Digital Preservation: Logical and bit-stream preservation using Plato, EPrints and the Cloud

Hannes Kulovits
Andreas Rauber

Department of Software Technology and Interactive Systems

Vienna University of Technology
kulovits@ifs.tuwien.ac.at
rauber@ifs.tuwien.ac.at

David Tarrant
Adam Field

School of Electronics and Computer Science

University of Southampton, UK
dct05r@ecs.soton.ac.uk
af05v@ecs.soton.ac.uk

Vienna University of Technology

- **Vienna University of Technology**
  http://www.tuwien.ac.at
  - Faculty of Computer Science
    http://www.cs.tuwien.ac.at
    - Department of Software Technology and Interactive Systems (ISIS)
      http://www.isis.tuwien.ac.at
- People in DP
  - **Andreas Rauber**
  - Christoph Becker
  - Mark Guttenbrunner
  - Rudolf Mayer
  - Florian Motlik
  - Michael Kraxner
  - Hannes Kulovits
  - Stephan Strodl

TU WIEN planets University of Southampton eprints

DP Activities in Vienna

- Web Archiving (AOLA)
  in cooperation with the Austrian National Library
- DELOS DPC (EU FP6 NoE)
- DPE: Digital Preservation Europe (EU FP6 CA)
- PLANETS (EU FP6 IP)
- eGovernment & Digital Preservation
  series of projects with Federal Chancellery
- National Working Group on Digital Preservation
  of the Austrian Computer Society, in cooperation with ONB
- Digital Memory Engineering: National research studio

TU WIEN planets UNIVERSITY OF Southampton eprints

University of Southampton, UK

- **University of Southampton**
  http://www.soton.ac.uk
  - School of Electronics & Computer Science
    http://www.ecs.soton.ac.uk
- **EPrints**
  http://www.epints.org
  - People in Preservation
    - Steve Hitchcock
    - **David Tarrant**
    - Chris Gutteridge
    - Tim Brody
    - Patrick McSweeny
  - EPrints Services
    - **Adam Field**
    - Tim Miles-Board

TU WIEN planets

UNIVERSITY OF Southampton

eprints

DP Activities in Southampton

- EPrints Preservation
  - KeepIt!
  - Preserv2
  - Preserv
- P2N – Preservation Network
  - Collabotarion with Oxford Univeristy
- P2-Registry
  - Linked Data for Digital Preservation
- Web Archiving
  - ECS project to archive old project websites and Wikis

TU WIEN planets University of Southampton eprints

Introductions

What will you know after this tutorial?

You will:

- Understand the challenges in digital preservation and
- Address them on both layers physical and logical.
- Understand why we need to *plan preservation activities*
- Know a *workflow to evaluate preservation strategies*
- Be familiar with Plato and EPrints
- Be *able to develop a specific preservation plan* that is optimized for
  - the objects in your institution
  - the users of your institution
  - the institutional requirements

TU WIEN planets UNIVERSITY OF Southampton eprints

Schedule

(1) Introduction
- What is Digital Preservation?
- EPrints
- Preservation Planning and Plato

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing

TU WIEN planets UNIVERSITY OF Southampton eprints

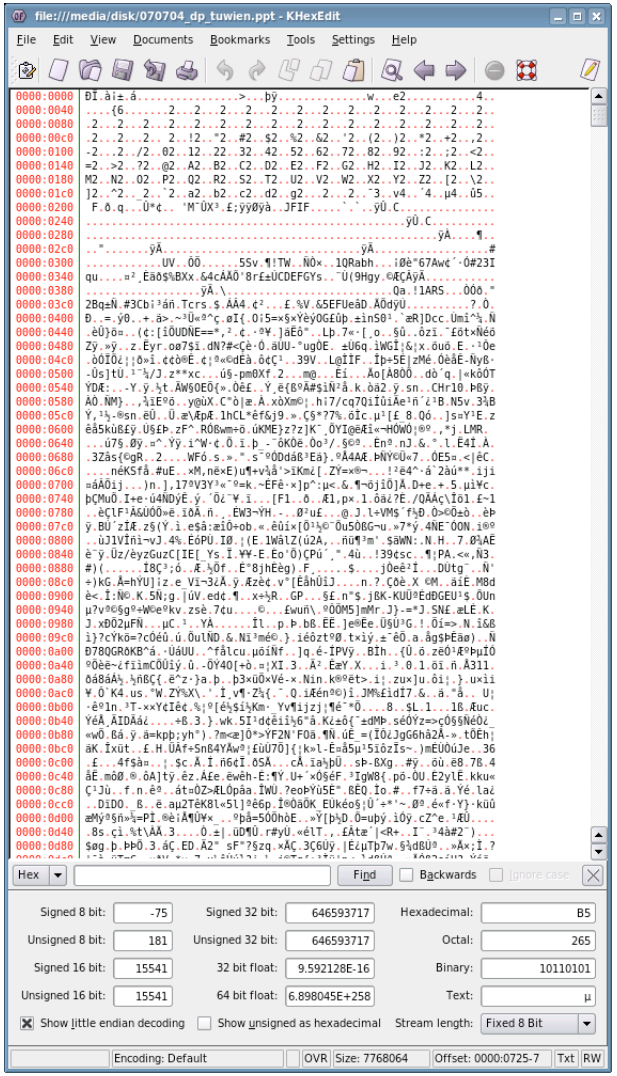
Overview

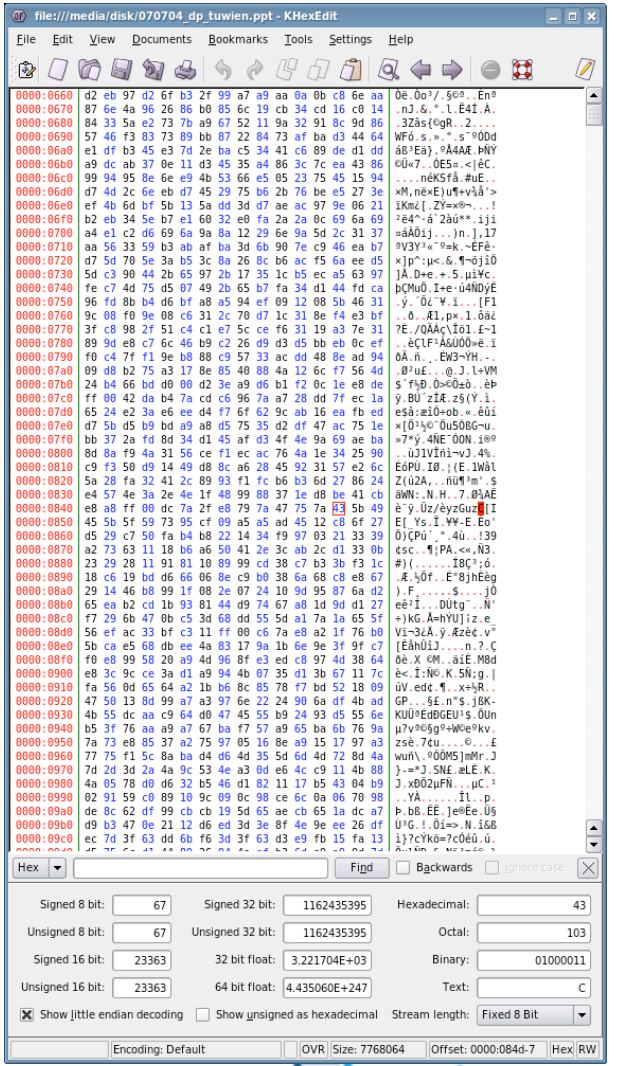
Part 1: Introduction

- What is Digital Preservation?
- What is the OAIS Reference model?
- Physical preservation with EPrints
- Logical preservation with Plato

Why do we need Digital Preservation?

Why do we need Digital Preservation?





TU WIEN

planets

UNIVERSITY OF Southampton

eprints

Why do we need Digital Preservation?

- Digital Objects require specific environment to be accessible :
  - Files need specific programs
  - Programs need specific operating systems (-versions)
  - Operating systems need specific hardware components
- SW/HW environment is not stable:
  - Files cannot be opened anymore
  - Embedded objects are no longer accessible/linked
  - Programs won‘t run
  - Information in digital form is lost
    (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

TU WIEN planets UNIVERSITY OF Southampton eprints

Why do we need Digital Preservation?

- Essential for all digital objects
  - Office documents, accounting, emails, …
  - Scientific datasets, sensor data, metadata, …
  - Applications, simulations,…

- All application domains
  - Cultural heritage data
  - eGovernment, public administration
  - Science / Research
  - Industry
  - Health, pharmaceutical industry
  - Aviation, control systems, construction, …
  - Private data
  - …

TU WIEN planets UNIVERSITY OF Southampton eprints

Strategies for Digital Preservation

Strategies

(grouped according to Companion Document to UNESCO Charter *http://unesdoc.unesco.org/images/0013/001300/130071e.pdf*)

- Investment strategies:
  - Standardization, Data extraction, Encapsulation, Format limitations
- Short-term approaches:
  - Museum, Backwards-compatibility, Version-migration, Reengineering
- Medium- / long-term approaches:
  - Migration, Viewer, Emulation
- Alternative approaches:
  - Non-digital Approaches, Data-Archeology

- No single optimal solution for all objects

Migration

- Transformation into different format, continuous or on-demand (Viewer)
+ Wide-spread adoption
+ Possibility to compare to un-migrated object
+ Immediately accessible
- Unintended changes, specifically over sequence of migrations
- Cannot be used for all objects
- Requires continuous action to migrate

Emulation

- Emulation of hardware or software (operating system, applications)
+ Concept of emulation widely used
+ Numerous emulators are available
+ Potentially complete preservation of functionality
+ *Object is rendered identically*
- *Object is rendered identically*
- Requires detailed documentation of system
- Requires knowledge on how to operate current systems in the future
- Complex technology
- Emulators must be emulated or migrated themselves
- Emulators potentially erroneous/incomplete

Strategies for Digital Preservation

Strategies

(grouped according to Companion Document to UNESCO Charter *http://unesdoc.unesco.org/images/0013/001300/130071e.pdf*)

- Investment strategies:
  - Standardization, Data extraction, Encapsulation, Format limitations
- Short-term approaches:
  - Museum, Backwards-compatibility, Version-migration, Reengineering
- Medium- / long-term approaches:
  - Migration, Viewer, Emulation
- Alternative approaches:
  - Non-digital Approaches, Data-Archeology

- No single optimal solution for all objects

Digital Preservation

- Is a complex task
- Requires a concise understanding of the objects, their intellectual characteristics, the way they were created and used and how they will most likely be used in the future
- Requires a continuous commitment to preserve objects to avoid the „digital dark hole“
- Requires a solid, trusted infrastructure and workflows to ensure digital objects are not lost
- Is essential to maintain electronic publications & data accessible
- Will become more complex as digital objects become more complex
- Needs to be defined in a preservation plan

Digital Preservation

- Reference Models
  - Records Management, ISO 15489:2000
  - OAIS: Open Archival Information System, ISO 14721:2003
- Audit & Certification Initiatives
  - RLG- National Archives and Records Administration Digital Repository Certification Task Force:
    Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)
  - NESTOR:
    Catalogue of Criteria of Trusted Digital Repositories
  - DCC/DPE:
    DRAMBORA: Digital Repository Audit Method Based on Risk Assessment

Overview

Part 1: Introduction

- What is Digital Preservation?
- What is the OAIS Reference model?
- Physical preservation with EPrints
- Logical preservation with Plato

OAIS

- NASA: National Space Science Data Center
  - NASA's first digital archive
  - Experienced many technological changes since 1966
- Consultative Committee for Space Data Systems
  - International group of space agencies
  - Developed range of discipline-independent standards
  - Evolved into ISO TC 20/ SC 13 working group around 1990
  - TC20: Aircraft and Space Vehicles
  - SC13: Space Data and Information Transfer Systems

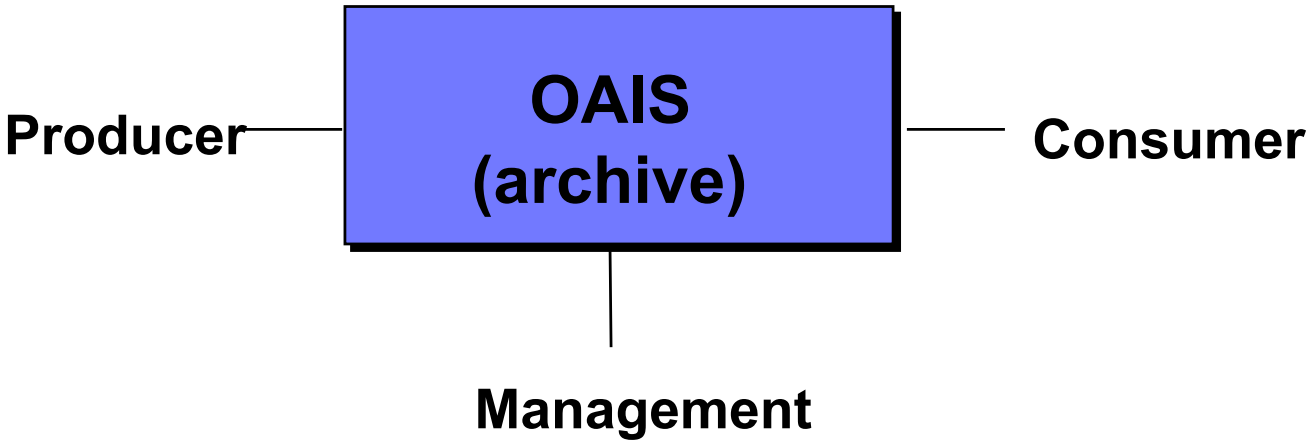
OAIS

- Reference Model for an Open Archival Information System (OAIS), Blue Book, CCSDS 650.0-B-1, January 2002
- ISO 14721:2003
- slides based on Blue Book and:
  - Don Sawyer, Lou Reich: ISO Reference Model for an Open Archival Information System (OAIS) Tutorial Presentation, LOC, June 13 2003
- http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html

OAIS

- Framework for understanding and applying concepts needed for long-term digital information preservation
  - Long-term: long enough to be concerned about changing technologies
  - Starting point for model addressing non-digital information
- Provides set of minimal responsibilities to distinguish an OAIS from other uses of ‘archive’
- Framework for comparing architectures and operations of existing and future archives
- Addresses a full range of archival functions
- Applicable to all long-term archives and those organizations and individuals dealing with information that may need long-term preservation
- Does NOT specify an implementation

OAIS



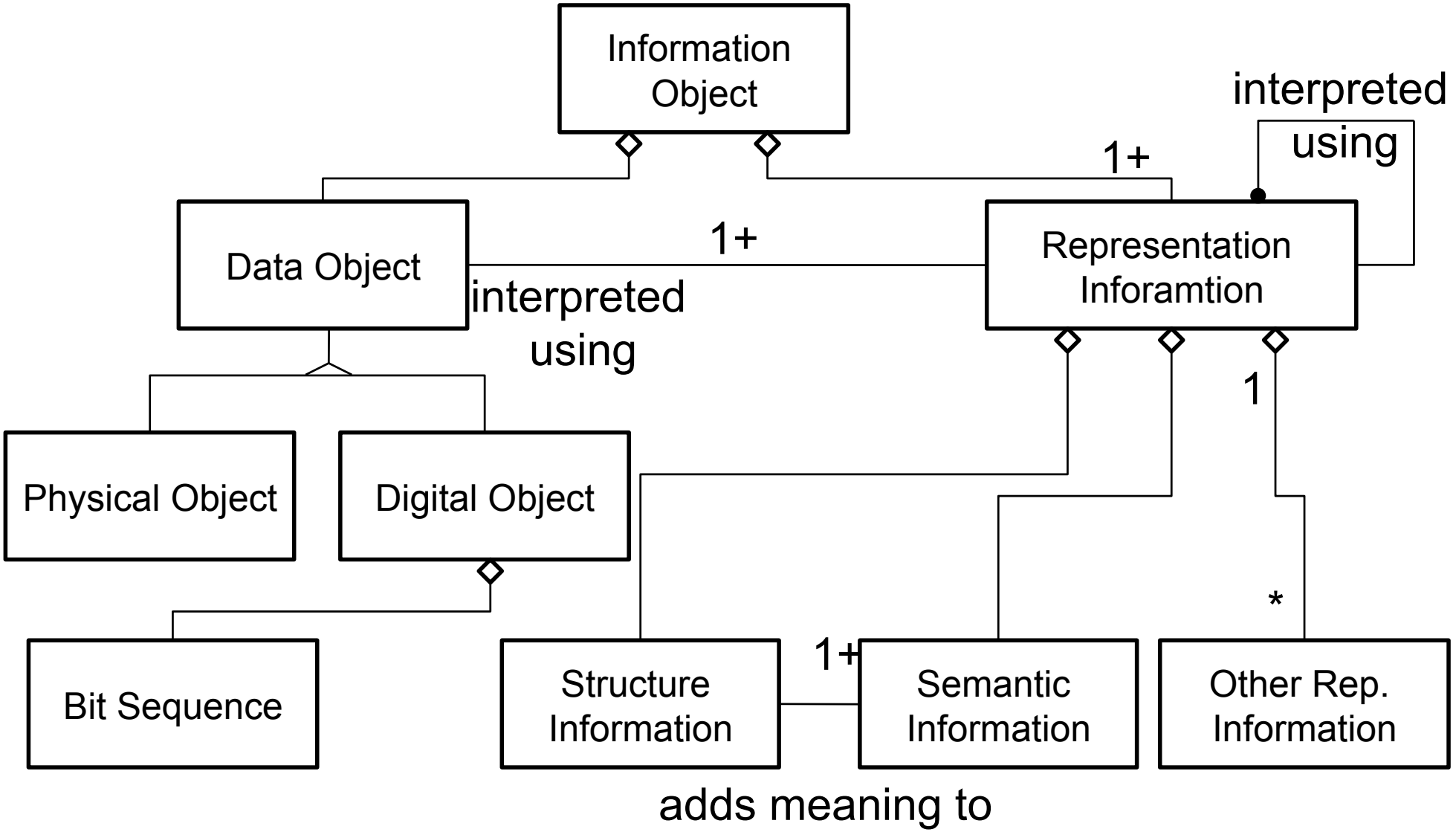
- Producer is the role played by those persons, or client systems, who provide the information to be preserved
- Management is the role played by those who set overall OAIS policy as one component in a broader policy domain
- Consumer is the role played by those persons, or client systems, who interact with OAIS services to find and acquire preserved information of interest

