Preservation planning 1

What to decide and how

March 24, 2014

Hannes Kulovits

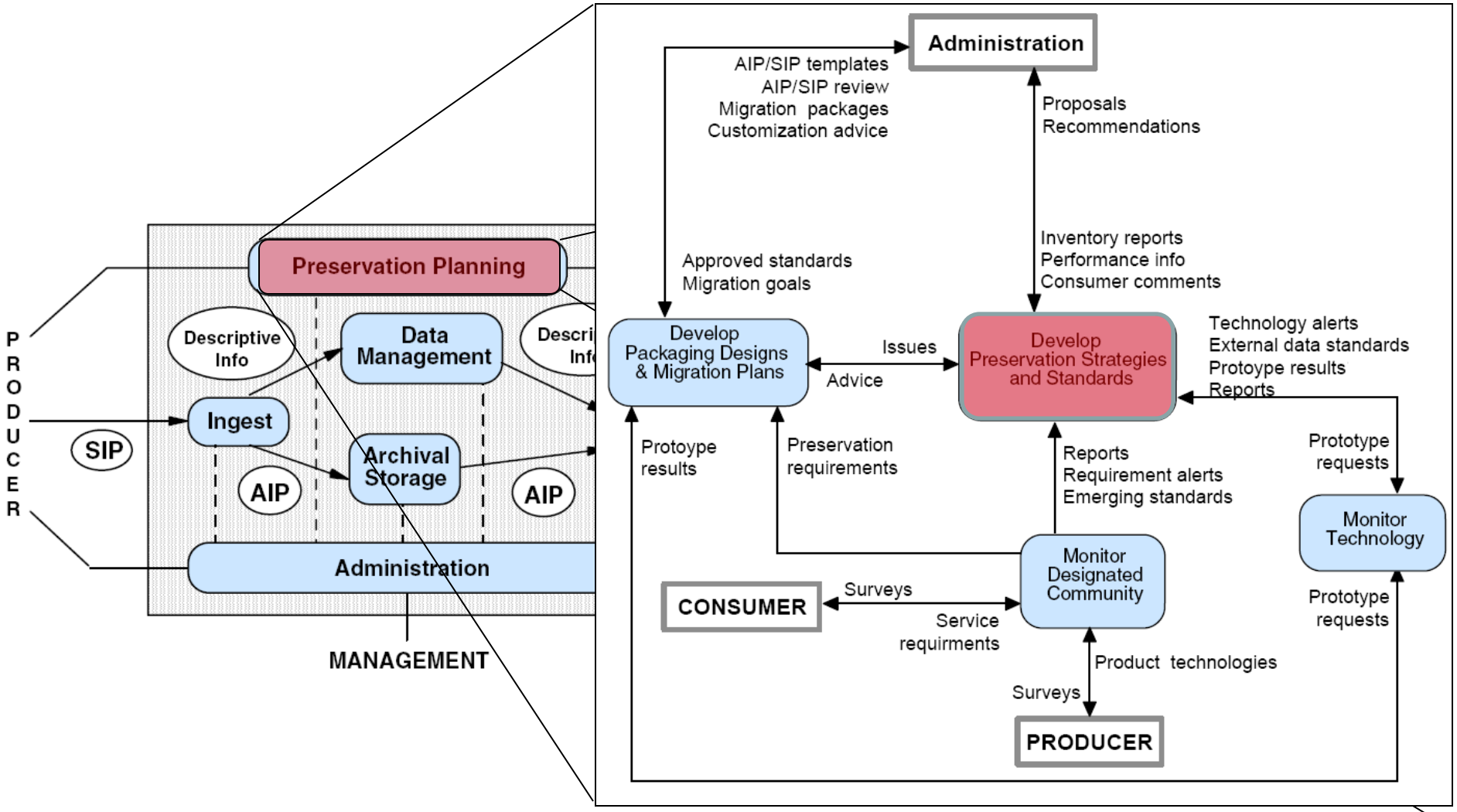
Digital preservation and planning

- The mission of digital preservation
  - keep content authentic and understandable for a user community over time
- The mission of peservation *planning*
  - What to do and how
  - Defining the right courses of actions
  - Questions:
    - How to select the right action in a given scenario?
    - How to ensure trust?
    - How to enable scalability?
    - How to ensure continuous alignment over time?

Preservation Planning 1: Agenda

- The Problem: What to decide and how?
  - Decision problems in digital preservation
- The context: Goals, drivers, constraints
- From goals to actions: Preservation planning overview
  - What is preservation planning?
  - What is a preservation plan?
  - How to create a preservation plan, part 1

Preservation Planning



Preservation Operations

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

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

- Systematic software component evaluation

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

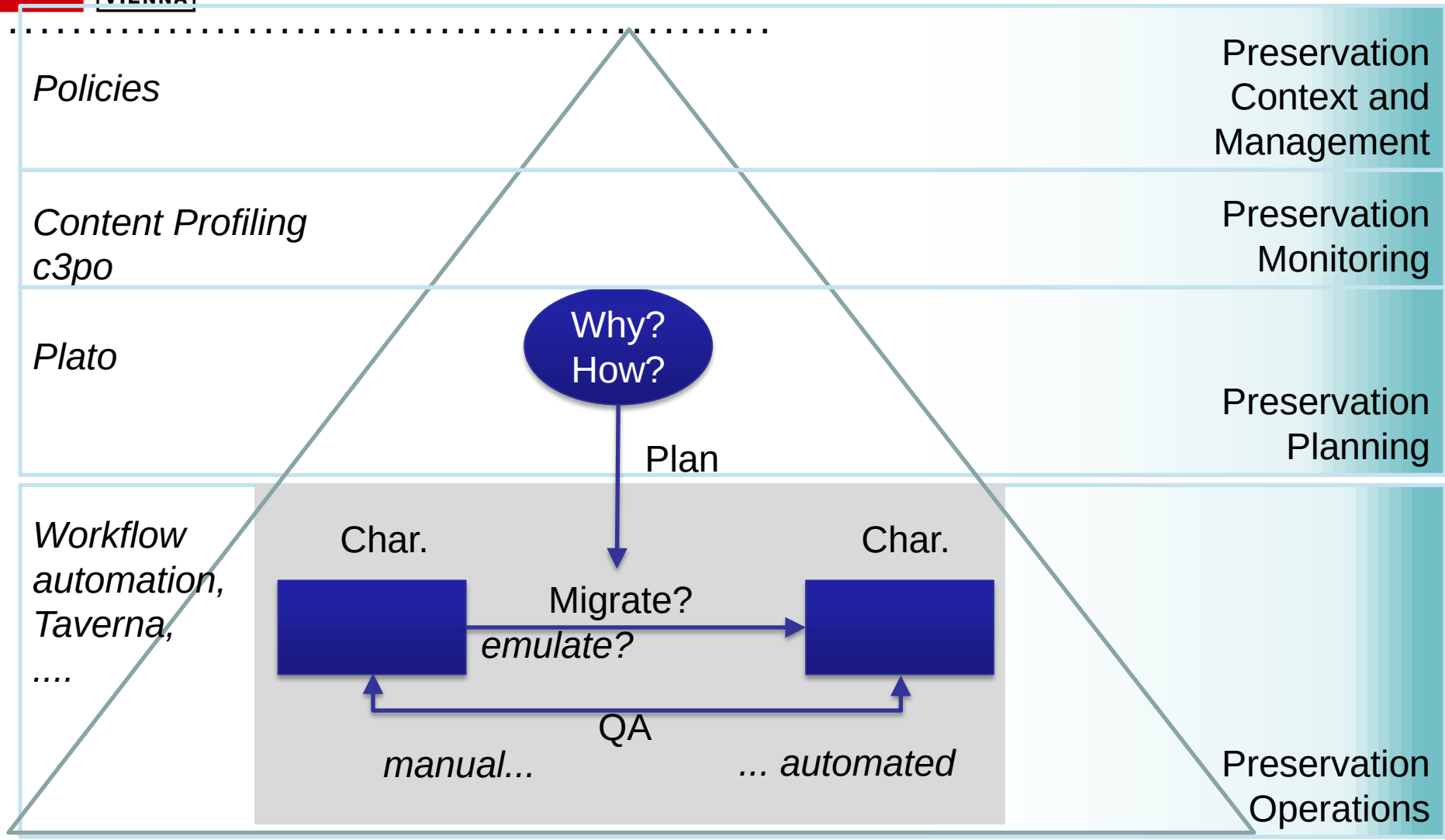
- **How can we ensure that decision processes scale up?**
  - How can we automate decision making?
  - How can we integrate continuous monitoring?
  - Which properties can be measured automatically, and how?
  - How can we create a controlled environment for observing the behaviour of components in a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

