Long-Term Preservation of Electronic Theses and Dissertations: A Case Study in Preservation Planning

Christoph Becker
Vienna University of Technology
Pereslavl, October 2007

Digital Preservation

- Everything is digital
- Digital objects have a short life span
  - Hardware stops working
  - Decay of media
  - Format obsolescence
  - Loss of metadata
- Digital preservation: Long-term storage and access to digital objects of all kinds
- Dominant strategies:
  - Migration
  - Emulation

planets TU VIENNA

Preservation Planning

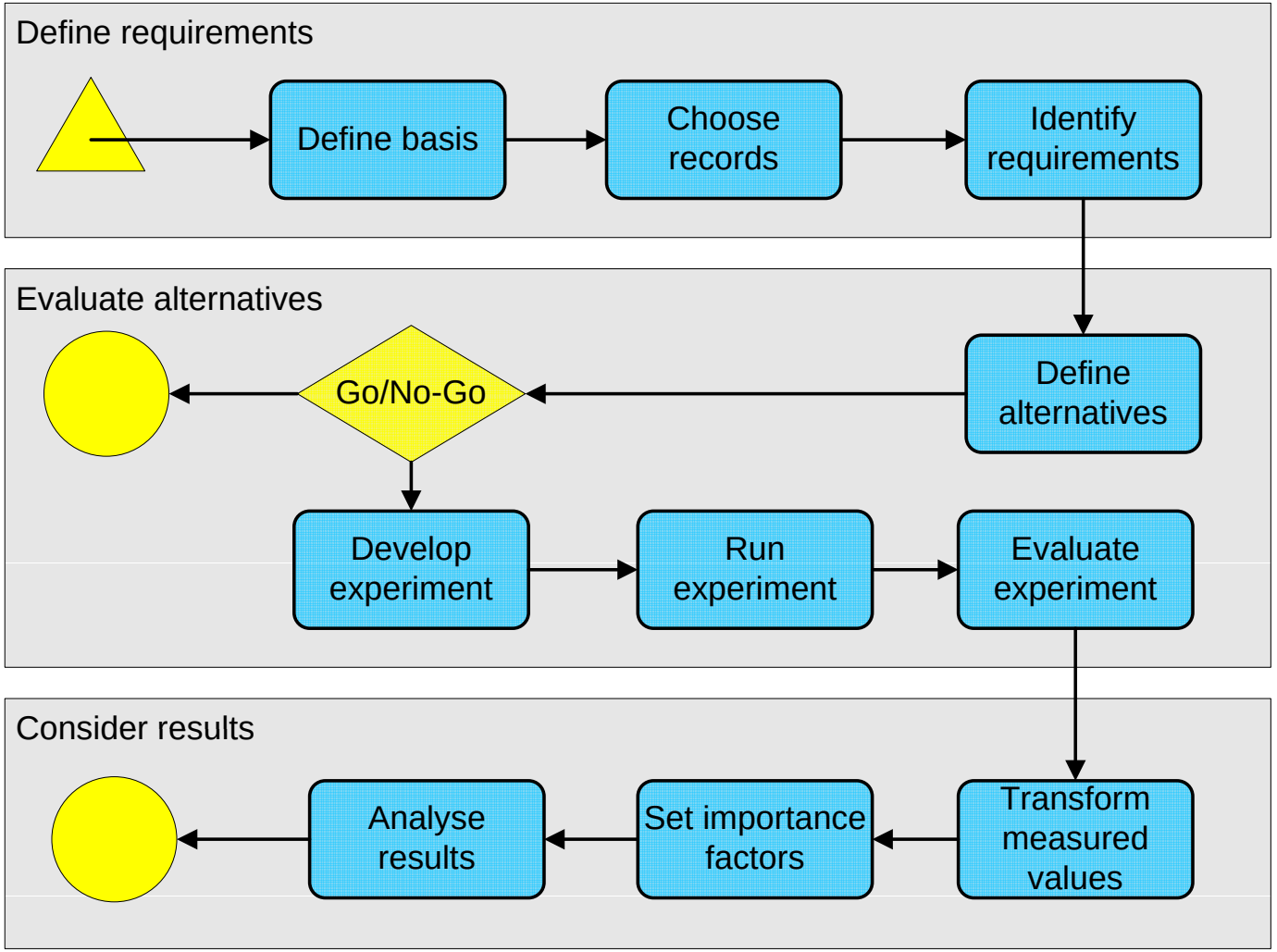
- For electronic documents, a variety of solutions exist
- All have specific strengths and weaknesses
- Individual requirements, obligations and constraints in every institution
- Decision between tools is complex
- Documentation and accountability is essential in decision-making

