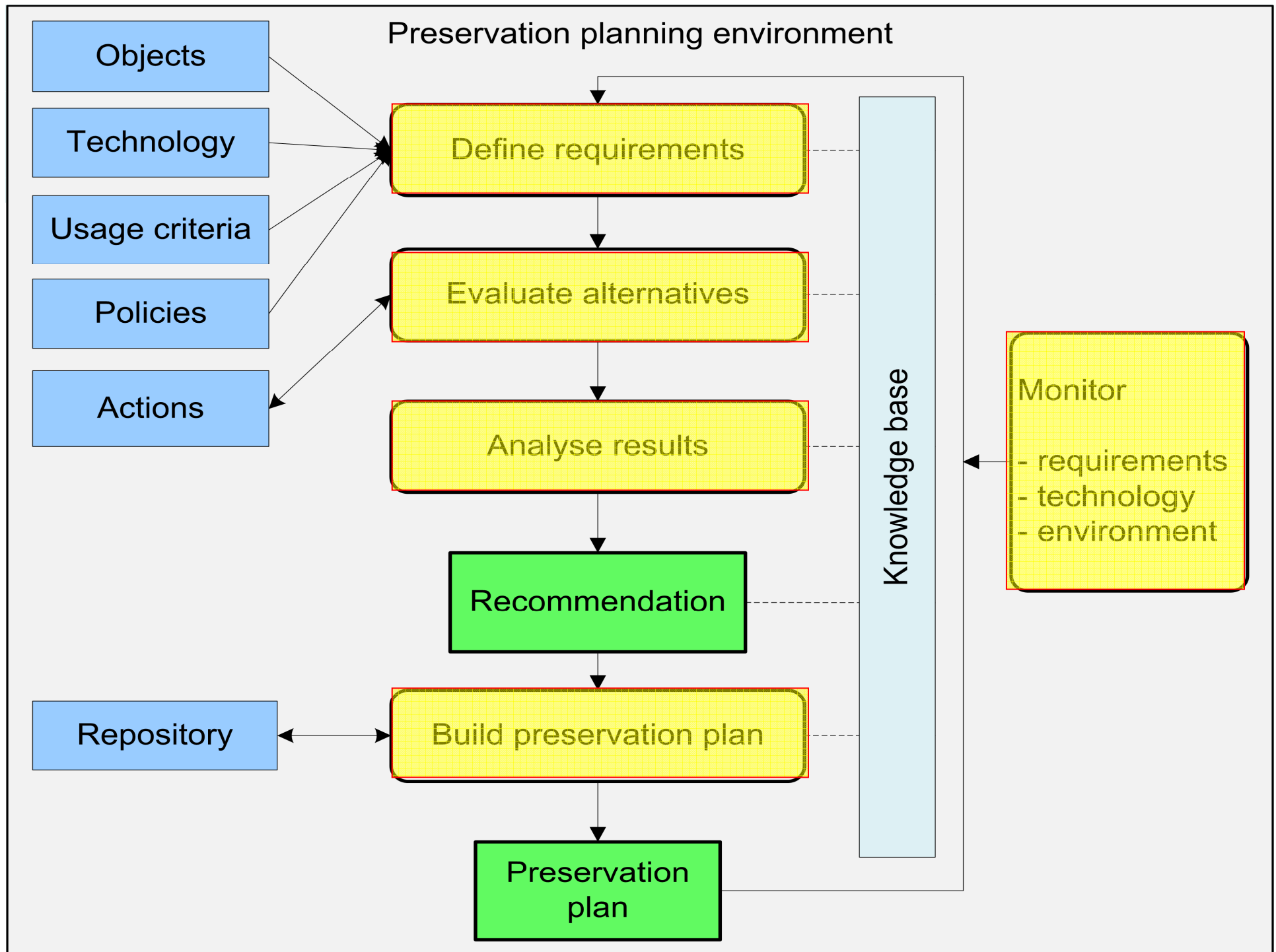


Ownership  
Awareness  
Responsibility

# 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

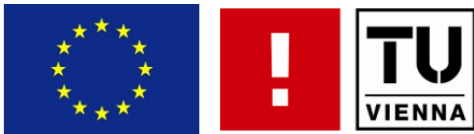


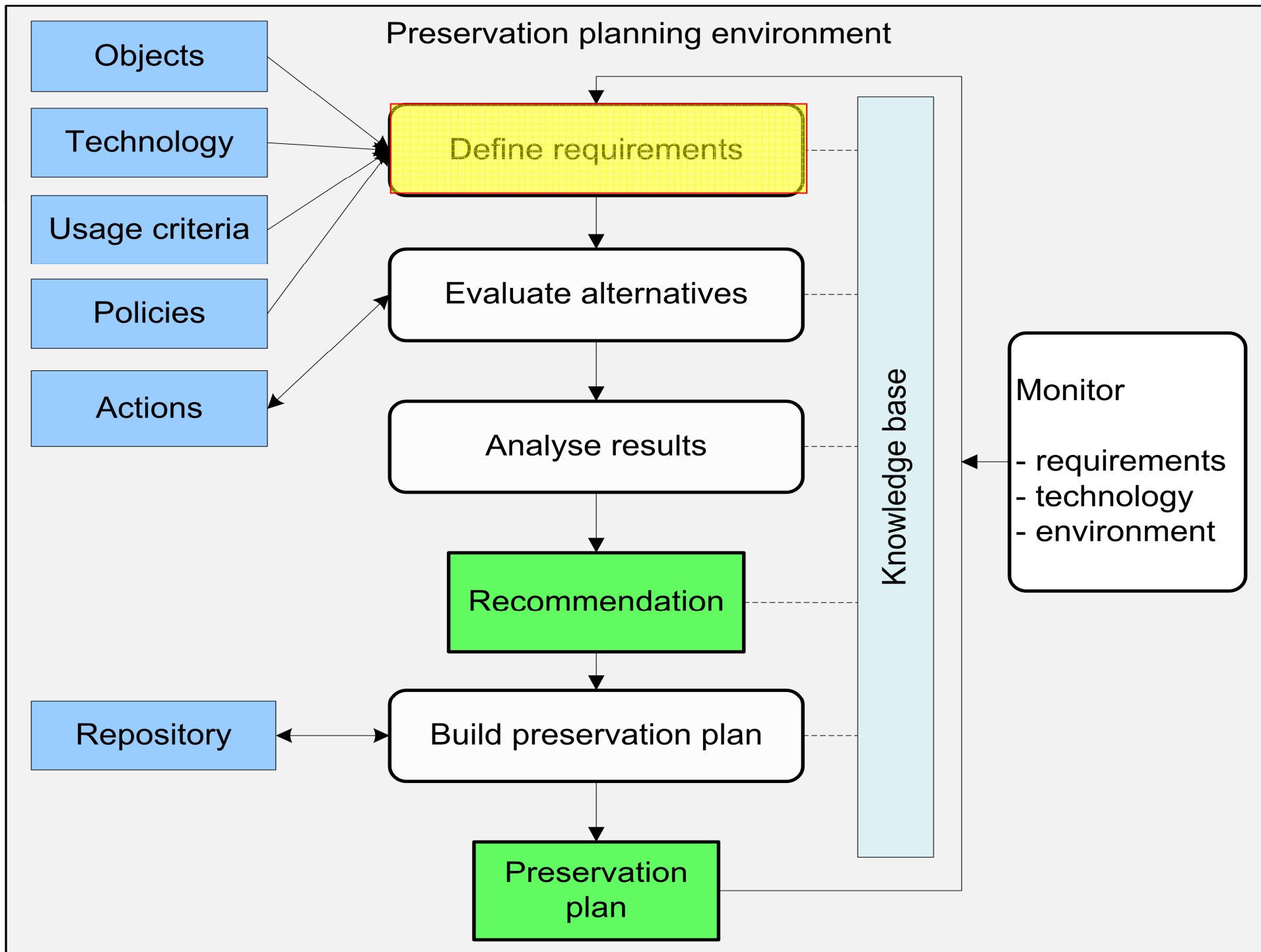




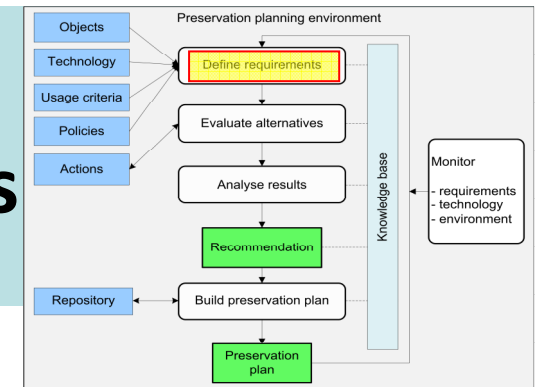
# Preservation Planning in Plato

- Web based planning tool implementing the Planets preservation planning workflow
- Publicly available
- Automation of the planning process
  - Integration of registries and services for
    - File format identification
    - Preservation action (migration, emulation...)
    - Characterisation and comparison
- Knowledge base to support planning
- Upcoming new release!
- <http://www.ifs.tuwien.ac.at/dp/plato>



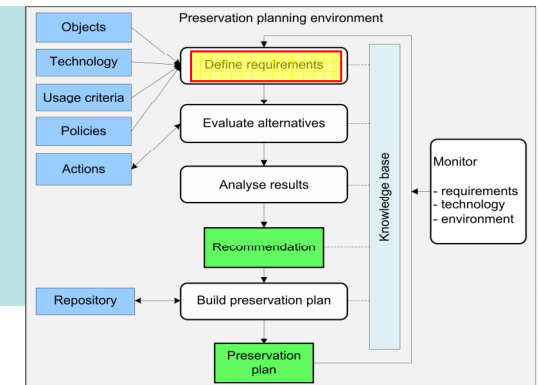


# Define basis and samples



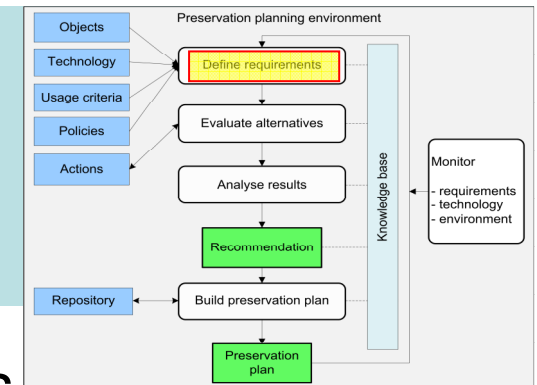
- Document basic assumptions and constraints
  - Mandate, objects, and designated community
  - Purpose of planning
  - Applying policies and constraints
  - Reasons for starting the planning process
- Collection
  - Size, type of objects, original environment, usage
  - Sample objects

