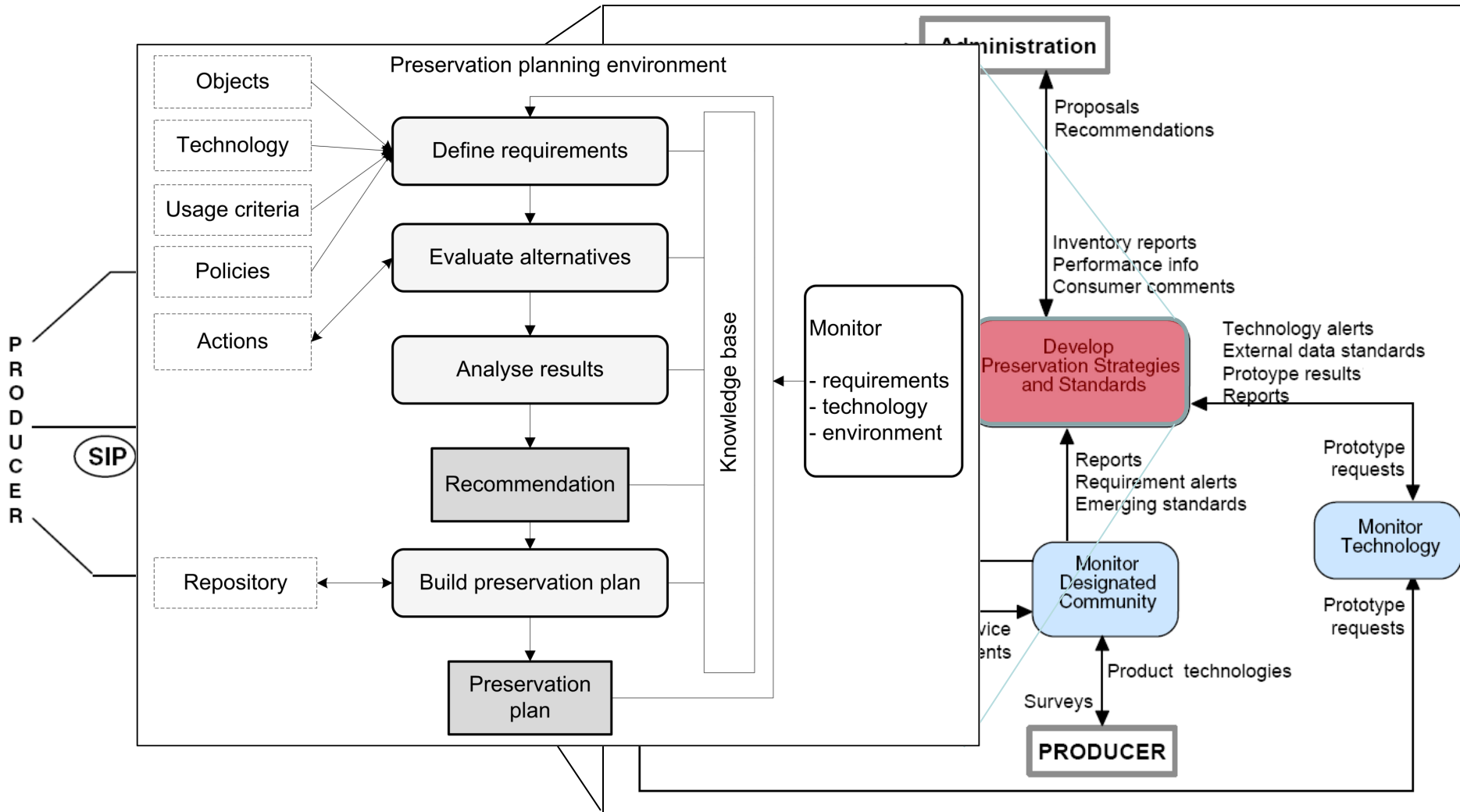# Preservation Planning 2

April 15, 2013

Hannes Kulovits
Vienna University of Technology

- Several preservation strategies developed

  - For each strategy: several tools available

    - For each tool: several parameter settings available

- How do you know which one is most suitable?

- What are the needs of your users? Now? In the future?

- Which aspects of an object do you want to preserve?

- What are the requirements?

