




Preservation planning

and Bitstream Segment Graphs
...and the exercises

April 30, 2009

Christoph Becker
Vienna University of Technology
www.ifs.tuwien.ac.at/~becker


FACULTY OF INFORMATICS



Agenda

- Preservation Planning 1: Overview
 - What is a preservation plan?
 - How to create a preservation plan
 - High-level requirements
 - Exercise
- Break
- Bitstream Segment Graphs
- DP-UE Exercises presentation, Q&A


FACULTY OF INFORMATICS



Motivation

- Variety of solutions and tools exist
- Each strategy has unique strengths and weaknesses
- Requirements vary across settings
- Decision on which solution to adopt is very complex
- Documentation and accountability is essential
- Preservation planning assists in decision making
- Evaluation of strategies on representative sample content


FACULTY OF INFORMATICS



What is a preservation plan?

- Definition of scope
 - What to preserve
- Set of actions
 - How to preserve it
- Evaluation of actions, recommendation for one
 - How to do it and why do it this way
- Documentation of actions and reasons
 - Why did we decide that
- Conditions for QA and monitoring
 - What to look out for


FACULTY OF INFORMATICS



What is a preservation plan?

- *A preservation plan defines a series of preservation actions to be taken by a responsible institution due to 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 and describes the preservation context, the evaluated preservation strategies and the resulting decision for one strategy, including the reasoning for the decision.*

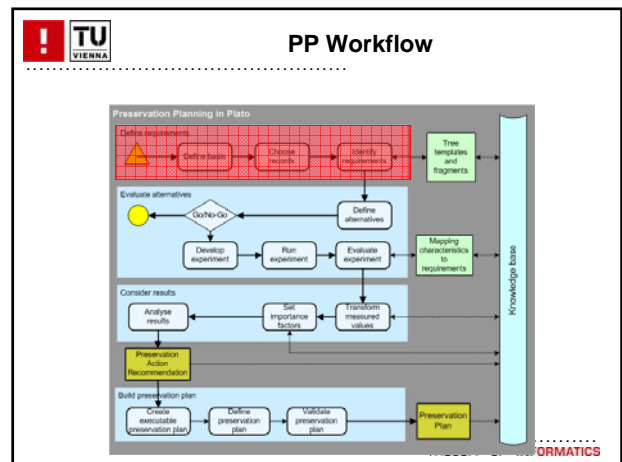
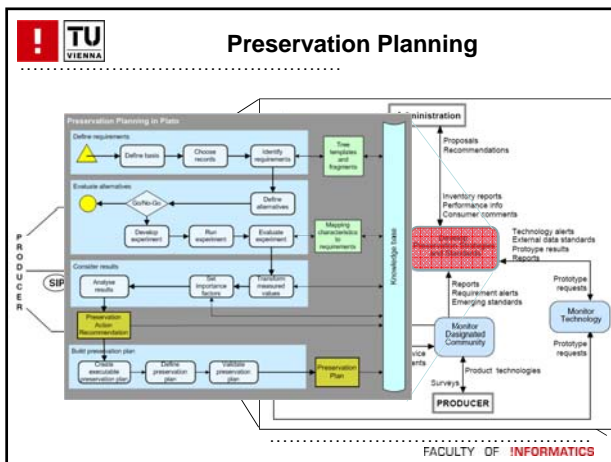
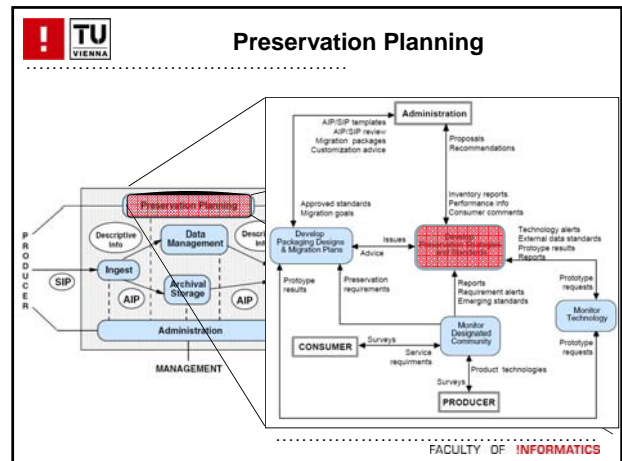
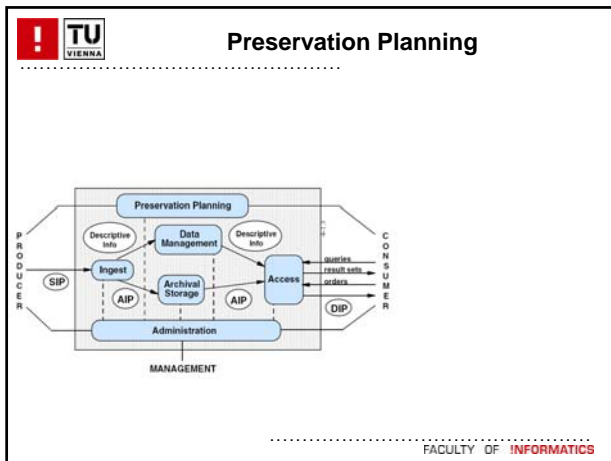
FACULTY OF INFORMATICS