# Choose sample objects/records



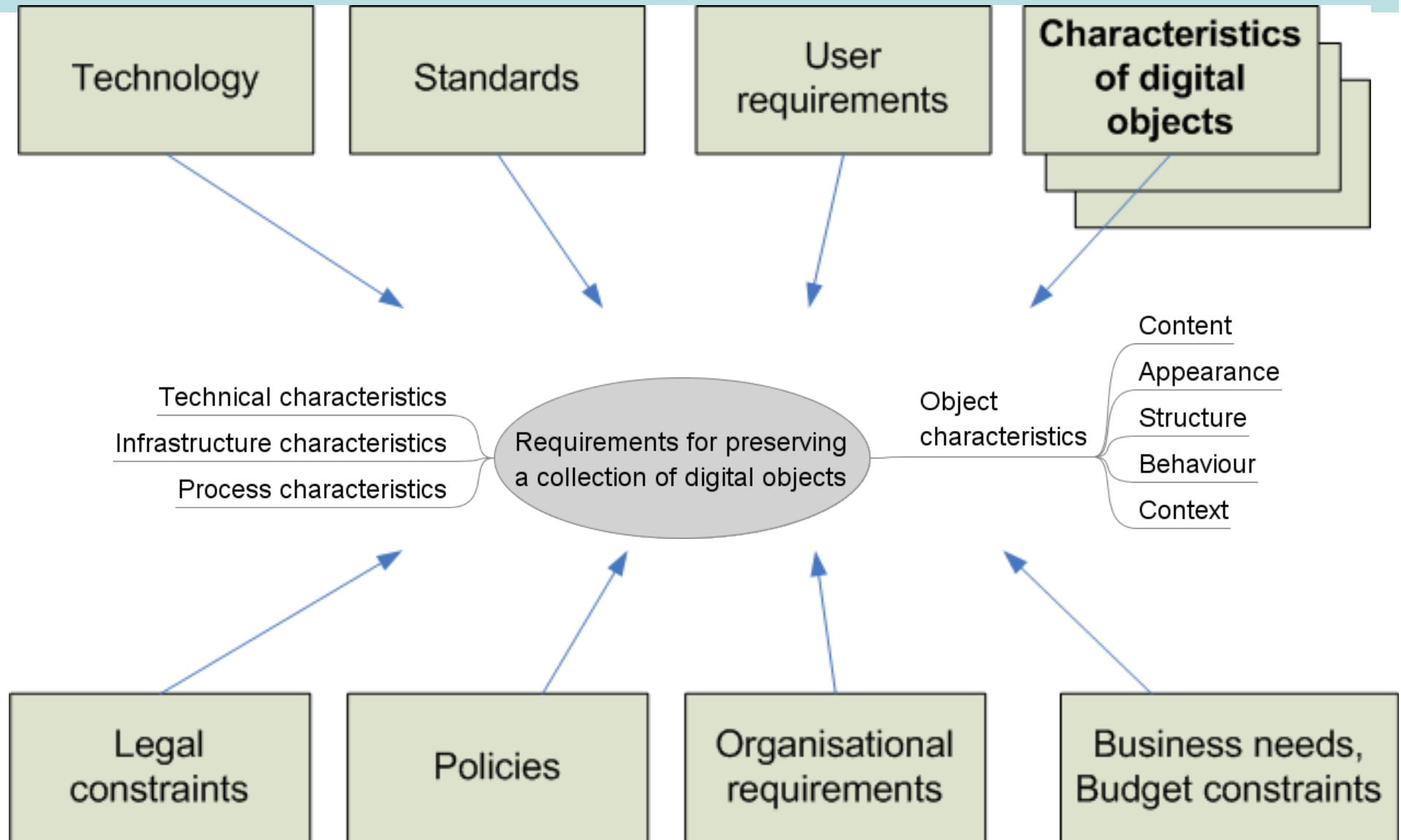
- Define the set of objects that are the subject of preservation planning
  - Size of the collection
  - Growth rate
  - Object format
  - ...
- Specify representative sample objects that cover the variety of significant properties and technical characteristics

# Identify requirements



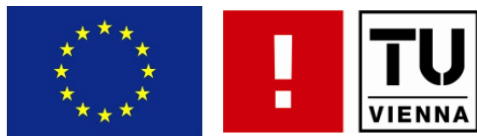
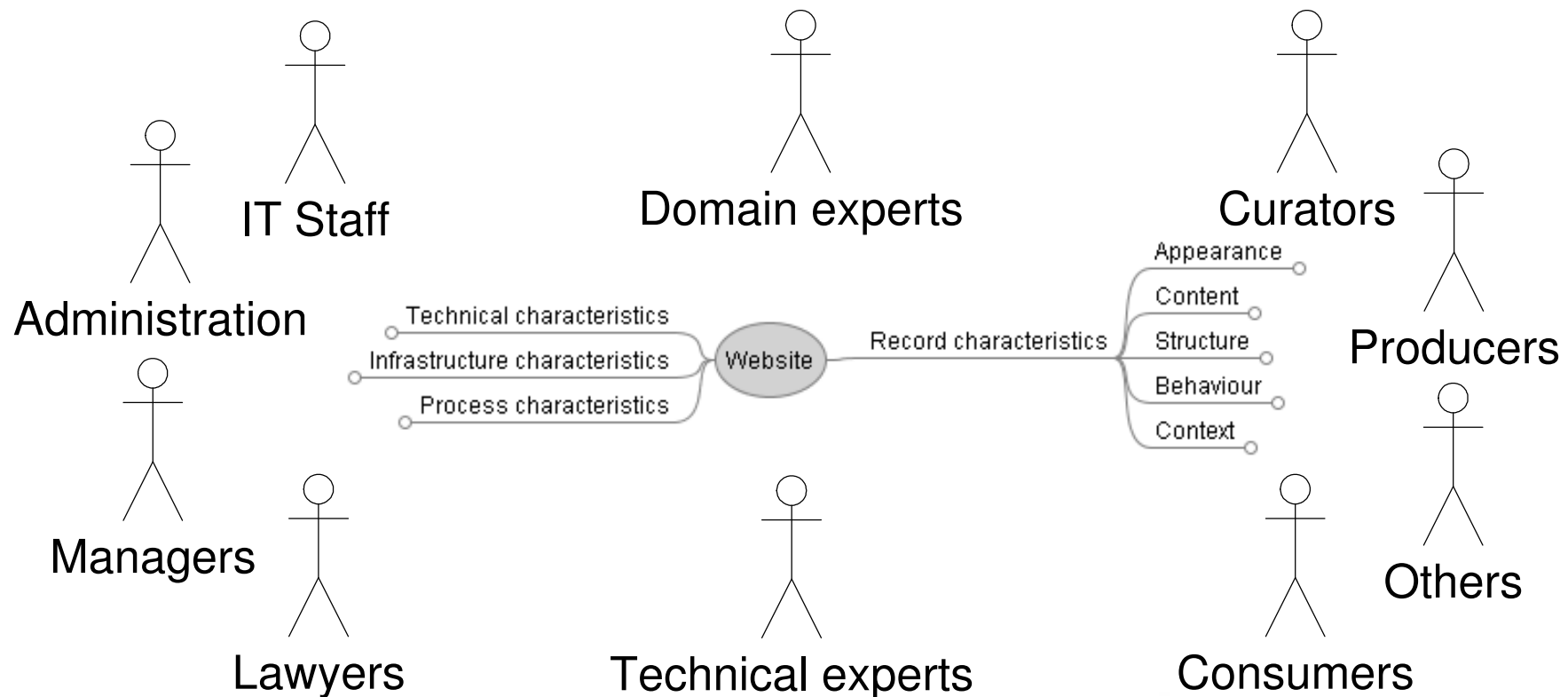
- Define all relevant goals and characteristics (high-level, detail) for the situation
- Usually four major groups:
  - object characteristics (content, metadata ...)
  - record characteristics (context, relations, ...)
  - process characteristics (scalability, error detection, ...)
  - costs (set-up,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