OAIS

OAIS Information Definition

- Information is always expressed (i.e., represented) by some type of data
- Data interpreted using its Representation Information yields Information
- Information Object preservation requires clear identification and understanding of the Data Object and its associated Representation Information

OAIS



TU WIEN planets

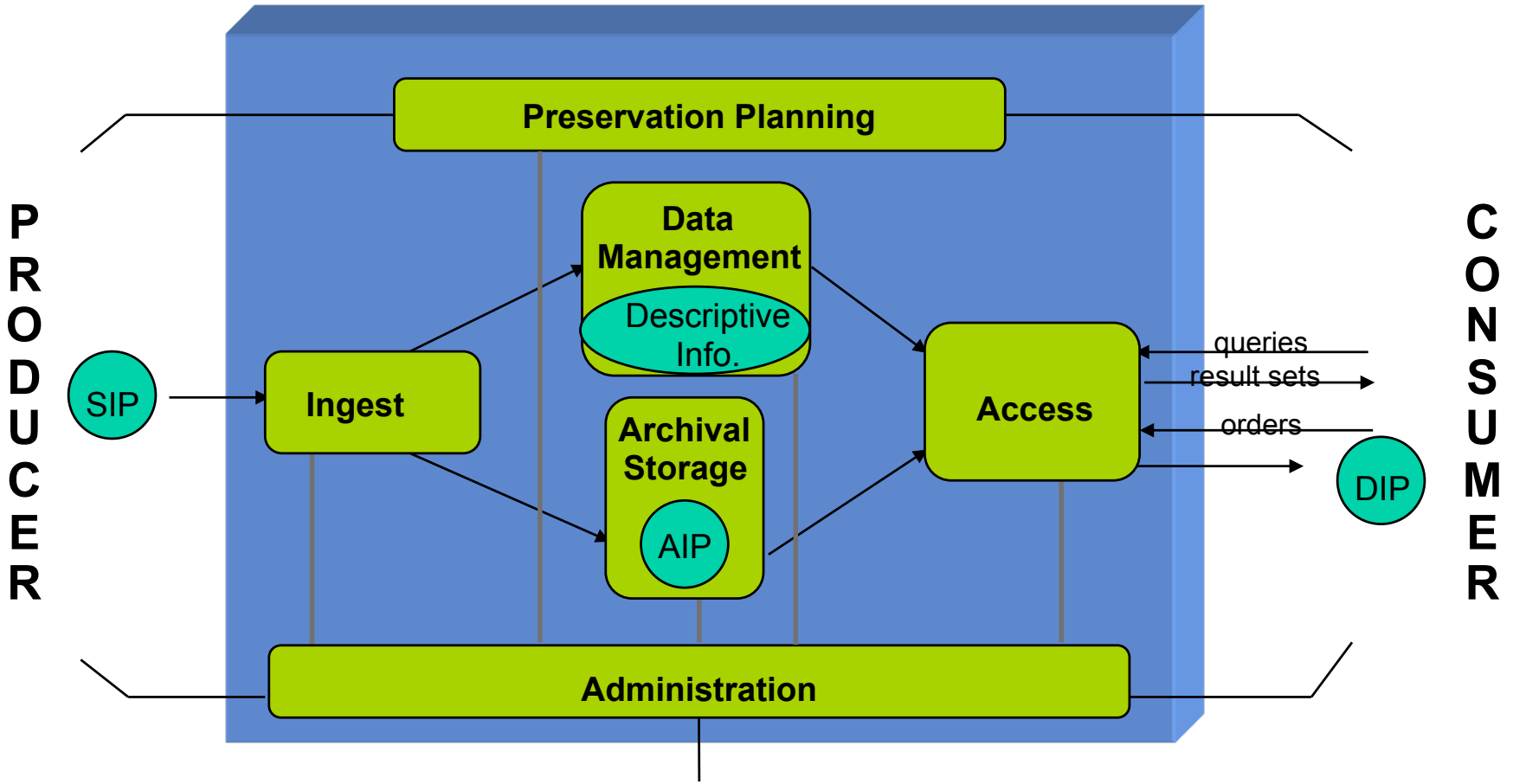
UNIVERSITY OF Southampton eprints

OAIS

Information Package Variants

- **SIP**: Submission Information Package
  - Negotiated between Producer and OAIS
  - Sent to OAIS by a Producer
- **AIP**: Archival Information Package
  - Information Package used for preservation
  - Includes complete set of Preservation Description Information (PDI) for the Content Information
- **DIP**: Dissemination Information Package
  - Includes part or all of one or more Archival Information Packages
  - Sent to a Consumer by the OAIS

OAIS

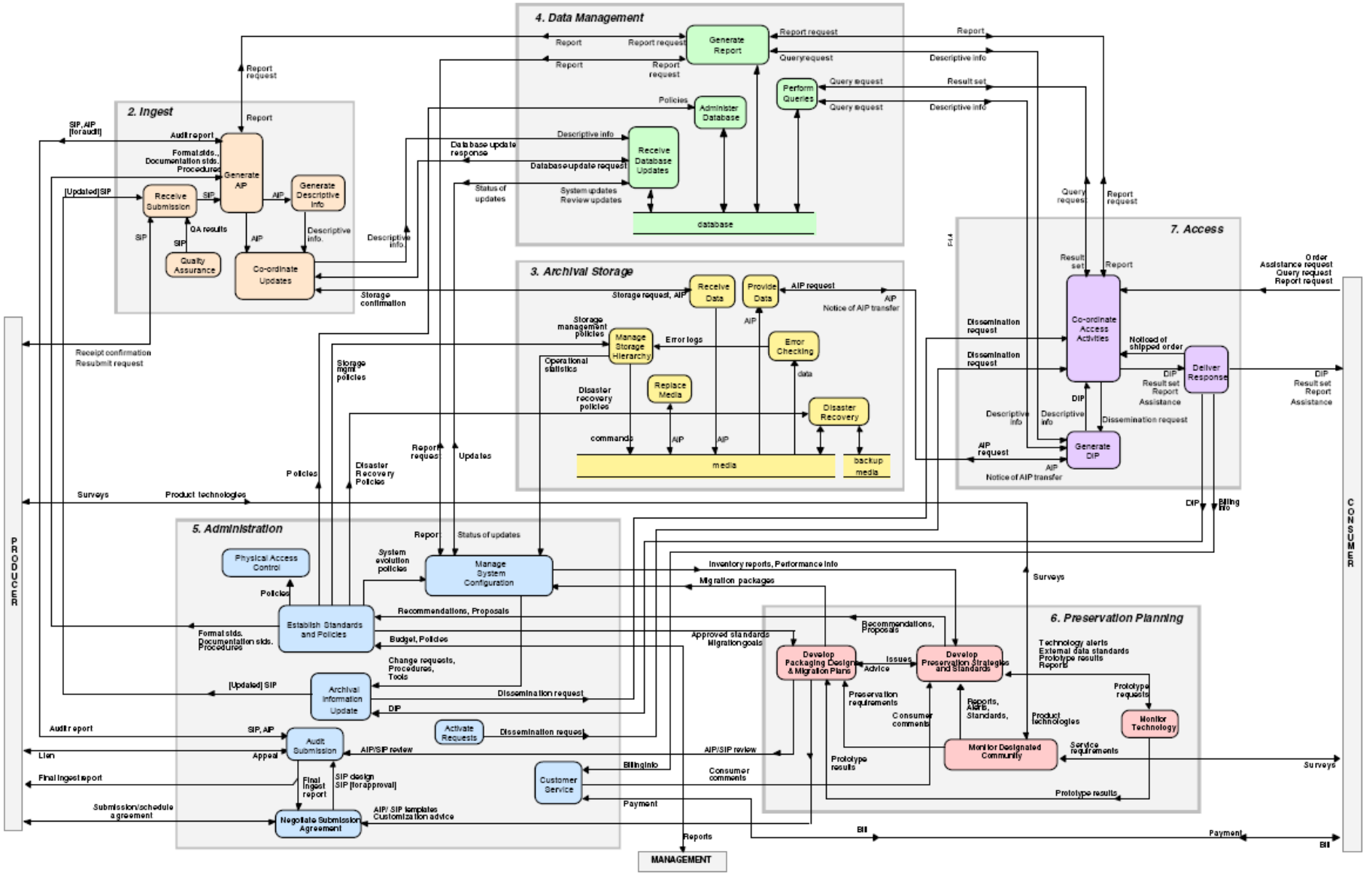


SIP = Submission Information Package

AIP = Archival Information Package

DIP = Dissemination Information Package

OAIS



Overview

Part 1: Introduction

- What is Digital Preservation?
- What is the OAIS Reference model?
- Intro to EPrints
- Logical preservation with Plato

What is EPrints For?

- EPrints offers a safe, open and useful place to store, share and manage material in the pursuit of research and educational agendas.

administrative reporting, collaboration, data sharing, digital profile enhancement , e-learning, e-publishing, e-research, marketing, open access, preservation, publicity, research assessment, research management, scholarly collections

TU WIEN planets

UNIVERSITY OF Southampton eprints

An EPrints repository is

- a valuable part of the researcher's information environment
  - directly integrating with the research desktop
  - offering sustainable storage and open access

- a competent and mature component of the institution's information environment
  - providing management and curation support for core business research data
  - leveraging information about research outputs to inform management strategy

TU WIEN planets UNIVERSITY OF Southampton eprints

Research Information Systems

- A repository needs to interoperate with management information systems
- Create reports based on research project activities as well as
  research outputs
- EPrints will support CERIF standard for Current Research Information Systems

TU WIEN planets UNIVERSITY OF Southampton eprints

Overview

Part 1: Introduction

- What is Digital Preservation?
- What is the OAIS Reference model?
- Physical preservation with EPrints
- Logical preservation with Plato

Preservation Planning

Why Preservation Planning?

- 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?

Preservation Planning

- Consistent workflow leading to a preservation plan
- Analyses, which solution to adopt
- 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

Digital Preservation

What is a preservation plan?

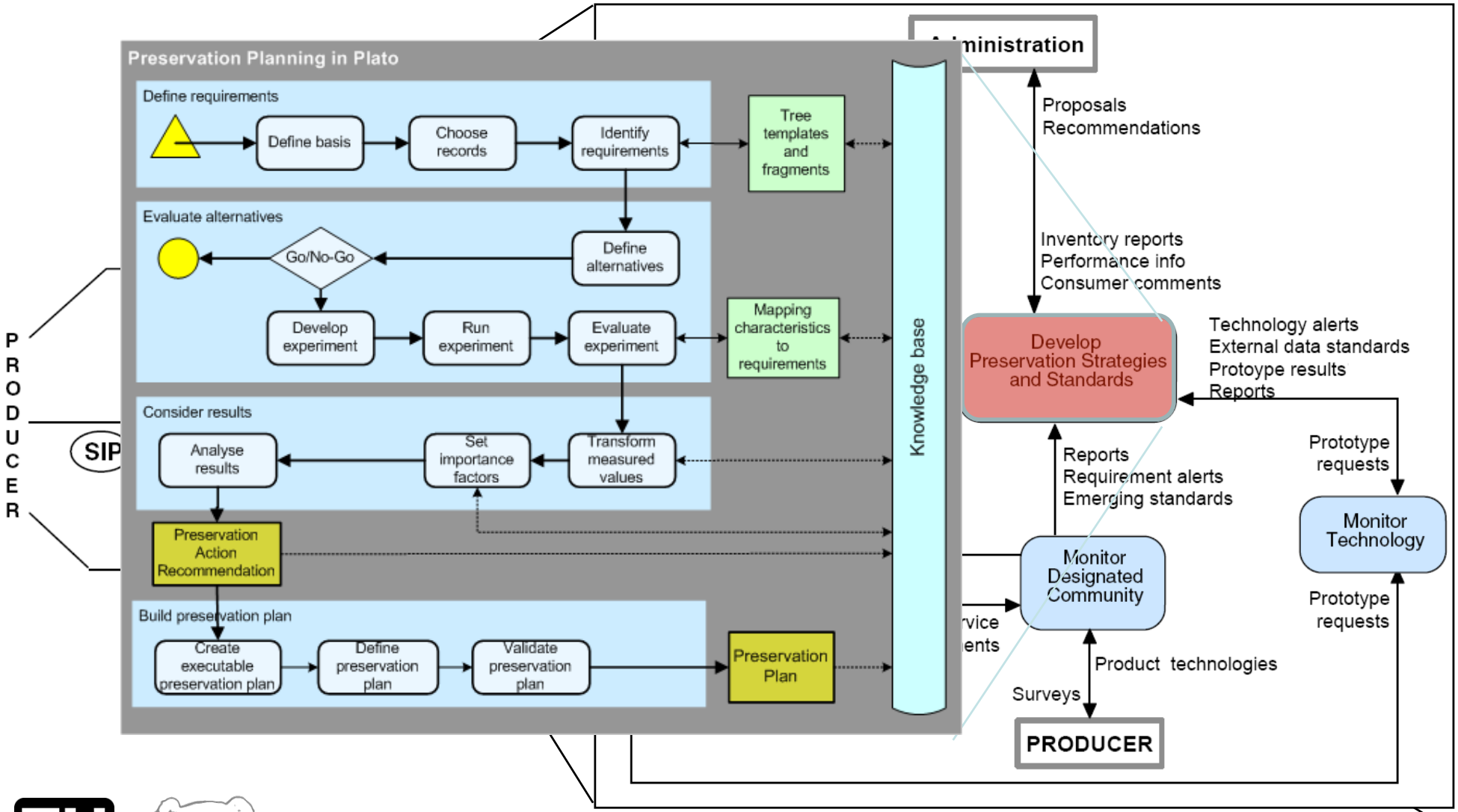
- 10 Sections
  - Identification
  - Status
  - Description of Institutional Setting
  - Description of Collection
  - Requirements for Preservation
  - Evidence for Preservation Strategy
  - Cost
  - Trigger for Re-evaluation
  - Roles and Responsibilities
  - Preservation Action Plan

Preservation Plan Template

Preservation Planning Workflow

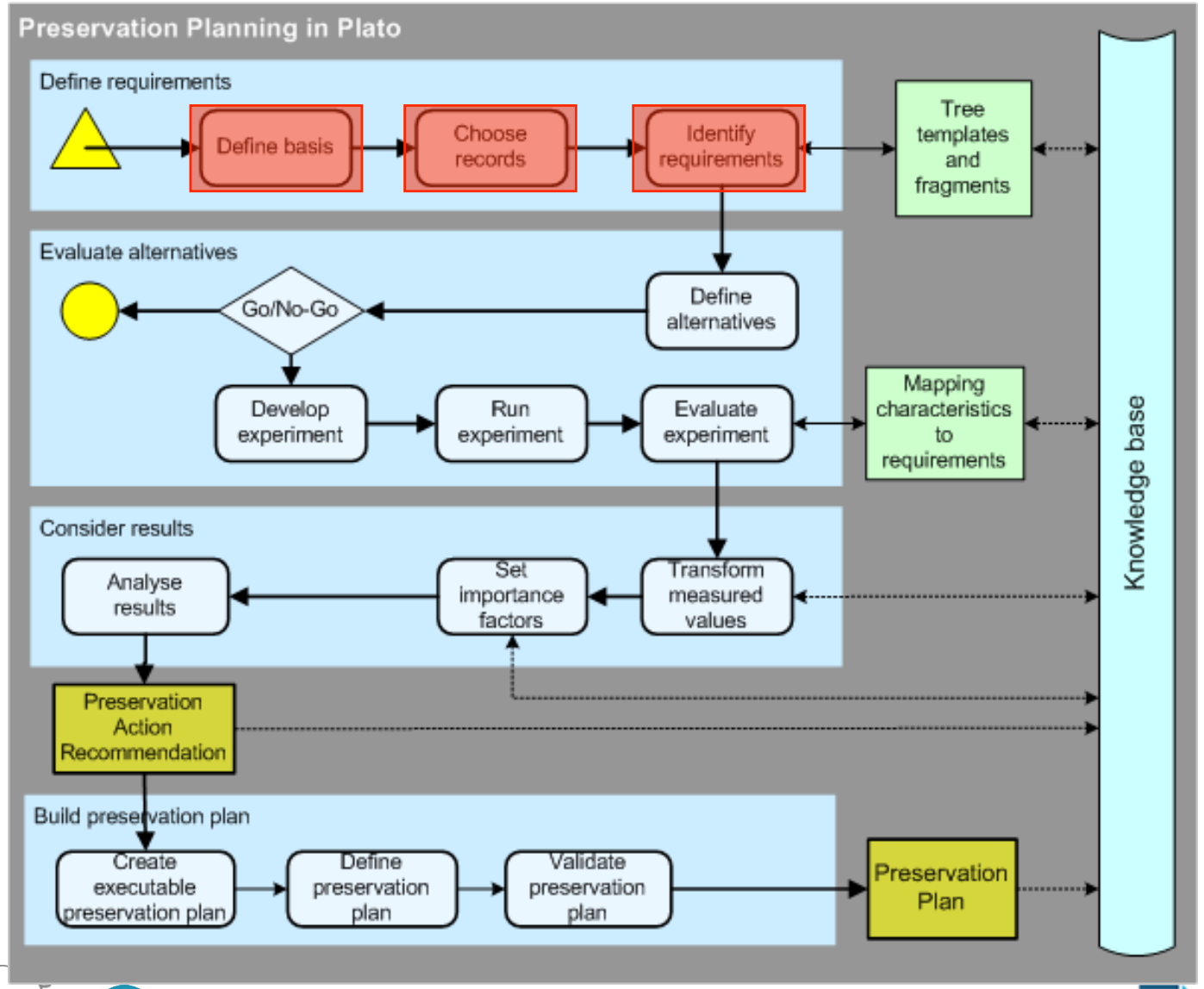
- Originally developed within the DELOS DP Cluster
  now refined and integrated within PLANETS
- Based on
  - Preservation Planning approach based on Utility Analysis, developed at TU Vienna
  - Testbed/lab for evaluation developed at Nationalarchief, The Netherlands
- Follows the OAIS model
- Consistent with requirements specified by ORLC/TRAC and Nestor criteria catalogue