Agenda

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

FACULTY OF INFORMATICS



> 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

FACULTY OF INFORMATICS

> Representative for the objects in the collection
 > Right choice of samples is essential
 > They should cover all essential features and characteristics of the collection in question
 > As few as possible, as many as needed
 > Often between 3-10

FACULTY OF INFORMATICS

TU VIENNA Identify requirements



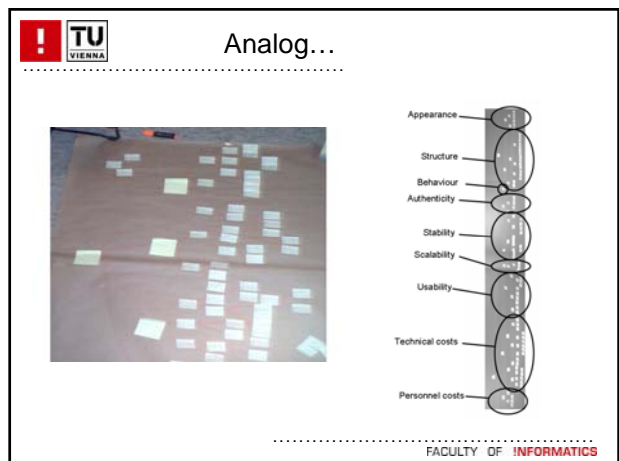
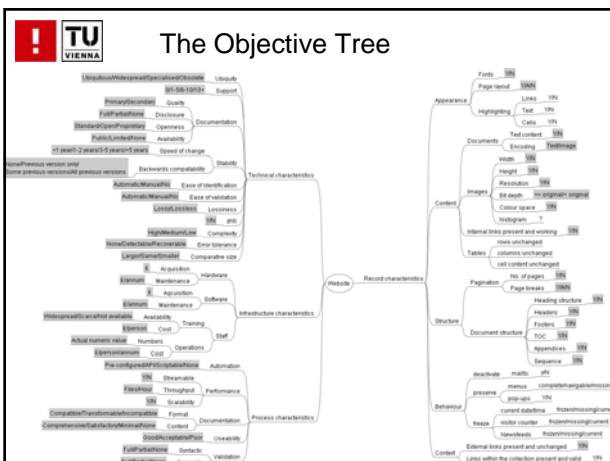
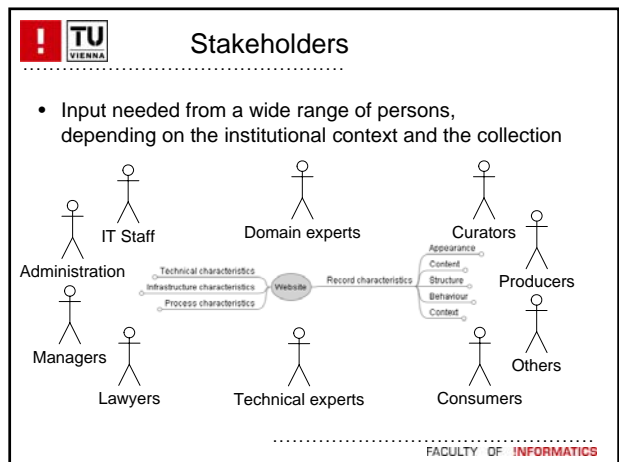
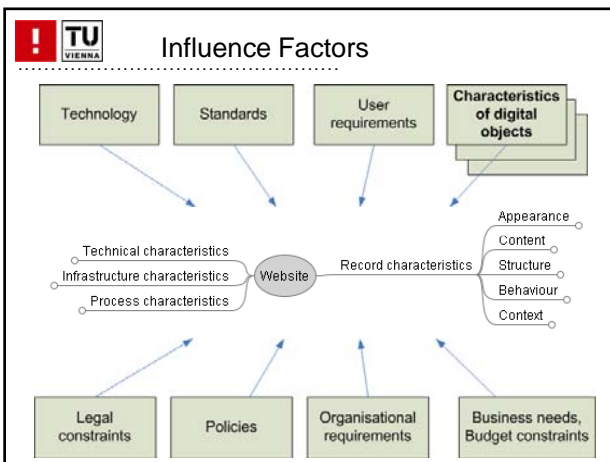
- Define all relevant goals and characteristics (high-level, detail) with respect to a given application domain
- Usually four major groups:
 - object characteristics (content, metadata ...)
 - record characteristics (context, relations, ...)
 - process characteristics (scalability, error detection,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