# Influence Factors

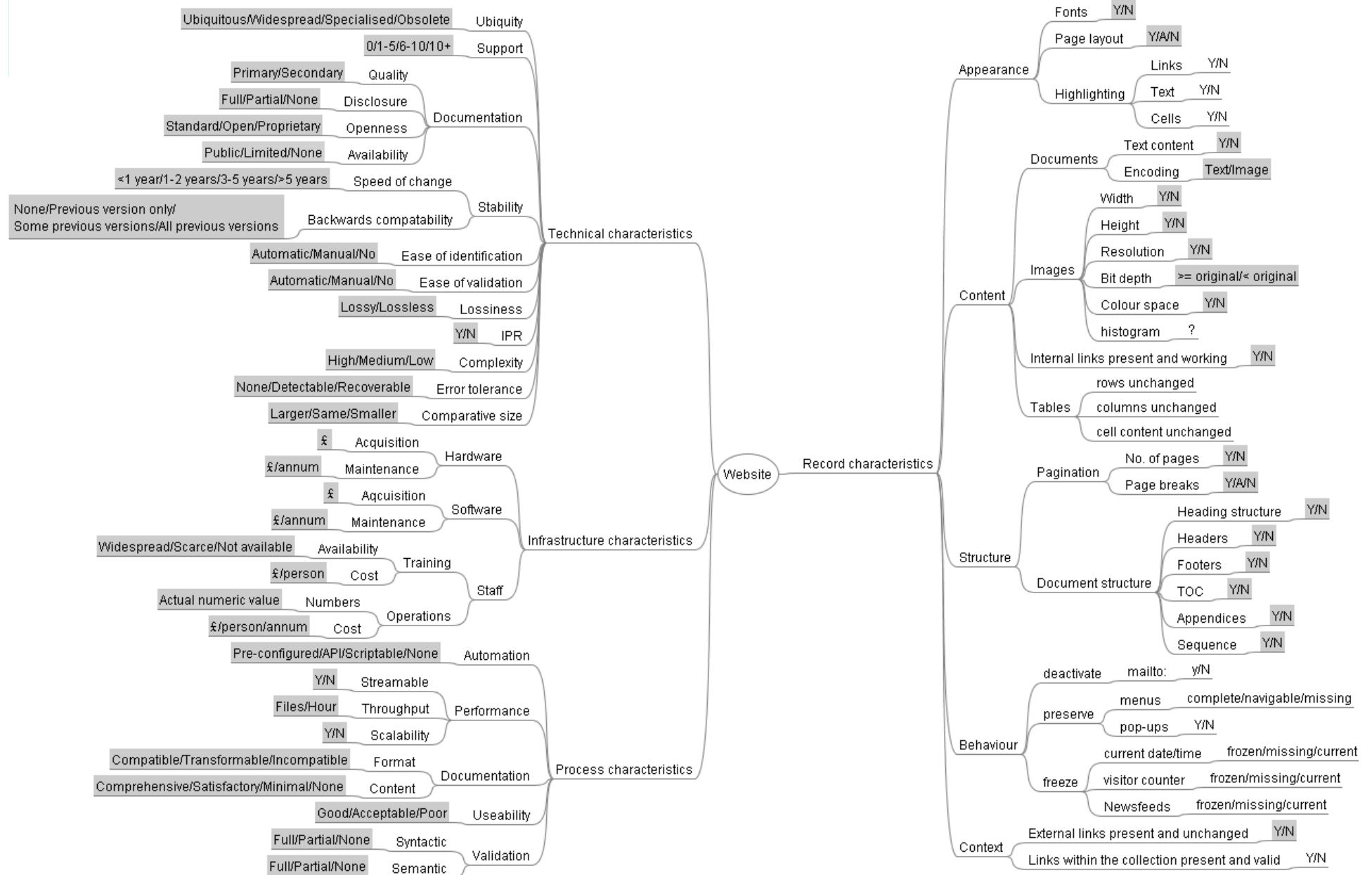


# Stakeholders

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

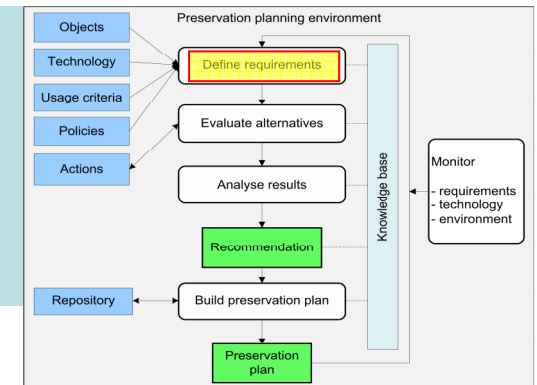


# An Objective Tree





# Types of requirements

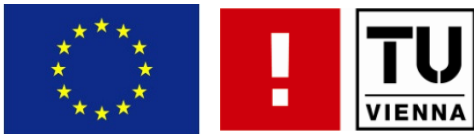


- Requirements on the outcome of actions
  - Access
  - Risks incurred
  - Format should be open, documented...
  - The objects should be
    - Authentic
    - Reliable
    - ...
- Requirements on the action
  - Fast
  - Reliable
  - Well supported
  - ...



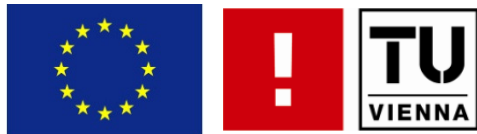
# User perspective

- 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 for 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)
  - ...



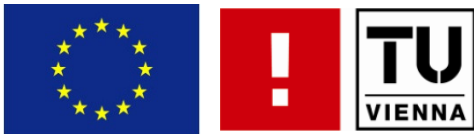
# Requirements for objects

- Authenticity
- Reliability
- Integrity
- Usability
- Accuracy



# Essential characteristics of 'digital objects'

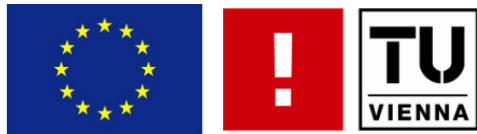
- What needs to be preserved?
  - Content
  - Context
  - Structure
  - Appearance
  - Behaviour



# 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



# Objective Tree



PLANETS Preservation Planning Tool (*Plato*)  
Institute of Software Technology and Interactive Systems

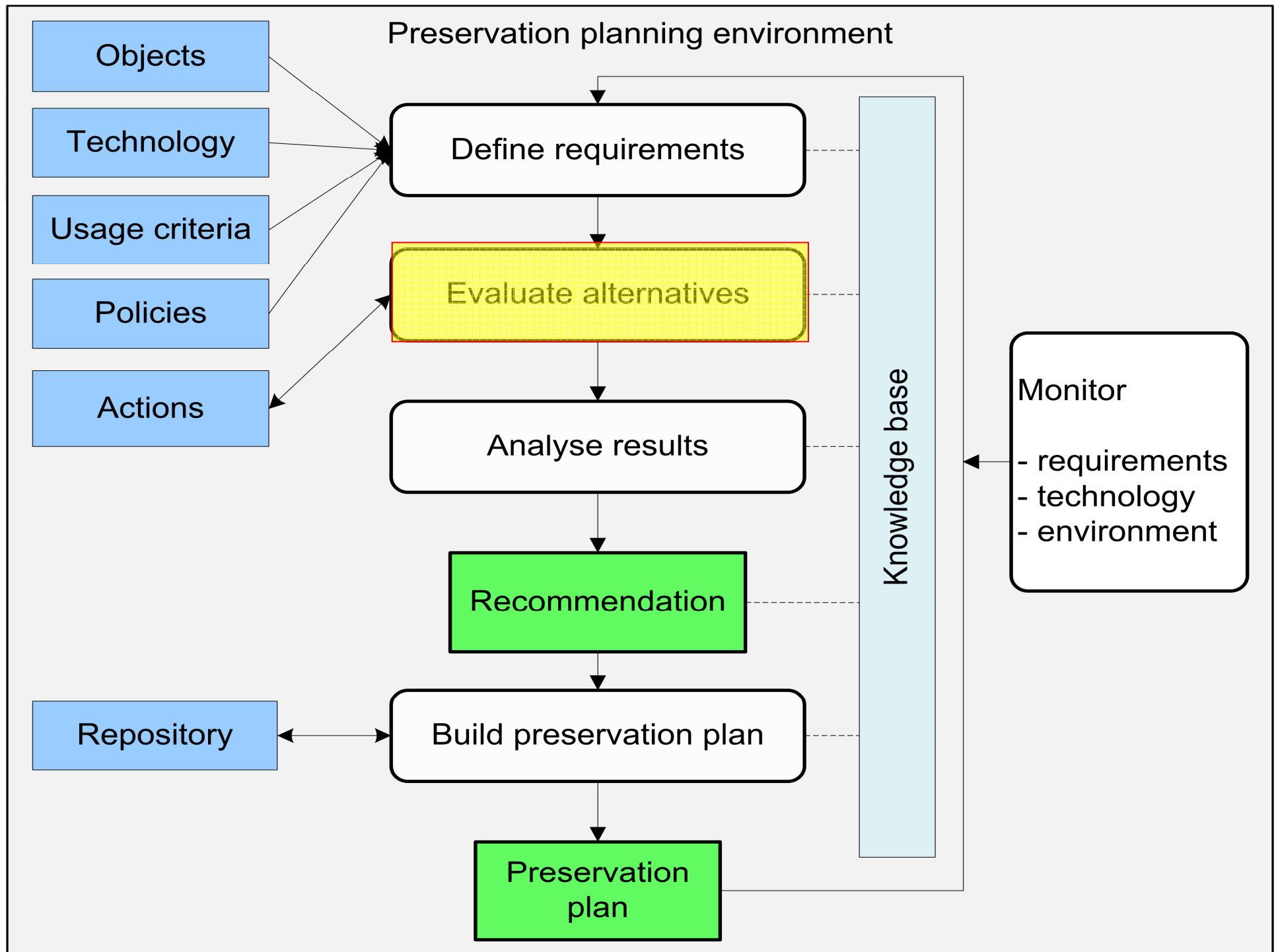


Project | Define Requirements | Evaluate Requirements | Consider Results | Loaded project: PP4 workshop - The National Archive

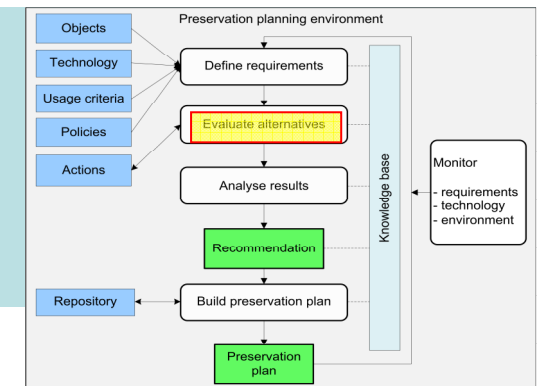
## Identify Requirements

[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:			📁	<input type="checkbox"/>	Boolean	Yes/No	
X	▼ preserve	🌳	✳️	📁				
X	▶ menus			📁	<input type="checkbox"/>	Ordinal	complete/navigable/missing	
X	▶ pop-ups			📁	<input type="checkbox"/>	Boolean	Yes/No	
X	▼ freeze	🌳	✳️	📁				
X	▶ current date/time			📁	<input type="checkbox"/>	Ordinal	frozen/missing/current	
X	▶ visitor counter			📁	<input type="checkbox"/>	Ordinal	frozen/missing/current	
X	▶ Newsfeeds			📁	<input type="checkbox"/>	Ordinal	frozen/missing/current	
X	▶ Context	🌳	✳️	📁				
X	▼ Technical characteristics	🌳	✳️	📁				
X	▶ Ubiquity			📁	<input type="checkbox"/>	Ordinal	Ubiquitous/Widespread/Specialised/Obs	
X	▶ Tool Support			📁	<input type="checkbox"/>	Positive Number		Number of tools
X	▶ Documentation	🌳	✳️	📁				
X	▶ Stability	🌳	✳️	📁				
X	▶ Ease of identification			📁	<input type="checkbox"/>	Ordinal	Automatic/Manual/No	
X	▶ Ease of validation			📁	<input type="checkbox"/>	Ordinal	Automatic/Manual/No	
				📁		Ordinal	Lossy/Lossless	



# Evaluate alternatives

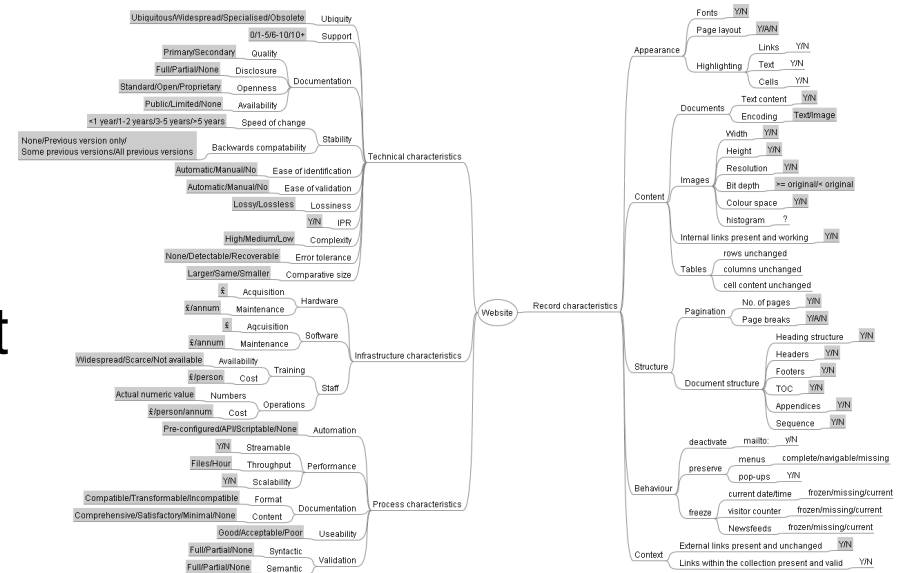


## ➤ List applicable actions

- Migration
- Emulation
- Both, other...