Preservation Planning



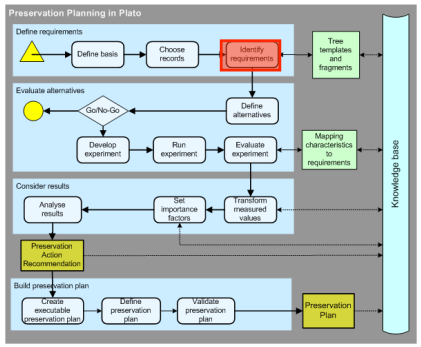
TU WIEN planets UNIVERSITY OF Southampton eprints

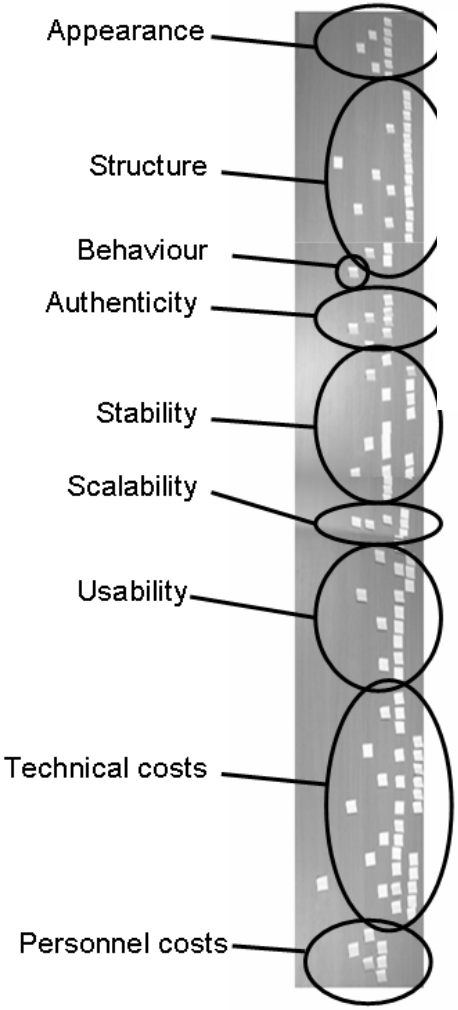
Preservation Planning Workflow

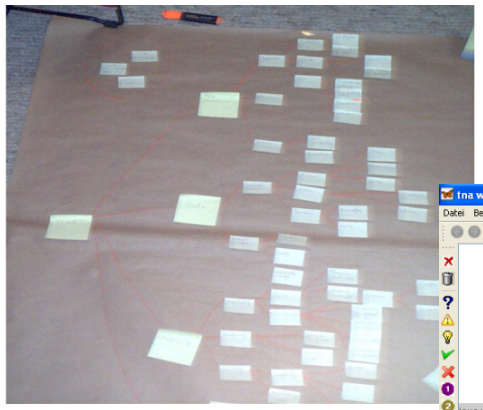


TU WIEN planets UNIVERSITY OF Southampton eprints

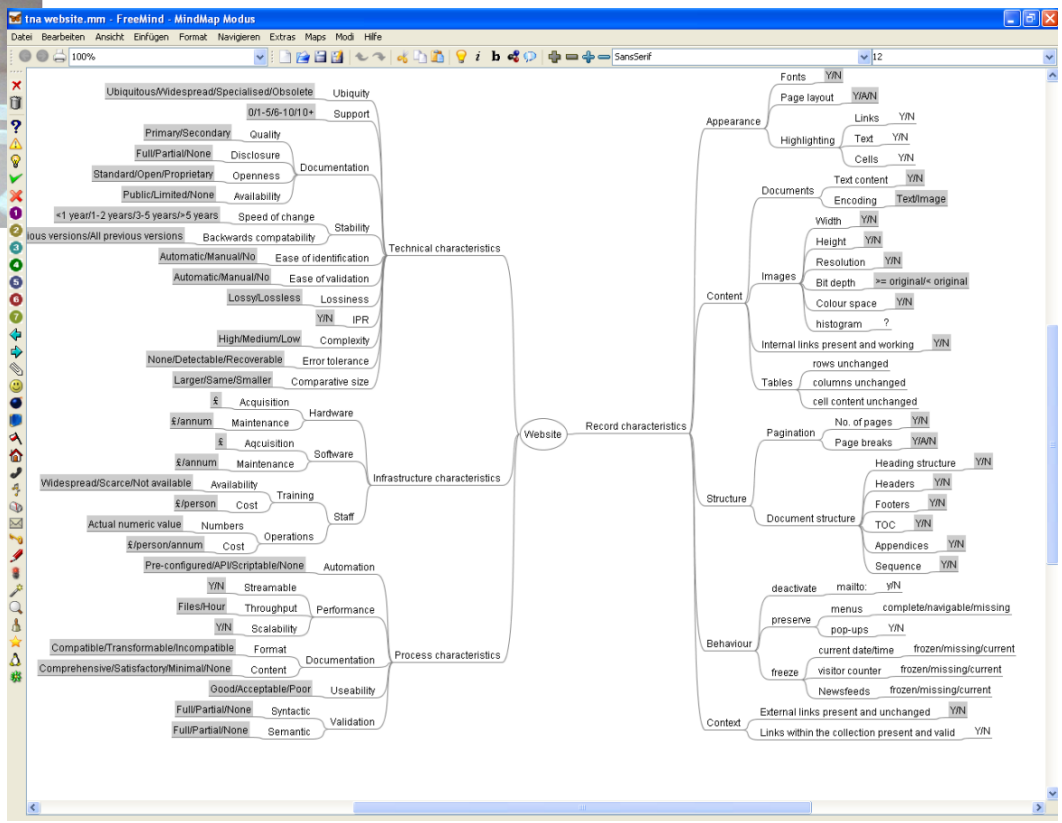
Identify requirements







Analog…

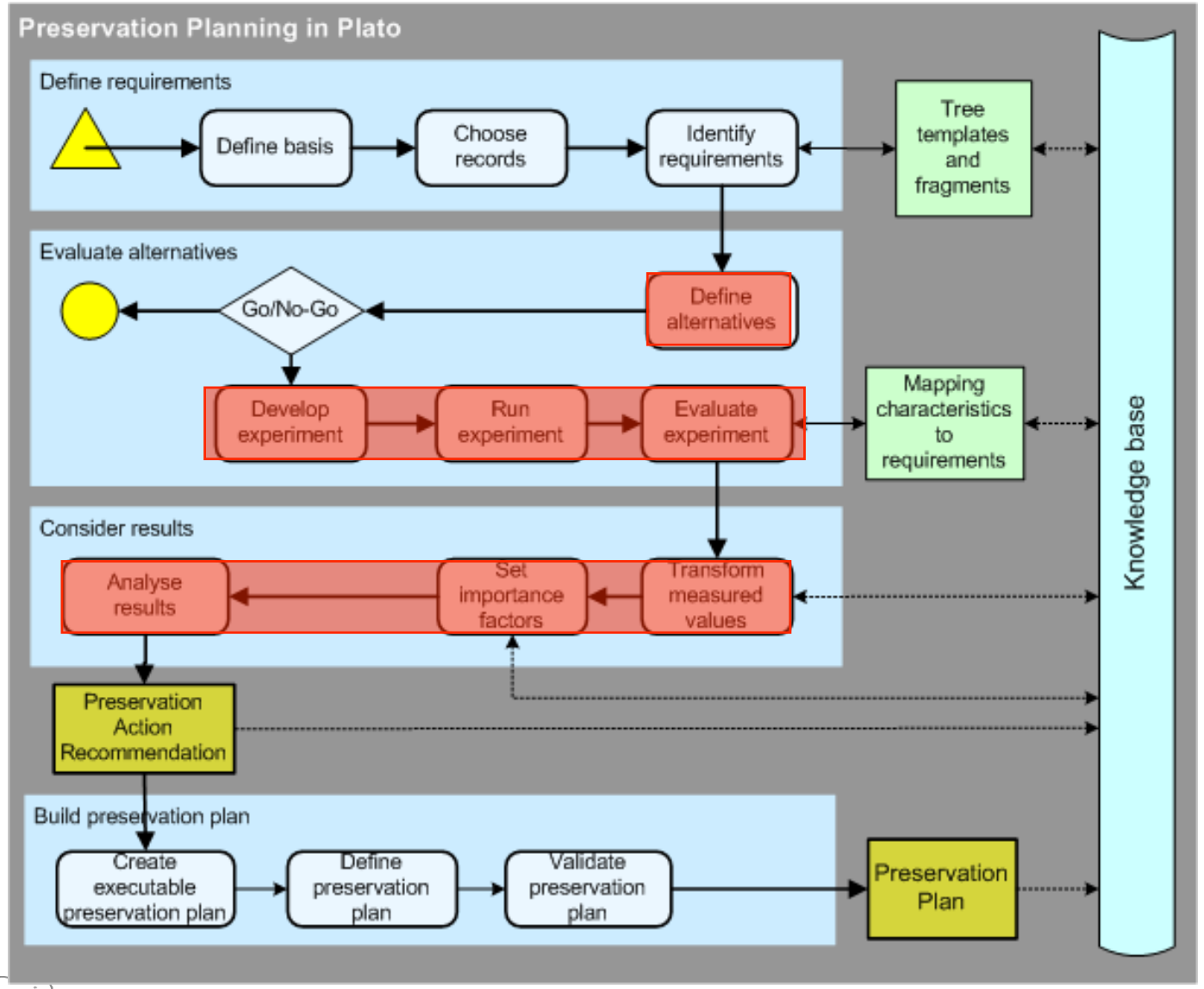


… or born digital

TU WIEN planets

UNIVERSITY OF Southampton eprints

Preservation Planning Workflow



TU WIEN planets UNIVERSITY OF Southampton eprints

Preservation Planning with Plato

Plato

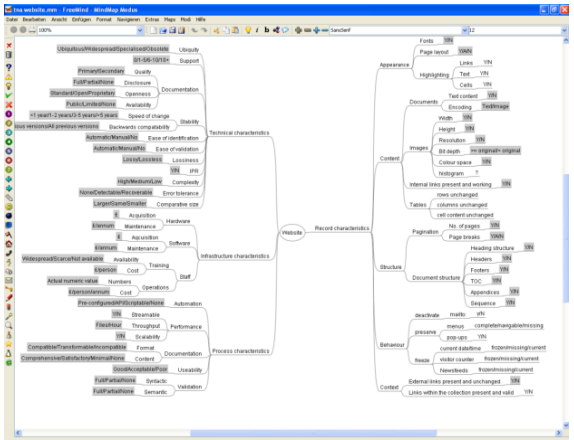
- Preservation Planning Tool
- Reference implementation of planning workflow
- Web-based application, release 2.0 Nov. 12 2008
- Documents the process and ensures that all steps are considered
- Automates several steps
- Creates a preservation plan (XML, PDF)
- Technical basis:
  - Java Enterprise Beans, EJB 3 (Hibernate)
  - Based on JBoss Application Server
  - JBoss Seam Integration Framework
  - Java Server Faces with Facelets
  - XML Import/Export

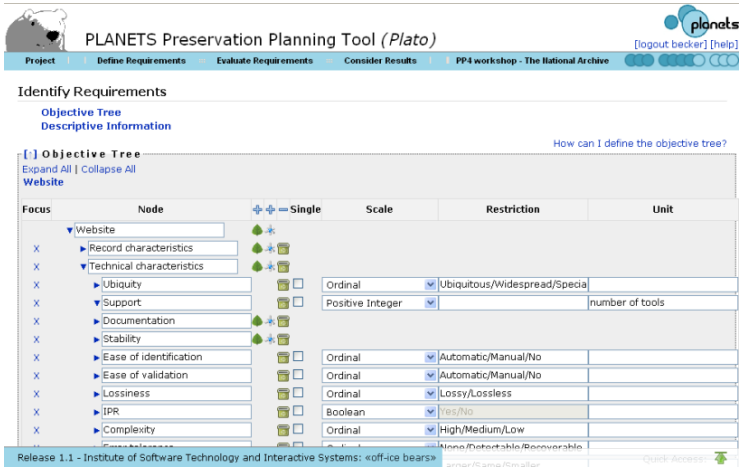
TU WIEN planets UNIVERSITY OF Southampton eprints

Preservation Planning with Plato

Plato

- 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



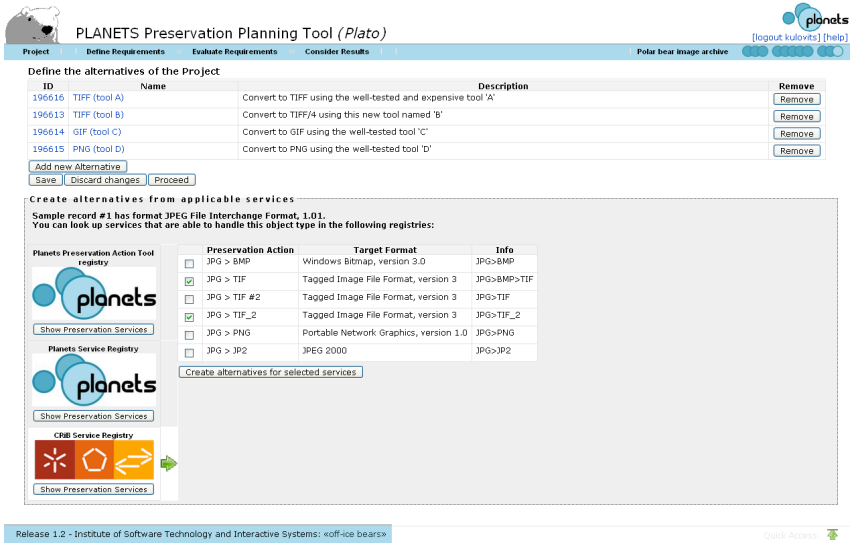


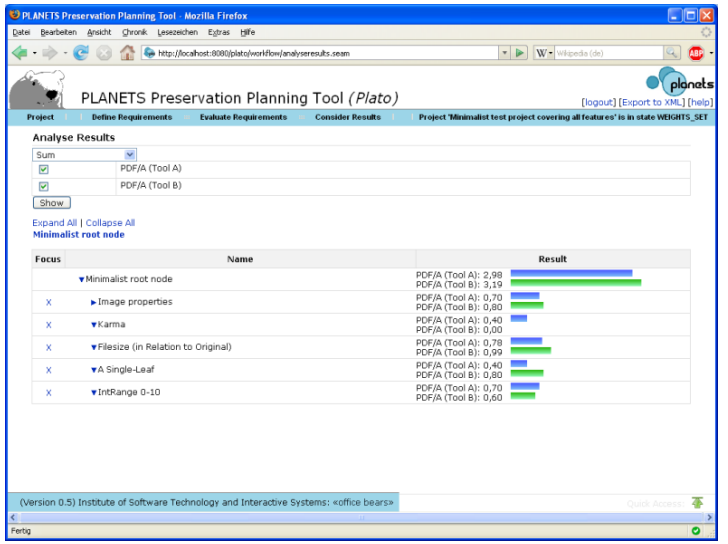
TU WIEN planets UNIVERSITY OF Southampton eprints

Preservation Planning with Plato

Plato

- Runs experiments and documents results
- Allows definition of transformation rules, weightings
- Performs evaluation, sensitivity analysis,
- Provides recommendation (ranks solutions)





Preservation Planning with Plato

What Preservation Planning produces:

- Basic Preservation Plan:
  - *PDF: Preservation Plan.pdf*
  - *XML: Preservation Plan.xml*
- That was developed in a solid, repeatable and documented process
- That is optimal for the needs of a given institution and for the data at hand

TU WIEN planets UNIVERSITY OF Southampton eprints

Conclusions

- Preservation Planning to ensure “optimal” preservation
- A simple, methodologically sound model to specify and document requirements
- Repeatable and documented evaluation
- Basis for well-informed, accountable decisions
- Concretization of OAIS model
- Follows recommendations of TRAC and nestor
- Generic workflow that can easily be integrated in different institutional settings
- **Plato**:
  - Tool support to perform solid, well-documented analyses
  - Creates core preservation plan

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

TU WIEN planets UNIVERSITY OF Southampton eprints

Schedule

(1) Introduction
- What is Digital Preservation?
- EPrints
- Preservation Planning and Plato

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing

TU WIEN planets UNIVERSITY OF Southampton eprints

Summary

1. Storage Ecosystem
   - Environmental study
2. Storage Controller
   - Interacting with your environment
3. Managing Stored Assets
   - Ensuring the future of your data

Where can we store data?