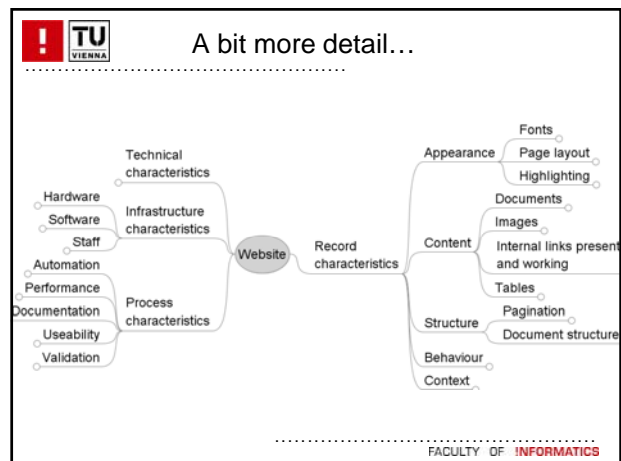
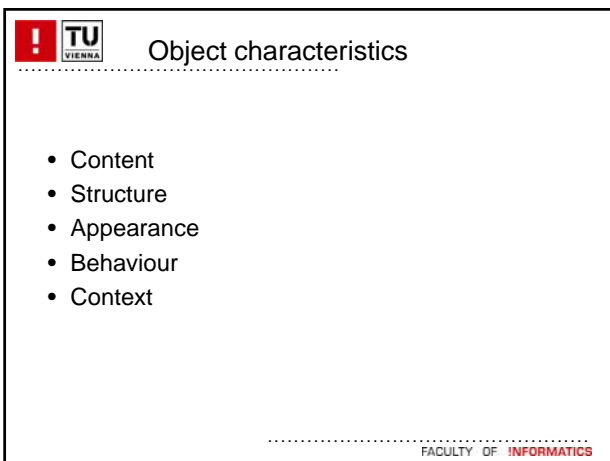
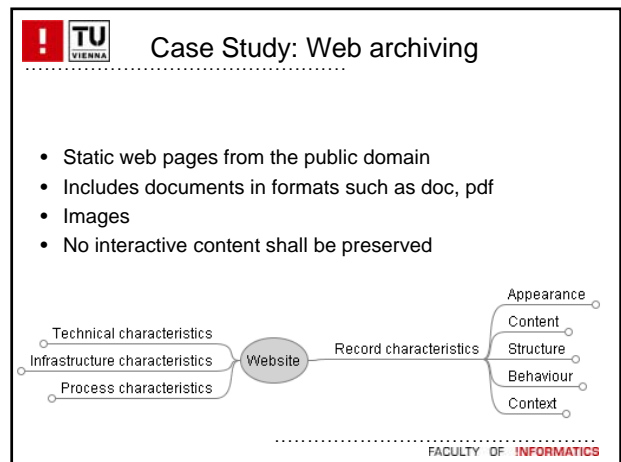