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

Preservation of electronic publications

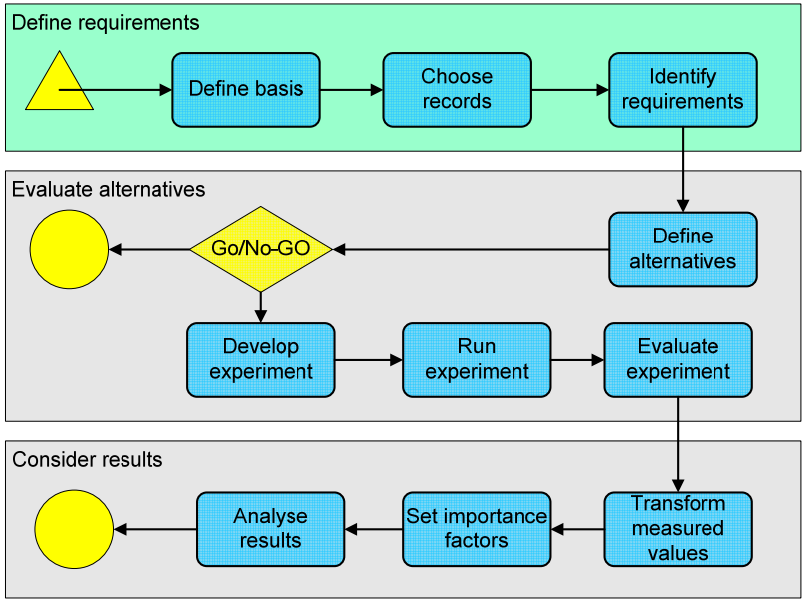
- Austrian National Library will collect and preserve Austrian theses and dissertations digitally
  - Legal obligation
  - Little control over submission (PDF)
- PLANETS: Case study evaluating different solutions
- Goals
  - Validating preservation planning methodology
  - Evaluate possible target formats
  - Document reasons
- Agenda:
  - Preservation Planning Methodology (tool support)
  - Case study results

Preservation Planning Workflow



Phase 1: Define requirements

1. Define basis
   - Describe Collection
   - Institutional settings
2. Choose sample objects/records
   - Representative for the objects in the collection
   - Right choice of samples is essential
3. Define requirements