STORAGE ECOSYSTEM

Local Disk Storage

- No local bandwidth costs
- Hard to expand
- Locally Managed
- High overheads cost
- Requires space and cooling
- Tied closely to the software

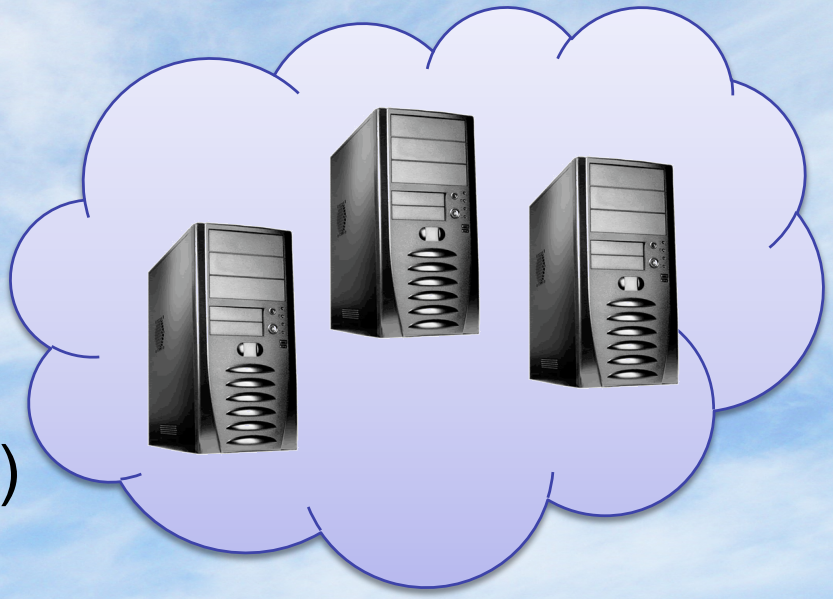
Local Archival Storage

- Specialist
- Expensive to purchase
- Locally Managed
- Space and running costs
- Expandable

TU WIEN planets UNIVERSITY OF Southampton eprints

Cloud Storage

- Scalable
- Externally controlled
- Known Costings
- Unclear retention policy
- Re-Useable (using simple APIs)
- Global Scale



But Clouds Blow Away

In the last 24 months:

- Yahoo Briefcase
- XDrive
- AOL Pictures
- HP Upline
- Sony Image Station

Source: Tom Spring - PCWorld

Why use Hybrid Storage

- Use the best features of each storage type
- Performance
  - Scaling-up bandwidth
- Optimisation
  - Large-file handling
  - Multimedia streaming
- Localised Delivery
  - Local delivery from the cloud

Which storage should we use?

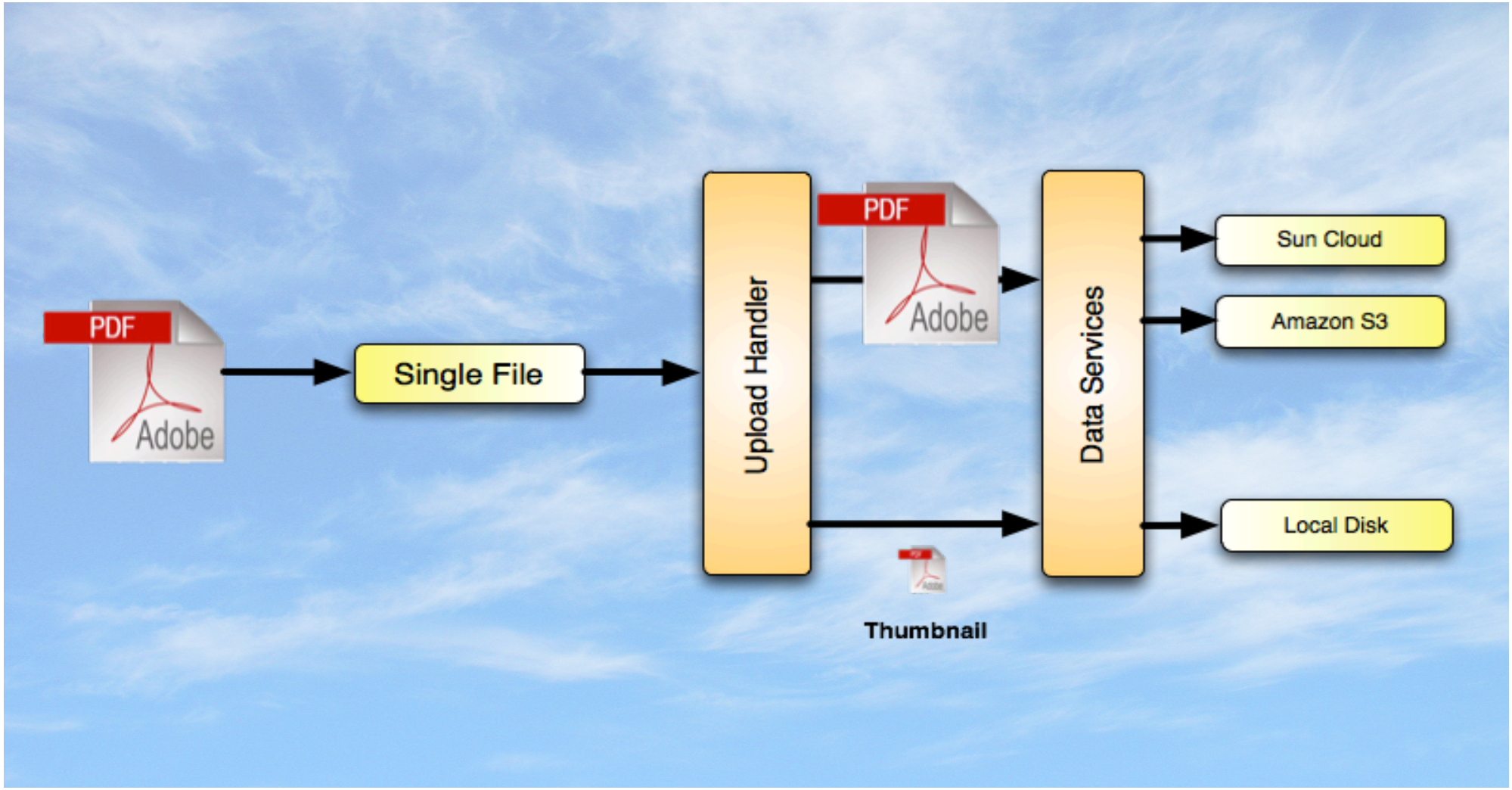
STORAGE CONTROLLER

EPrints Storage Controller

- The storage controller decides where to put a file.
- Uses rule based policy defined by simple configuration file (XML)
- Examples:
  - Large binary files of scientific data (raw machine result data) can be stored in a large disk (slower access) system and sent to a tape company for long term storage.
  - Processed results can be stored locally and in the cloud ready for rapid delivery to end points.

TU WIEN planets UNIVERSITY OF Southampton eprints

Hybrid Storage Policies



Desktop & Cloud Integration
Part 1: Hybrid Storage Policies

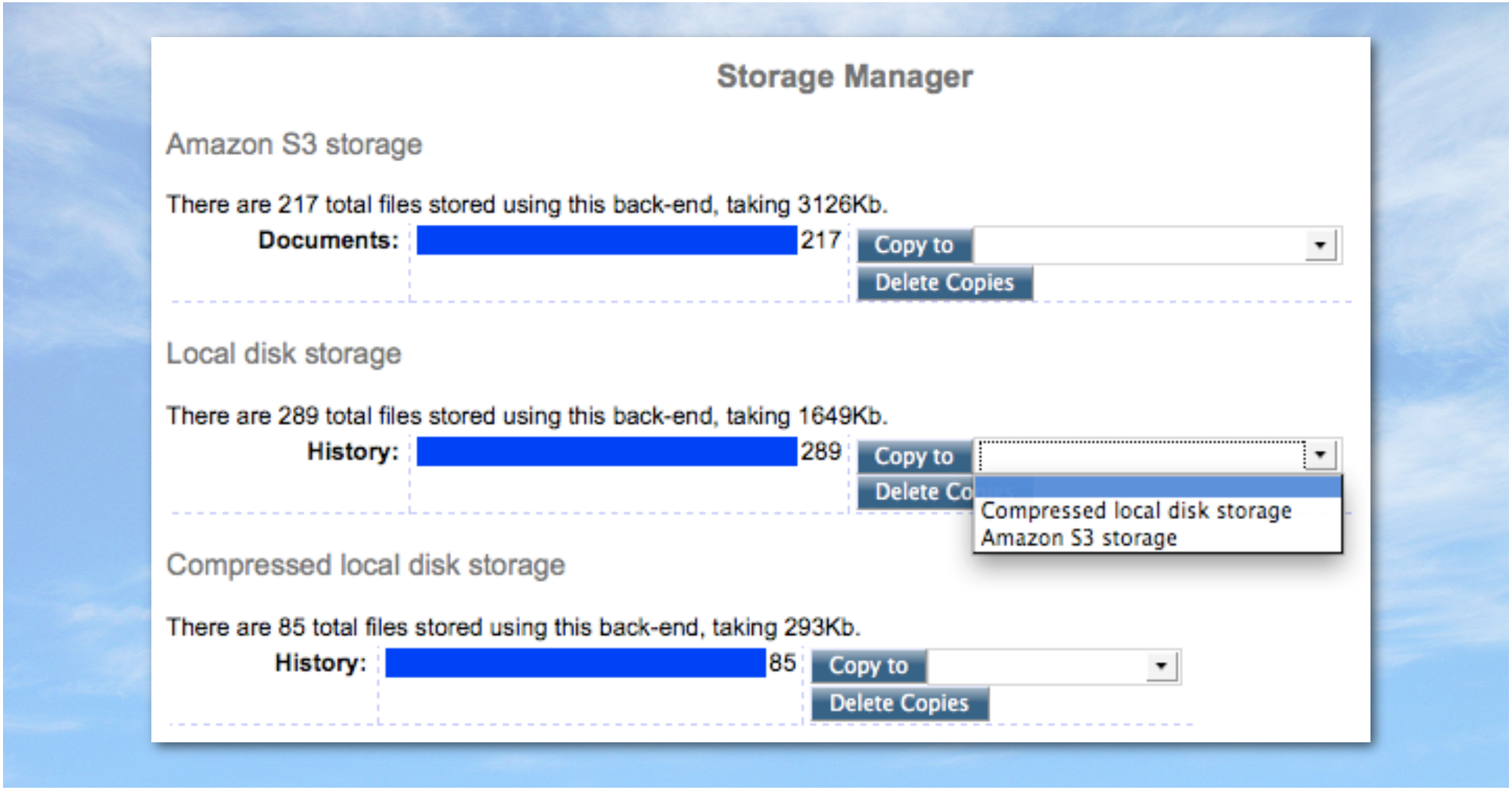
```
<choose>
    <when test="datasetid = 'document'">
        <choose>
            <when test="$parent{relation_type} = 'isVolatileVersionOf'">
                <plugin name="Local"/>
            </when>
            <otherwise>
                <plugin name="SunCSS"/>
                <plugin name="AmazonS3"/>
            </otherwise>
        </choose>
    </when>
    <otherwise>
        <plugin name="Local"/>
    </otherwise>
</choose>
```

TU WIEN planets UNIVERSITY OF Southampton eprints

How do I move data around?

MANAGING STORED ASSETS

EPrints Storage Manager

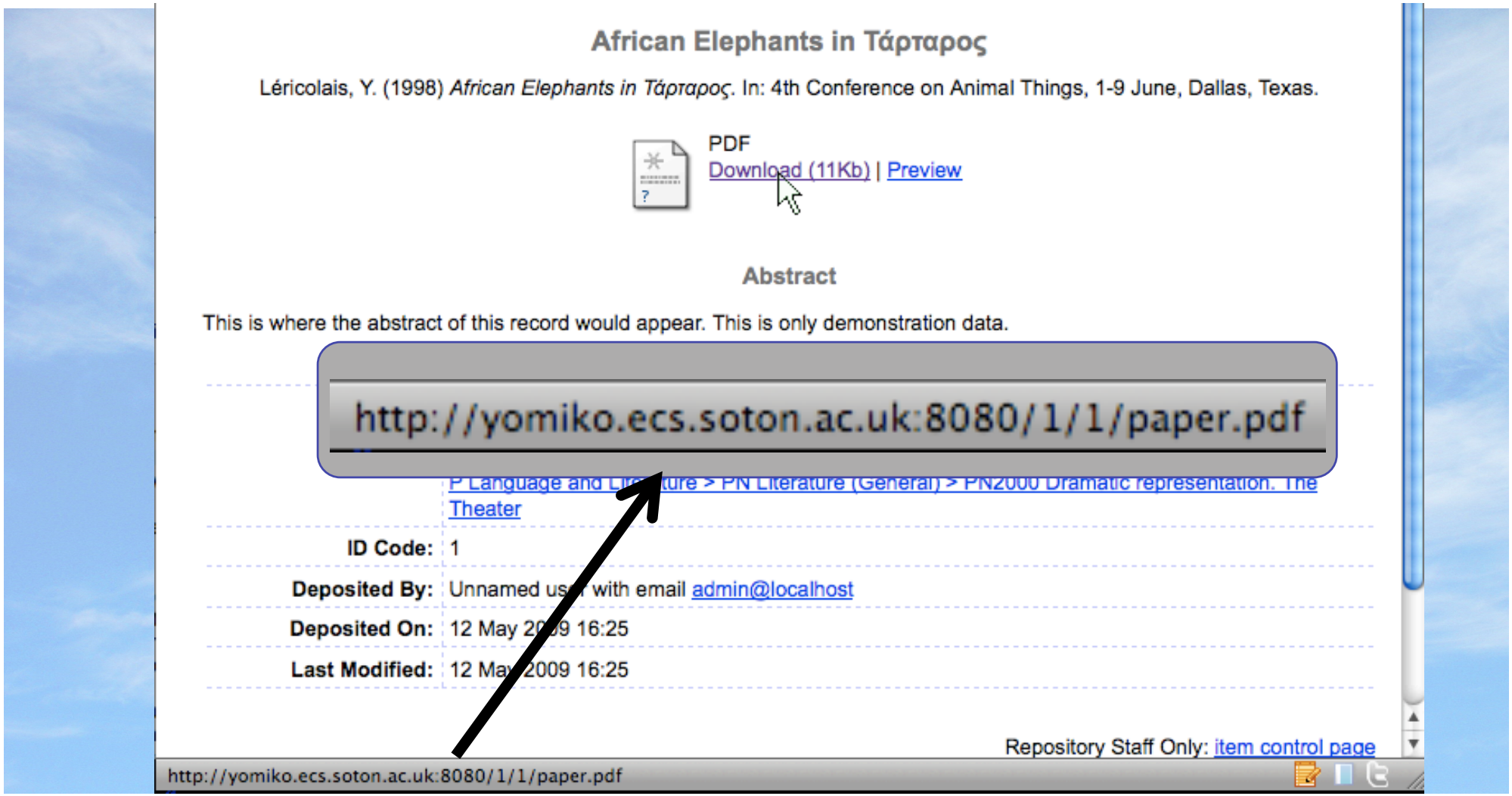


TU WIEN planets

UNIVERSITY OF Southampton

eprints

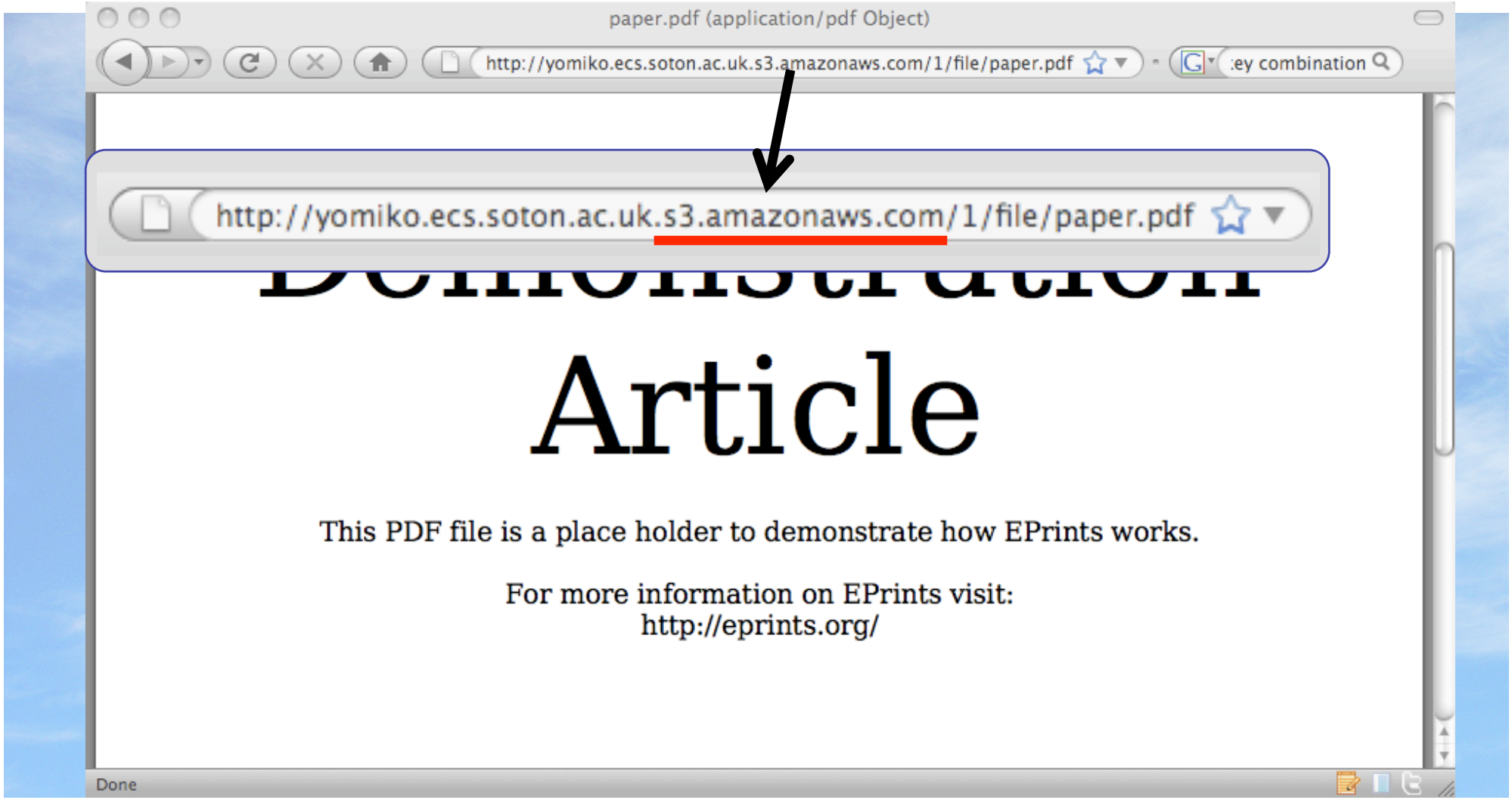
Amazon S3 Localisation (1)



TU WIEN planets

UNIVERSITY OF Southampton eprints

Amazon S3 Localisation (2)



TU WIEN planets

UNIVERSITY OF Southampton eprints

Recap

1. Storage Ecosystem
   - There are a great number of products and services available designed to protect your resources. Each is aimed at a market with different needs based on the type of content.
2. Storage Controller
   - Allows you to utilise a diverse range of storage services simultaneously. Take advantage of the current ecosystem.
3. Managing Stored Assets
   - If the ecosystem changes, moving of resources to a new service is a seamless operation.