Evaluating preservation actions

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

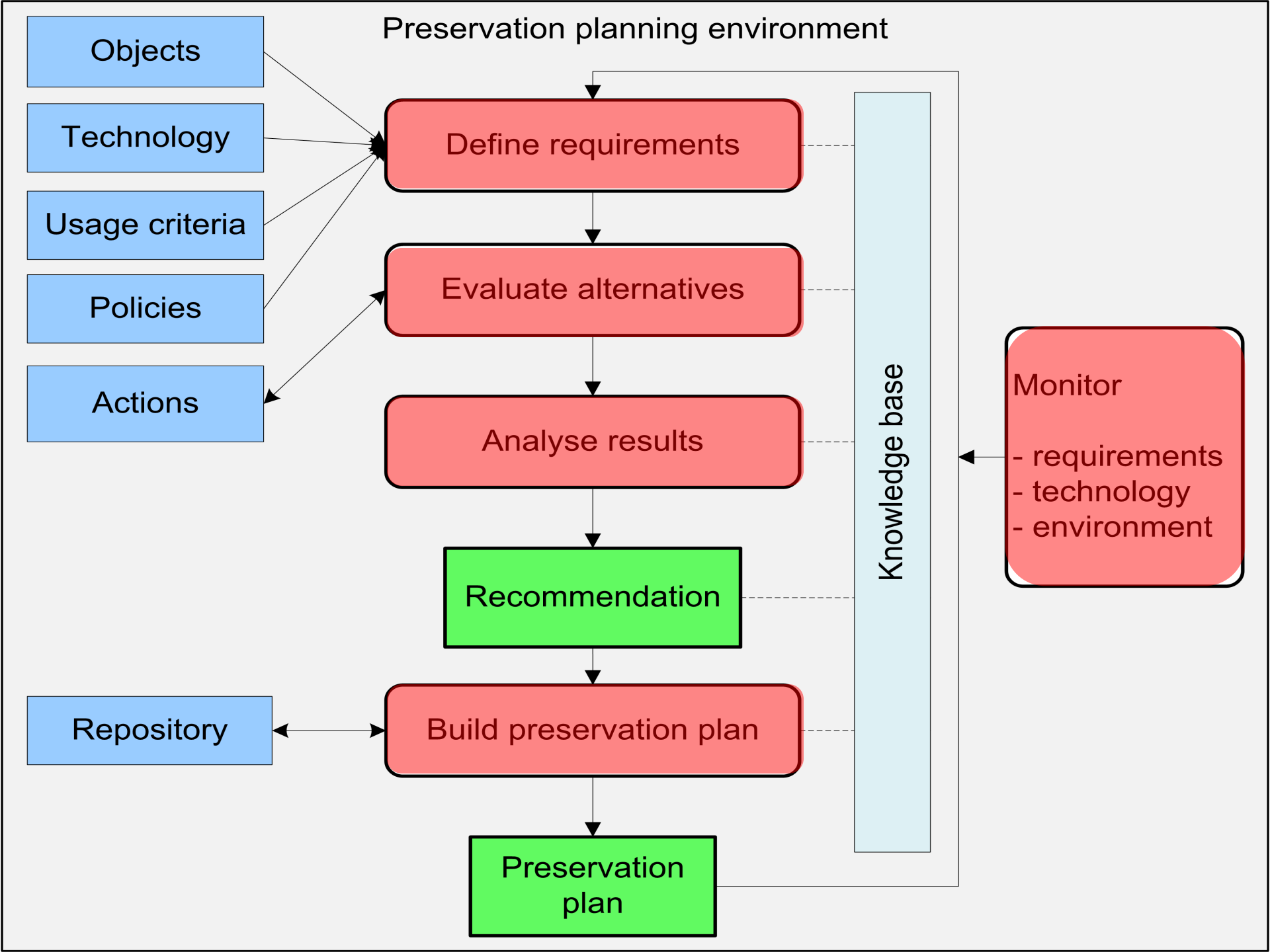
- Preservation planning approach
  - Evaluate potential actions objectively against scenario-specific requirements in a repeatable way
  - Sample-based experiments in controlled environment
  - Quantitative analysis of strengths and weaknesses
  - Evaluate suitability of each potential action

Operations in context



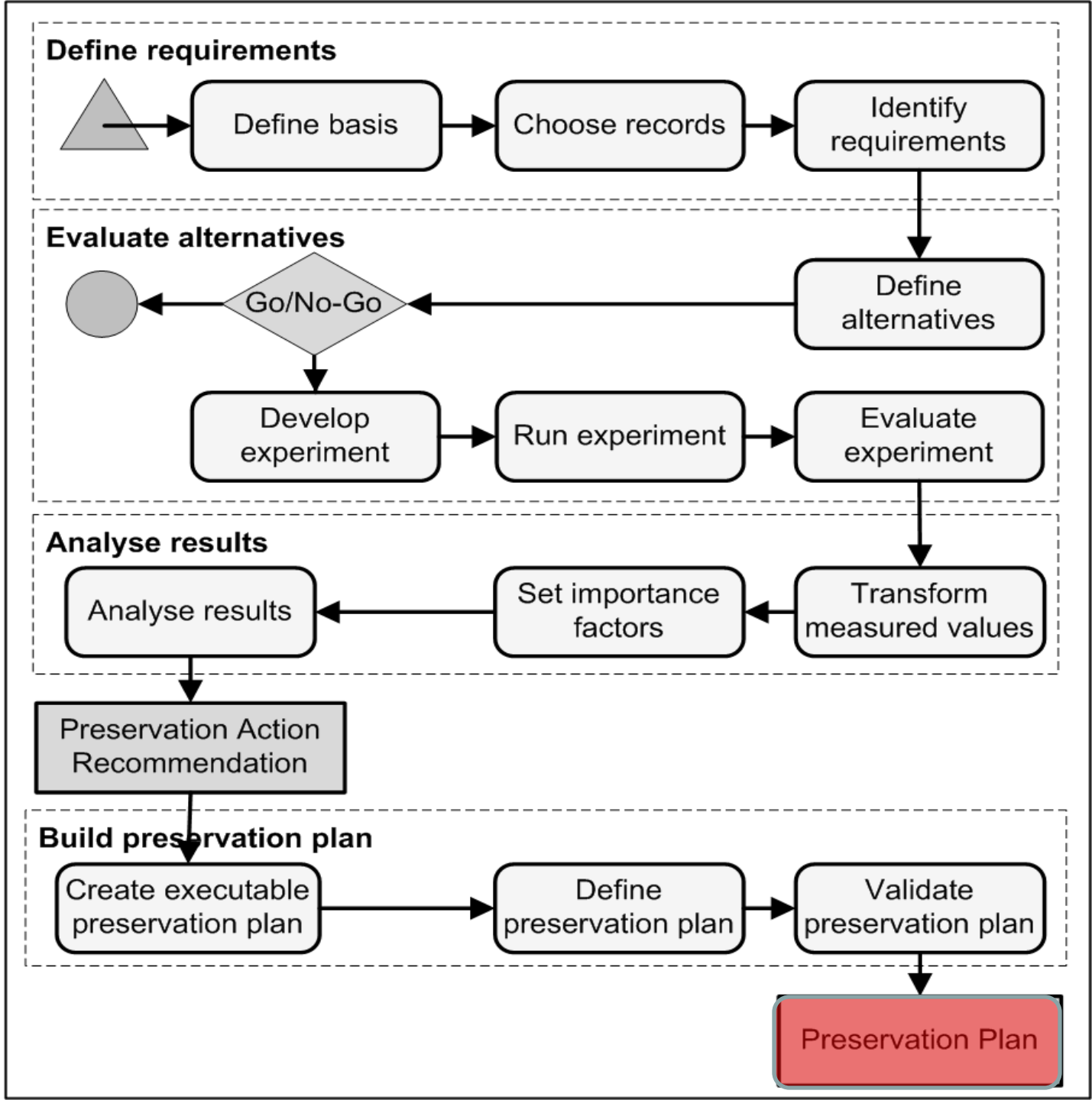
Trustworthy preservation planning

- Preservation planning as a capability:
  - the ability to assess the impact of influencers and specify actionable preservation plans that define concrete courses of actions and the directives governing their execution
  - the operative management of obsolescence to maximize expected value with minimal costs
- A preservation plan specifies actions
  - scope and what, how, when, who, why
- Trust requires evidence
  - Trust has to be evaluated in a realistic context
    - Documented evidence
    - Controlled experimentation
    - Scenario-specific requirements assessment


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

Preservation Planning: Key concepts

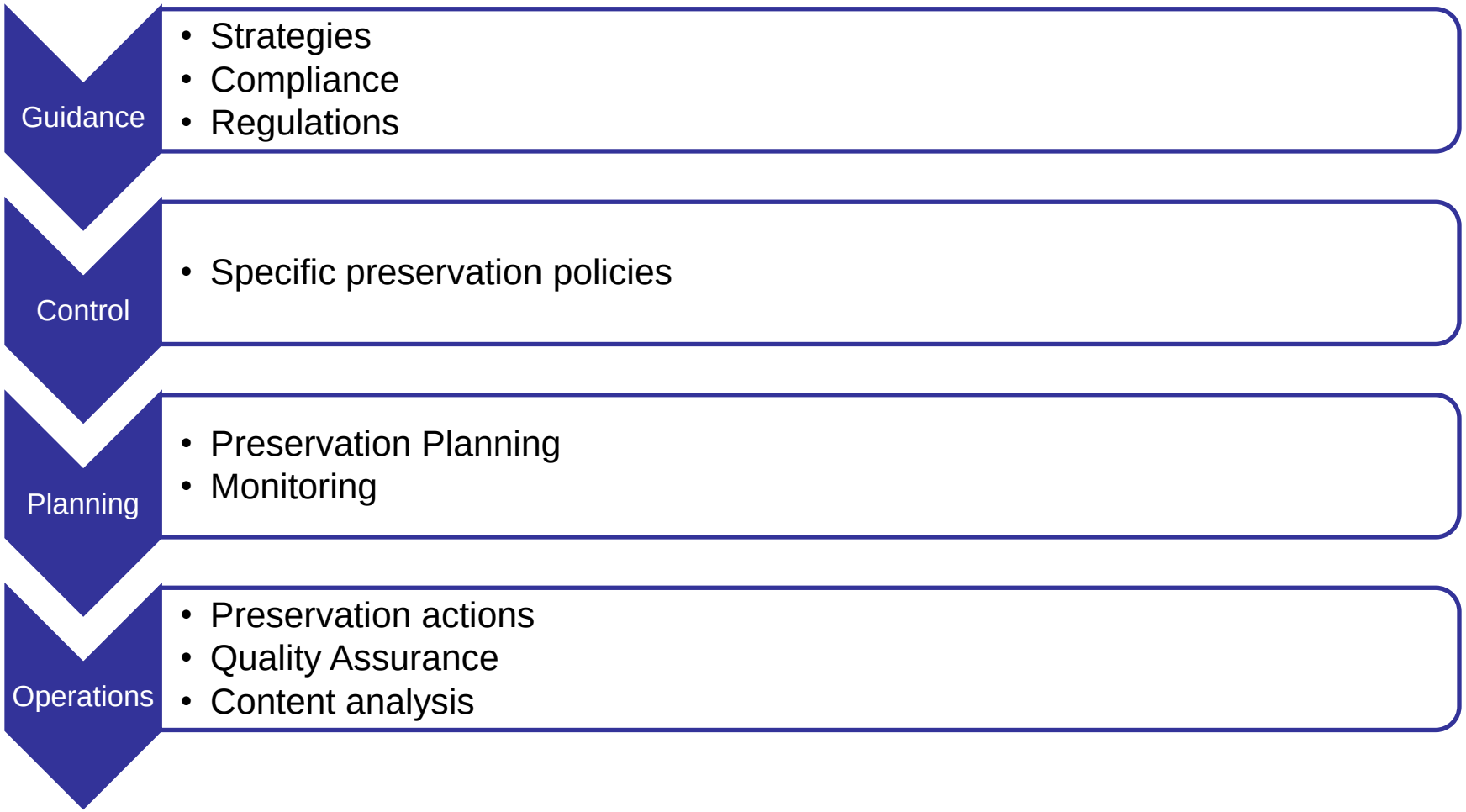
- Repeatable, standardized planning workflow
- A weighted hierarchy of objectives
  - Measurable criteria on the leaf level of the tree
  - Utility functions make criteria comparable
- Controlled experimentation on sample content
  - Evidence-based decision making
- Standardized structure for plan specification
  - Transparency and documentation
  - Comparability across scenarios
  - Integration with repository systems
- Planning tool Plato guides, validates, documents
  - www.ifs.tuwien.ac.at/dp/plato
- Automation: Reduce manual effort


Define requirements
Define basis
Choose records
Identify requirements
Evaluate alternatives
Define alternatives
Go/No-Go
Develop experiment
Run experiment
Evaluate experiment
Analyse 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

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.