- How to prove in 10, 20, 50, 100 years, that the decision was correct / acceptable at the time it was made?
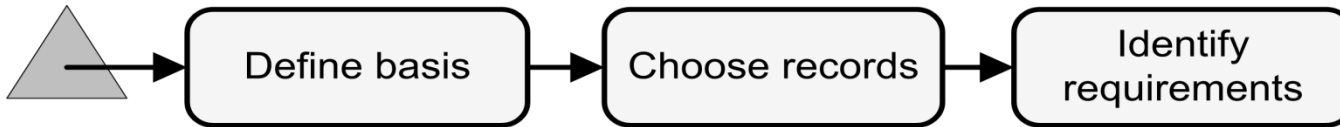
Evaluating preservation actions

- **Context: Trustworthy repositories**
  - Open Archival Information System (OAIS)
  - Trustworty repositories criteria (RAC)
  - 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
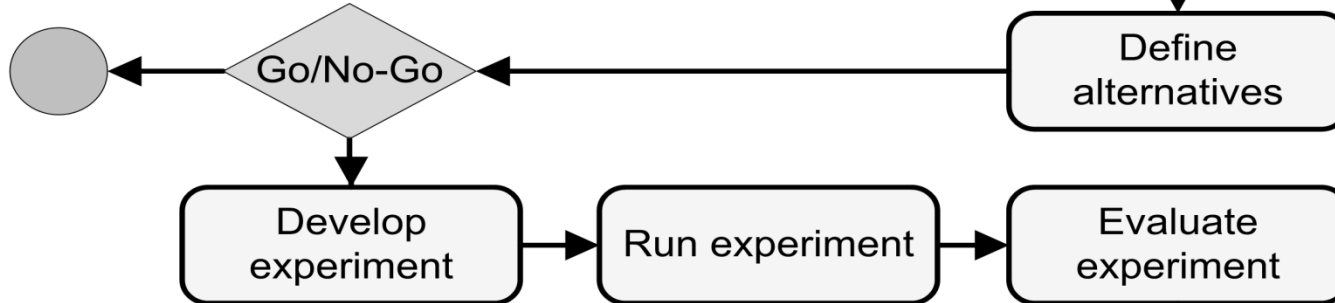 - Evaluate suitability of each potential action

Preservation Planning

Define requirements

Define basis → Choose records → Identify requirements

Evaluate alternatives

Go/No-Go ← Define alternatives

Develop experiment → Run experiment → Evaluate experiment

Analyse 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

- 'A preservation plan defines a series of preservation actions to be taken by responsible institution to address an identified risk for a given set of digital objects or records (called collection).'

- Considers
  - preservation policies
  - legal obligations
  - organisational and technical constraints
  - user requirements and preservation goals
- Describes the
  - preservation context
  - evaluated preservation strategies
  - resulting decision including the reasoning

- Repeatable, solid evidence

What is in a preservation plan?