## ➤ Develop and run an experiment

- Apply each action to each sample
- Measure effects
- Evaluate outcome






# 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**

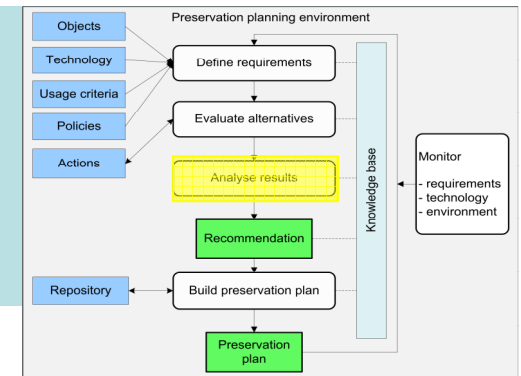


Show Preservation Services

	Preservation Action	Target Format	Info
<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

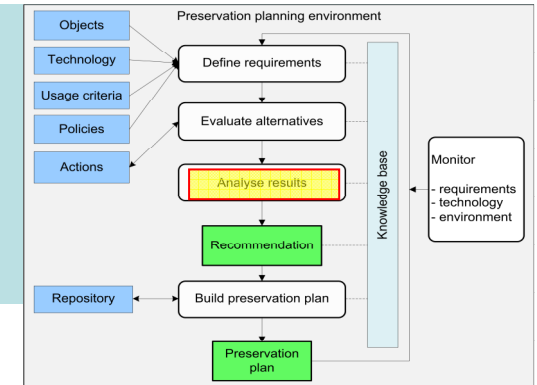
Create alternatives for selected services

# Develop and run experiment

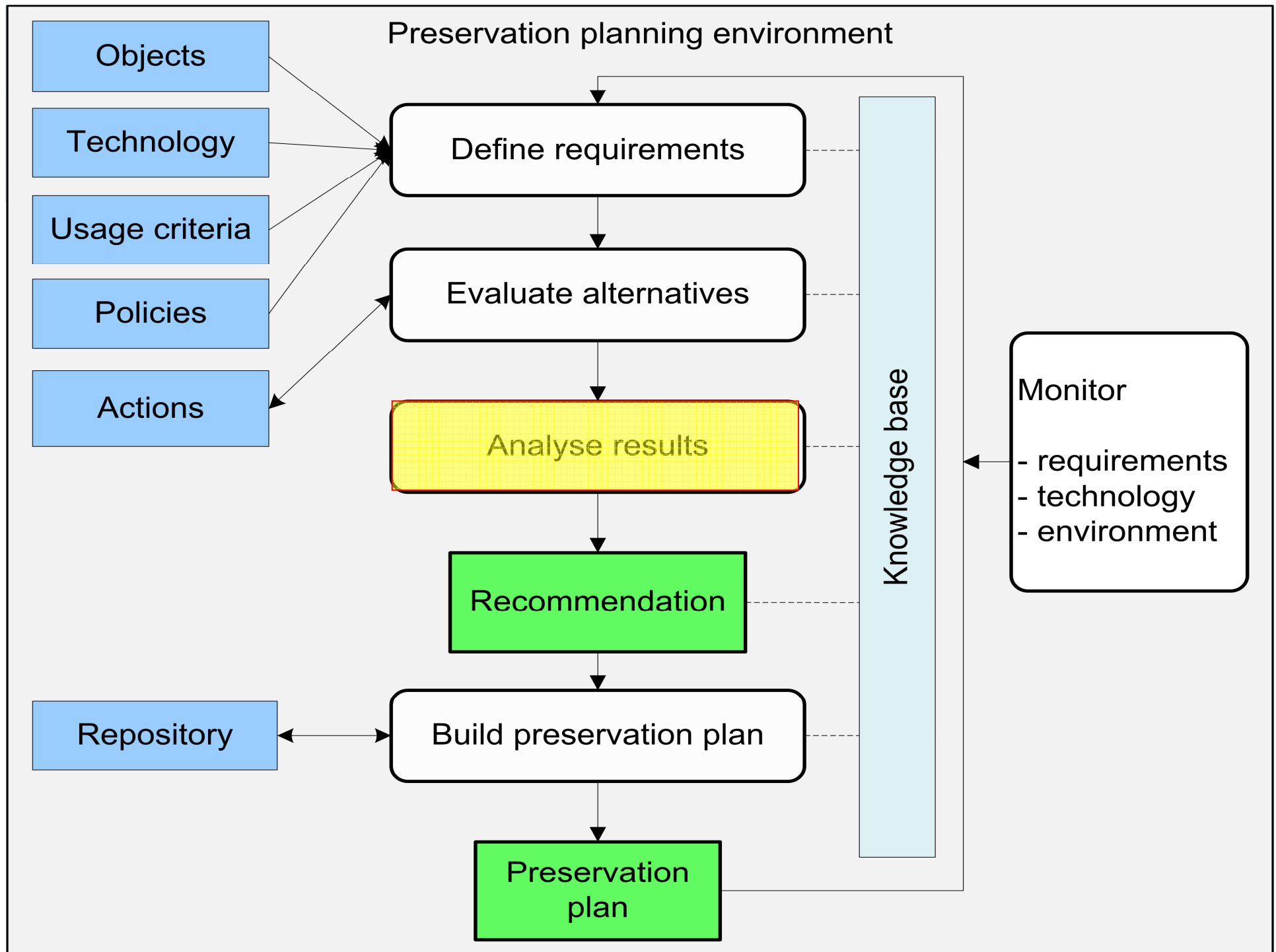


- Formulate for each experiment detailed
  - procedures and preparation
  - parameter settings for integrating preservation services
  - Evaluation/experiment plan (workflow/sequence of activities)
- Apply the selected potential preservation actions on the sample objects
  - Partly automated by web services
  - Partly manual

# Evaluate experiment

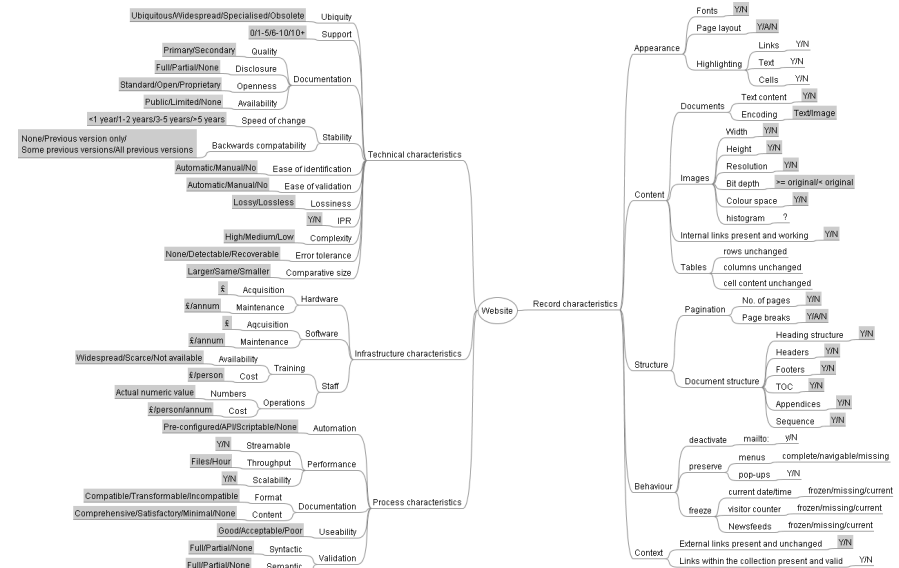
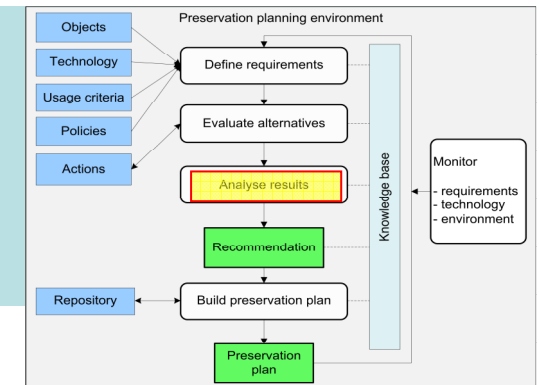


- Evaluate the outcome of each alternative for each leaf of the objective tree
- Partly automated by tool support
  - Comparing objects: XCL, Jhove, ImageMagick, ...
  - Measuring performance
  - Judging file formats
  - ...
- Result: evaluated tree

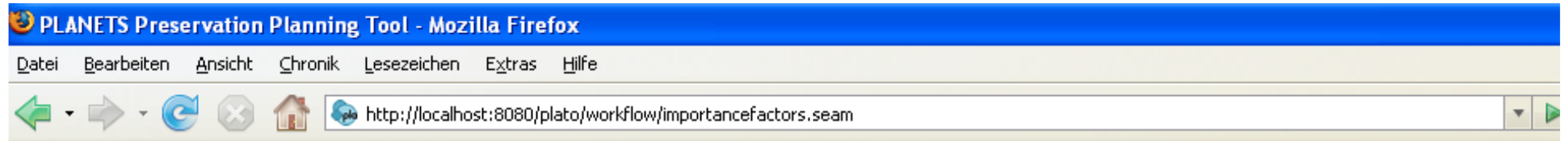
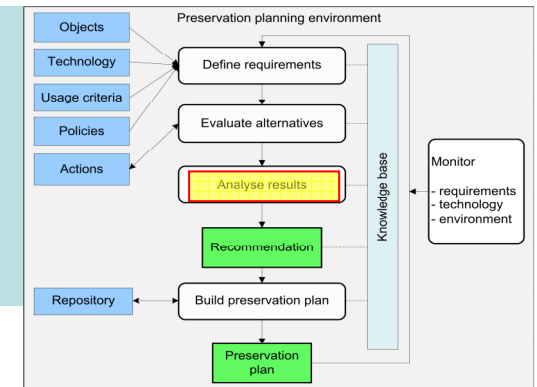