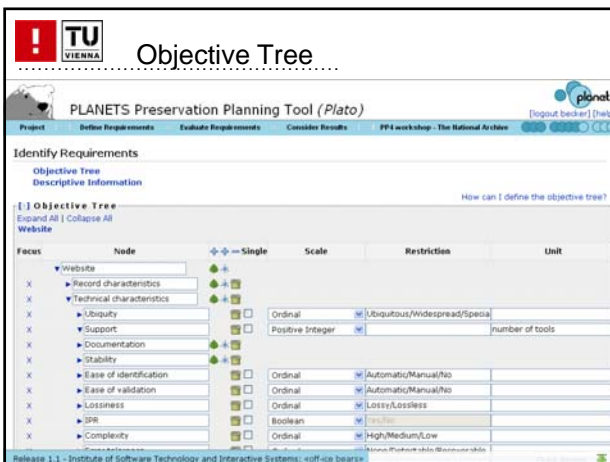
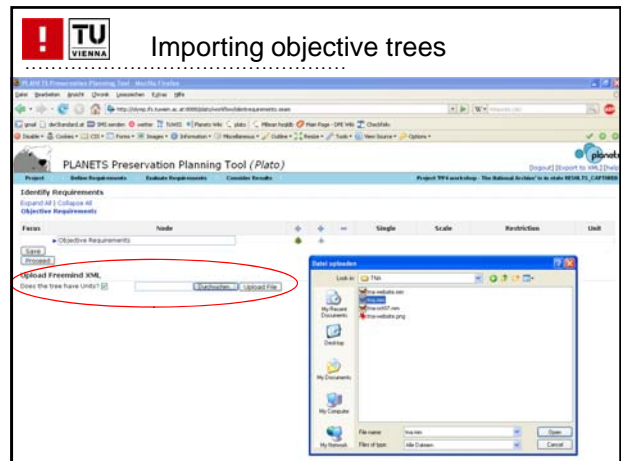
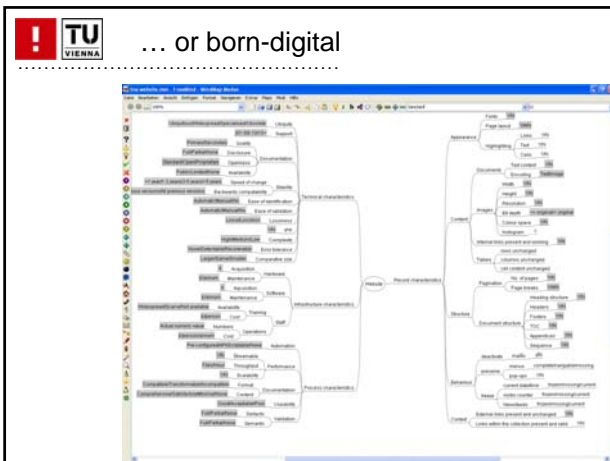
FACULTY OF INFORMATICS