Levels of control



Example scenarios for planning

- Digital photography archive with diverse camera RAW files from many cameras
  - Should we normalise to reduce risk and facilitate access? To DNG? To TIFF? How?
- Console video games
  - Use emulation? Which emulator and how?
- Large-scale digitized newspaper archive in non-standard TIFF
  - Convert to JPEG2000 or TIFF6 compressed to reduce costs? Leave unchanged and wait?
- Interactive digital art
  - What to do?

Digital Preservation drivers

Table 4 - DP drivers as specified in the SHAMAN Reference Architecture

Internal	Business Vision	Goals, Scope of designated community, etc.
	Resources	Infrastructure (e.g., operational costs, expertise needed), Hardware (e.g., operational costs, technological capability), Software (e.g., operational costs, technological capability), Staff (e.g., expertise and qualifications, commitment)
	Data	Volume, Structure, Representation, Semantics, etc.
	Processes	Dependencies, Responsibilities, Alignment, etc.
External	Producers	Demand satisfactions, Content, Technology, Trust and reputation
	User community	Technology, Knowledge, Demand satisfaction, Trust and reputation
	Contracts	Deposit, Supplier and service, Interoperability, Access, etc.
	Supply	Technology, Services, People
	Competition	Overlap of: Services, Content, User community, Producers, Technology, Mandate, Rights, Funding, Capabilities
	Regulation and mandate	Regulation/Legal constraints, Embedding organization regulation, Mandate, Rights and ownership, Certification, Funding

Preservation policy: examples

- Policies are elements of governance
- They “govern”, i.e. **guide, shape and control** preservation operations

- 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

- 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.

National Archives of Australia

- 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?**
- For example:
  - all records that are accepted should be preserved, regardless of file format, medium, application, etc.
  - transform to open standard + keep ‘original’ format

ISO/TR 18492:2005

- International standard: Long-term preservation of electronic *document-based i*nformation

- 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.

From guidance to action *(simplified)*

Guidance
- Open formats
- Migration

Control
- Formats must be ISO-standardised
- No lossy compression may be used

Planning
- Concrete criteria and decision

Operations
- Migrate to PDF/A
- Quality assure results

What is *in* a preservation plan?

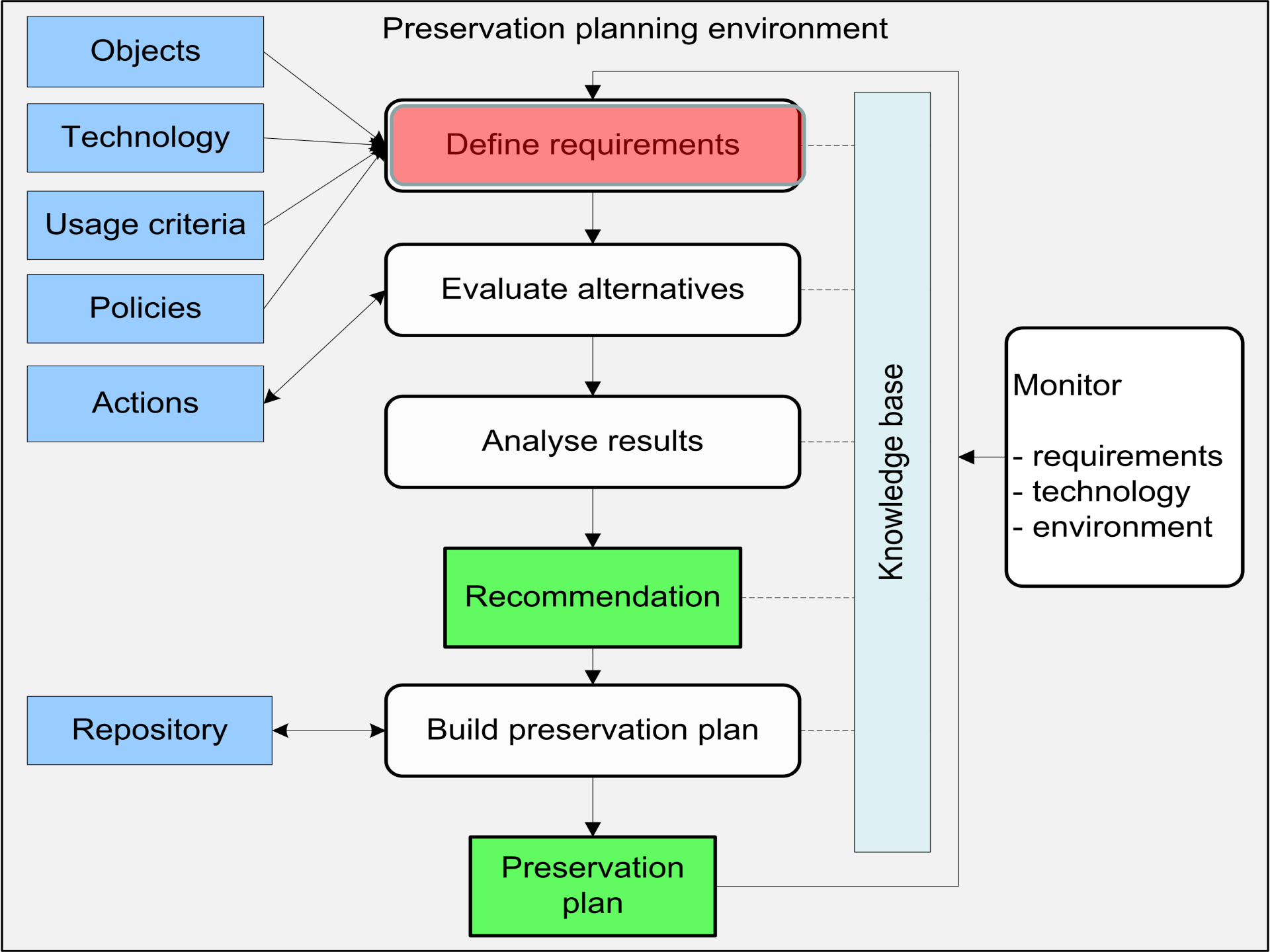
- Definition of scope and context
  - What to preserve and why
- Definition of objectives
  - What to achieve
- Set of actions, evaluation and recommendation
  - How to preserve it and why
- Documentation of actions and reasons
  - Why did we decide what
- Conditions for QA and monitoring
  - What to look out for

Characteristics of a preservation plan

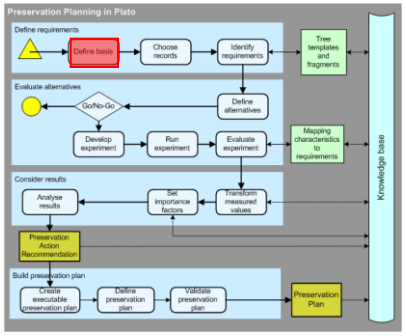
- Translation of a preservation policy
- Specification of how to treat a collection in a given setting
- Monitored for
  - ✓ changes in technology
  - ✓ changes in organisational setting
  - ✓ changes in user requirements
  - ✓ changes in available tools
  - ✓ changes in preservation methods
- Species 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

Agenda

- The Problem: What to decide and how?
  - Decision problems in digital preservation
- Preservation planning overview
  - What is preservation planning?
  - What is a preservation plan?
  - **How to create a preservation plan: Part 1**

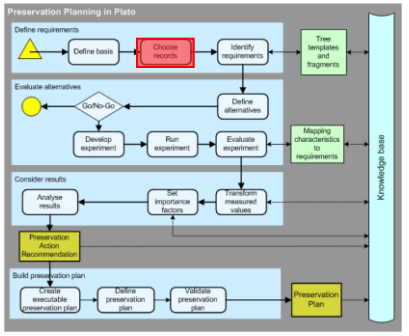

Preservation planning environment
Objects
Technology
Usage criteria
Policies
Actions
Define requirements
Evaluate alternatives
Analyse results
Recommendation
Build preservation plan
Preservation plan
Repository
Knowledge base
Monitor
- requirements
- technology
- environment

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

Select content type

- e.g.: Legal documents from the enterprise archive

Select properties

- Property set determined by type “documents” (page count, …)

High-level issue detection

- Object-level policy violations: Validity, encryption, …
- Collection-level: format normalisation…