# Transform measured values

- Measures come in seconds, euro, bits,...
- Need to make them comparable
- Transform measured values to uniform scale
- Target scale 0-5
- Two types of transformation
  - Numeric
  - Ordinal
- Result: tree ready for analysis



# Set importance factors



## Set Importance Factors

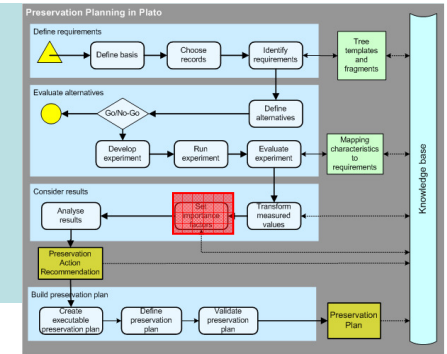
Balance weights automatically ☒

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

**Object characteristics**

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

# Analyse Results



- 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
- Rank alternatives according to overall performance value at root
- Performance of each alternative
  - overall
  - for each sub-criterion (branch)
- Comparison of different alternatives

## 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**

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

### Polar bear image preservation

Focus	Name	Result
	<input type="checkbox"/> Polar bear image preservation	TIFF (tool A): 4,78 TIFF (tool B): 4,28 PNG (tool D): 3,97
X	<input type="checkbox"/> Process	TIFF (tool A): 1,65 TIFF (tool B): 1,16 PNG (tool D): 0,74
	Complexity	TIFF (tool A): 2,50 TIFF (tool B): 2,50 PNG (tool D): 1,25
	Cost	TIFF (tool A): 2,50 TIFF (tool B): 1,00 PNG (tool D): 1,00
X	<input type="checkbox"/> Image properties	TIFF (tool A): 1,70 TIFF (tool B): 1,70 PNG (tool D): 1,70
	Bits of colour depth	TIFF (tool A): 5,00 TIFF (tool B): 5,00 PNG (tool D): 5,00
X	<input type="checkbox"/> Technical characteristics	TIFF (tool A): 1,43 TIFF (tool B): 1,43 PNG (tool D): 1,53
	Official standard	TIFF (tool A): 3,50 TIFF (tool B): 3,50 PNG (tool D): 3,50
	Filesize (in Relation to Original)	TIFF (tool A): 0,83 TIFF (tool B): 0,83 PNG (tool D): 1,12

# Analyse results

## Conclusion

### Recommendation

Recommendation:

Reasoning:

Effects of applying this strategy:

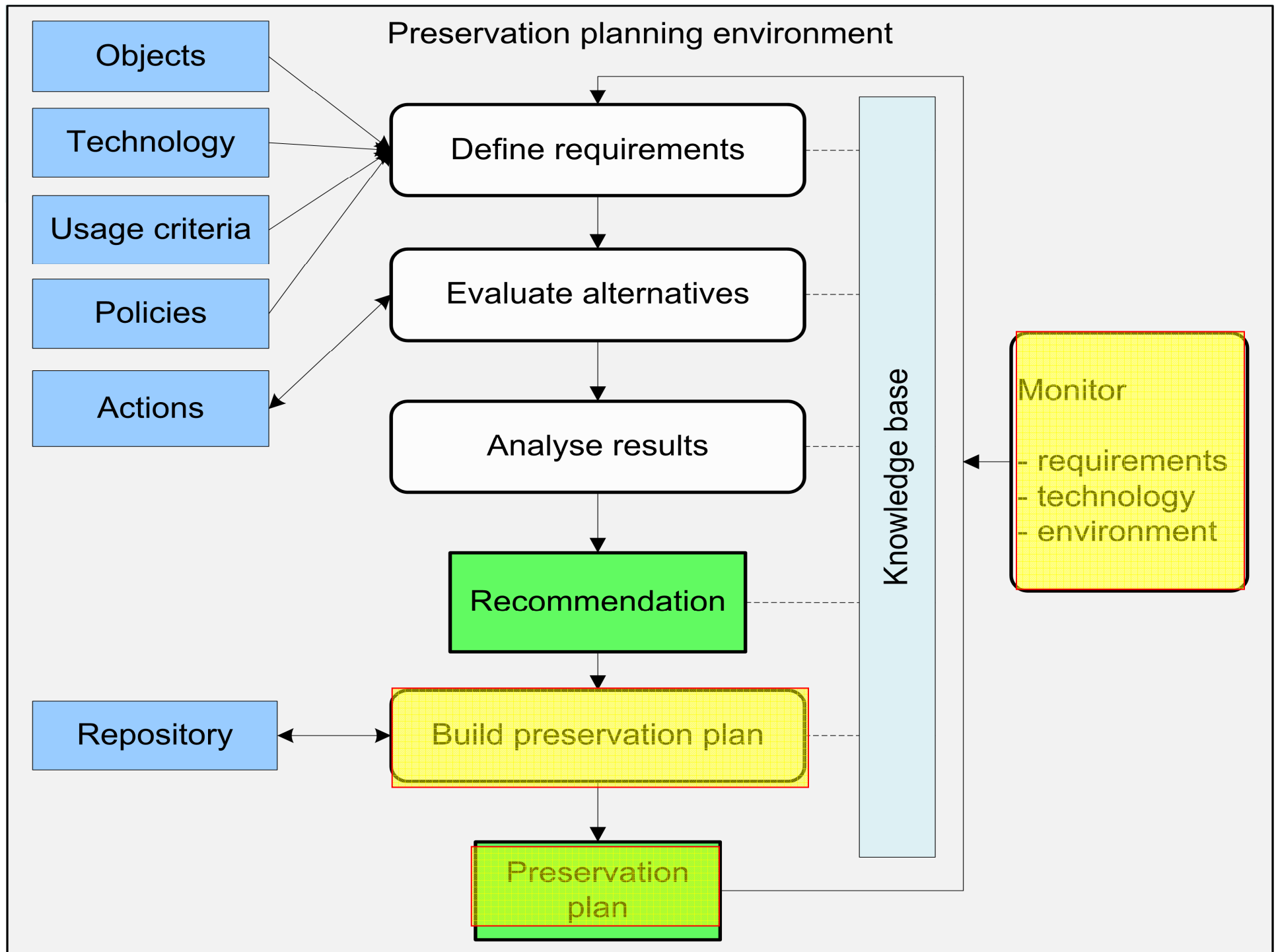


\*

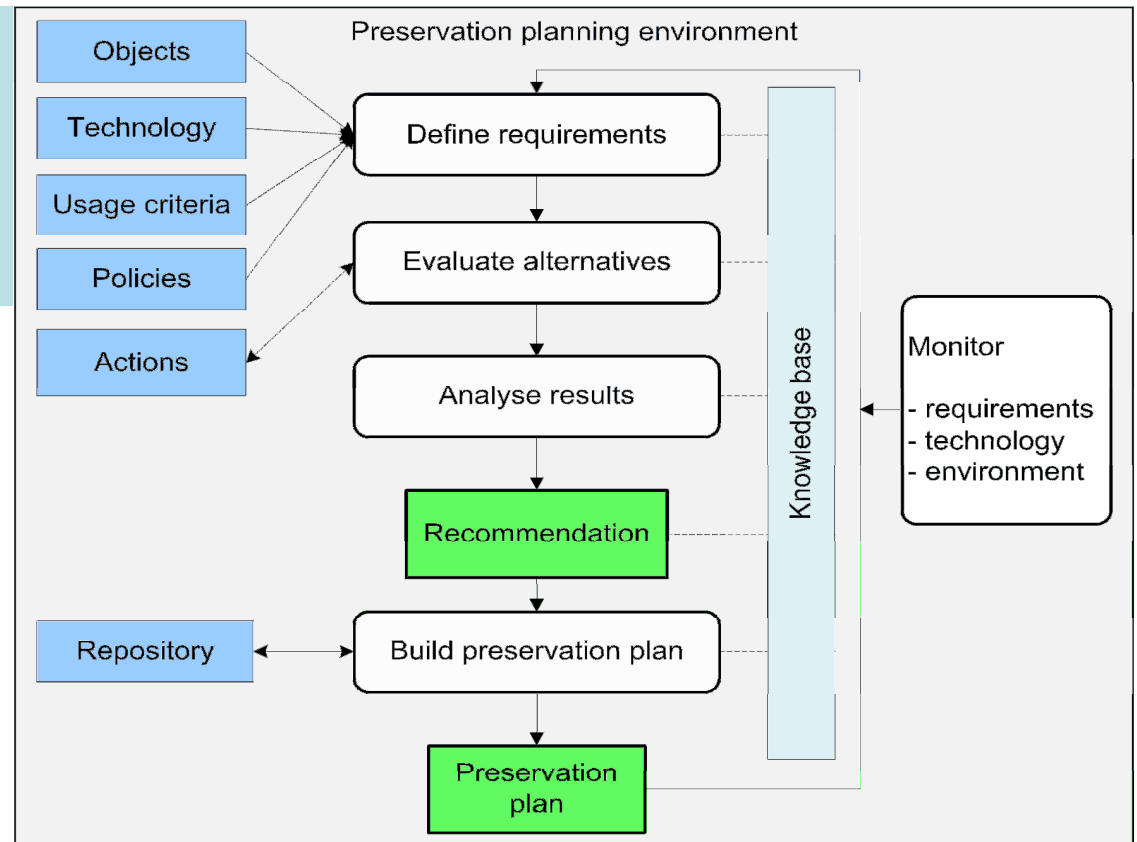


\*





# Questions?



[becker@ifs.tuwien.ac.at](mailto:becker@ifs.tuwien.ac.at)

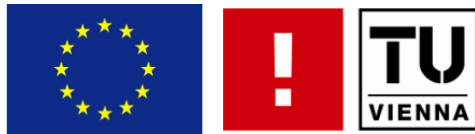
[www.ifs.tuwien.ac.at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)

[www.planets-project.eu](http://www.planets-project.eu)



## Scenario: We need a plan

- The purpose of planning is to find a strategy on how to preserve a collection for the future, i.e. choose a tool to handle our collection with.
- The tool must be compatible with our existing hardware and software infrastructure, to install it within our server and network environment.
- The files haven't been touched for several years now and no detailed description exists. However, we have to ensure their accessibility for the next years.
- ‘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).’