The Preservation Process

Preservation - Check

- Bit checking & checksum calculation

Preservation - Analyse

- What is the type of file, is the file valid?
- Is the file at risk of not having an editor/reader?
- Is there a better format available? Lossless or Lossy?

Preservation - Action

- File migration to avert risks found by analysis.
- Movement of file to new storage.

Analysis

Preservation - Analyse

- What is the type of file, is the file valid?
  - Droid is a good classification tool for this.

- Is the file at risk of not having an editor/reader?
  - Functionality is being developed in PRONOM technical registry.

The technical registry
PRONOM

- Is there a better format available? Lossless or Lossy?

.

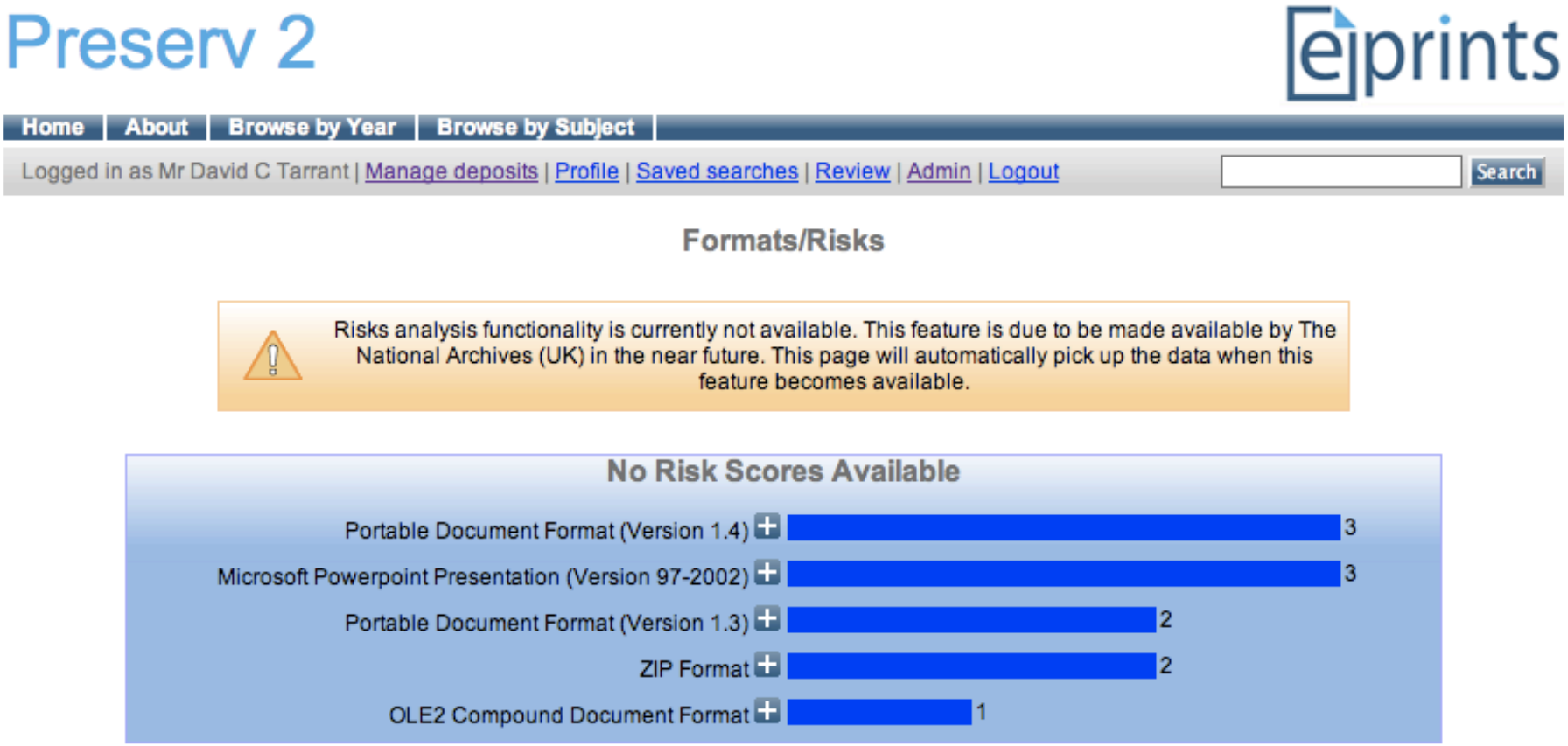
UNIVERSITY OF Southampton

eprints

File Format Analysis

Preservation - Analyse

EPrints File Classification



TU WIEN planets

University of Southampton eprints

Risk Analysis

Preservation - Analyse

The technical registry PRONOM

- Is the file at risk of not having an editor/reader?
  - Functionality is being developed in PRONOM technical registry.

- Simple SOAP web service

- Takes file format identification id's, hands back risk score.
- Breakdown of risk score may also be available in future releases.

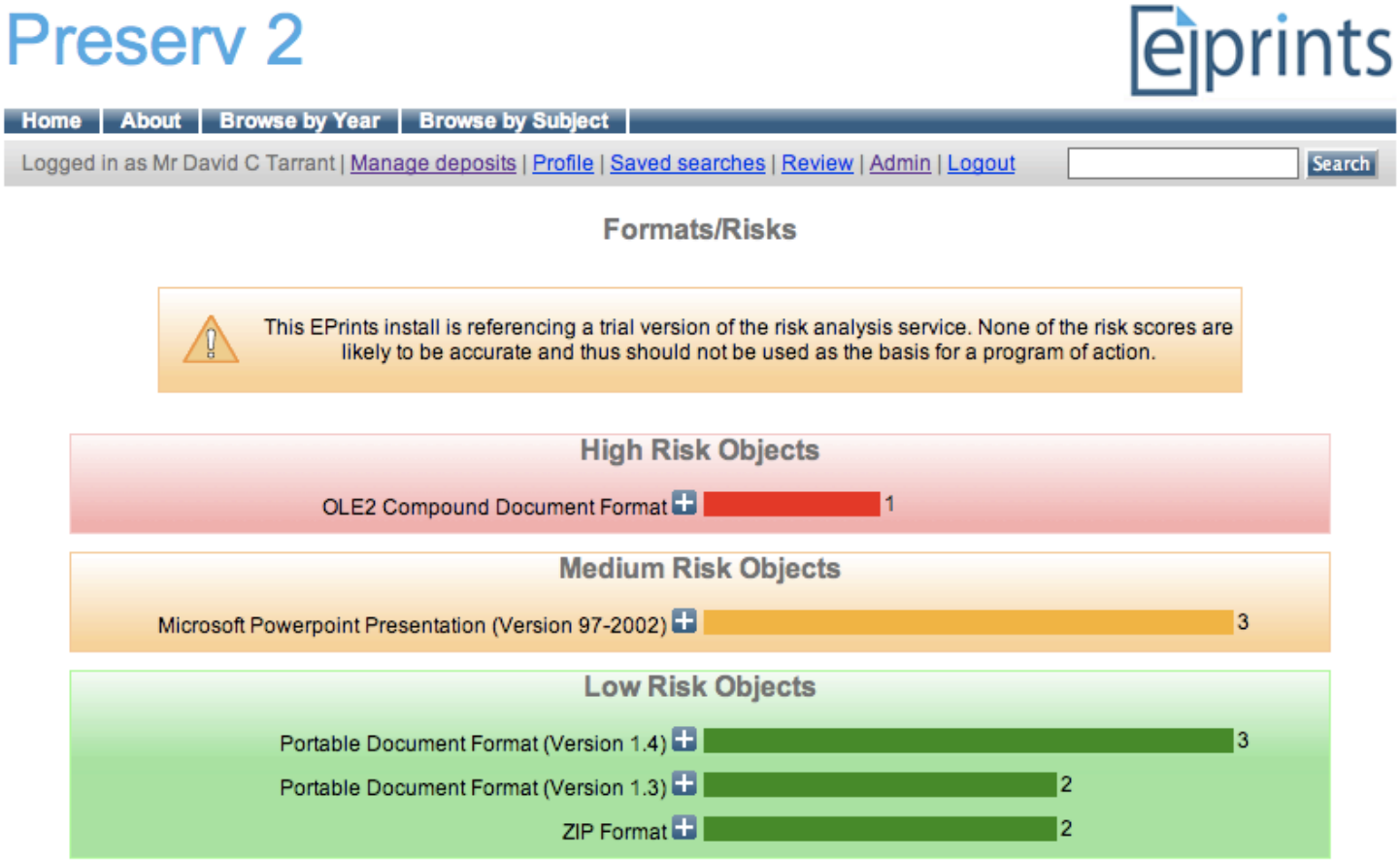
- A stub you can download and run providing this functionality before the official release with mock up risk scores is available at http://preserv2.googlecode.com

TU WIEN planets University of Southampton eprints

Risk Analysis In EPrints

Preservation - Analyse

EPrints File Classification + Risk Analysis



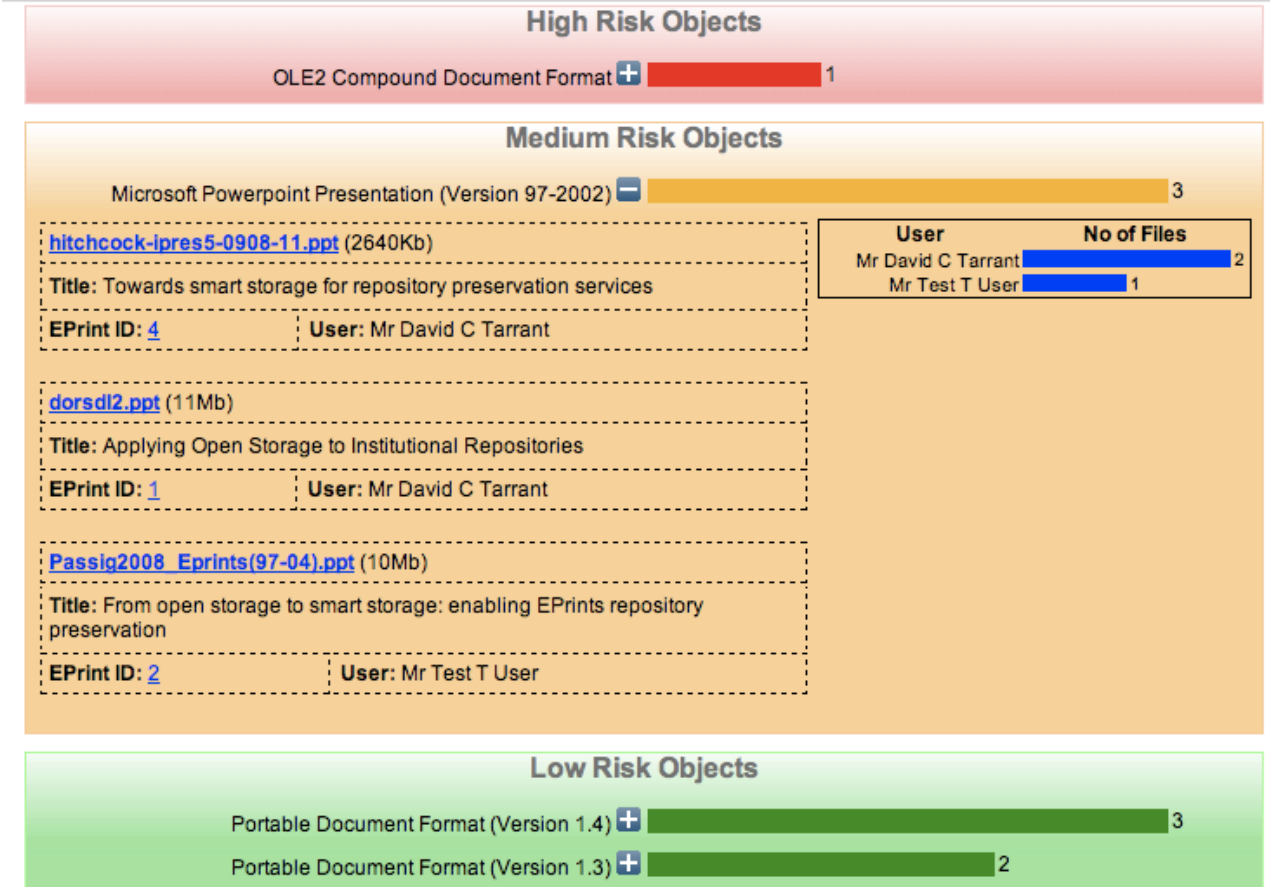
TU WIEN planets

UNIVERSITY OF Southampton eprints

Risk Analysis In EPrints
Detail View

Preservation - Analyse

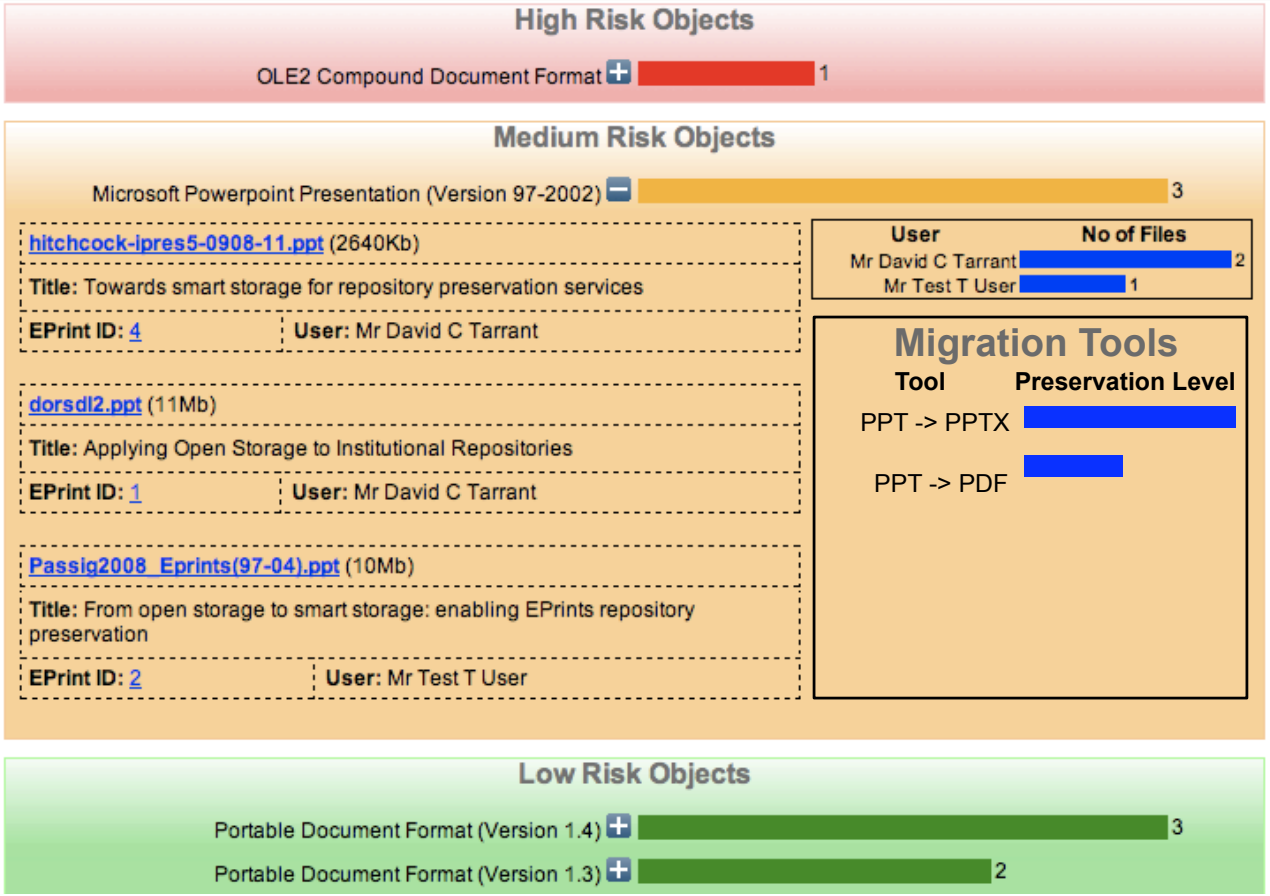
EPrints File Classification + Risk Analysis



Risk Analysis In EPrints Migration?