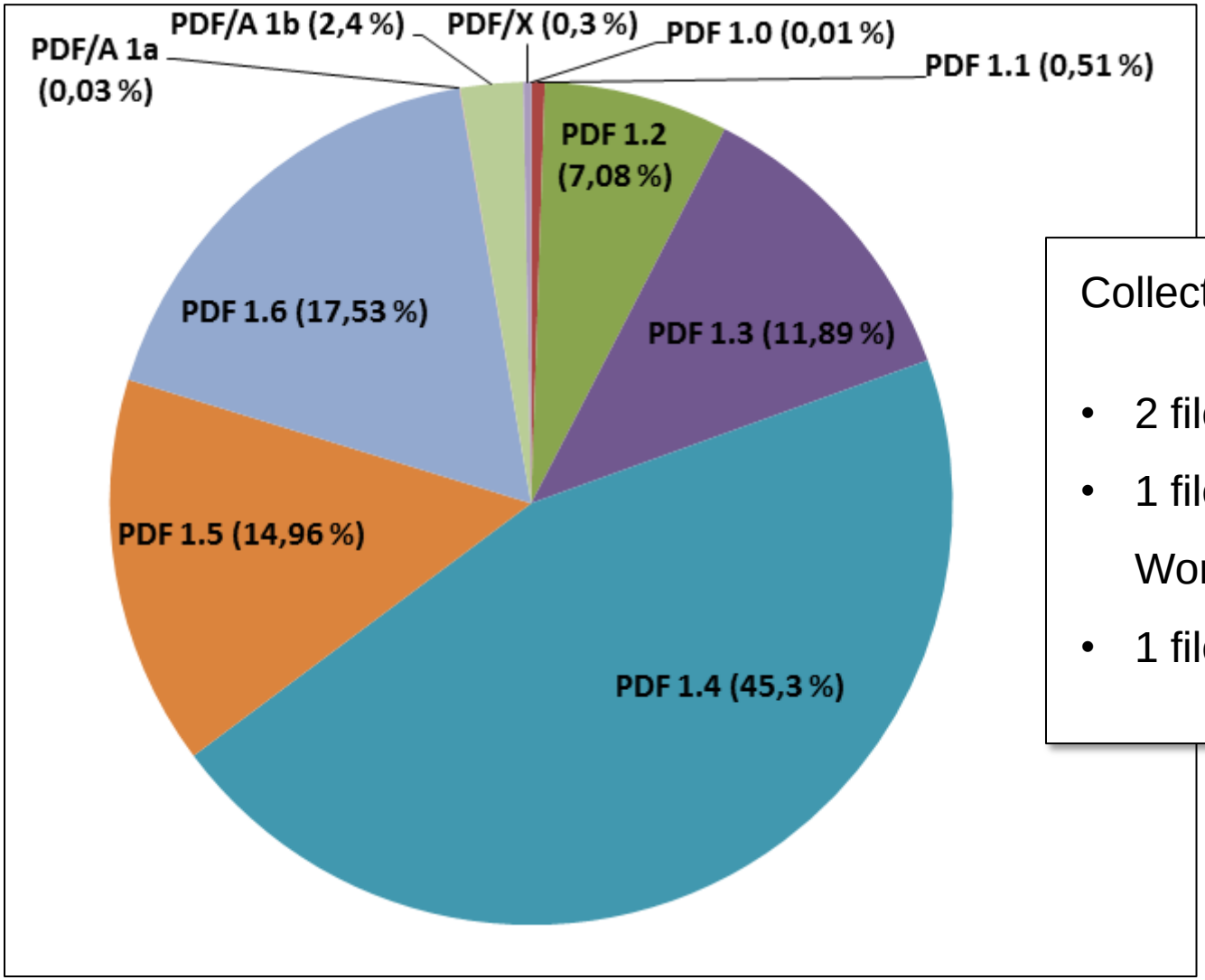
Select scoping property

- Subformat: PDF 1.2…
- Other properties: all protected documents….

Select samples

- Single dimension: page count, size, age, validity…
- Multiple dimensions: Largest invalid, oldest protected …

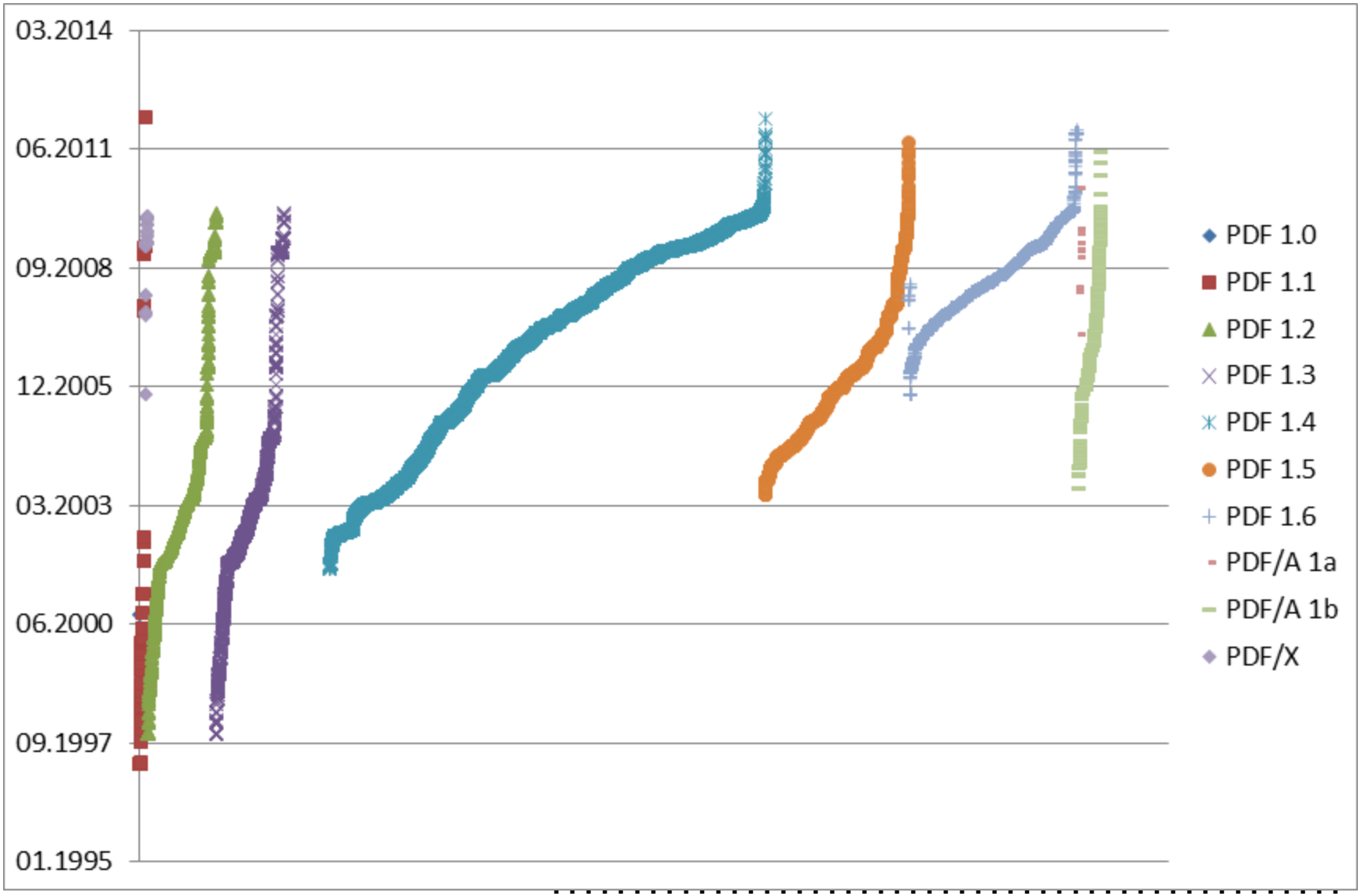
An example collection of PDF documents



Collection size: ~40.000 files

- 2 files „Unknown Binary“
- 1 file exposed as MS Word DOC
- 1 file with size 0

Creation date per PDF file type



Which two files are similar?

- Consider three files A, B, C

	A	B	C
Format	PDF 1.2	PDF 1.2	PDF 1.4

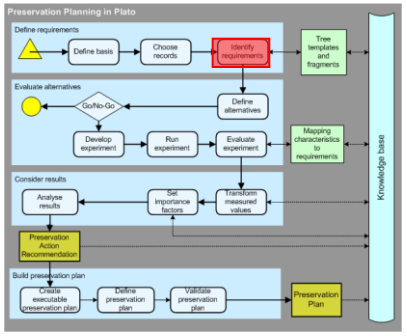
Which two files are similar?

- Consider three files A, B, C

	A	B	C
Format	PDF 1.2	PDF 1.2	PDF 1.4
Page count	20	20.000	40
Encryption	Yes	No	Yes
File size	1MB	120 MB	2 MB
Valid	no	yes	No
Well-formed	Yes	yes	Yes
Digital signature	no	yes	no

- … file format is just one property
- Careful selection is critical

Requirements definition



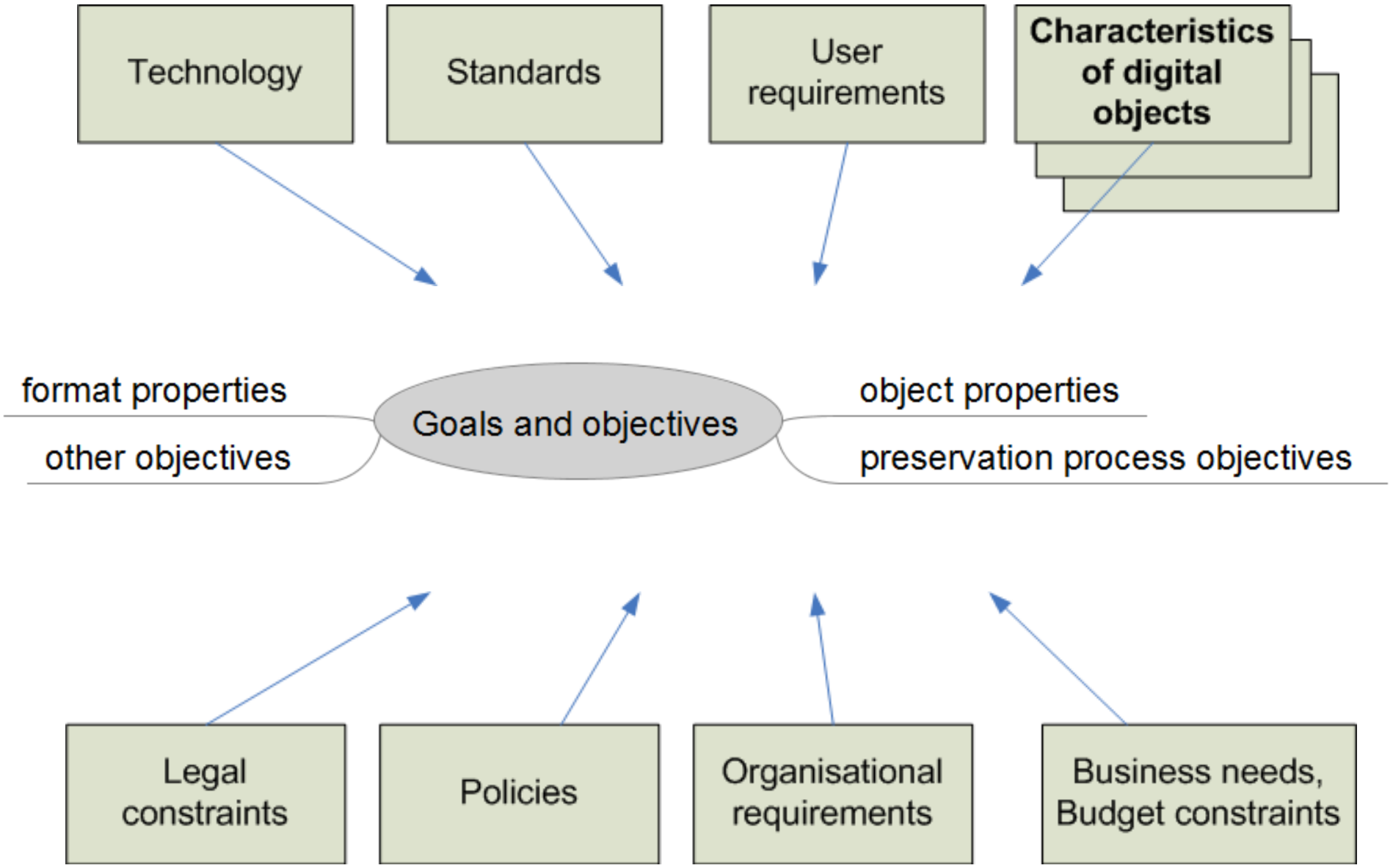
- What are our goals and objectives?
- How do we measure achievement of our goals?
- Which drivers have an influence on which objectives?
- Define complementary criteria for all objectives
  - Trade-offs between objectives might eventually be necessary
    - Usability vs. authenticity
    - Structure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