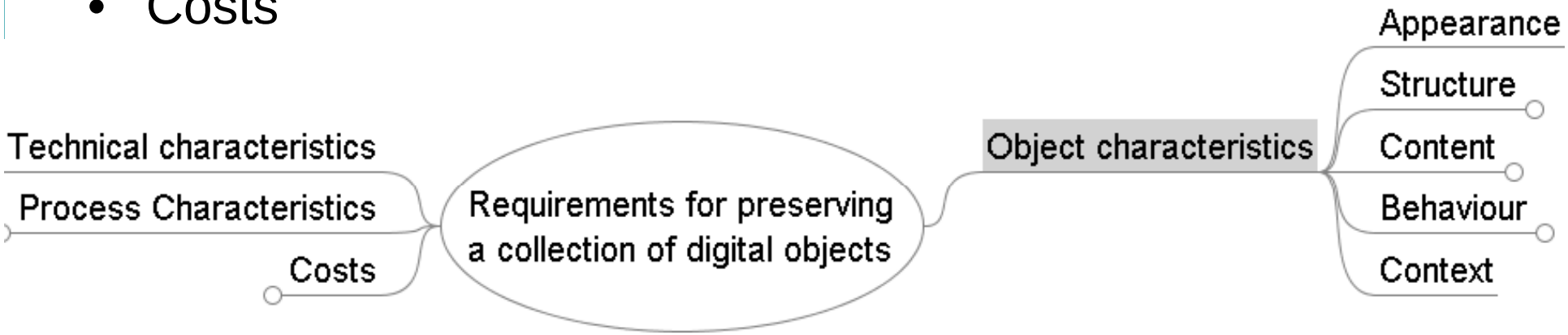


Defining requirements

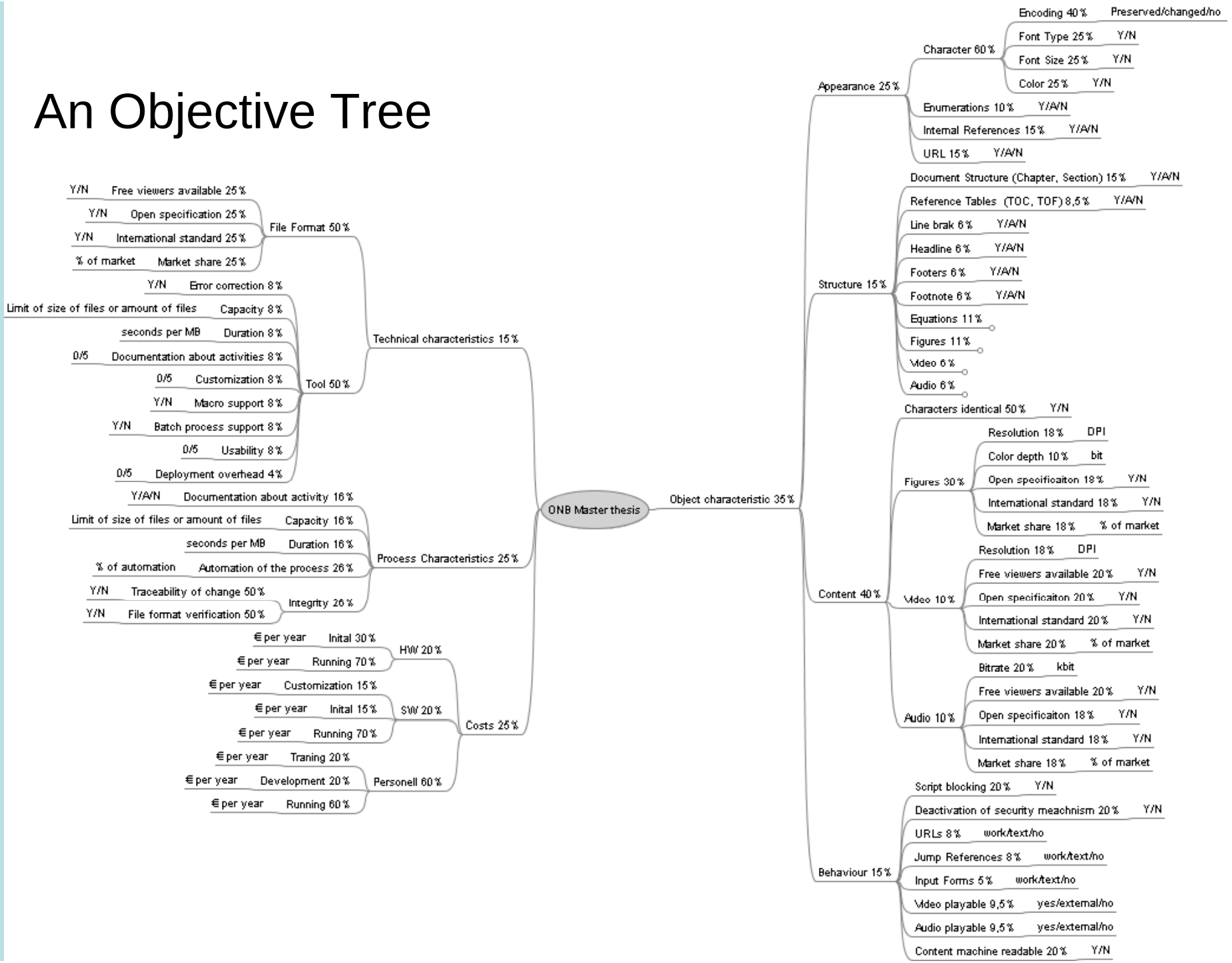
- Identify requirements and goals
  - Influence factors
  - Input from different stakeholders
  - Workshop setting
- Tree structure called ‘objective tree’
  - Utility analysis
- Top-down or bottom-up
  - Start from high-level goals and break down to specific criteria
  - Start from low-level criteria and organize in tree structure

Requirements include…

- Object characteristics
  - Content
  - Structure
  - Appearance
  - Behaviour
  - Context
- Technical characteristics
- Process characteristics
- Costs