- **Definition of scope**
  - What to preserve
- **Set of actions**
  - How to preserve it
- **Evaluation of action, 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 quality assurance and monitoring**
  - What to look out for
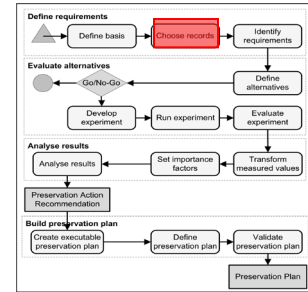
Define Basis
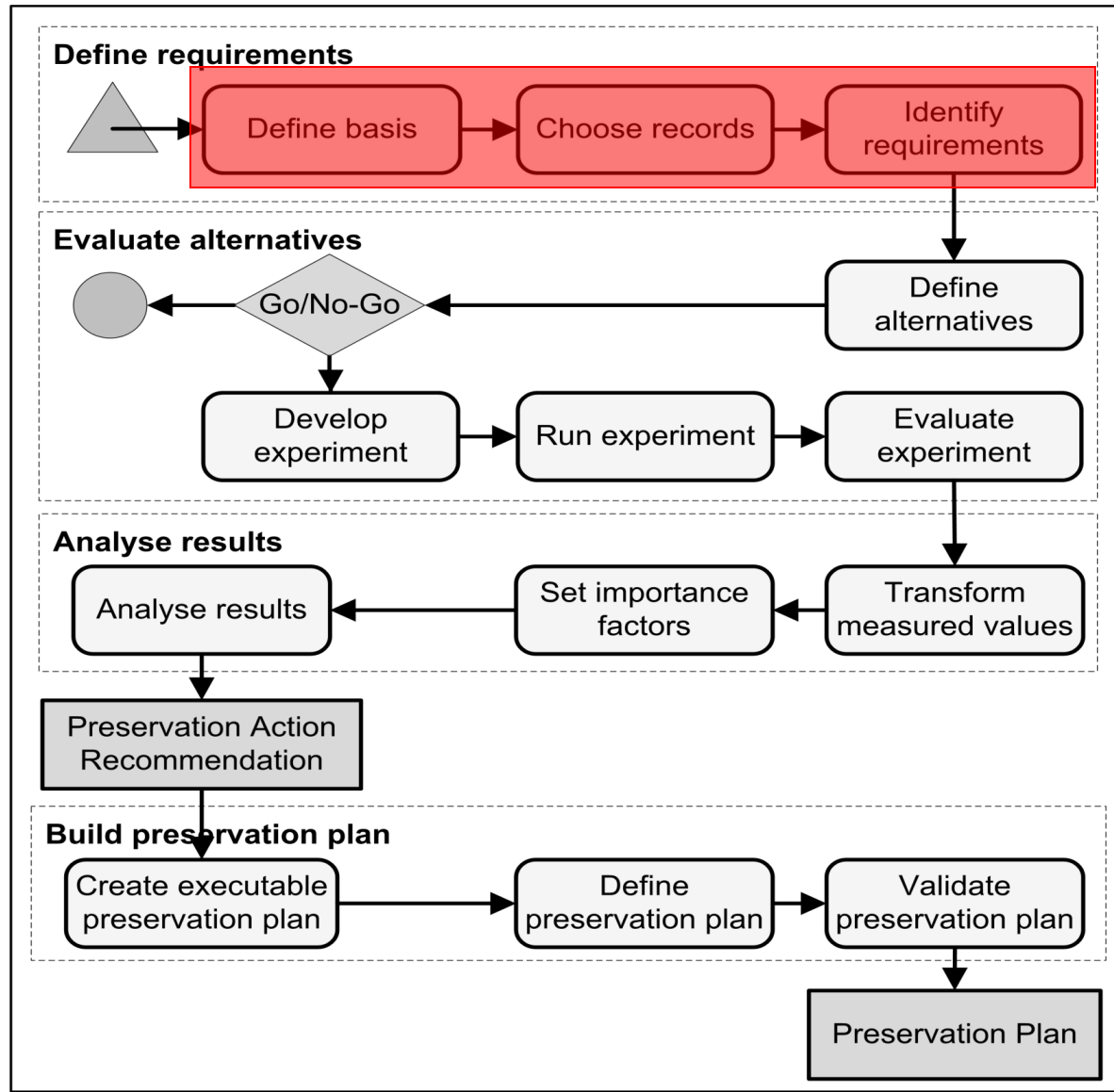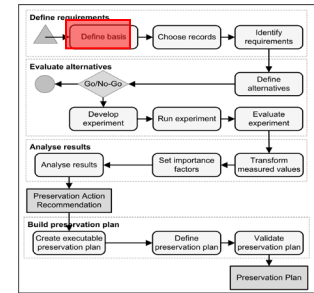
- **Description of the context**
  - Institutional settings
  - Legal obligations
  - User groups, target community
  - Organisational constraints