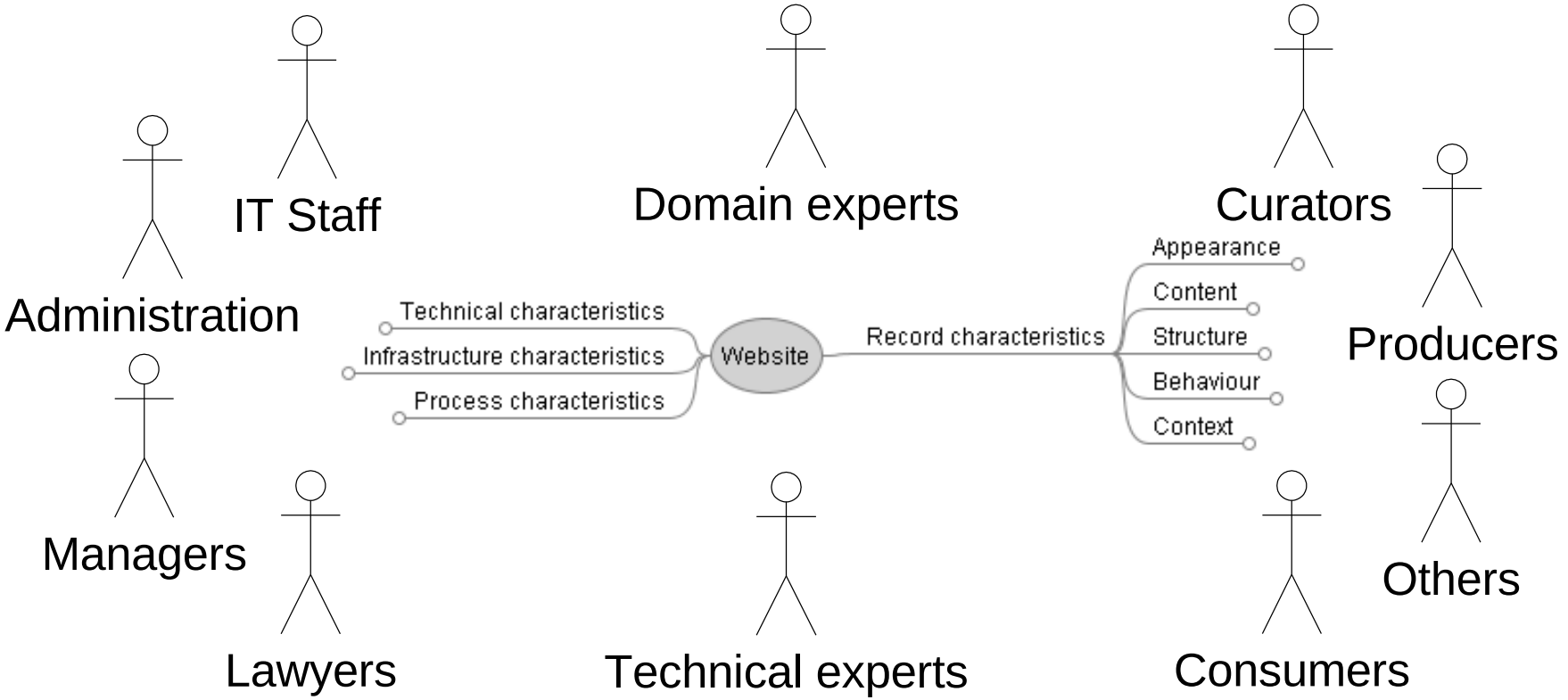
- Created top-down or bottom-up
  - Start from high-level goals and break down to specific criteria
  - Collect criteria and organize in tree structure

Influence Factors

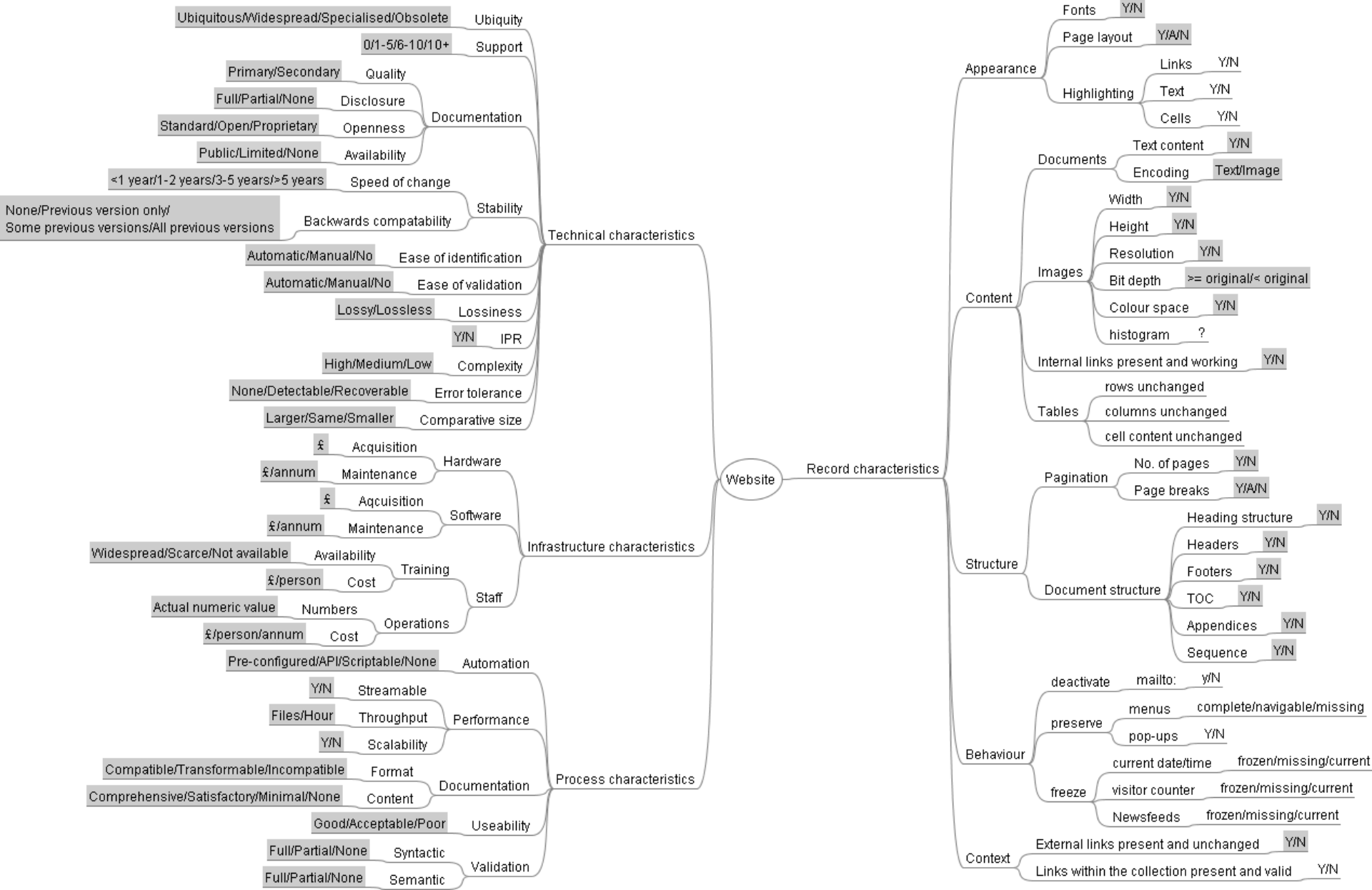


Stakeholders

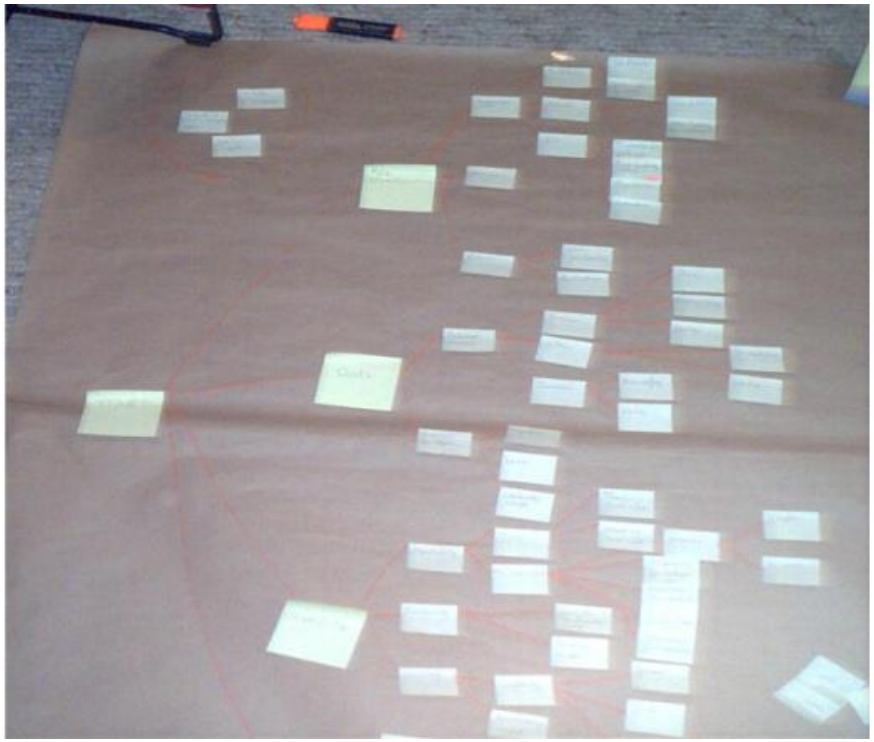
- Input needed from a wide range of persons, depending on the institutional context and the collection