An Objective Tree



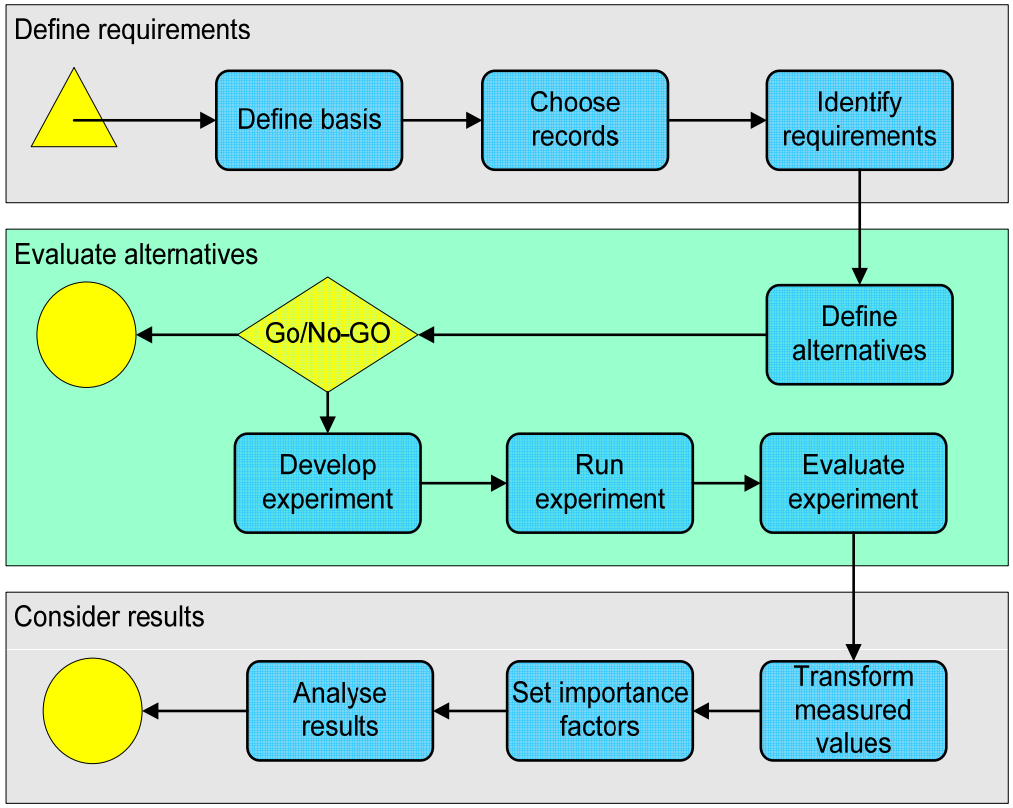
Assign Measurable Units

- 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

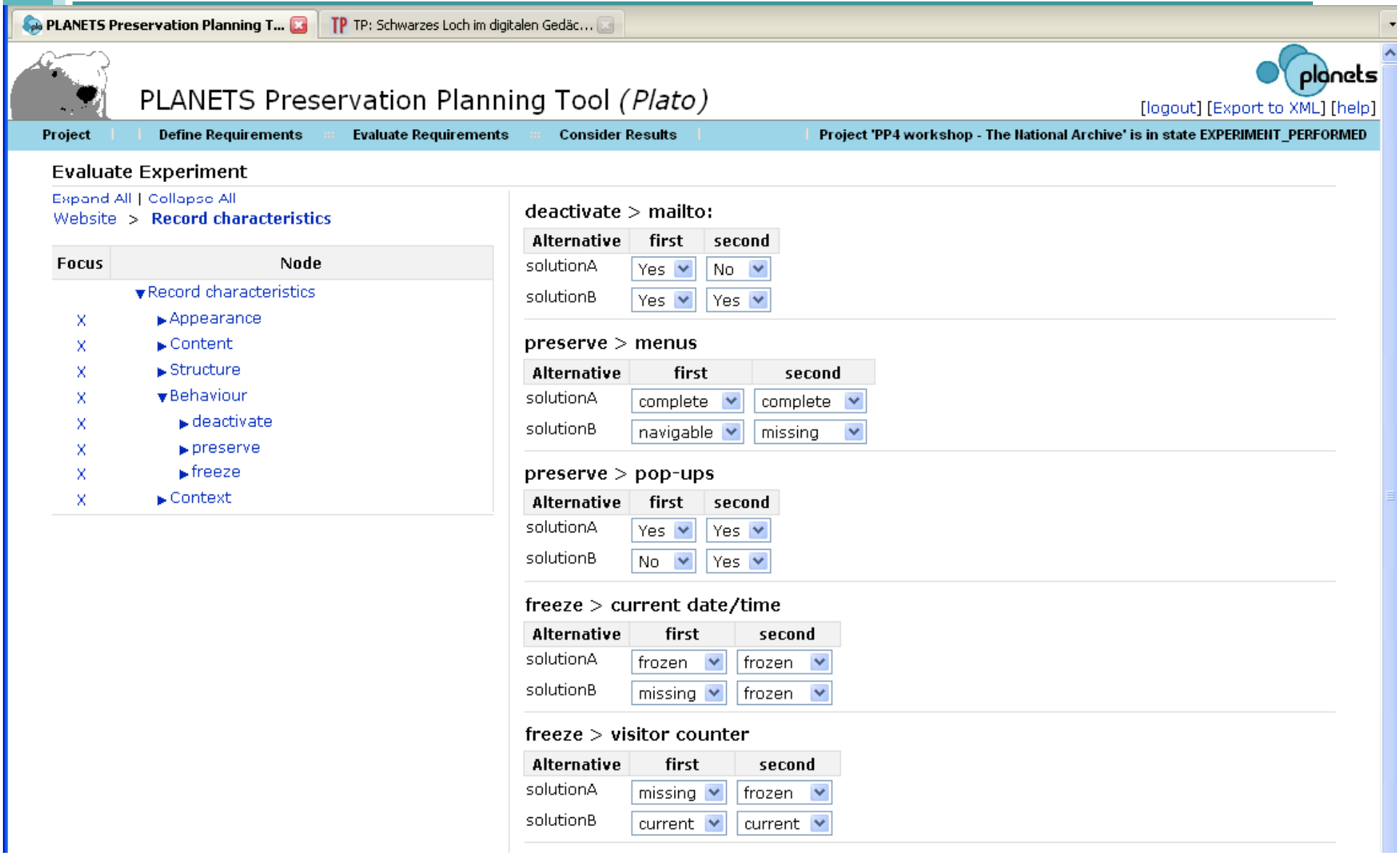