- **What are the objects?**

- **What are the fundamental requirements?**
  - Authenticity, reliability, integrity, usability
  - Metadata (for different purposes)
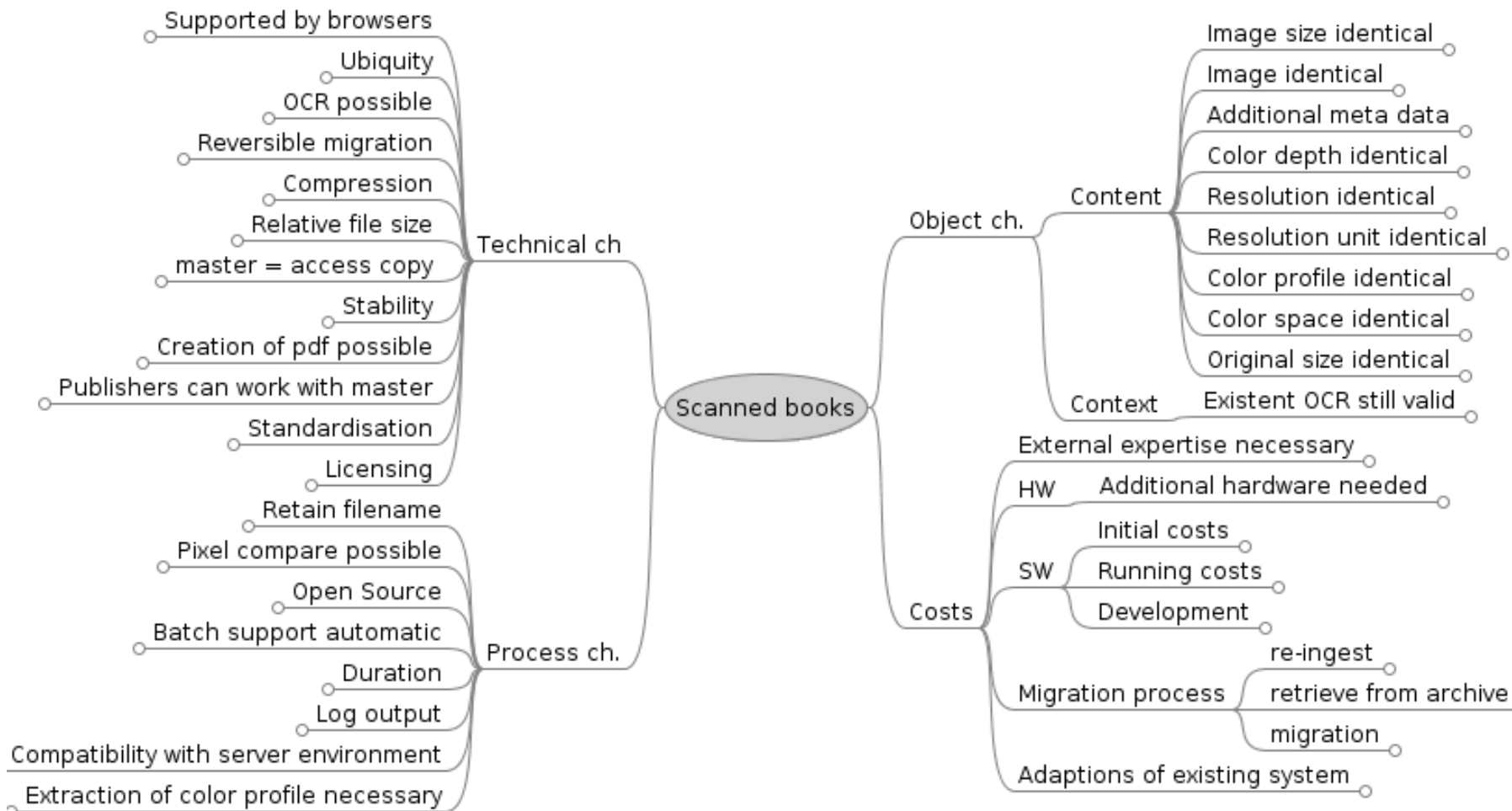
Choose Sample Objects

- **Identify consistent (sub-)collections**
  - Homogeneous type of objects (format, use)
  - To be handled with a specific (set of) tools

- **Describe the collection**
  - What types of objects?
  - How many?
  - Which format(s)?