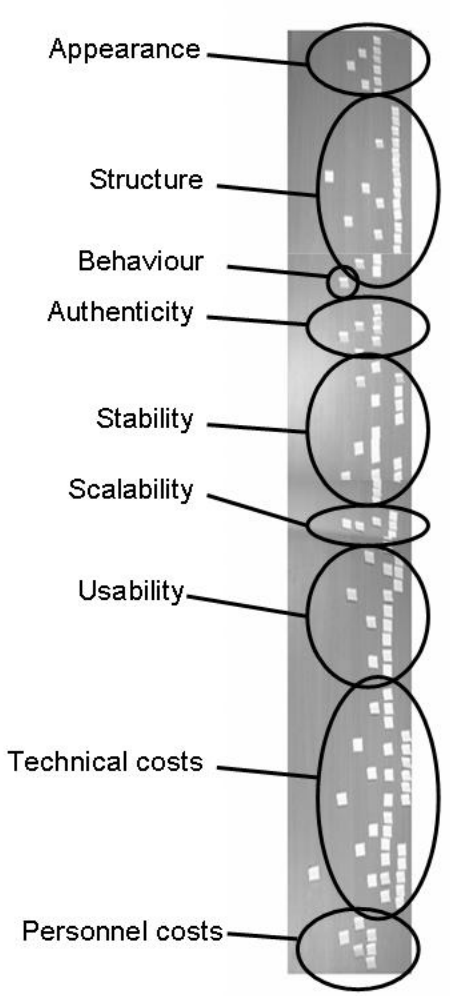


An Objective Tree

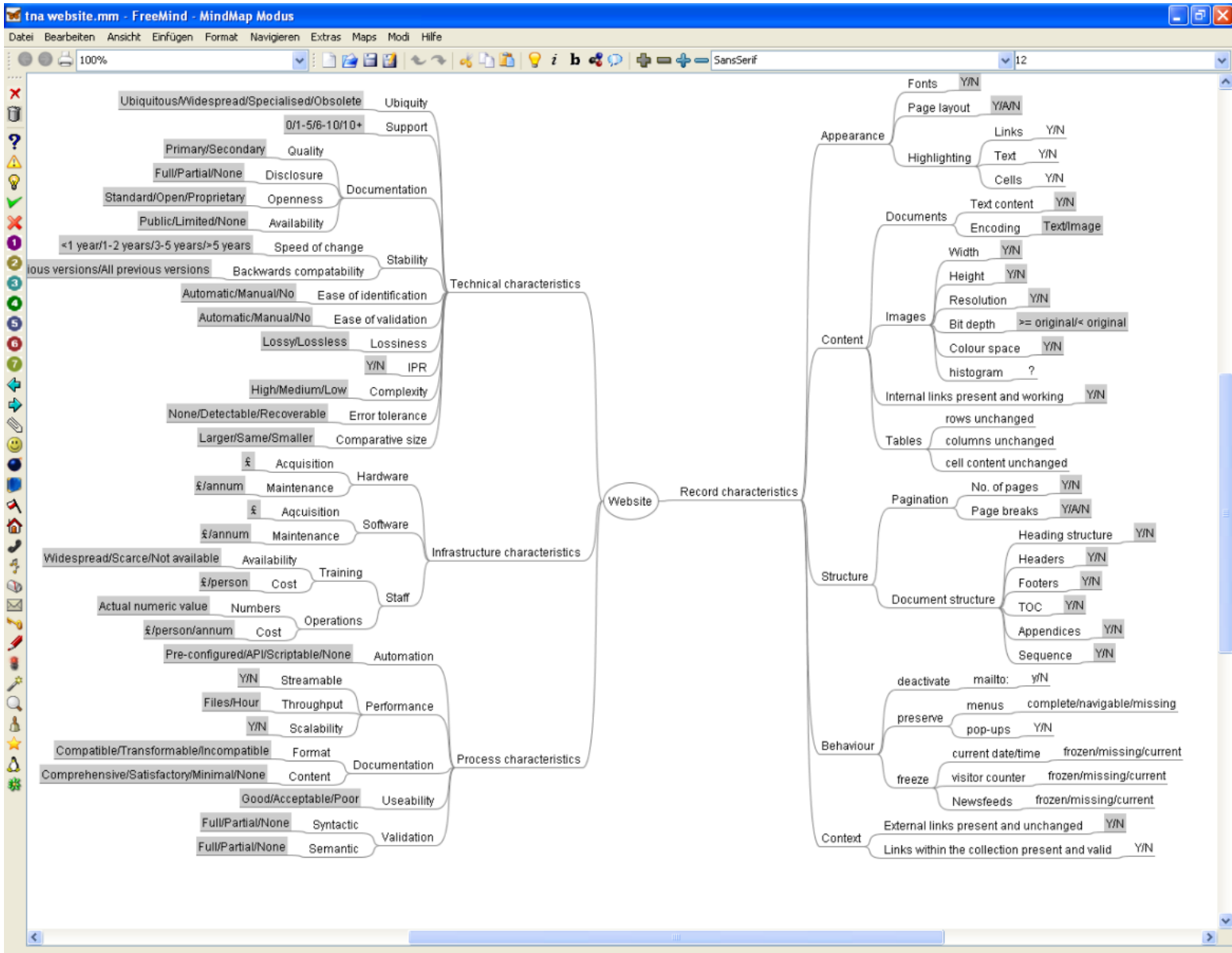


Analog…




Appearance
Structure
Behaviour
Authenticity
Stability
Scalability
Usability
Technical costs
Personnel costs

… or born-digital



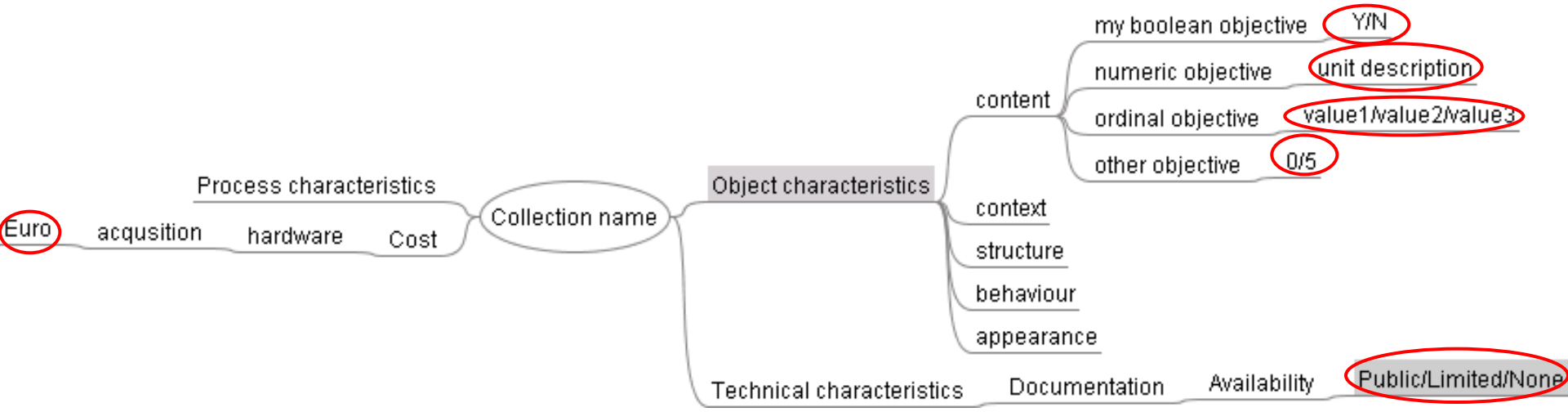
Assign Measurable Units

- 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)
- Ordinal: define the possible values
- *(Subjective 0-to-5)*



Models for decision factors

- Format Properties
  - Library of Congress format evaluation http://www.digitalpreservation.gov/formats/
  - PRONOM format evaluation
- Software Quality
  - ISO 25010 SQUARE: Standardised software quality model
- Object properties
  - Formats and their properties
  - Representation Instances
  - Authenticity: Significant Properties

User perspective

- Goal of digital preservation is to serve (future) users in providing usable and authentic information
  - What is the preservation intent? http://www.dlib.org/dlib/january13/webb/01webb.html
- 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)
  - …

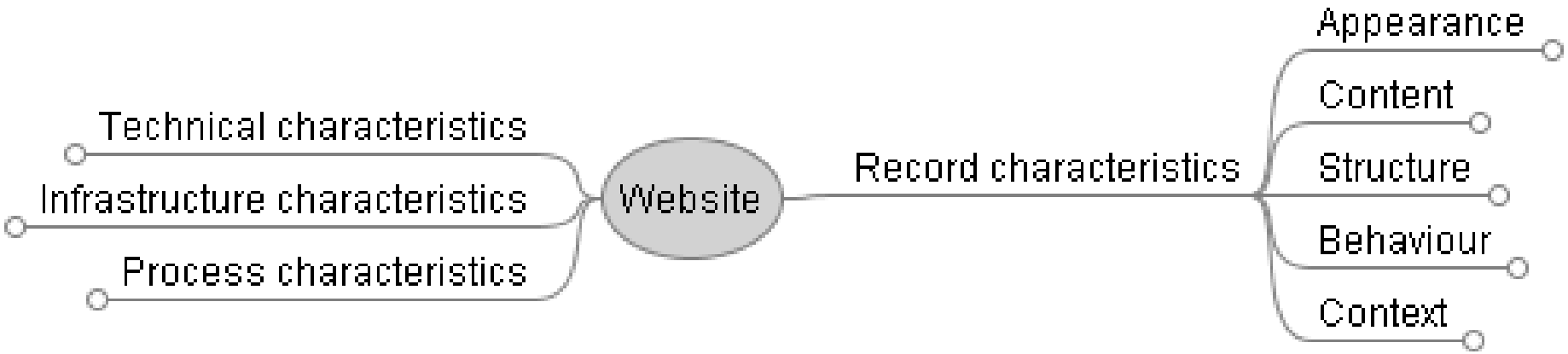
Significant properties of ‘digital objects’