Preservation - Action

Mock up Transformation Interface



Recap

Preservation - Check

- Handled by our storage manager and reported back via the preservation interface.

Preservation - Analyse

- Parallels can be drawn with storage, in that we are integrating with and utilising currently available services to perform our analysis.
- Processing of the results leads to a powerful interface which tells us many things about the repository ecosystem and it's future.

Preservation - Action

- Future plan is to utilise further web based services to ensure information remains comprehensive and up to date set, 0day digital preservation.

Schedule

(1) Introduction
- What is Digital Preservation?
- EPrints
- Preservation Planning and Plato

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing

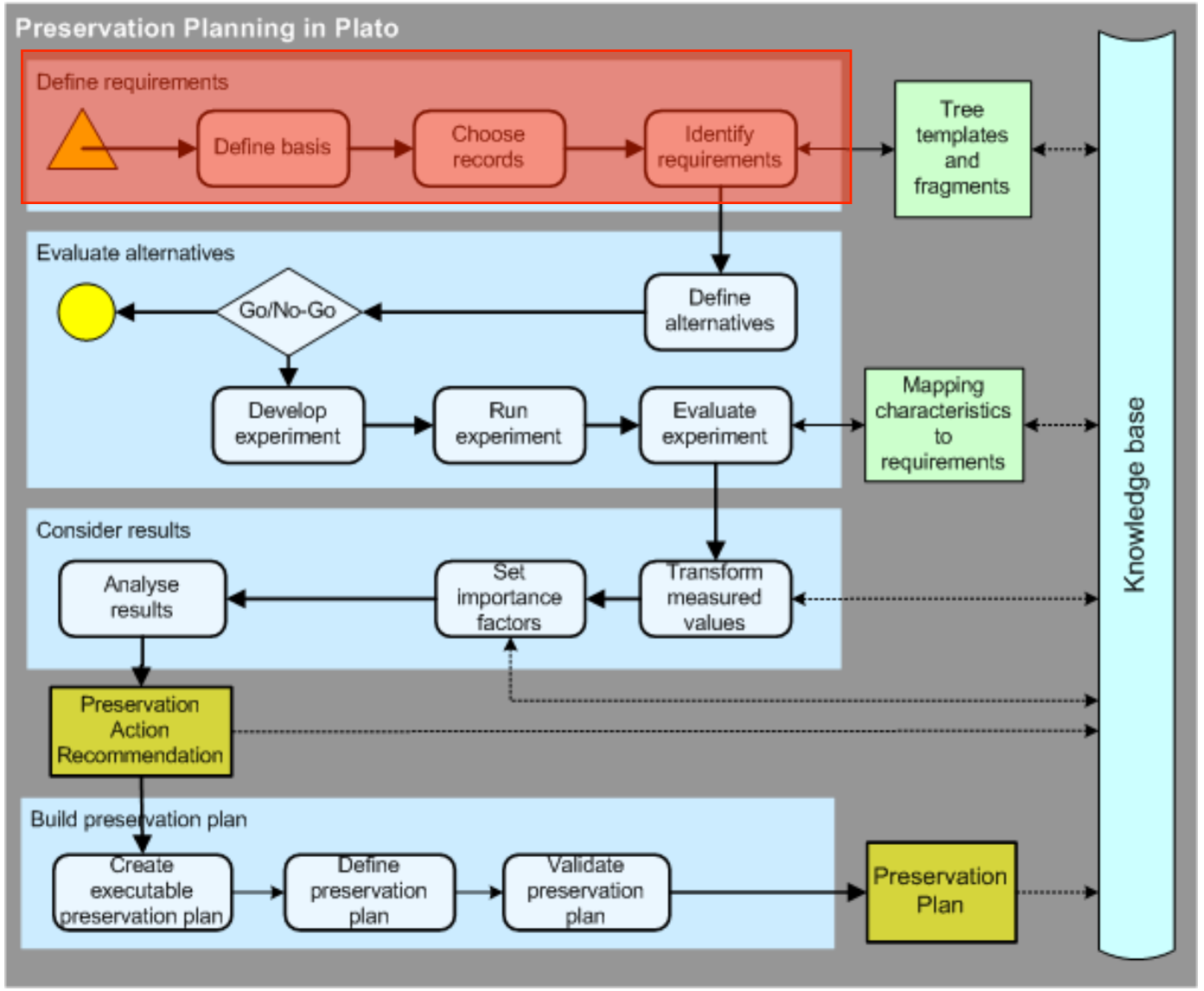
TU WIEN planets UNIVERSITY OF Southampton eprints

Overview

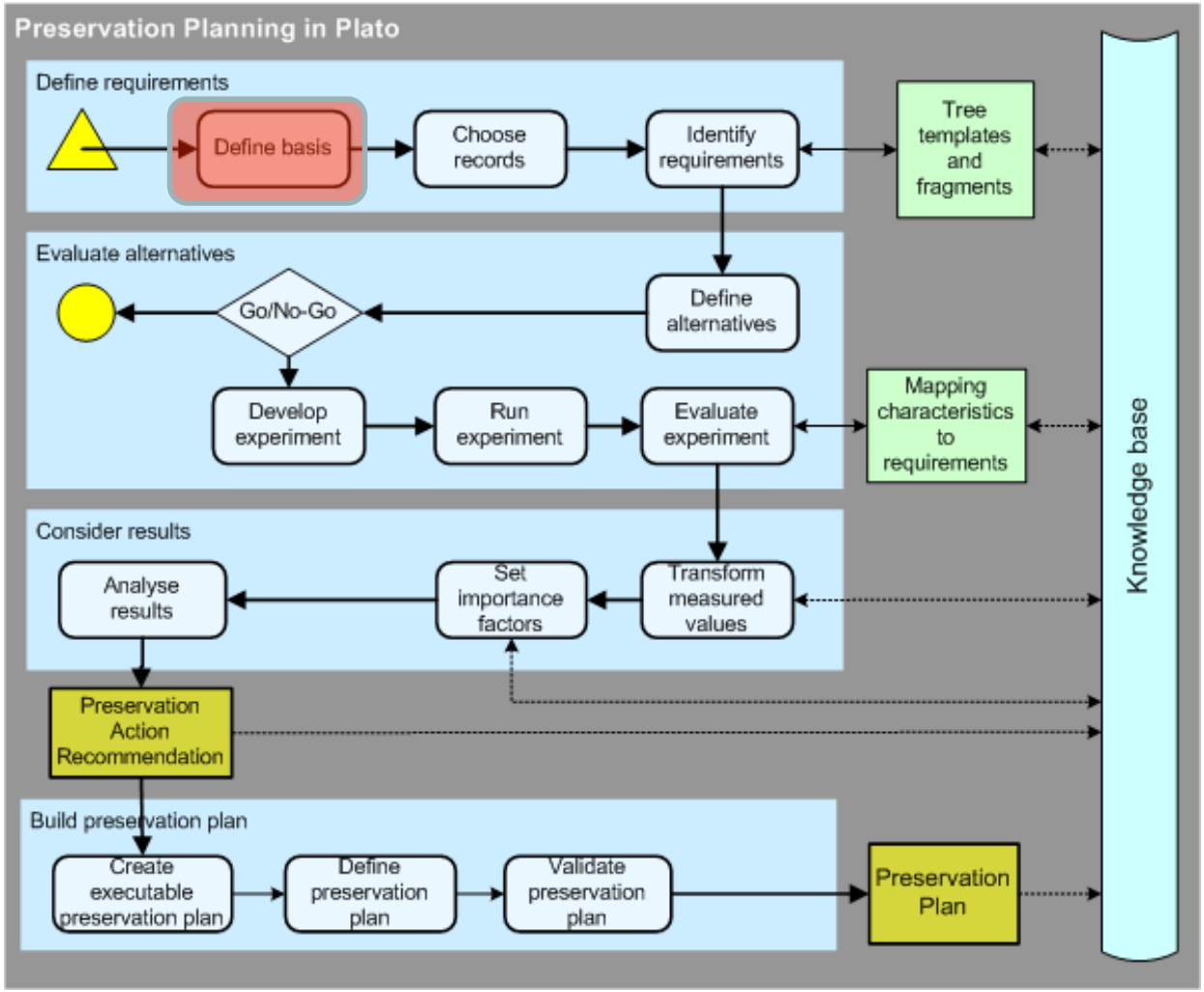
Part 3: Preservation Planning with Plato

- Preservation planning workflow
- Exercises

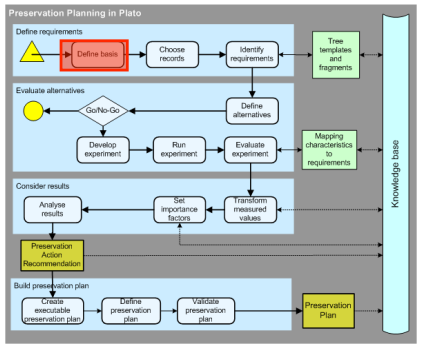
PP Workflow


Preservation Planning in Plato
Define requirements
Define basis
Choose records
Identify requirements
Tree templates and fragments
Evaluate alternatives
Go/No-Go
Define alternatives
Develop experiment
Run experiment
Evaluate experiment
Mapping characteristics to requirements
Consider 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
Knowledge base

Orientation



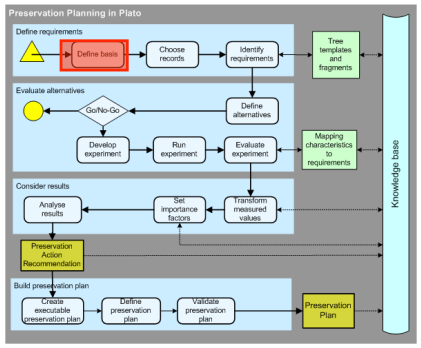
Define Basis



- Basic preservation plan properties
- Describe the context
  - Institutional settings
  - Legal obligations
  - User groups, target community
  - Organisational constraints
- 5 triggers
  - New Collection Alert (NCA)
  - Changed Collection Profile Alert (CPA)
  - Changed Environment Alert (CEA)
  - Changed Objective Alert (COA)
  - Periodic Review Alert (PRA)

TU WIEN planets UNIVERSITY OF Southampton eprints

Define Basis

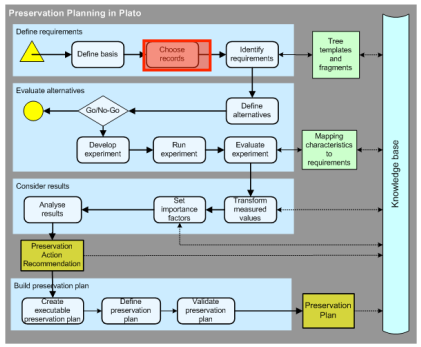


Organizational structure

- Mandate, Mission Statement
  - Provide reliable, long-term access to digital objects
  - Internet Archive: *"The Internet Archive is working to **prevent the Internet** […] and other 'born digital' materials **from disappearing** into the past. Collaborating with institutions including the Library of Congress and the Smithsonian, we are working to preserve a record for generations to come."*
    http://www.archive.org/about/about.php
  - Oxford Digital Library: *"Like traditional collection development **long-term sustainability** and **permanent availability** are major goals for the Oxford Digital Library."*
    http://www.odl.ox.ac.uk/principles.htm

TU WIEN planets UNIVERSITY OF Southampton eprints

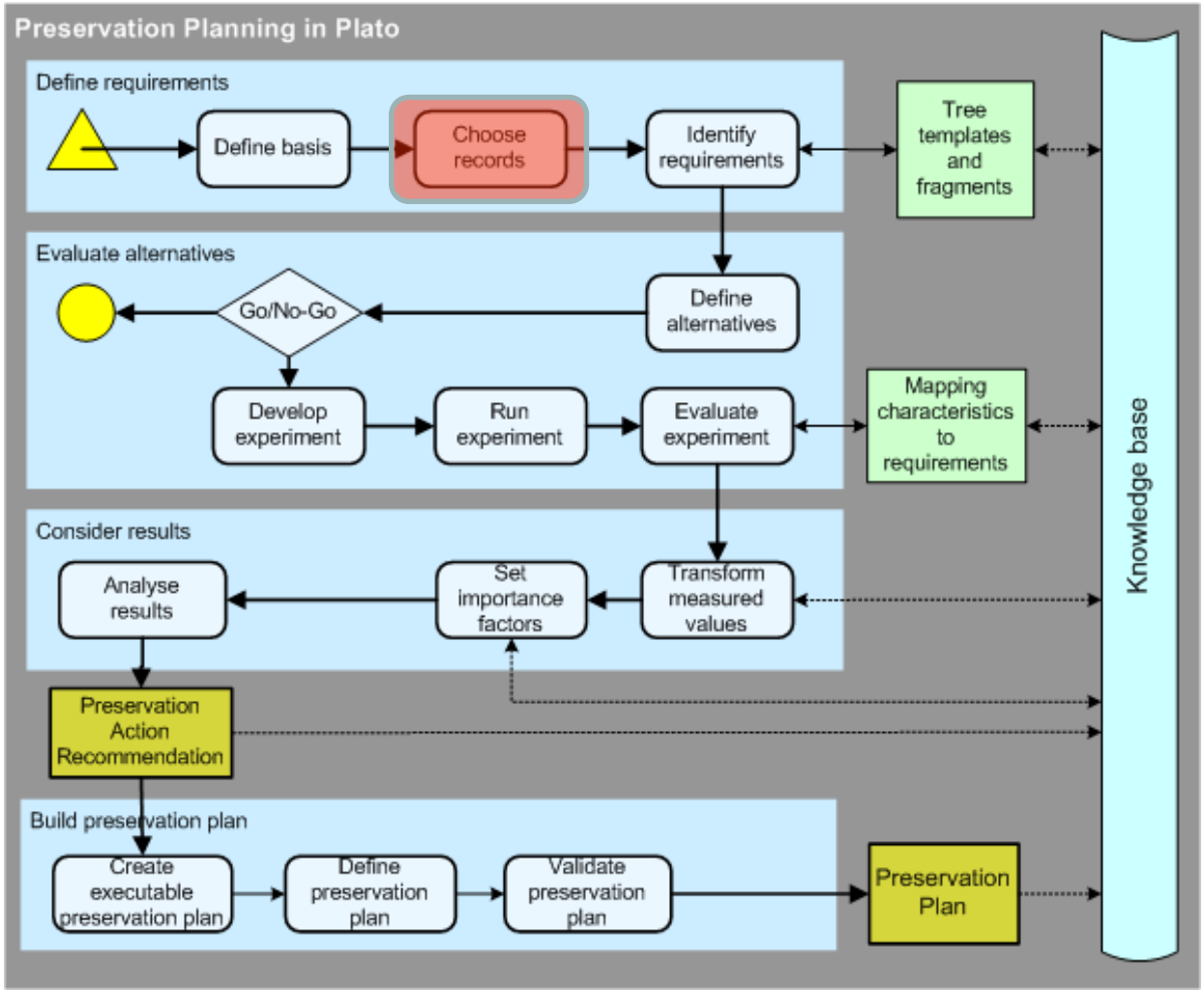
Define Basis



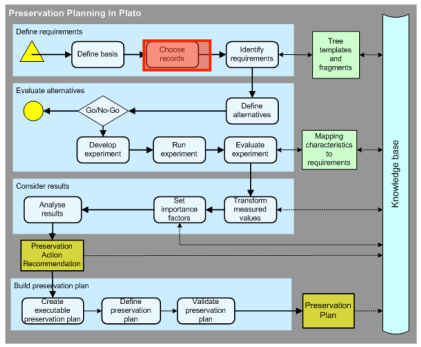
[↑] Identification

Identification Code:	
Number of Objects:	all electronic theses submitted from 2008 on
Document types:	electronic publications in PDF format. Origin ranging from MS Word to Latex and other systems
Plan name:	ONB: electronic theses and dissertations
Plan description:	PLANETS PP4 case study on preserving Austrian electronic theses and dissertations
Responsible planners:	Christoph Becker, Stephan Strodl
Organisation:	Vienna University of Technology\Austrian National Library

Orientation



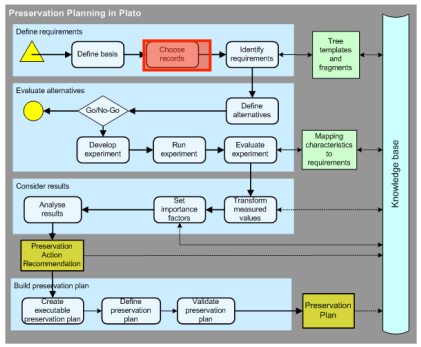
Choose Sample Objects



- 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
  - Often between 3 – 10

TU WIEN planets UNIVERSITY OF Southampton eprints

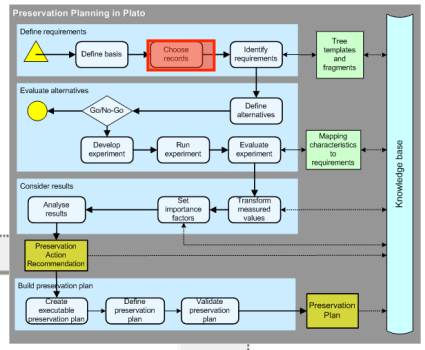
Choose Sample Objects



- Stratification – all essential groups of digital objects should be chosen according to their relevance
- Possible stratification strategies
  - File type
  - Size
  - Content (e.g. document with lots of images, including macros)
  - Time (objects from different periods of times)
- File Format Identification
  - DROID
  - PRONOM

TU WIEN planets UNIVERSITY OF Southampton eprints

Define Sample Objects



[↑] Sample Records

Description of sample records: several samples of electronic theses ?