- **Selection**
  - Representative for the objects in the collection
  - Right choice of sample is essential
  - They should cover all essential features and characteristics of the collection in question
  - As few as possible, as many as needed
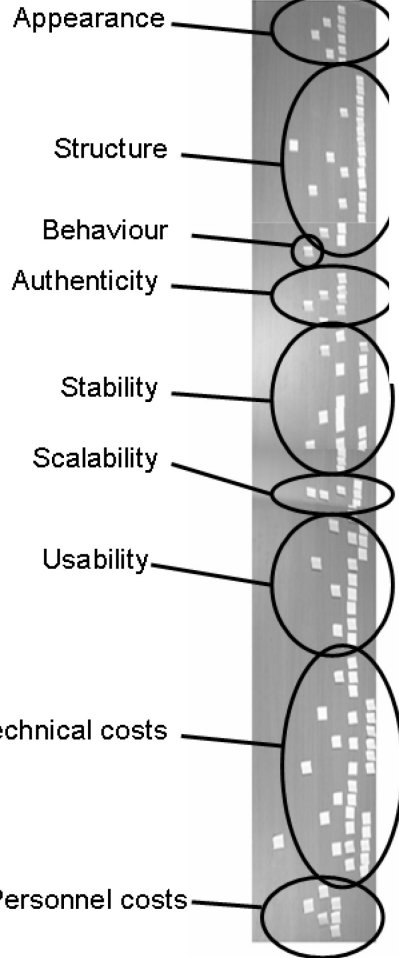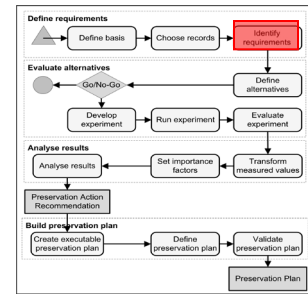 - Stratification across file type, size, content, time

Identify Requirements
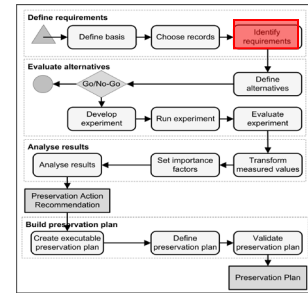
- Define all relevant goals and characteristics
  (high-level, detail) with respect to a certain application domain

- 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 requirements in relation to each other (hierarchical)

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

Scanned books

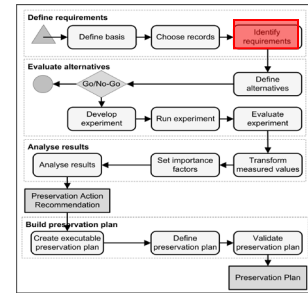## Technical ch
- Supported by browsers
- Ubiquity
- OCR possible
- Reversible migration
- Compression
- Relative file size
- master = access copy
- Stability
- Creation of pdf possible
- Publishers can work with master
- Standardisation
- Licensing

Process ch.
- Retain filename
- Pixel compare possible
- Open Source
- Batch support automatic
- Duration
- Log output
- Compatibility with server environment
- Extraction of color profile necessary

Object ch.
- Content
  - Image size identical
  - Image identical
  - Additional meta data
  - Color depth identical
  - Resolution identical
  - Resolution unit identical
  - Color profile identical
  - Color space identical
  - Original size identical
- Context
  - Existent OCR still valid

Costs
- External expertise necessary
- HW
  - Additional hardware needed
- SW
  - Initial costs
  - Running costs
  - Development
- Migration process
  - re-ingest
  - retrieve from archive
  - migration
- Adaptions of existing system

Identify requirements

Appearance

Structure

Behaviour

Authenticity

Stability

Scalability

Usability

Technical costs

Personnel costs

analogue…

… or digital

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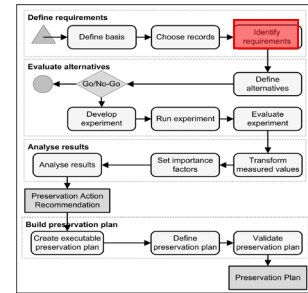
- **What needs to be preserved?**
  - Authenticity
  - Reliability
  - Integrity
  - Usability
  - Accuracy

  - Content
  - Context
 - Structure
  - Appearance
  - Behaviour

Some file format requirements

- Specifications available (syntax + semantics)
- Standardized (ISO, ANSI, ITEF, ...)
- Accepted and widely used
- Not covered by patent (license fees)
- Free of any cryptographical techniques (risk of losing keys)
- Free of compression

Assign Measurable Units
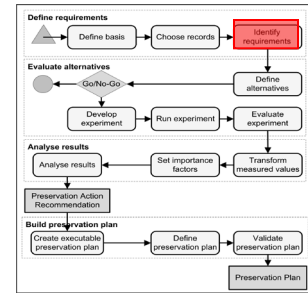
- **Leaf criteria should be objectively measurable**
  - Seconds per object
  - Euro per object
  - Bits of colour depth