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

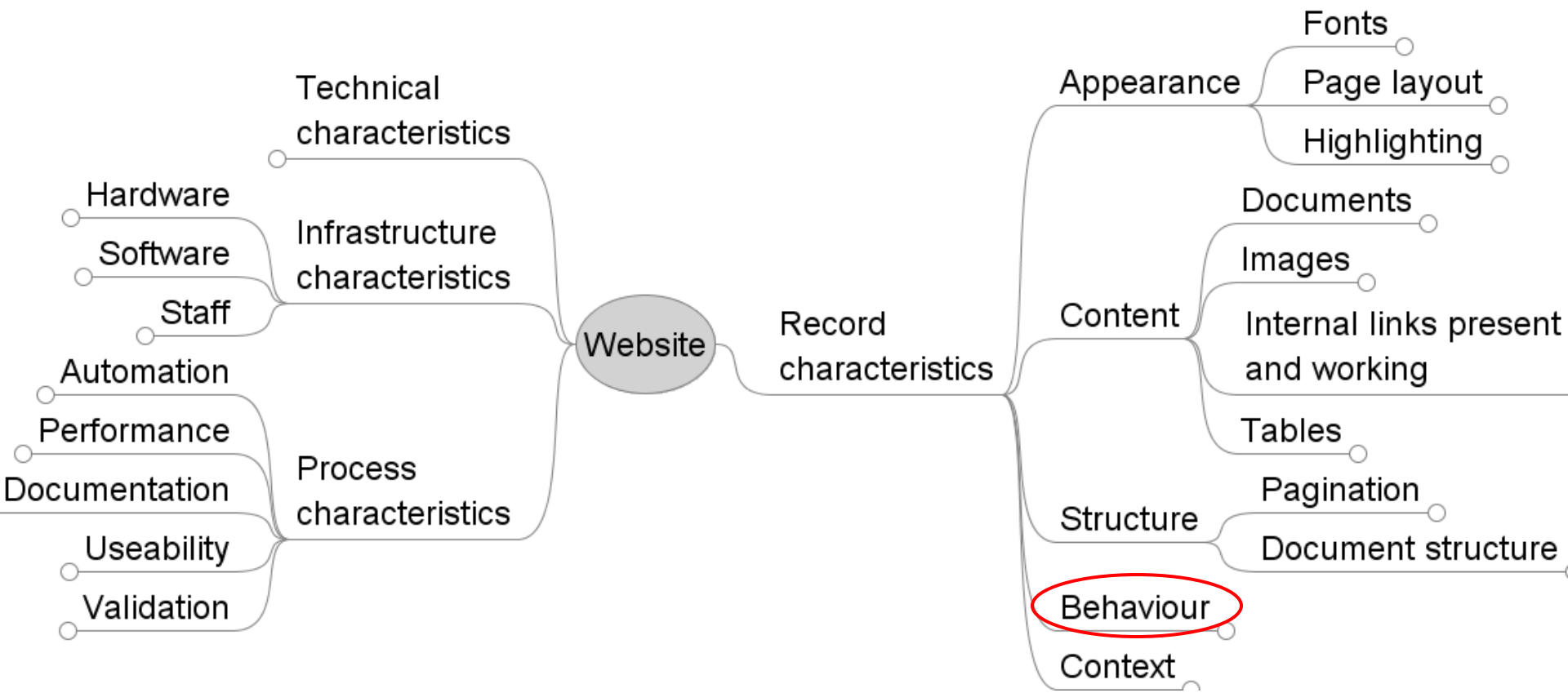
  - Content
  - Context
  - Structure
  - Appearance
  - Behaviour