➢ Quantitative results

Phase 2: Evaluate Alternatives

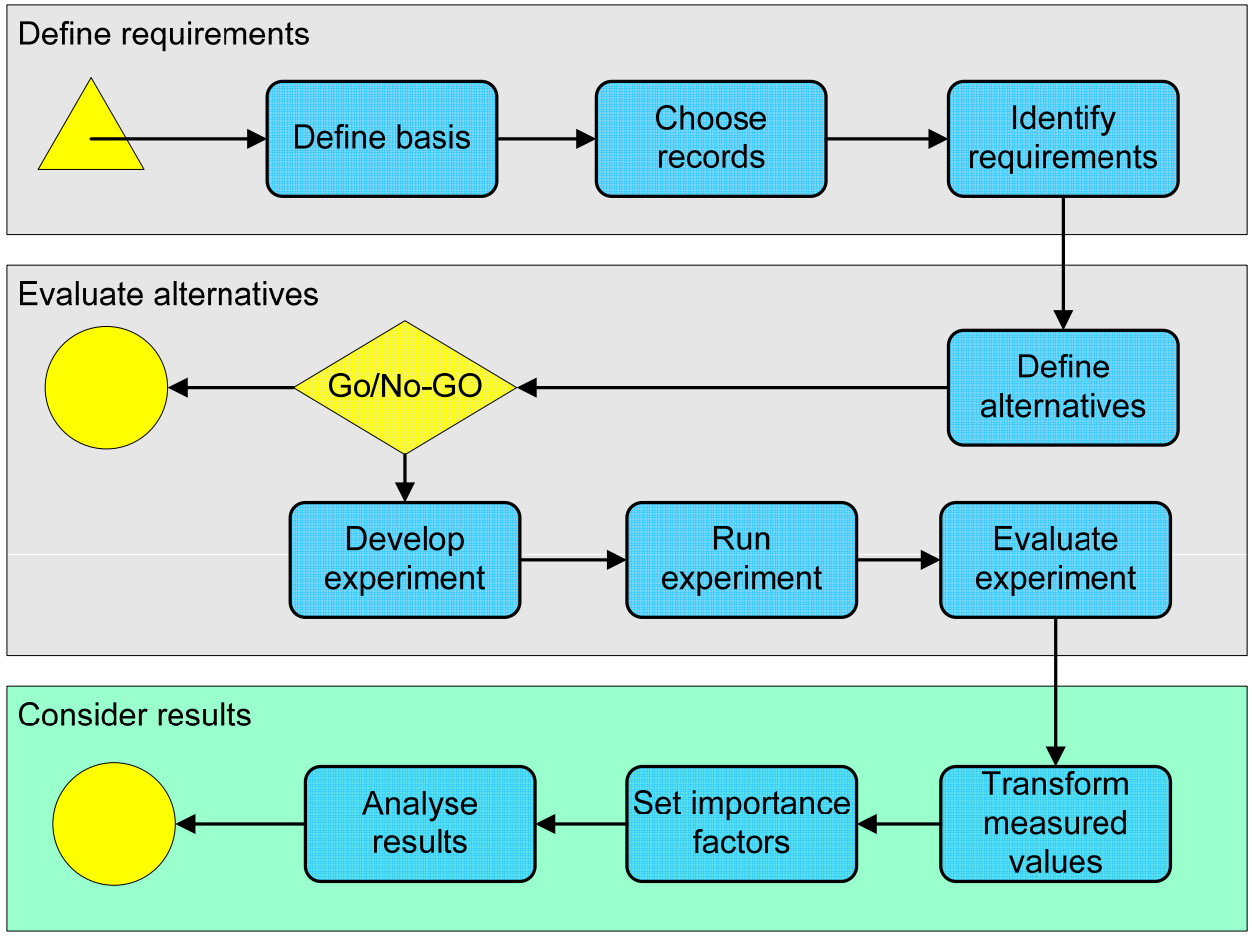
4. Define Alternatives
5. Go/No-Go decision
6. Develop experiment
7. Run experiment
8. Evaluate experiment