TU VIENNA Identify Requirements: The Objective Tree

- Identify requirements and goals
- 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

FACULTY OF INFORMATICS





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

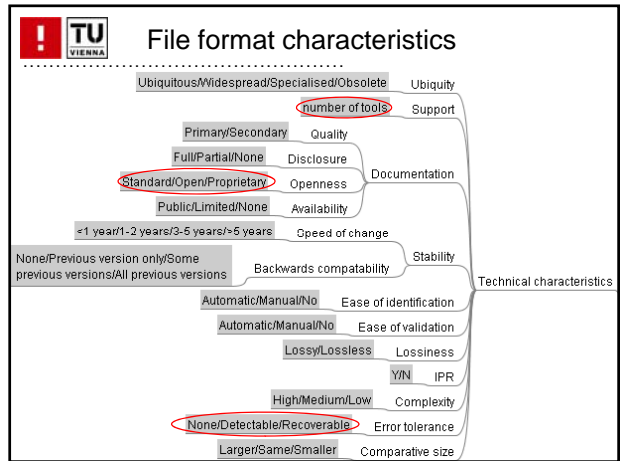
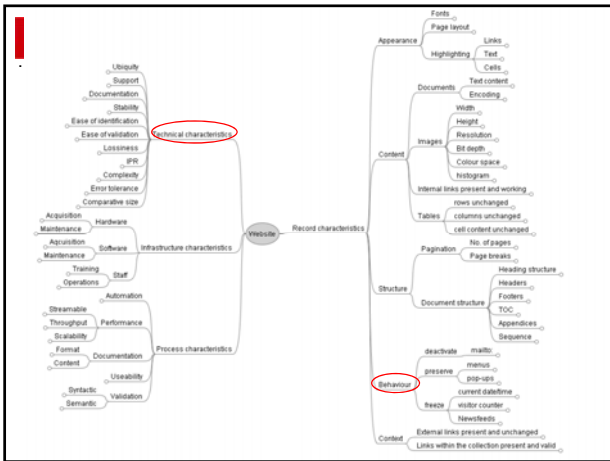
FACULTY OF INFORMATICS

Types of scales

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

A mind map centered on 'Object characteristics' with branches for 'Process characteristics', 'Collection name', 'Technical characteristics', 'Documentation', and 'Availability'. 'Object characteristics' further branches into 'content', 'context', 'structure', 'behaviour', and 'appearance'. 'Content' includes 'my boolean objective', 'numeric objective', 'ordinal objective', and 'other objective'. 'Context' includes 'my boolean objective', 'numeric objective', 'ordinal objective', and 'other objective'. 'Structure' includes 'my boolean objective', 'numeric objective', 'ordinal objective', and 'other objective'. 'Behaviour' includes 'my boolean objective', 'numeric objective', 'ordinal objective', and 'other objective'. 'Appearance' includes 'my boolean objective', 'numeric objective', 'ordinal objective', and 'other objective'. 'Collection name' includes 'acquisition', 'hardware', and 'cost'. 'Technical characteristics' includes 'Public/Limited/None'. 'Documentation' includes 'Public/Limited/None'. 'Availability' includes 'Public/Limited/None'.

FACULTY OF INFORMATICS

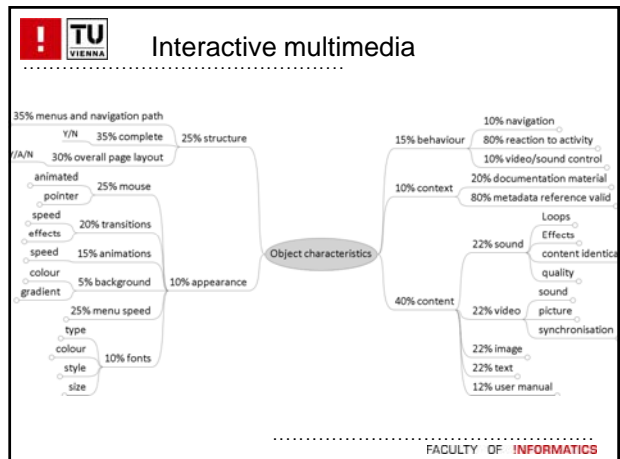


Behaviour

A mind map centered on 'Behaviour' with branches for 'deactivate', 'mailto', 'Y/N', 'menus', 'complete/navigable/missing', 'preserve', 'pop-ups', 'Y/N', 'current date/time', 'frozen/missing/current', 'freeze', 'visitor counter', 'frozen/missing/current', and '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?)