# Scenario: Scanned images

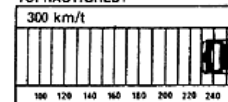
- Discussion scenario for today: Scanned images
- Specific exercise scenario tomorrow:  
Create a preservation plan for a collection of scanned images
- General characteristics of this scenario
- Mission statement
- High-level requirements

## LAMBORGHINI COUNTACH LP 500

Pris: ca. 748.000

ITALIENSK

TOPHASTIGHED:



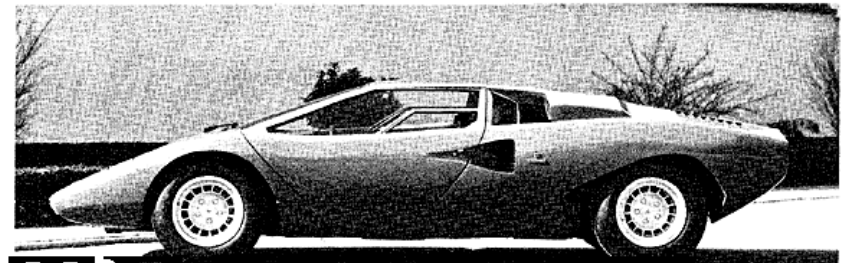
ACCELERATION:

0-100 km/t: 4,5 sek.



NORMALT BENZINFORBRUG:

3-5 km/l



## TEST DATA

Bevægelse er en af de mest udfordrende og fascinerende aspekter af livet. Den er en del af vores fremtid, men i realiteten er den skinbarlige nutid – eller fortid. For hvor findes den landevej, der kan leve op til de enorme ressourcer af fart og køreegenskaber, der er gemt i de hidsende linjer? Men hvor er det skønt, at den slags legetøj stadig fremstilles. Desværre er den håndværksmæssige kvalitet ikke helt i klasse med prisen.

**Opbygning:** Medbærende karrosseri og rør-ramme. 2 døre. Bagagerum bagtil. Centermotor. Træk på baghjulene.

**Motor:** 12-cylindret, 4-takts V-motor med 4 overliggende knastakler. Vandrilling. Boring: 85 mm. Slaglængde: 73 mm. Slagvolumen: 4971 cm<sup>3</sup>. Kompression: 10,5:1. HK: 440 DIN ved 7400 o/m. Motorudnyttelse: 89 hk pr. 1000 cm<sup>3</sup>. Vægt/kraftforhold: 3,0 kg/hk. Maks. drejningsmoment: 50,5 kpm ved 5000 o/m. 6 dobbelte Weber karburatorer (vandret), type 45 DCOE.

**Transmission:** Tor enkeltpladekobling. Gulvgear-

stang, 5 gear med synkronisering.

**Bremser:** Skivebremser for og bag. Vakuumforstærker, 2 kredse.

**Forhjulsophæng:** Dobbelte triangler. Skruefjedre. Krængningsstabilisator.

**Baghjulsophæng:** Dobbelte triangler. Skruefjedre. Krængningsstabilisator.

**Støddæmpere:** Teleskop.

**Styretilbehør:** Tandstang.

**El-system:** 12 volt. Vekselstrømsgenerator: 980 watt. Akkumulator: 72 amperetimer.

**Mål og vægt:** Længde: 414 cm. Bredde: 189 cm.

Højde: 107 cm. Frihøjde: 14 cm. Akselafstand: 245 cm. Sporvidde for/bag: 150/152 cm. Vendediameter: 11,2 m. Dæk: for/bag: 205/70/215/70 VR 14. Tankindhold: 120 l. Oliesump: 15 l. Kolerindhold: 16 l. Egenvægt: 1300 kg.

**Reklamationsfrist:** 12 mdr. eller 20.000 km.

**Olieskift/serviceeftersyn:** 5.000/10.000 km.

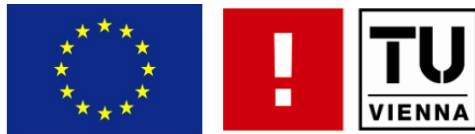
**Øvrige modeller:** Ingen.

**Importer:** As Nielsen Holst's Eff., Edwin Rahrs Vej 52, 8220 Brabrand. Tlf. (06) 262244.



# High level requirements 1/2

- Formats
  - must/shall be standardised...
  - Compression?
- Tools
  - must/shall be open source,
  - Must not cost more than...
- Bit-stream preservation costs...
  - Depend on the file size and other factors
  - Must not exceed ... (per object)
- Strategy
  - consider migration ,
  - consider emulation (copyright?)

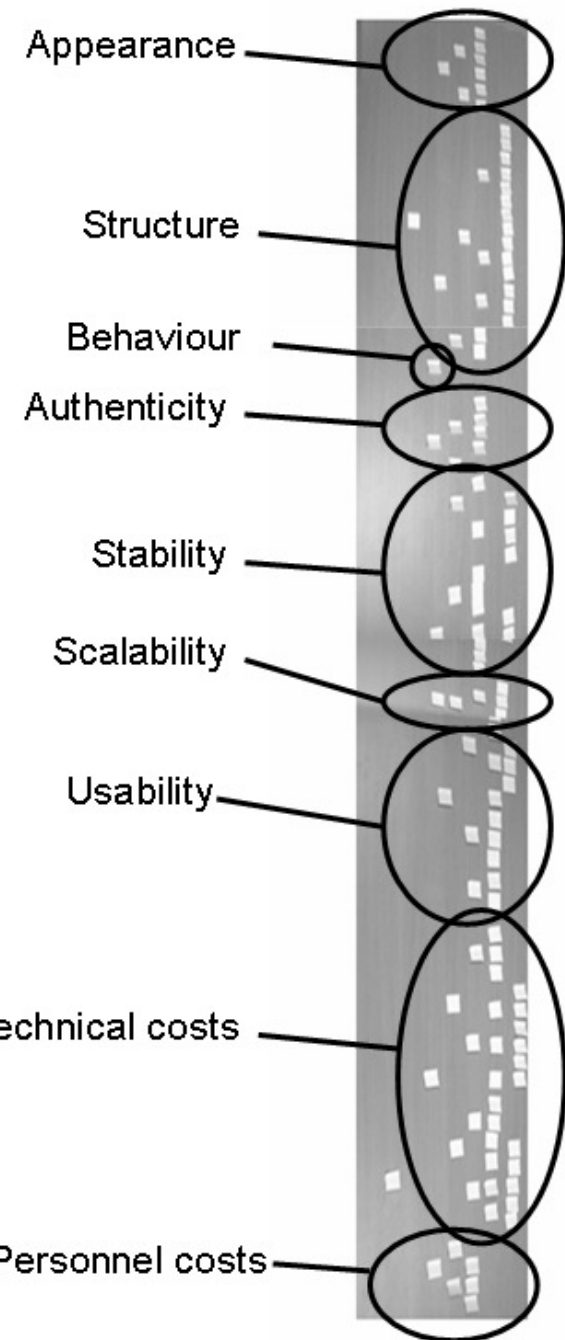
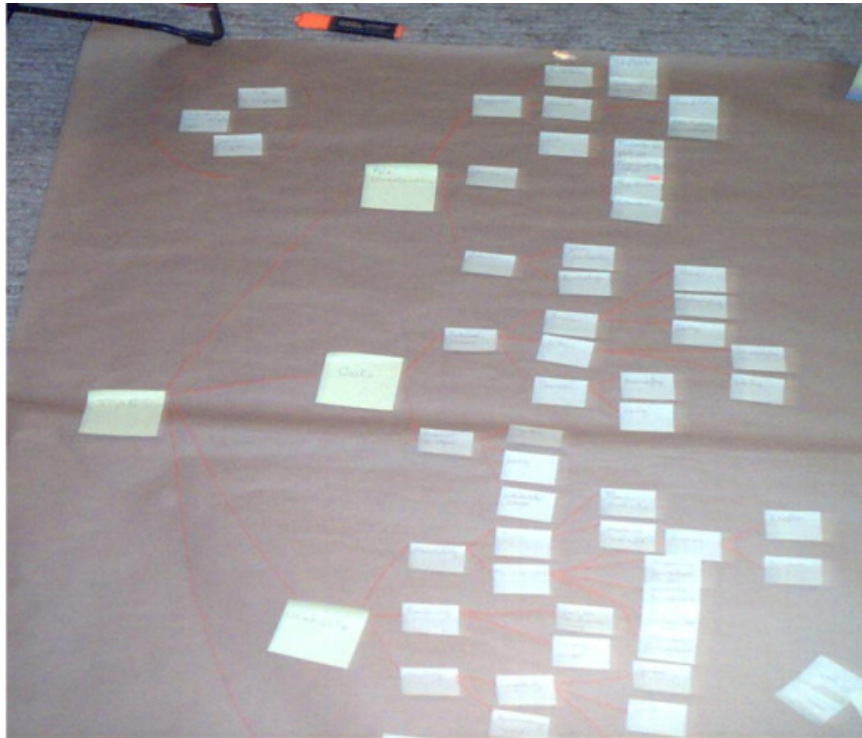


## High level requirements 2/2

- Objects must be
  - “the same” – “unchanged” – “authentic” ...
  - Significant properties need to be defined and measured
  - Content, context, structure, appearance, behaviour
- Trade-offs might be necessary
  - Usability vs. authenticity
  - Structure vs. independency
  - Access vs. costs
  - ...