Sample Record		Object Format	
Full name:	sample thesis 1 ?	PUID:	?
Short name:	DA1 ?	Name:	?
Has data:	☑ download	Version:	?
Original technical environment:	?	Mime-type:	?
Description:	?		Identify format

Remove record

Save | Discard changes | Proceed

Add new record without file

Add record

Upload new record

Durchsuchen... | Upload File ?

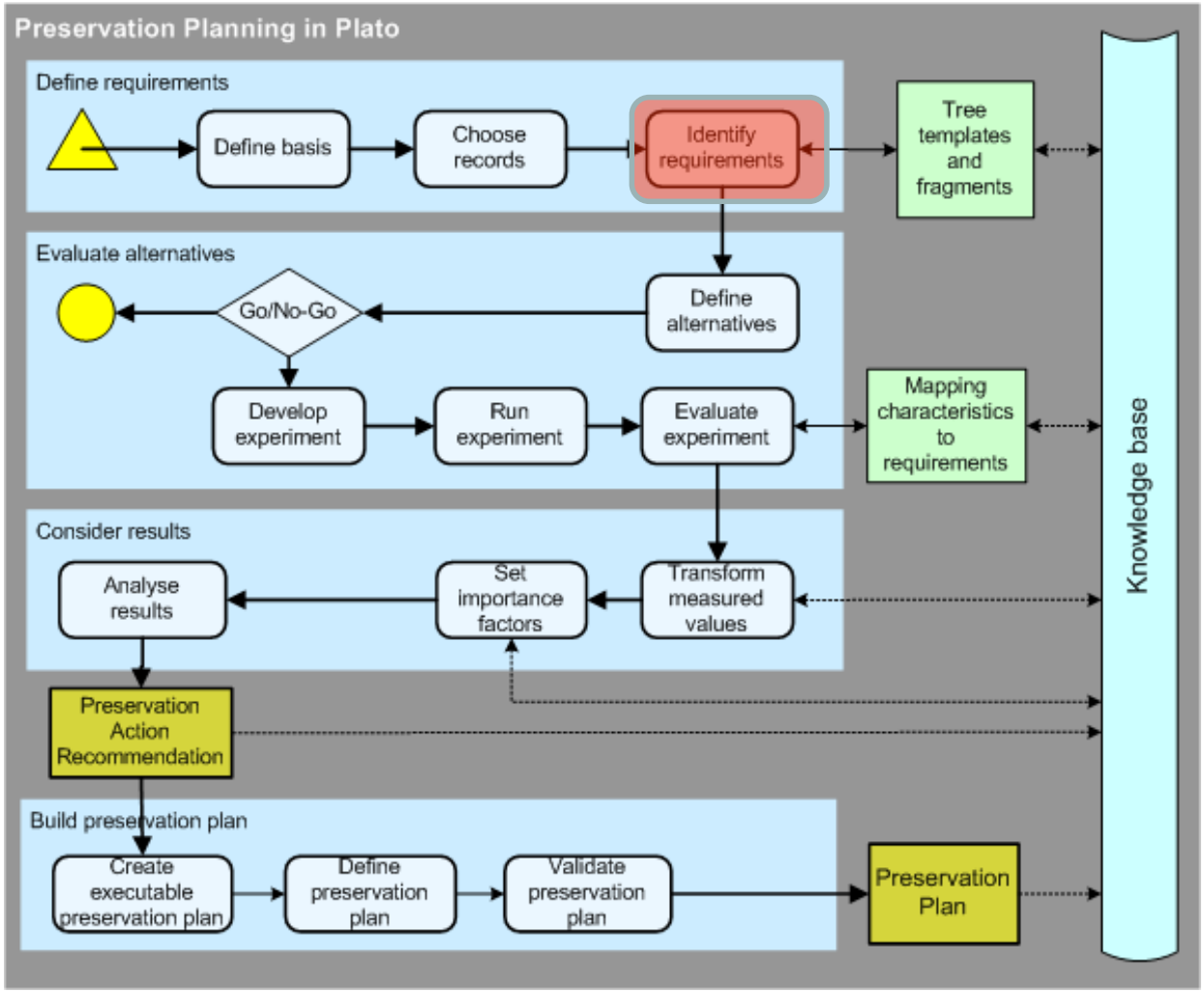
TU WIEN planets

UNIVERSITY OF Southampton eprints

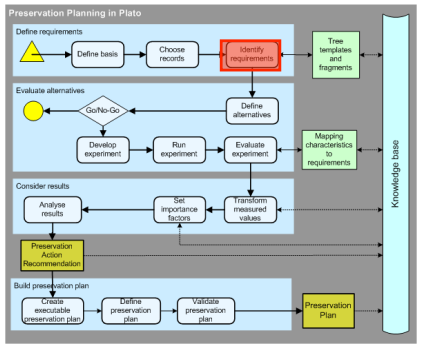
Practise time!

- Public institution – State and University Library
- Mission to preserve the state's cultural heritage in the form of any publication
- Scanned collection of yearbooks, 9000 objects
  - One file per page
  - Scans are black and white
- Copyright held for the physical material, same for digital content
- Objects are provided

Orientation

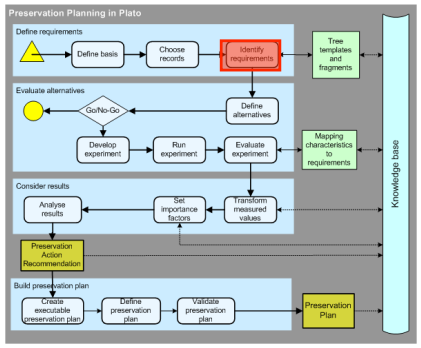


Identify Requirements

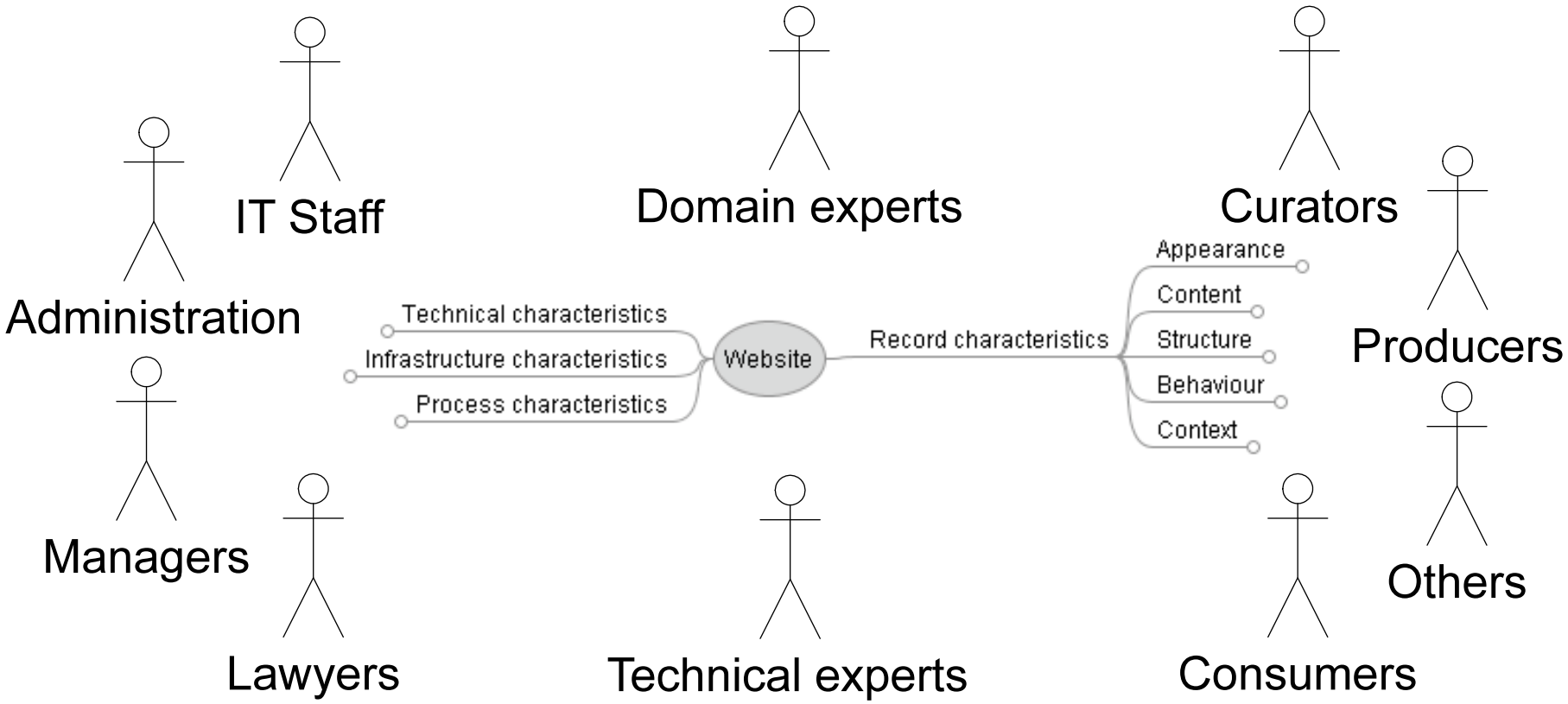


- Define all relevant goals and characteristics (high-level, detail) with respect to a given application domain
- Put the requirements in relation to each other
  → Tree structure
- Top-down or bottom-up
  - Start from high-level goals and break down to specific criteria
  - Collect criteria and organize in tree structure

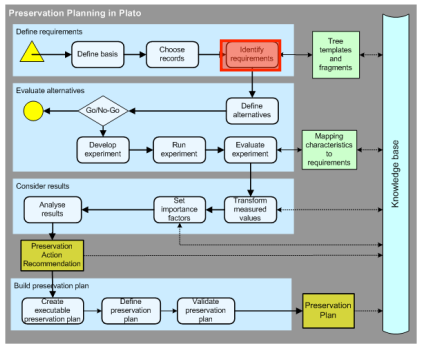
Identify Requirements



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

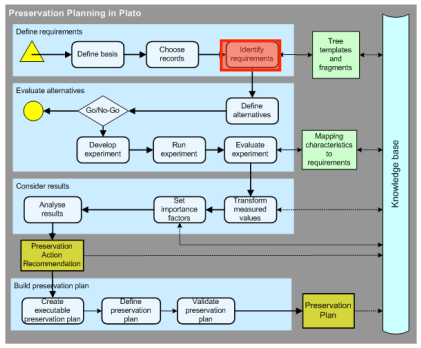


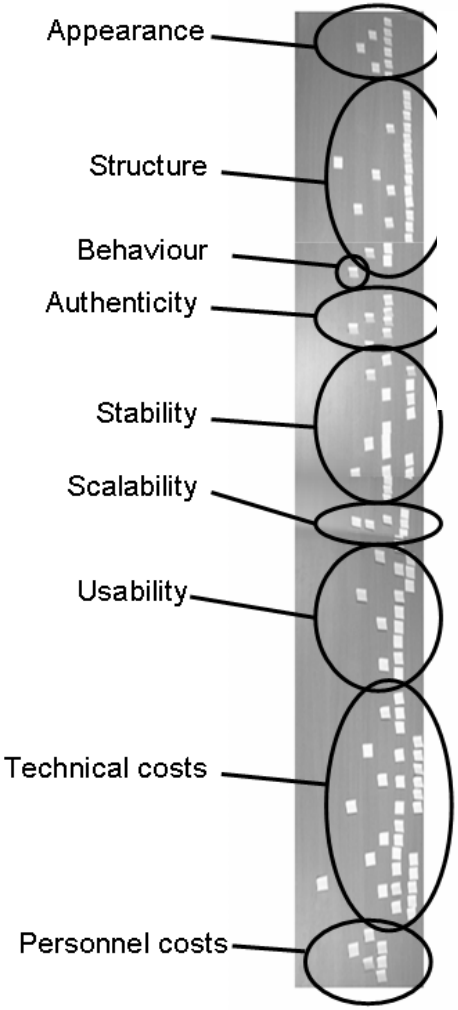
Identify requirements



- Core step in the process
- Define all relevant goals and characteristics (high-level, detail) with respect to 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,…)

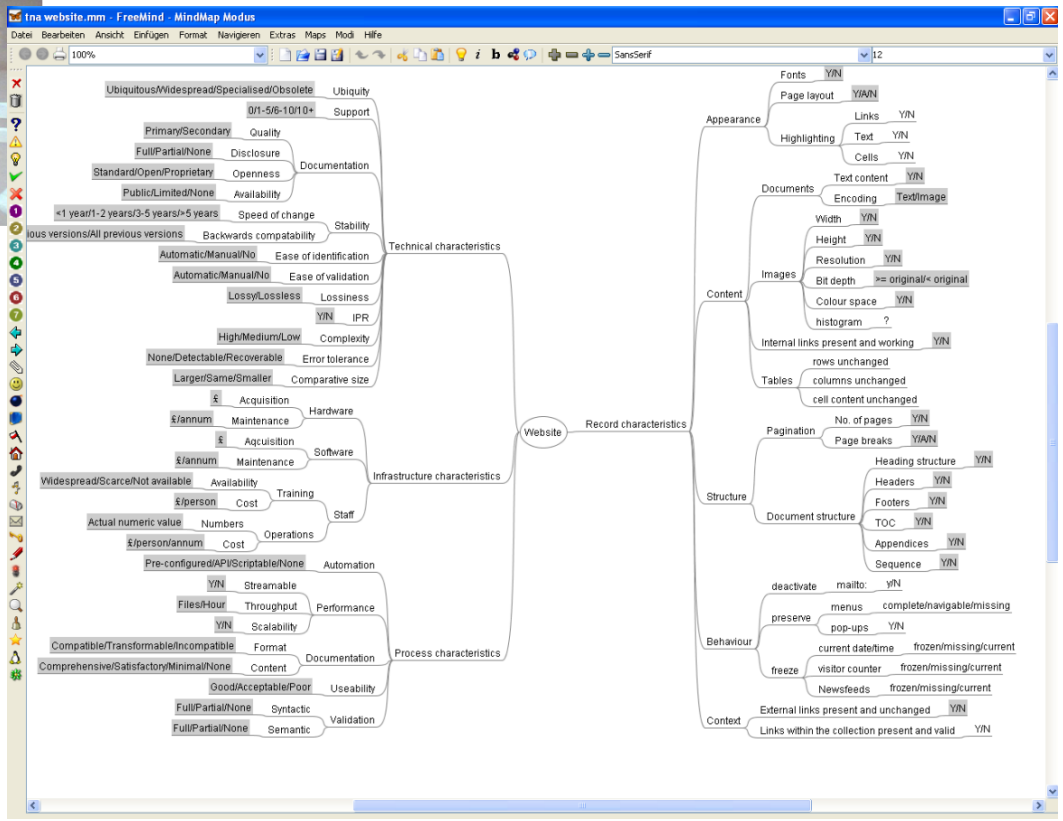
Identify requirements



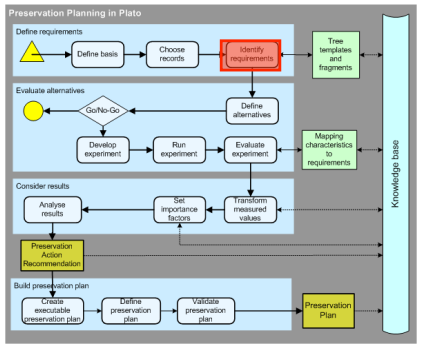


analogue…

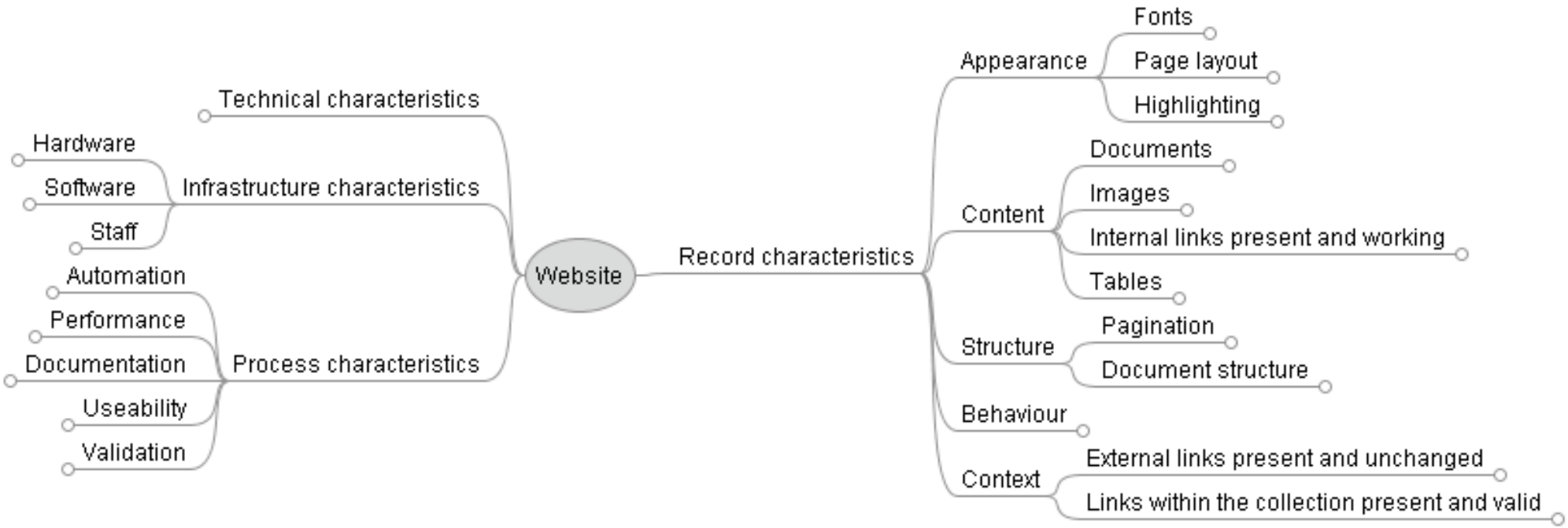
… or digital



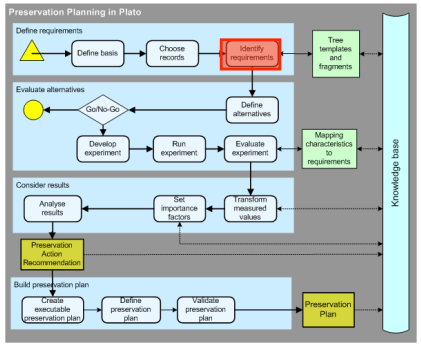
Identify requirements



Example: Webarchive



Identify requirements



- Creation within PLATO with Tree-Editor

PLANETS Preservation Planning Tool *(Plato)*

[logout becker] [help]

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

Identify Requirements

Objective Tree
Descriptive Information

How can I define the objective tree?