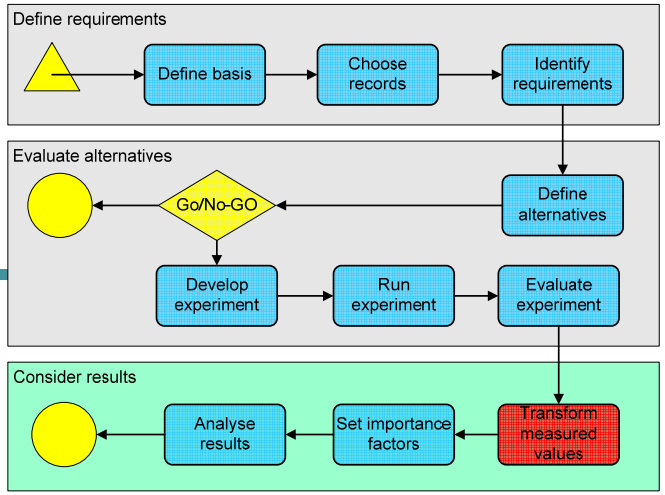
Evaluating results



Phase 3: Consider Results

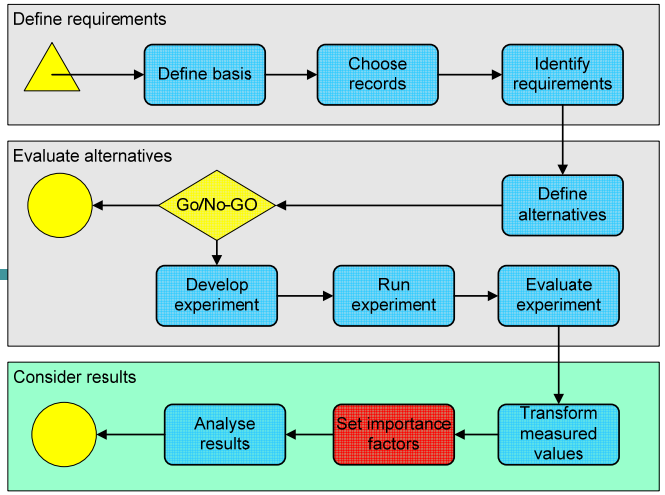


Transform measured values



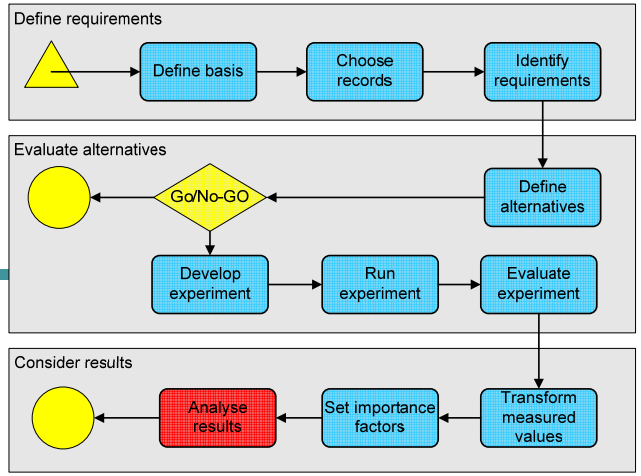
- Measures come in seconds, euro, bits, goodness values,…
- Need to make them comparable
- Transform measured values to uniform scale
- Transformation tables for each leaf criterion
- Scale 0-5 (0 is *unacceptable*)

Set importance factors



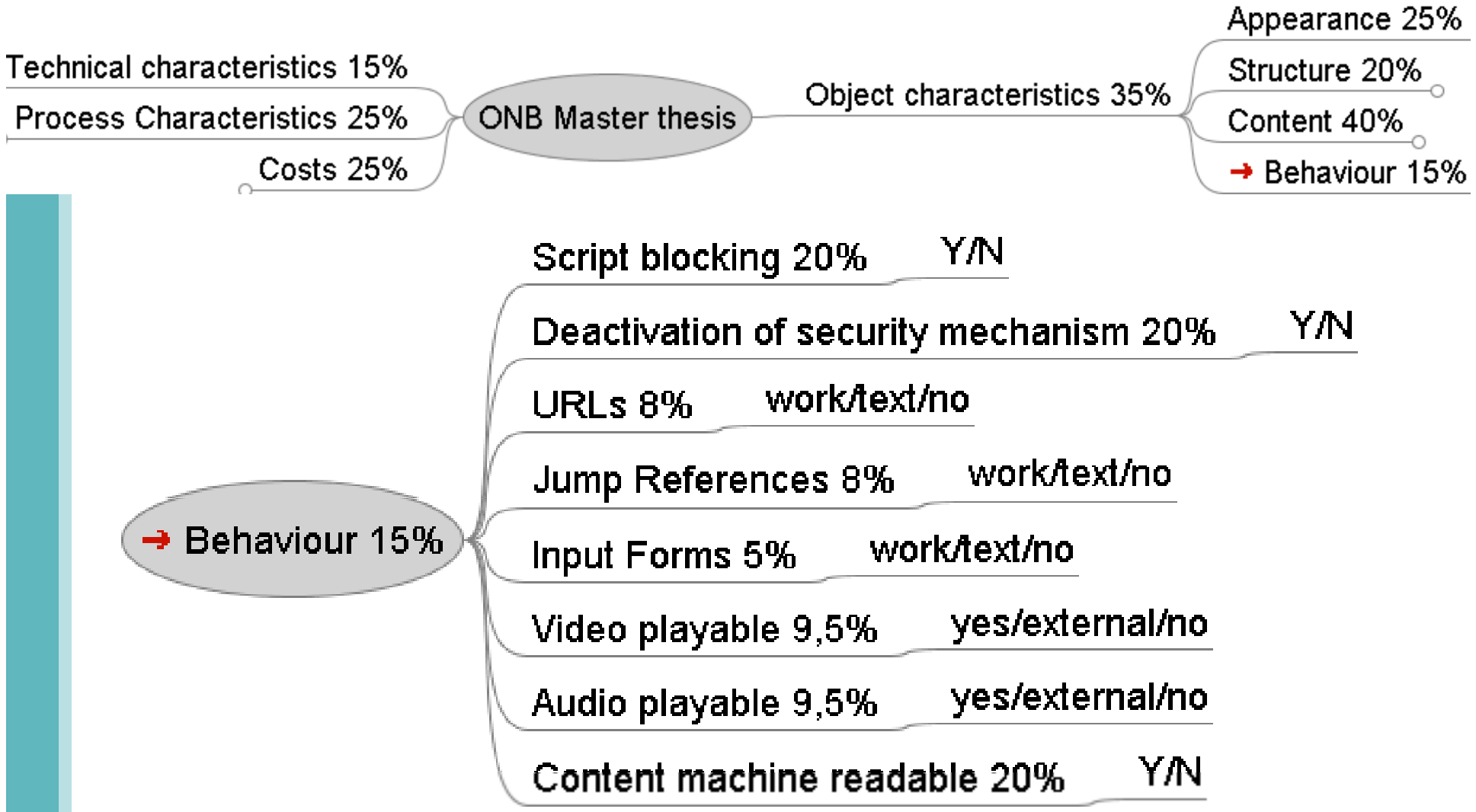
- Branches are weighted equally by default
- Not all leaf criteria are equally important
- Adjust relative importance of all siblings in a branch
- Weights are propagated down the tree to the leaves