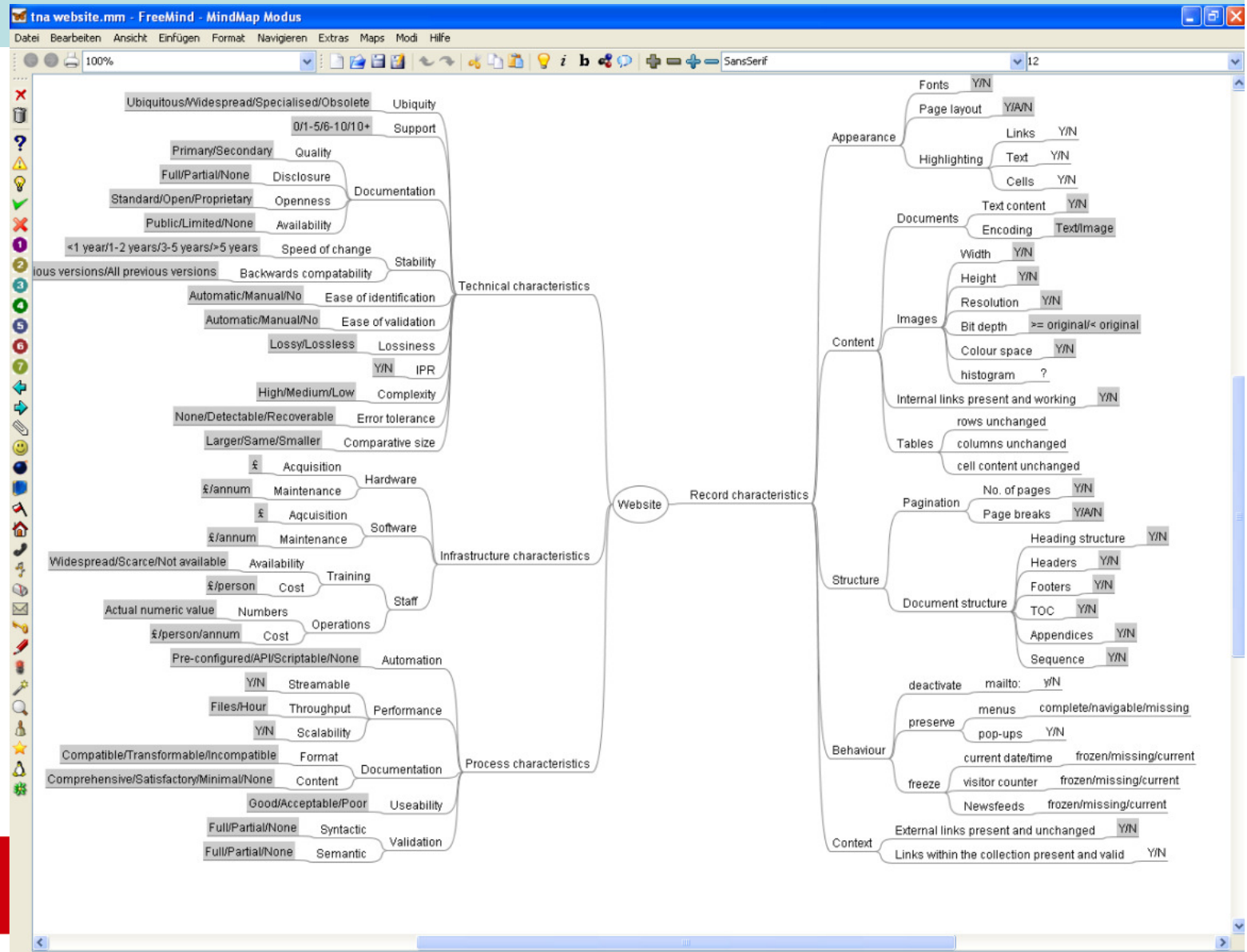


# Analog...





# ... or born-digital





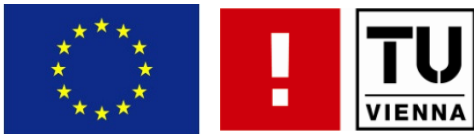
# Requirements for objects

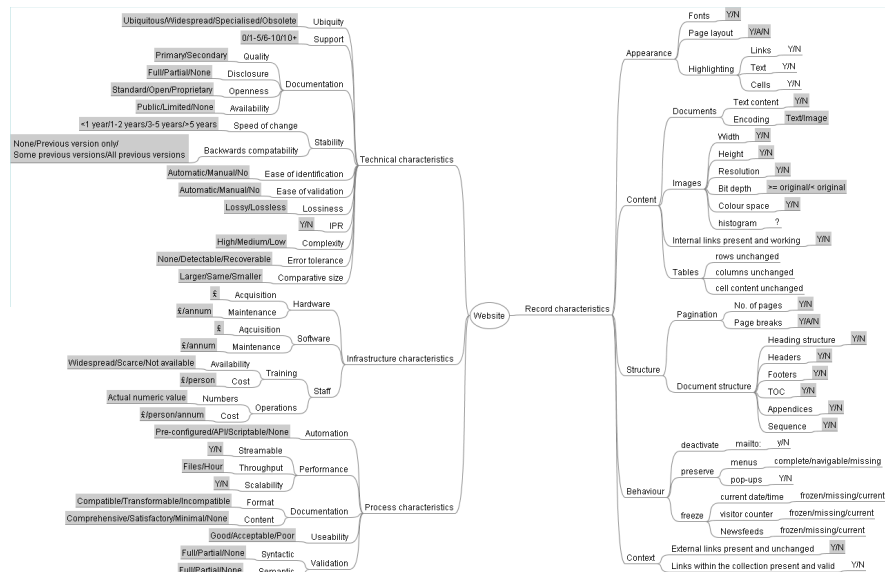
- **Authenticity**
  - to be what it purports to be,
  - to have been created or sent by the person purported to have created or sent it, and
  - to have been created or sent at the time purported
- **Reliability**
  - contents can be trusted as a full and accurate representation of the transactions, activities or facts to which they attest and can be depended upon in the course of subsequent transactions or activities
- **Integrity**
  - being complete and unaltered
- **Usability**
  - can be located, retrieved, presented and interpreted
- **Accuracy**
  - the degree to which data, information, documents or records are precise, correct, truthful, free of error or distortion or pertinent to the matter.



# Essential characteristics of 'digital objects'

- What needs to be preserved?
  - Content
  - Context
  - Structure
  - Appearance
  - Behaviour





## 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  
Expand All Collapse All

Focus	Name	Result
	Polars bear image preservation	TIFF (tool A): 4,78 TIFF (tool B): 4,28 PNG (tool D): 3,97
X	Process	TIFF (tool A): 1,65 TIFF (tool B): 1,16 PNG (tool D): 0,74
	Complexity	TIFF (tool A): 2,50 TIFF (tool B): 2,50 PNG (tool D): 1,25
	Cost	TIFF (tool A): 2,50 TIFF (tool B): 1,00 PNG (tool D): 1,00
X	Image properties	TIFF (tool A): 1,70 TIFF (tool B): 1,70 PNG (tool D): 1,70
	Bits of colour depth	TIFF (tool A): 5,00 TIFF (tool B): 5,00 PNG (tool D): 5,00
X	Technical characteristics	TIFF (tool A): 1,43 TIFF (tool B): 1,43 PNG (tool D): 1,53
	Official standard	TIFF (tool A): 3,50 TIFF (tool B): 3,50 PNG (tool D): 3,50
	Filesize (in Relation to Original)	TIFF (tool A): 0,83 TIFF (tool B): 0,83 PNG (tool D): 1,12

## Conclusion

**Recommendation**

Recommendation:

Reasoning:

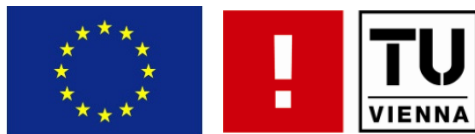
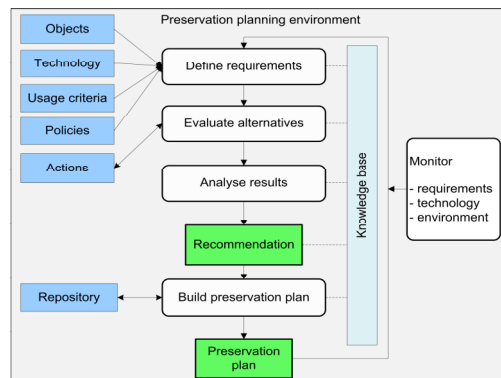
Effects of applying this strategy:

# Questions?

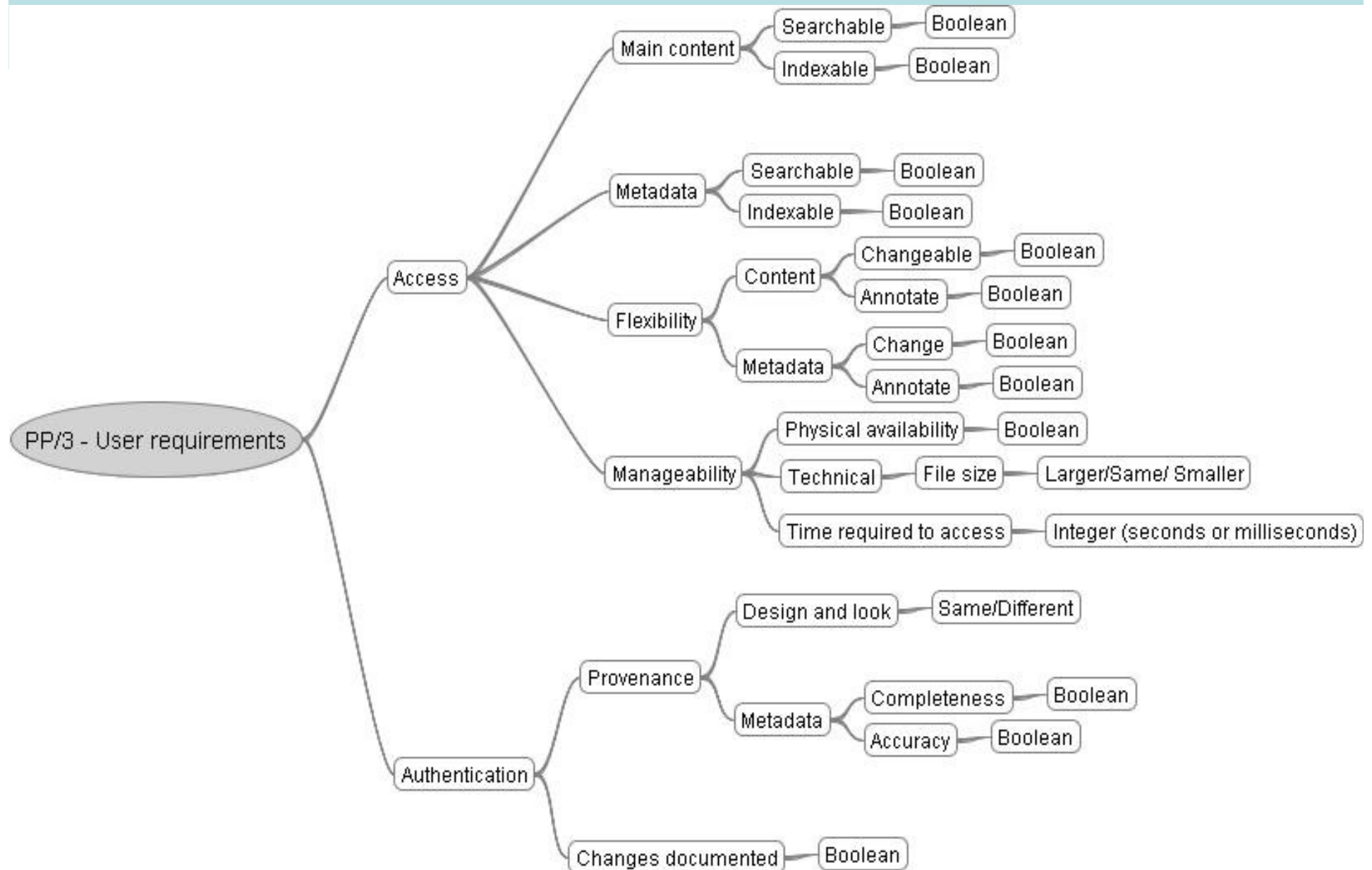
[becker@ifs.tuwien.ac.at](mailto:becker@ifs.tuwien.ac.at)