FACULTY OF INFORMATICS



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/non
reaction to activity		
mouse		
position	Boolean	
clicks	Boolean	
keyboard	Boolean	
video/sound control		
structure		
menus and navigation path	Ordinal	complete and free/partial (linear)/non
complete	Boolean	
overall page layout	Ordinal	Y/A/N

FACULTY OF INFORMATICS

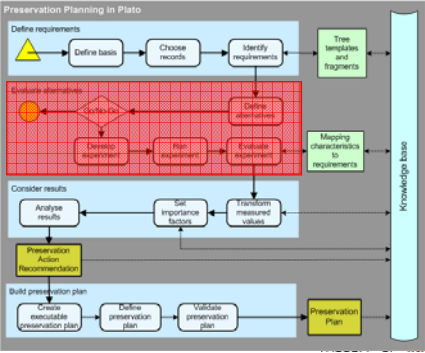
Results of Phase 1

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



FACULTY OF INFORMATICS


PP Workflow



FACULTY OF INFORMATICS

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



FACULTY OF INFORMATICS


Service discovery



FACULTY OF INFORMATICS


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,...)



FACULTY OF INFORMATICS


TU VIENNA **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?

FACULTY OF **INFORMATICS**


TU VIENNA **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)

FACULTY OF **INFORMATICS**


TU VIENNA **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

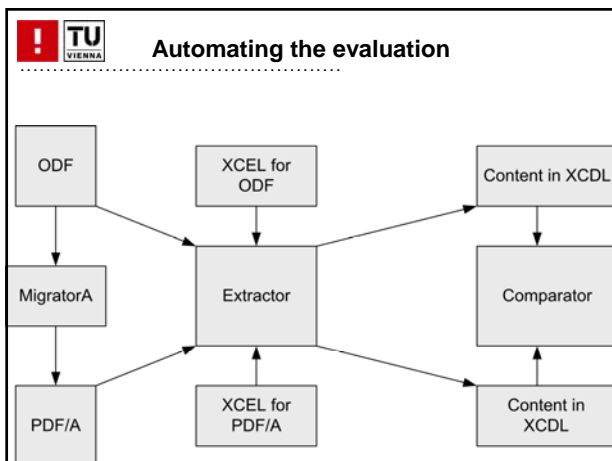
FACULTY OF **INFORMATICS**

TU VIENNA **Evaluate experiment**



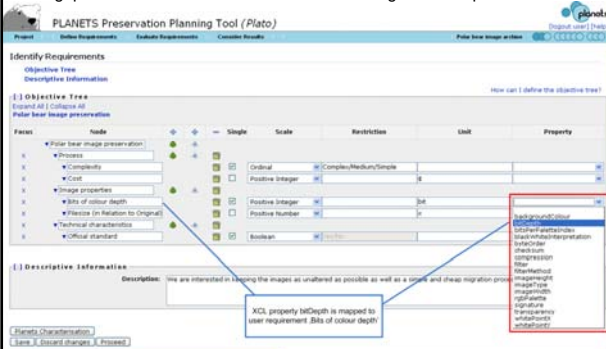
- Evaluate the outcome of each alternative for each leaf of the objective tree
- The evaluation will identify
 - Need for repeating the process
 - Unexpected (or undesired) results
- Includes both technical and intellectual aspects