Analyse Results



- Aggregate values
  - Weighted sum and weighted multiplication over all branches of the tree
  - Performance values for each alternative
- Rank alternatives according to overall performance value at root
- Performance of each alternative
  - overall
  - for each sub-criterion (branch)
- Comparison of different alternatives

Case study: Some requirements



Results

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 (ConvertDoc 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 is a knock-out criterion (PDF)
- Image formats do not provide full-text search
- RTF tools show major weaknesses in appearance and structure
- Plain text fails appearance, structure and content requirements

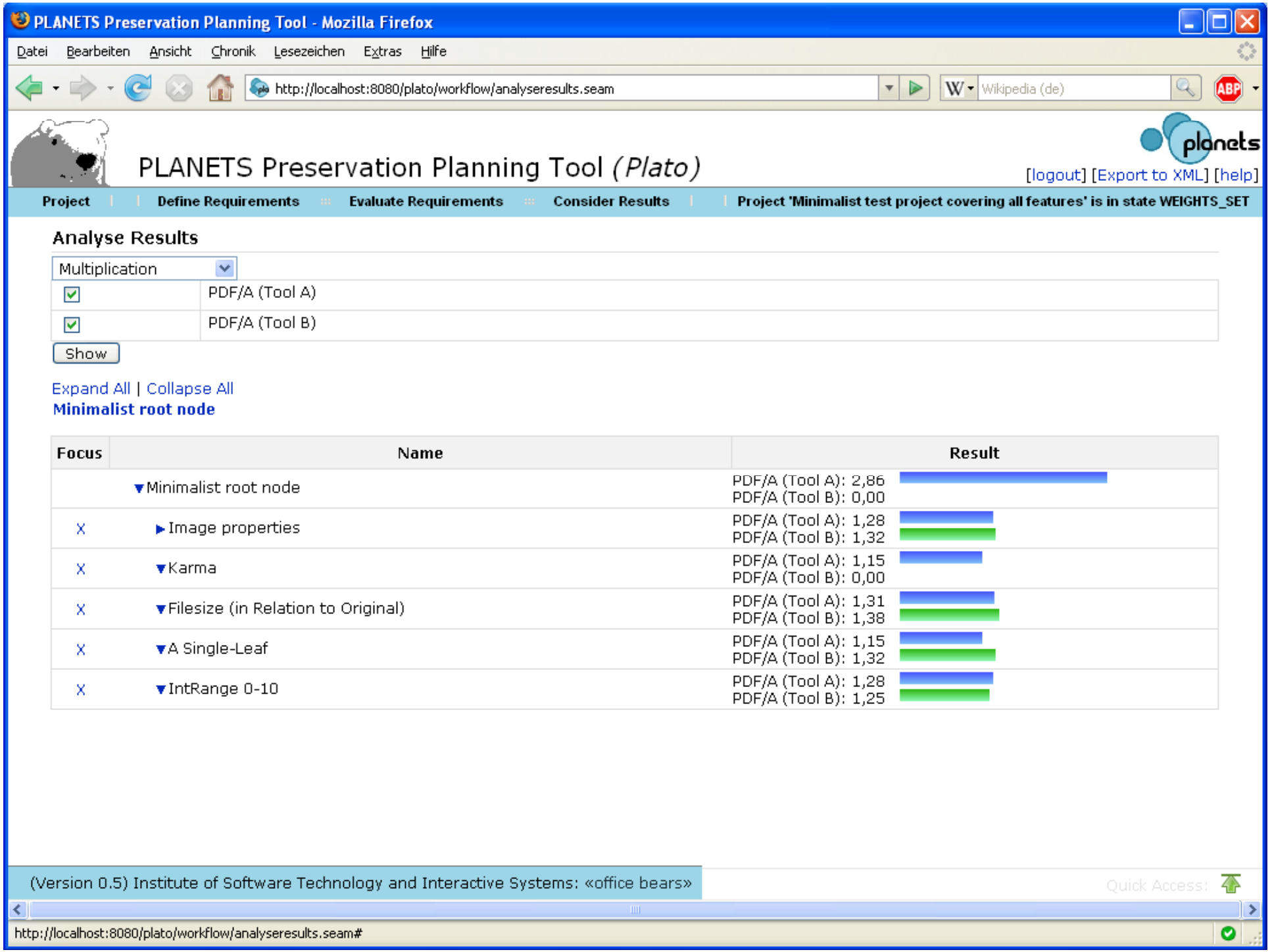
☑	TIFF
☑	EPS
☑	JPEG2000
☑	RTF-acrobat
☑	RTF-convertdoc
☑	TXT