[www.ifs.tuwien.ac.at/dp/plato](http://www.ifs.tuwien.ac.at/dp/plato)

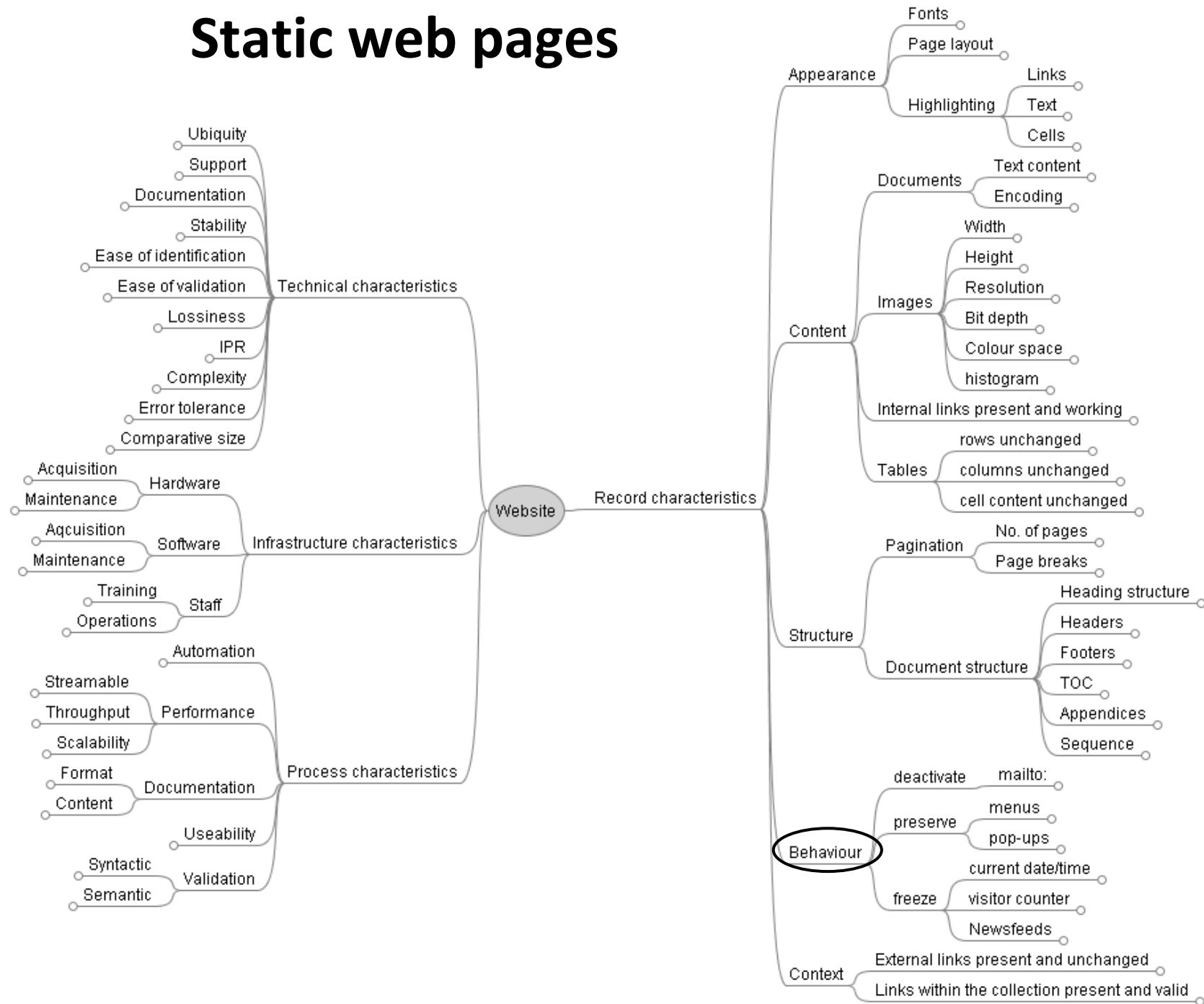
[www.planets-project.eu](http://www.planets-project.eu)



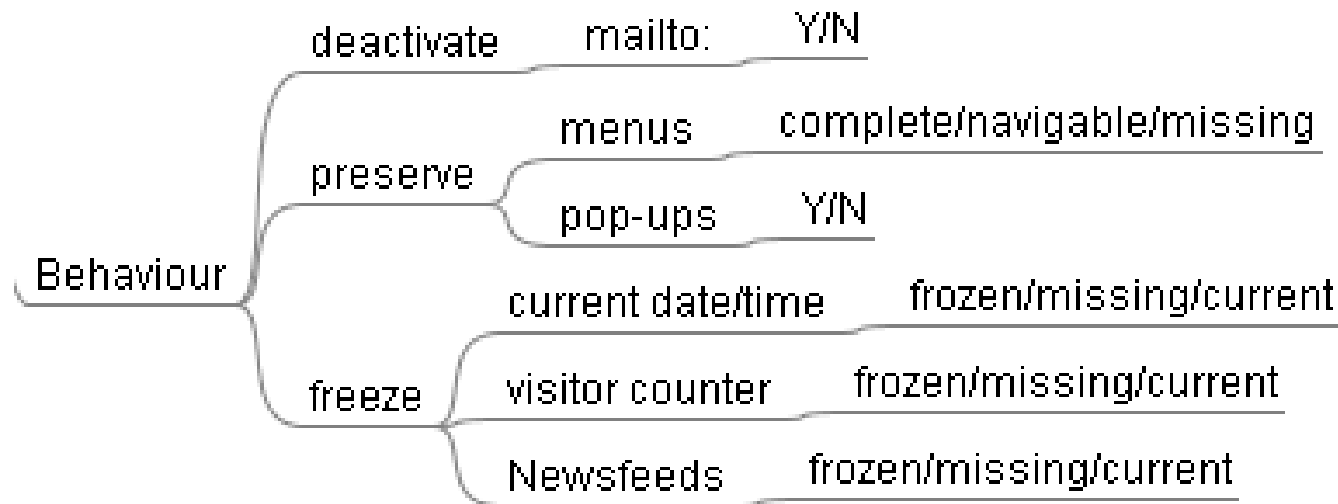
# Usage



# Static web pages

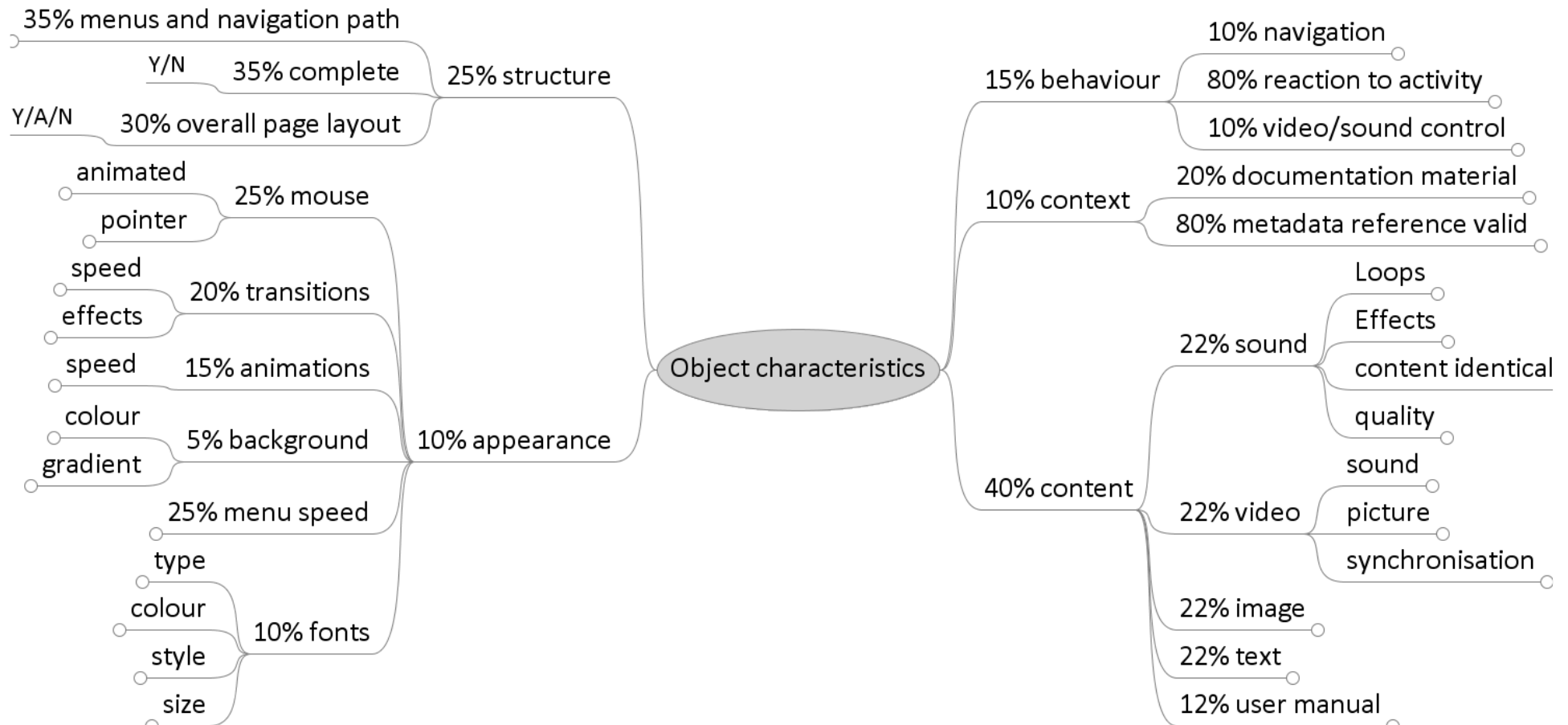


# Behaviour



- 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





# 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 P-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