FACULTY OF **INFORMATICS**

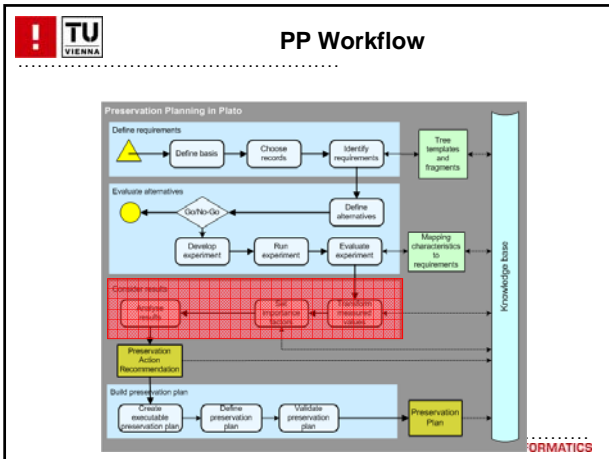


TU VIENNA **Automating the evaluation**

Close gap between technical characteristics and high-level requirements



Process	Node	Single	Scale	Restriction	Unit	Property
✖	Process					
✖	Complexity		Ordinal	Complex/Medium/Simple		
✖	Cost		Positive Integer			
✖	Image properties					
✖	Bits of colour depth		Positive Integer		bit	
✖	Filesize (in Relation to Original)		Positive Number			
✖	Technical characteristics					
✖	Official standard		Boolean			



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"

FACULTY OF INFORMATICS

Set importance factors

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

FACULTY OF INFORMATICS

Balancing weights

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

Buttons: Save, Proceed

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
 - Creates performance values for each alternative on each of the sub-criteria identified

FACULTY OF INFORMATICS

Analyse results

Focus	Name	Result
	Minimalist root node	PDF/A ToolA: 2.88 PDF/A ToolB: 3.19
X	Image properties	PDF/A ToolA: 0.60 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	InfRange 0-10	PDF/A ToolA: 0.10 PDF/A ToolB: 0.60


! TU VIENNA Analyse results



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

FACULTY OF INFORMATICS


! TU VIENNA Analyse results



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

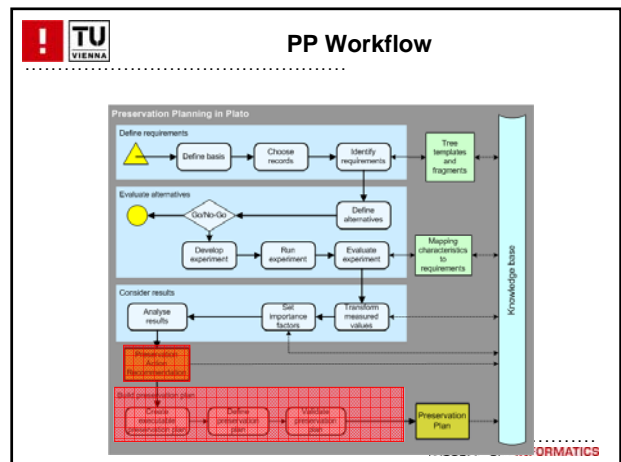
FACULTY OF INFORMATICS

! TU VIENNA Consider results




- The review of the results may help to refine
 - The evaluation process/procedure
 - The preservation planning environment itself
 - The evaluation metrics
 - Understanding of the essential characteristics of the objects,
 - and identify further evaluations, experiments
- The review should take into account all previous work done in the preservation planning environment
- The review should look at both the technical and intellectual aspects of digital objects

FACULTY OF INFORMATICS

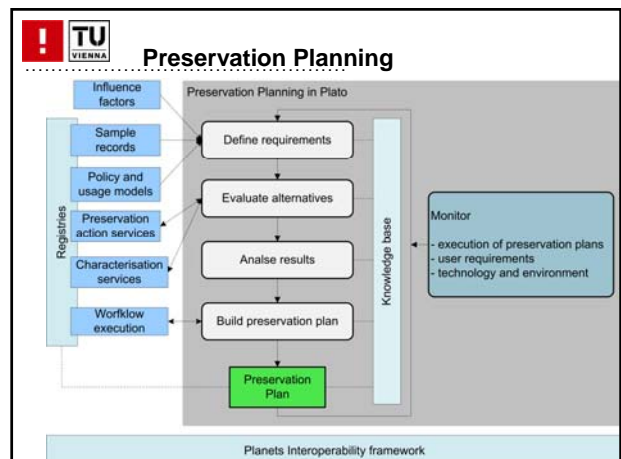



! TU VIENNA Build Preservation Plan



- 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 PP
 - Management activity according to OAIS
 - Sign and deploy