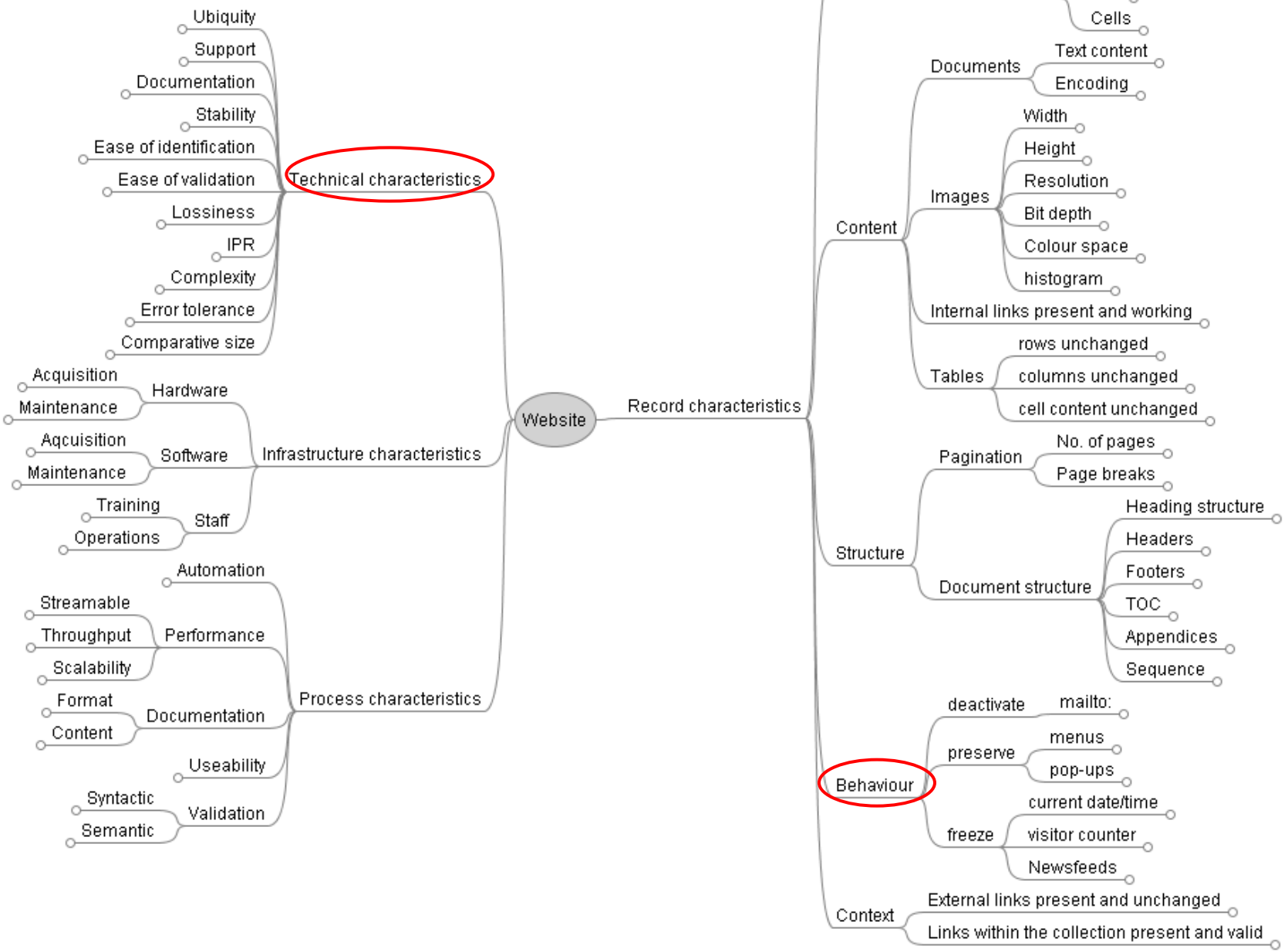
Case Study: Web archiving

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

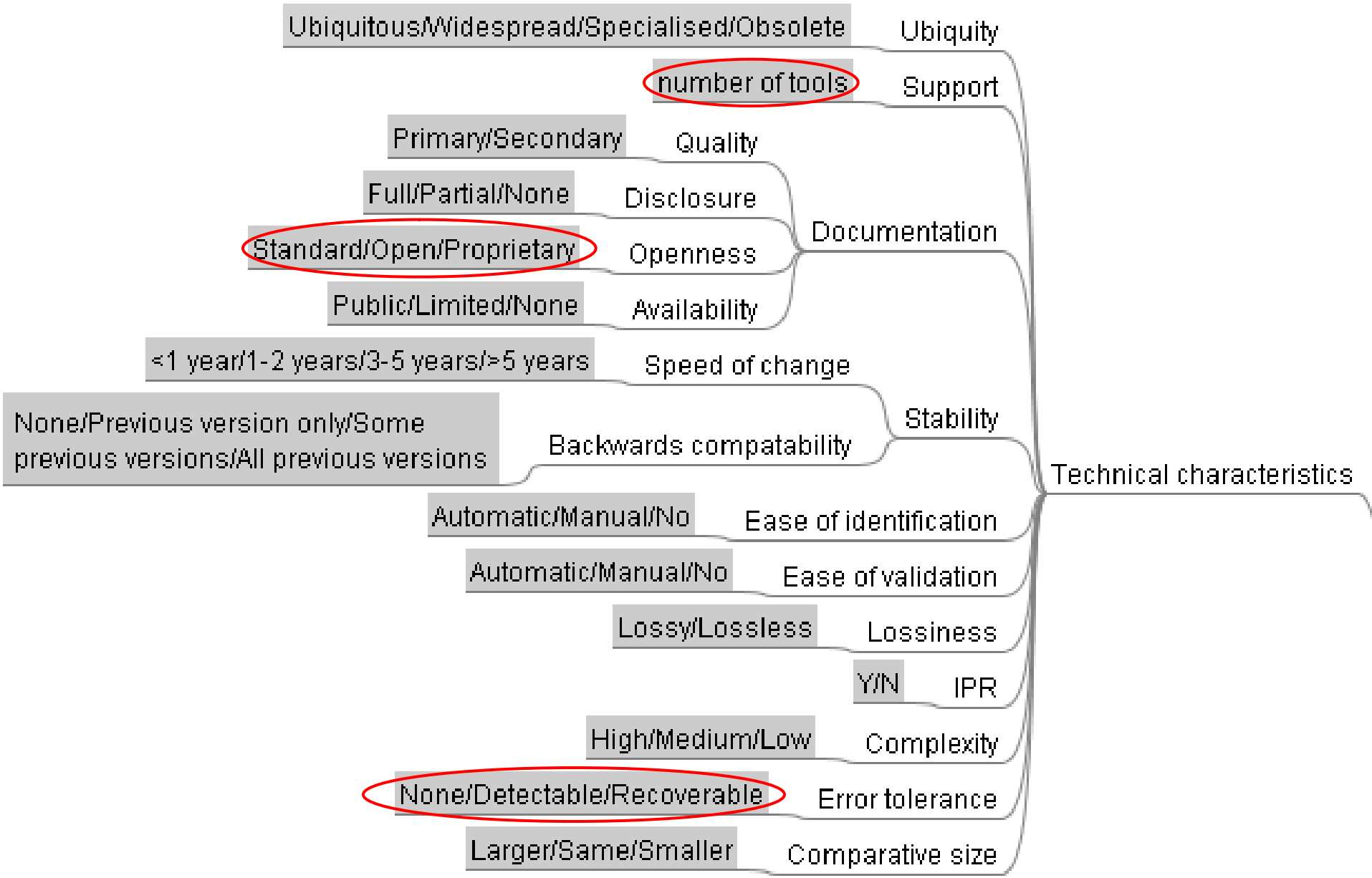


A bit more detail…

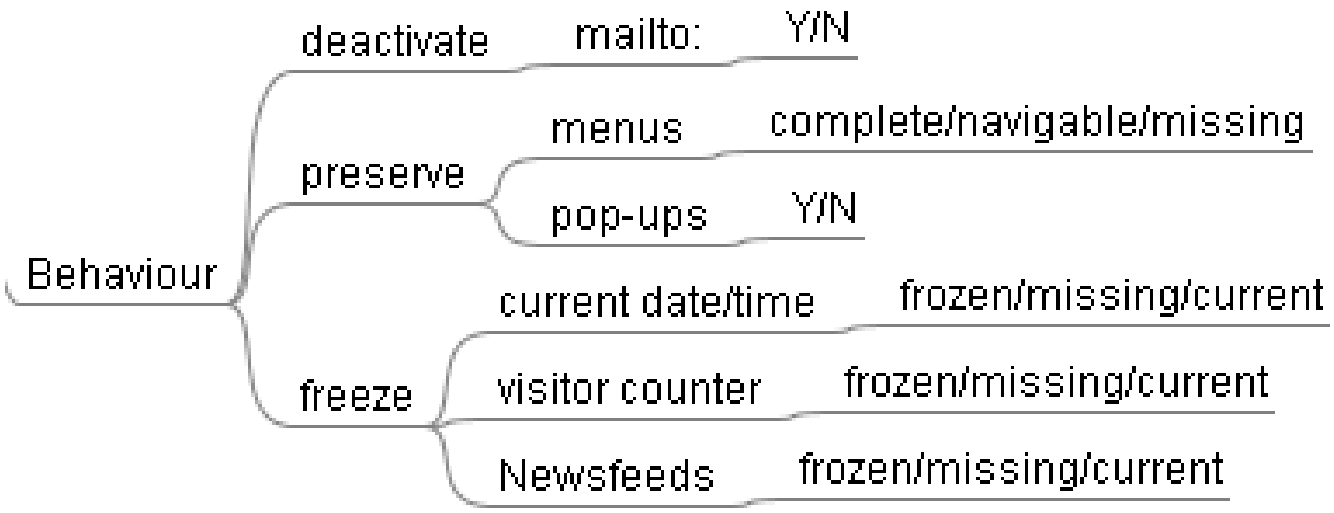



Website
Record characteristics
Technical characteristics
Ubiquity
Support
Documentation
Stability
Ease of identification
Ease of validation
Lossiness
IPR
Complexity
Error tolerance
Comparative size
Infrastructure characteristics
Hardware
Acquisition
Maintenance
Software
Aqcuisition
Maintenance
Staff
Training
Operations
Process characteristics
Automation
Performance
Streamable
Throughput
Scalability
Documentation
Format
Content
Useability
Validation
Syntactic
Semantic
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

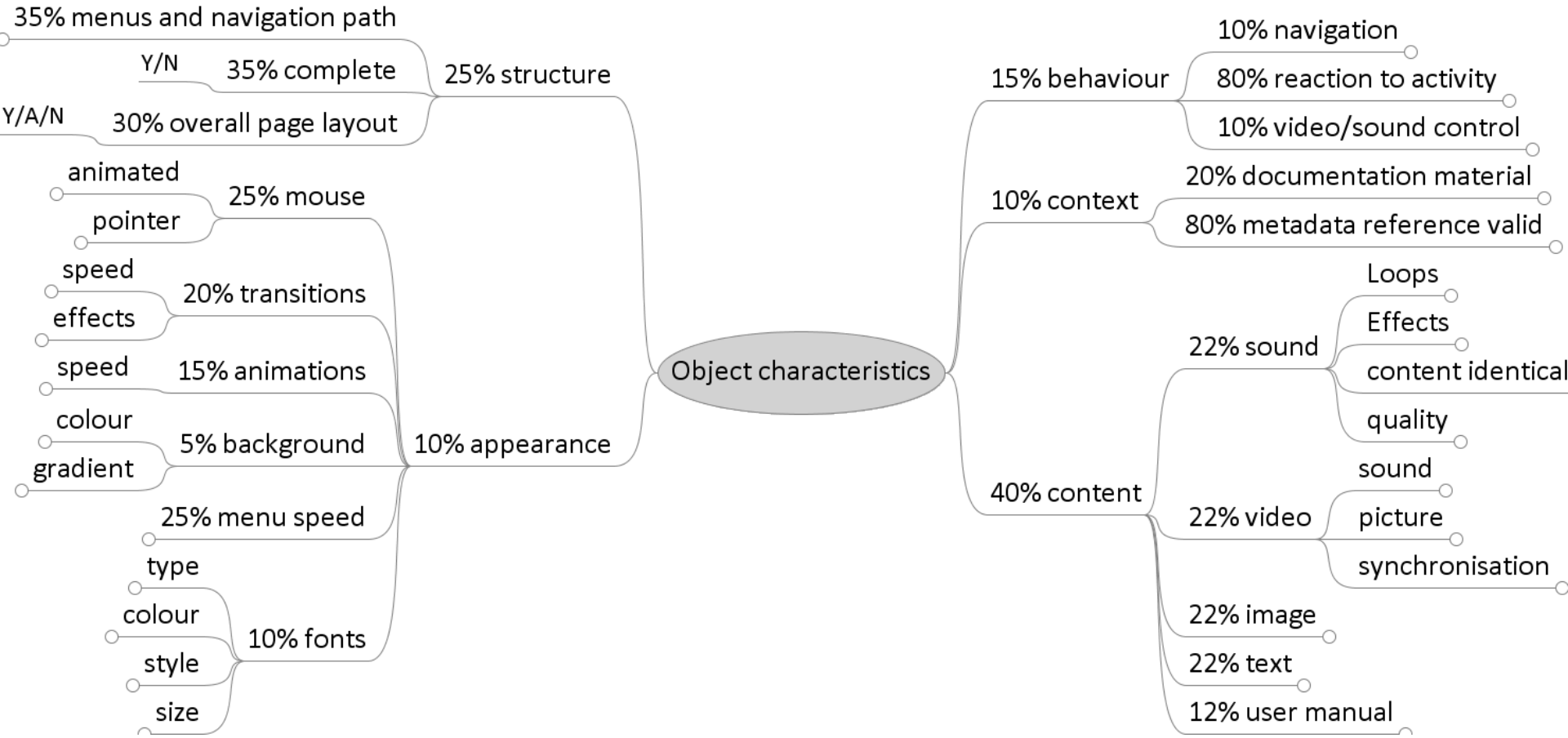


Behaviour



- 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

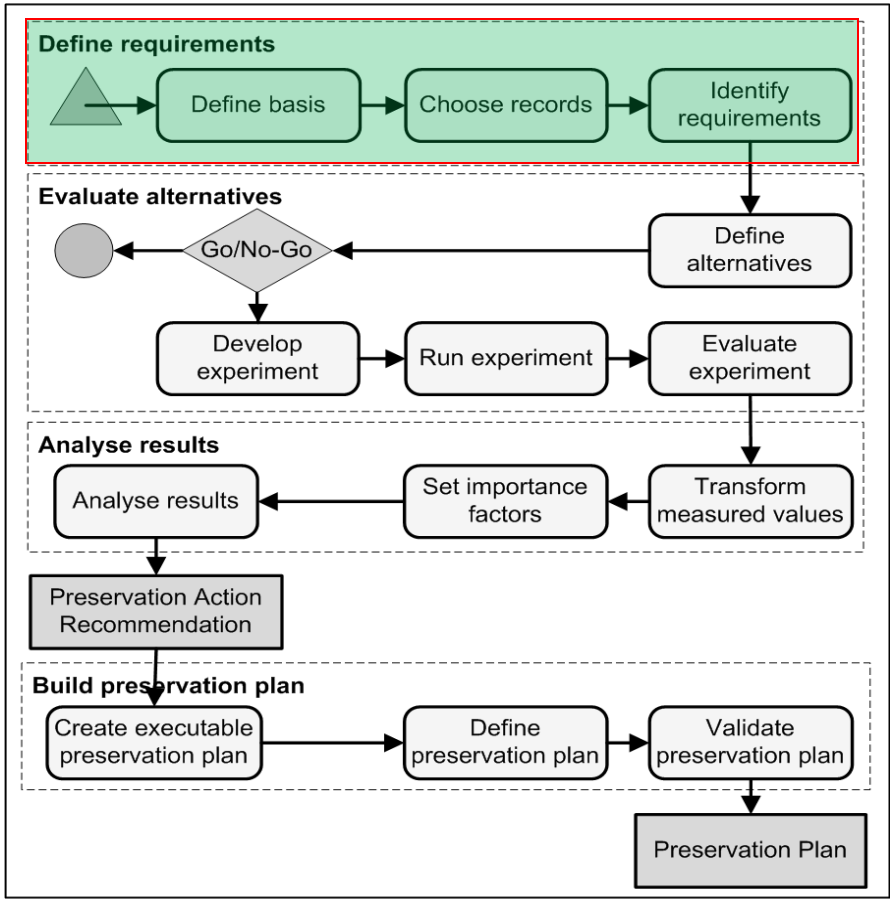
Node	Scale	Restriction
Object characteristics		
behaviour		
navigation	Ordinal	interactive and integrated/navigatable/none
reaction to activity		
mouse		
position	Boolean	
clicks	Boolean	
keyboard	Boolean	
video/sound control		
structure		
menus and navigation path	Ordinal	complete and free/partial (linear)/none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

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…)

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



Thank you for your attention!

- Next week:
  - How to plan: evaluation in Plato (hands-on)
  - Case studies: Examples of planning