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

➢ Quantitative results

Result of Phase 1
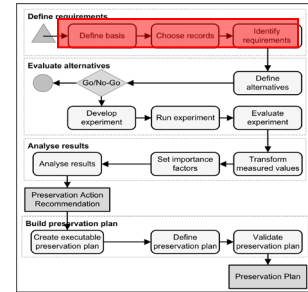
- **Defined and documented the context of a preservation problem**
  - Which types of objects
  - Which environment
  - What are the obligations and constraints
- **Defined and documented representative samples for performing experiments**
- **Defined and documented goals and requirements**

Define Alternatives
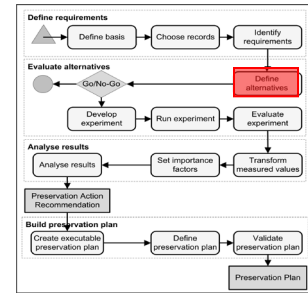
- **Given the type of object and requirements, what strategies are possible and which is most suitable**
  - Migration, emulation, other?

- **For each alternative, precise definition of**
  - Which tool (OS, version)
  - Which functions of the tool
  - Which parameters
 - Resources that are needed (human, technical, time and cost)

Go/No-Go
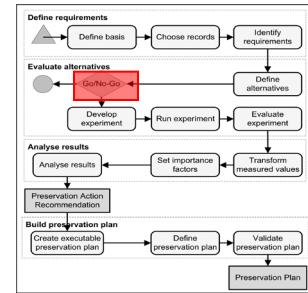
- Deliberate step for taking a decision whether it will be useful and cost-effective to continue the procedure, given

    - 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?

- Decision per alternative: go / no-go / deferred-go

Develop experiment
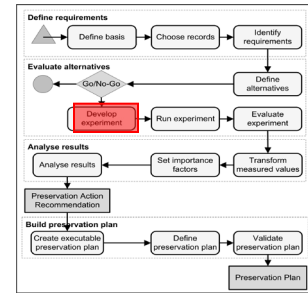
- **Plan for each experiment**

  - steps to build and test SW components

  - HW set-up

  - Procedures and preparation

  - Parameter settings, capturing measurements (time, logs...)

- **Ensures that results are comparable and repeatable**

Run experiment
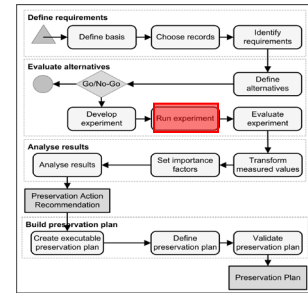
- Before conducting an evaluation or running an experiment, the experiment process as designed has to be tested
  - It may lead to re-design or even termination of the evaluation/ experiment process

- Capture process measurements
  (Start-up time, time per object, throughput, ...)

- The whole process needs to be documented

- Capture resulting objects, system logs, error messages, …

Evaluate experiment
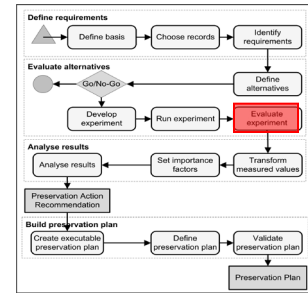
- Evaluate the outcome of each alternative for each leaf of the objective tree

- Evaluation analyses

  - Experiment measurements, results

  - Necessity to repeat an experiment

  - Undesired / unexpected results

- Includes both technical and intellectual aspects
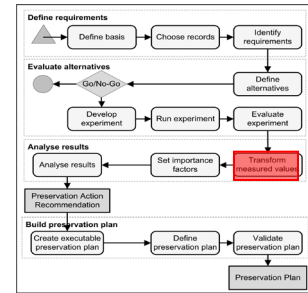
- Measures come in seconds, euro, bits, goodness values, …