FACULTY OF INFORMATICS




 **Conclusions**

- A simple, methodologically sound model to specify and document requirements
- Repeatable and documented evaluation for informed and accountable decisions
- Set of templates to assist institutions
- Generic workflow that can easily be integrated in different institutional settings
- **Plato:**
Tool support to perform solid, well-documented analysis
- Provides basic preservation plan

<http://www.ifs.tuwien.ac.at/dp/plato>

.....
FACULTY OF **INFORMATICS**

 **Questions?**


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

.....
FACULTY OF **INFORMATICS**

 **Practice time!**


A digital preservation scenario

.....
FACULTY OF **INFORMATICS**

 **Context: National library**


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

.....
FACULTY OF **INFORMATICS**

 **Scanned newspapers archive**

- One of the (many) collections we have is a set of newspaper scans
- ~800.000 GIF images, mostly black-and-white or grayscale
- Different size, resolution, age, created with different tools,...
- Should be preserved in an openly specified format, cheap but safe, and easily accessible to the public (online)

.....
FACULTY OF **INFORMATICS**

 **Create a preservation plan**

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

.....
FACULTY OF **INFORMATICS**

TU VIENNA

Create a preservation plan


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

FACULTY OF INFORMATICS

TU VIENNA

Create a preservation plan

- The planning tool **Plato** supports our exercise



FACULTY OF INFORMATICS

TU VIENNA

Questions?

...10 minutes break

- BSGs
- DP-UE examples

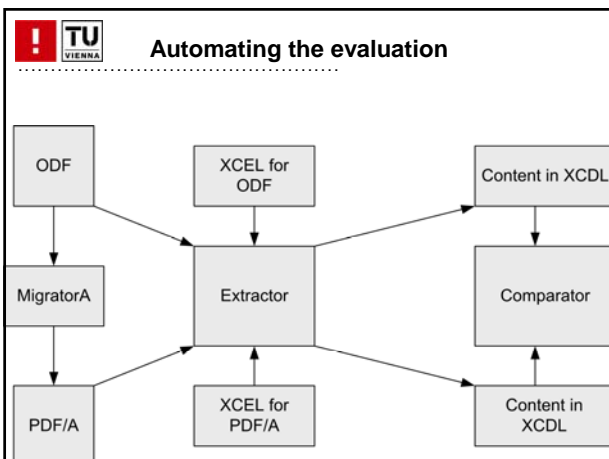
FACULTY OF INFORMATICS

TU VIENNA

In-depth characterisation approaches

- XCL...
 - eXtensible Characterisation Languages
 - XCDL, the description language
 - XCEL, the extraction language
- Bitstream Segment Graphs (BSG)
- Alternative, new approach based on reasoning and rules

FACULTY OF INFORMATICS

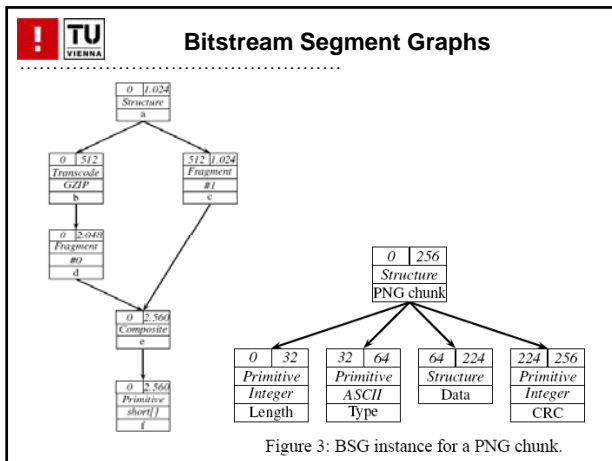


TU VIENNA

Bitstream Segment Graphs

- Use a graph to describe the structure of a file
- Define sets of rules used a reasoner to create such a graph
- General rule base, format-specific rules
- Reasoner calculates BSG and “coverage” of a file
- BSG editor allows construction and exploration of the map

FACULTY OF INFORMATICS



TU VIENNA Bitstream Segment Graphs

➤ DEMO

FACULTY OF INFORMATICS

TU VIENNA

Questions?

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

FACULTY OF INFORMATICS