Show

Expand All | Collapse All

ONB Master thesis > **Object characteristic**

Focus	Name	Result
	▼Object characteristic	PDF-A: 1,66 PDF-unchanged: 0,00 TIFF: 1,62 EPS: 1,62 JPEG2000: 1,60 RTF-acrobat: 0,00 RTF-convertdoc: 1,36 TXT: 0,00
X	▶Appearance	PDF-A: 1,50 PDF-unchanged: 1,50 TIFF: 1,50 EPS: 1,50 JPEG2000: 1,50 RTF-acrobat: 1,26 RTF-convertdoc: 1,19 TXT: 0,00
X	▶Structure	PDF-A: 1,27 PDF-unchanged: 1,27 TIFF: 1,27 EPS: 1,26 JPEG2000: 1,26 RTF-acrobat: 0,00 RTF-convertdoc: 1,18 TXT: 0,00
X	▶Content	PDF-A: 1,84 PDF-unchanged: 1,84 TIFF: 1,84 EPS: 1,84 JPEG2000: 1,84 RTF-acrobat: 0,00 RTF-convertdoc: 1,43 TXT: 0,00
X	▶Behaviour	PDF-A: 1,21 PDF-unchanged: 0,00 TIFF: 1,13 EPS: 1,14 JPEG2000: 1,09 RTF-acrobat: 1,18 RTF-convertdoc: 1,19 TXT: 1,19


PLANETS Preservation Planning Tool - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://localhost:8080/plato/workflow/analyseresults.seam
Wikipedia (de)
PLANETS Preservation Planning Tool (Plato)
planets
[logout] [Export to XML] [help]
Project | Define Requirements ::: Evaluate Requirements ::: Consider Results | Project 'Minimalist test project covering all features' is in state WEIGHTS_SET
Analyse Results
Multiplication
PDF/A (Tool A)
PDF/A (Tool B)
Show
Expand All | Collapse All
Minimalist root node
Focus
Name
Result
▼Minimalist root node
PDF/A (Tool A): 2,86
PDF/A (Tool B): 0,00
X
▶Image properties
PDF/A (Tool A): 1,28
PDF/A (Tool B): 1,32
X
▼Karma
PDF/A (Tool A): 1,15
PDF/A (Tool B): 0,00
X
▼Filesize (in Relation to Original)
PDF/A (Tool A): 1,31
PDF/A (Tool B): 1,38
X
▼A Single-Leaf
PDF/A (Tool A): 1,15
PDF/A (Tool B): 1,32
X
▼IntRange 0-10
PDF/A (Tool A): 1,28
PDF/A (Tool B): 1,25
(Version 0.5) Institute of Software Technology and Interactive Systems: «office bears»
Quick Access:
http://localhost:8080/plato/workflow/analyseresults.seam#

Tool support

- First internal version in December
- First public version next year
  - Integration of Planets services
- Technical
  - Java Enterprise application
  - Planets Application Server based on JBoss 4.0.5
  - JBoss Seam 1.2.1
  - Java Server Faces, Facelets
    - AJAX-enabled component libraries
      - Apache Trinidad
      - JBoss RichFaces, AJAX4JSF
    - EJB 3 (Hibernate)
    - Database: Apache Derby (exchangeable)
    - XML export and import

The PLANETS project

- **P**reservation and **L**ong-term **A**ccess through **NET**worked **S**ervices
- Distributed preservation infrastructure and services
- 4-year project funded under the 6th Framework Programme of the European Union (~15m EUR)
- 16 partners from 9 countries
  - National Libraries
  - National Archives
  - Universities
  - Research and technology companies

www.planets-project.eu

Something different…
the DPE Digital Preservation Challenge

- Digital Preservation Europe: coordinating EU project
- DPE Challenge: Competition with several tasks to solve
- Overcome the barriers hindering access to digital objects
- Open for all participants

Awards

1. First Prize 3000 Euros
2. Second Prize 1500 Euros
3. Third Prize 500 Euros

Next challenge online in January 2008, submission deadline in March

www.digitalpreservationeurope.eu/challenge

Thank you very much for your attention.

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