- 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
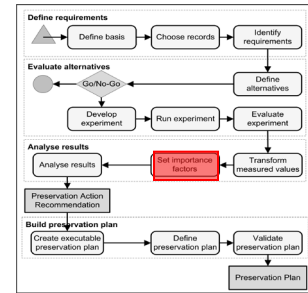
- Not all leaf criteria are equally important
- Depends on individual preferences and requirements
- By default, weights are distributed equally
- Adjust relative importance of all siblings in a branch
- Weights are propagated down the tree to the leaves
- Influence on the final ranking
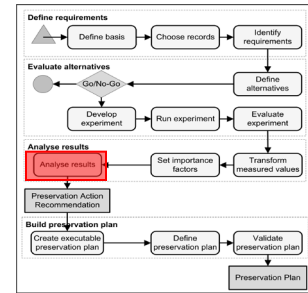
- **Aggregate values in Objective Tree**
  - Multiply the transformed measured values in the leaf nodes with the leaf weights
  - Sum up the transformed weighted values over all branches of the tree
  - Creates performance values for each alternative on each of the sub-criteria identified

- **Single performance value for each alternative**
  - to rank the alternatives
  - for each sub-set of criteria to identify the best combination of alternatives
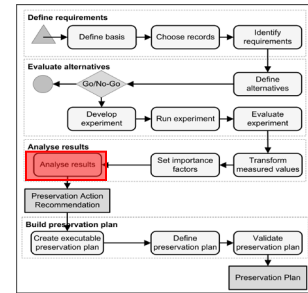
- Sensitivity Analysis: Analysis of the influence of small changes in the weight on the final value

- Basis for making Informed, well-documented, repeatable, accountable decisions

- 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
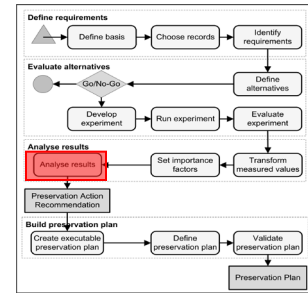
Example: Electronic documents

Alternative	Total Score Weighted Sum	Total Score Weighted Multiplication
PDF/A (Adobe Acrobat 7 prof.)	4.52	**4.31**
PDF (unchanged)	**4.53**	0.00
TIFF (Document Converter 4.1)	4.26	3.93
EPS (Adobe Acrobat 7 prof.)	4.22	3.99
JPEG 2000 (Adobe Acrobat 7 prof.)	4.17	3.77
RTF (Adobe Acrobat 7 prof.)	3.43	0.00
RTF (ConvertDoc 4.1)	3.38	0.00
TXT (Adobe Acrobat 7 prof.)	3.28	0.00

- Deactivation of scripting and security are knock-out criterium (PDF)
- RTF is weak in *Appearance* and *Structure*
- Plain text doesn't satisfy several minimum requirements

Build Preservation Plan
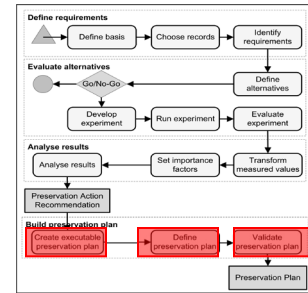
- **Create executable elements of preservation plan**
  - Sequence of preservation actions to call, parameters, …
  - Automatic steps + manual interventions where required
  - Automatic verification of results during deployment

- **Define preservation plan**
  - Create PP based on evidence produced during the PP process
  - Verify completeness of PP

- **Seek approval and validation of Preservation Plan**
  - Management activity according to OAIS
  - Sign and deploy

- 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

Preservation Planning with Plato

- Implements previously described planning methodology
- Assists in analyzing the collection
  - Profiling, analysis of sample objects via Pronom and other services
- Allows creation of objective tree
  - Within application or via import of mindmaps
- Allows the selection of Preservation action tools
- Runs experiments and documents results
- Easier re-evaluation of experiments
- Allows definition of transformation rules, weightings
- Performs evaluation, sensitivity analysis,
- Provides recommendation (ranks solutions)

Thank you very much for your attention!

Questions?

Activities and Roles/Stakeholders Involved

Core Activities
Control policy specification
Content profile generation
Document organisational context
Document relevant influence factors of the context
Define scope, select content
Specify user community and access requirements
Requirements specification
Select options to be considered
Diagnose options
Assess options against requirements
Assign responsibilites for putting plan into operation

Who is doing what?

Roles/Stakeholder Involved
Executive Management
Repository Manager
Technology Manager
Operational Manager
Repository Operator
Technology Operator
Solution Provider
Producer
Depositor
Consumer