[-] Objective Tree

Expand All | Collapse All

Website

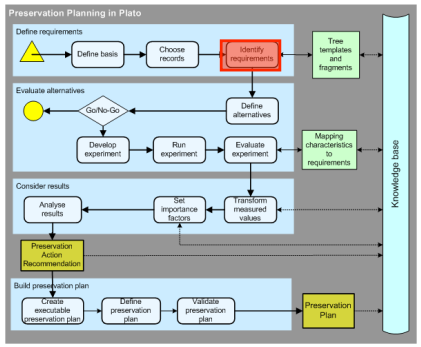
Focus	Node	Single	Scale	Restriction	Unit
	Website				
X	Record characteristics				
X	Technical characteristics				
X	Ubiquity	☐	Ordinal	Ubiquitous/Widespread/Specia	
X	Support	☐	Positive Integer		number of tools
X	Documentation				
X	Stability				
X	Ease of identification	☐	Ordinal	Automatic/Manual/No	
X	Ease of validation	☐	Ordinal	Automatic/Manual/No	
X	Lossiness	☐	Ordinal	Lossy/Lossless	
X	IPR	☐	Boolean	Yes/No	
X	Complexity	☐	Ordinal	High/Medium/Low	

Release 1.1 - Institute of Software Technology and Interactive Systems: «off-ice bears»

Quick Access:

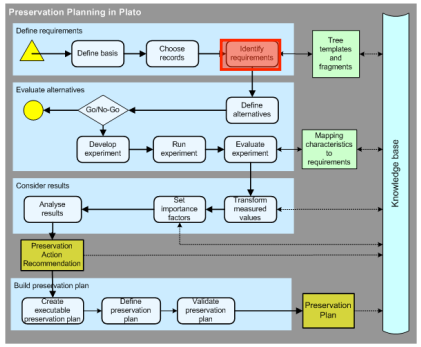
TU WIEN planets UNIVERSITY OF Southampton eprints

Identify requirements



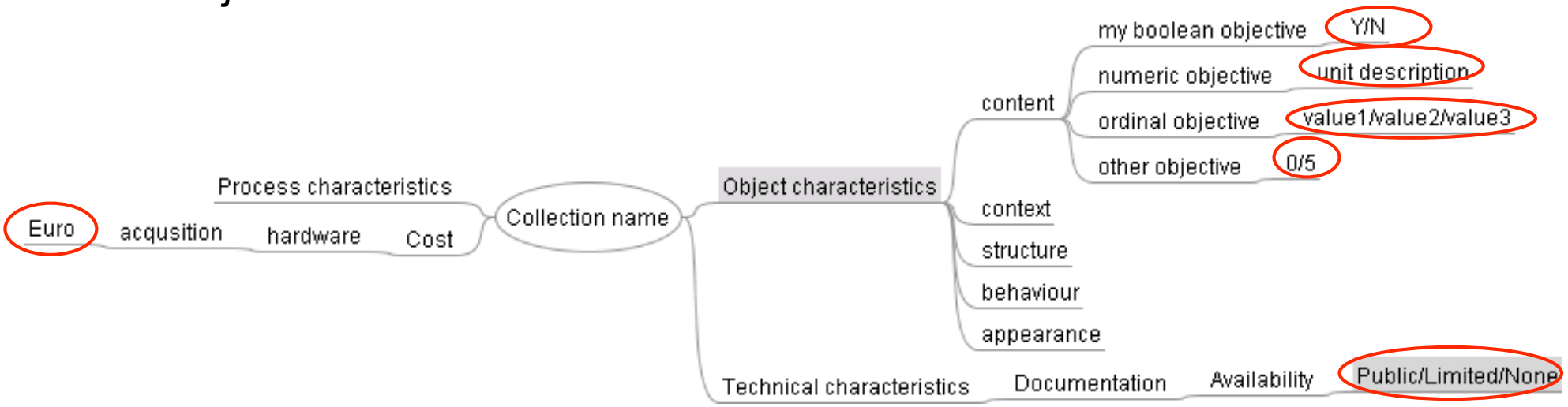
- Assign measurable unit to each leaf criterion
  - As far as possible automatically measurable
    - seconds / Euro per object
    - colour depth in bits
    - ...
  - Subjective measurement units where necessary
    - diffusion of file format
    - amount of expected support
    - ...
- No limitations on the type of scale used

Identify requirements

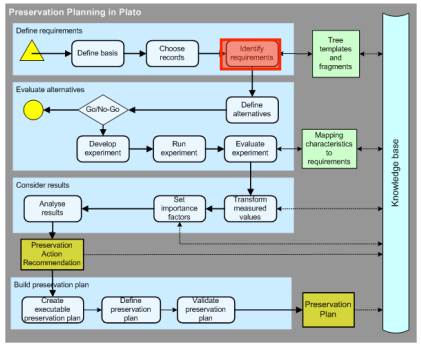


Types of scales

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



Identify requirements



- Creation within PLATO with Tree-Editor

PLANETS Preservation Planning Tool *(Plato)*

[logout becker] [help]

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

Identify Requirements

Objective Tree
Descriptive Information

How can I define the objective tree?

[-] Objective Tree

Expand All | Collapse All

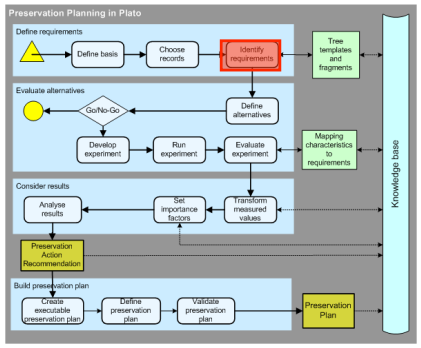
Website

Focus	Node	Single	Scale	Restriction	Unit
	Website				
X	Record characteristics				
X	Technical characteristics				
X	Ubiquity	☐	Ordinal	Ubiquitous/Widespread/Specia	
X	Support	☐	Positive Integer		number of tools
X	Documentation				
X	Stability				
X	Ease of identification	☐	Ordinal	Automatic/Manual/No	
X	Ease of validation	☐	Ordinal	Automatic/Manual/No	
X	Lossiness	☐	Ordinal	Lossy/Lossless	
X	IPR	☐	Boolean	Yes/No	
X	Complexity	☐	Ordinal	High/Medium/Low	

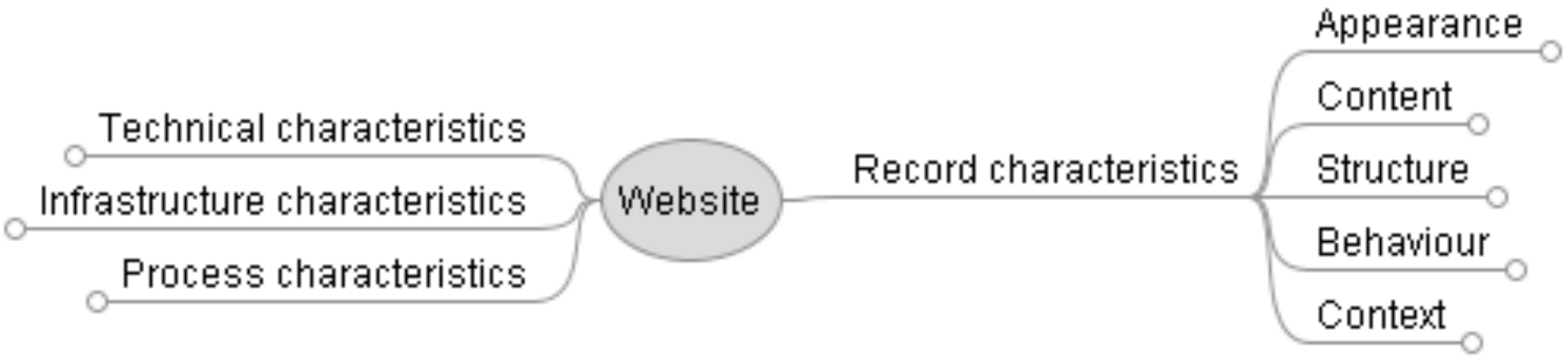
Release 1.1 - Institute of Software Technology and Interactive Systems: «off-ice bears»

Quick Access:

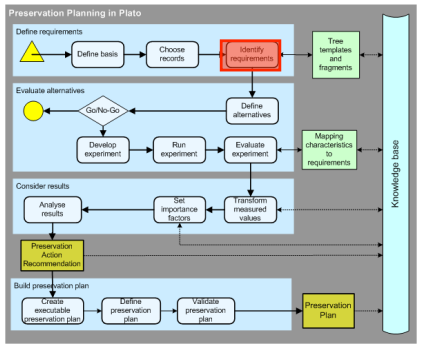
Identify Requirements: Example

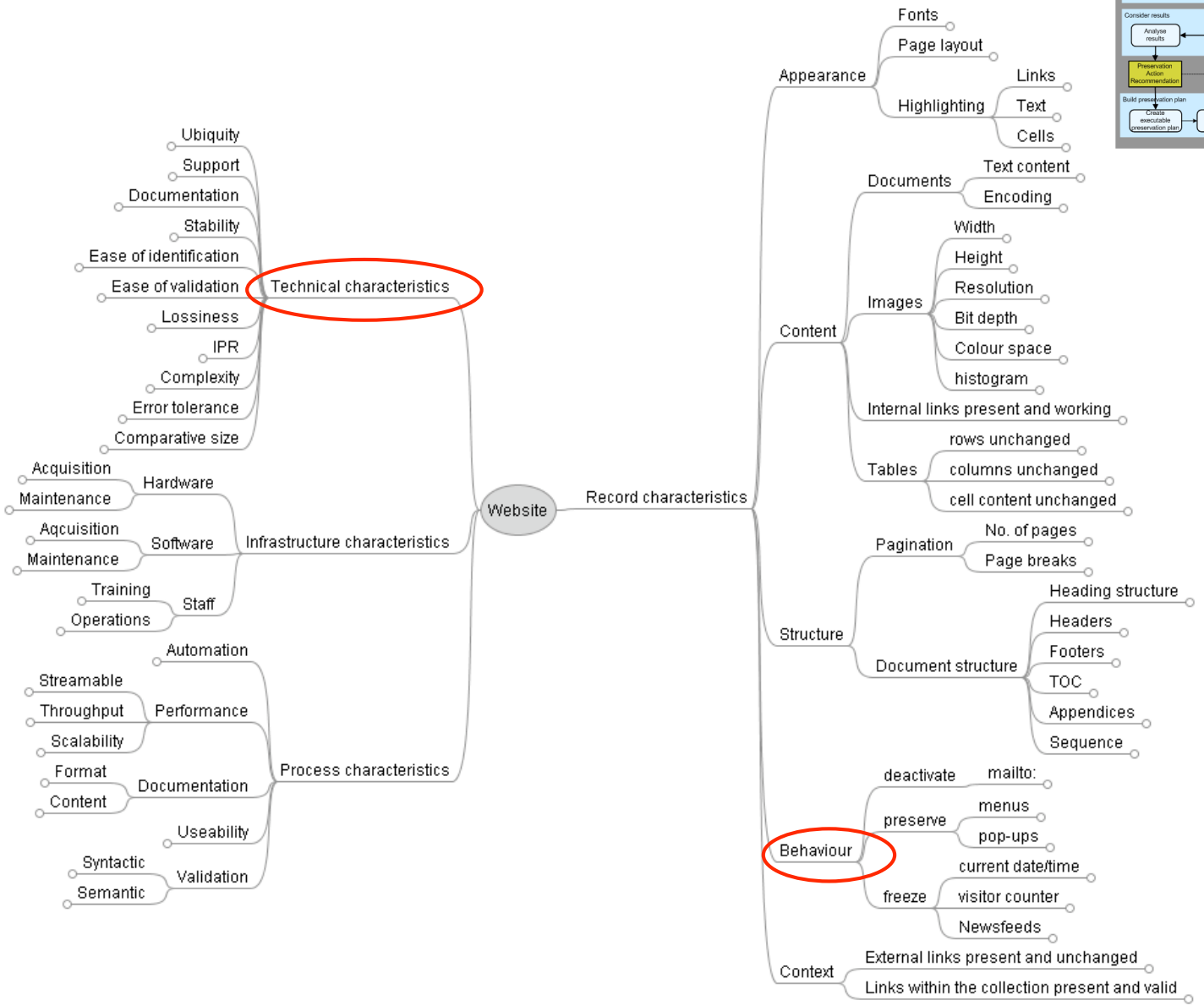


- Example Webarchiving:
  - Static Webpages
  - Including linked documents such as doc, pdf
  - Images
  - Interactive elements need not be preserved



Identify Requirements: Example



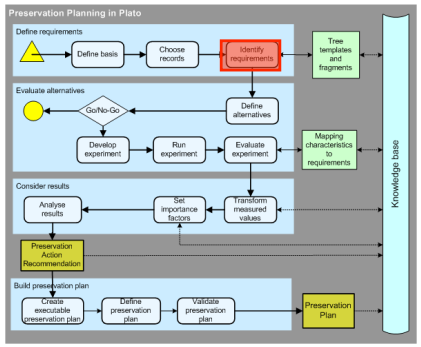


TU WIEN planets

UNIVERSITY OF Southampton

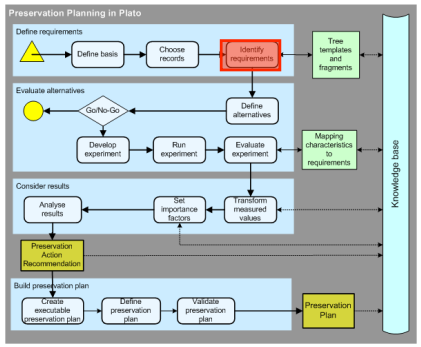
eprints

Identify Requirements: Example

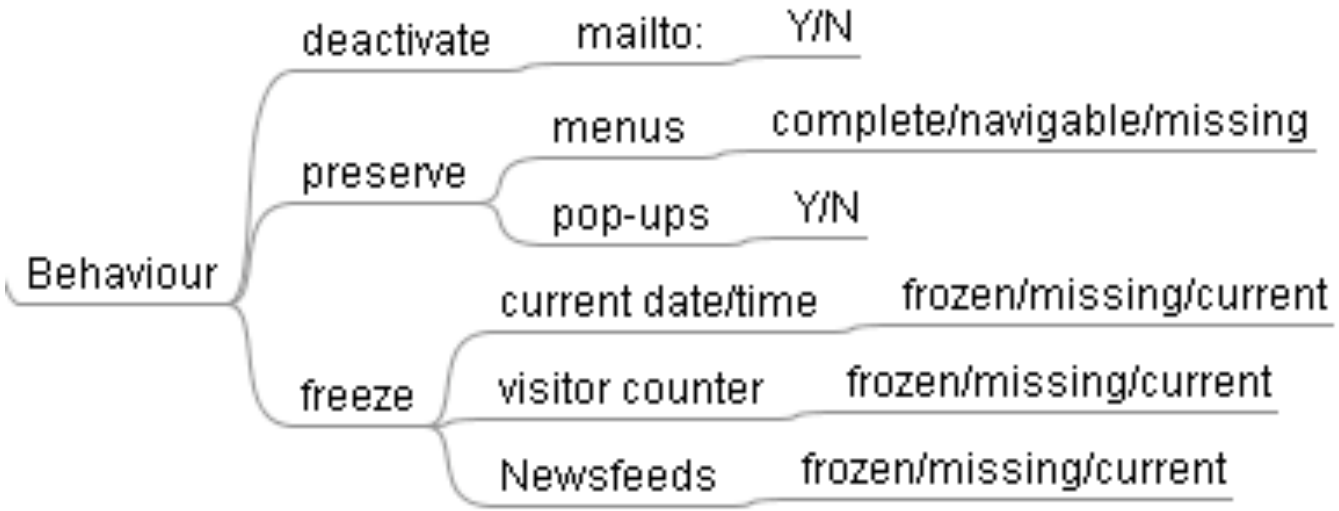


- Technical characteristics
  - Ubiquity: Ubiquitous/Widespread/Specialised/Obsolete
  - Support: number of tools
  - Documentation
    - Quality: Primary/Secondary
    - Disclosure: Full/Partial/None
    - Openness: Standard/Open/Proprietary
    - Availability: Public/Limited/None
  - Stability
    - Speed of change: <1 year/1-2 years/3-5 years/>5 years
    - Backwards compatability: None/Previous version only/Some previous versions/All previous versions
  - Ease of identification: Automatic/Manual/No
  - Ease of validation: Automatic/Manual/No
  - Lossiness: Lossy/Lossless
  - IPR: Y/N
  - Complexity: High/Medium/Low
  - Error tolerance: None/Detectable/Recoverable
  - Comparative size: Larger/Same/Smaller

Identify Requirements: Example



Behaviour

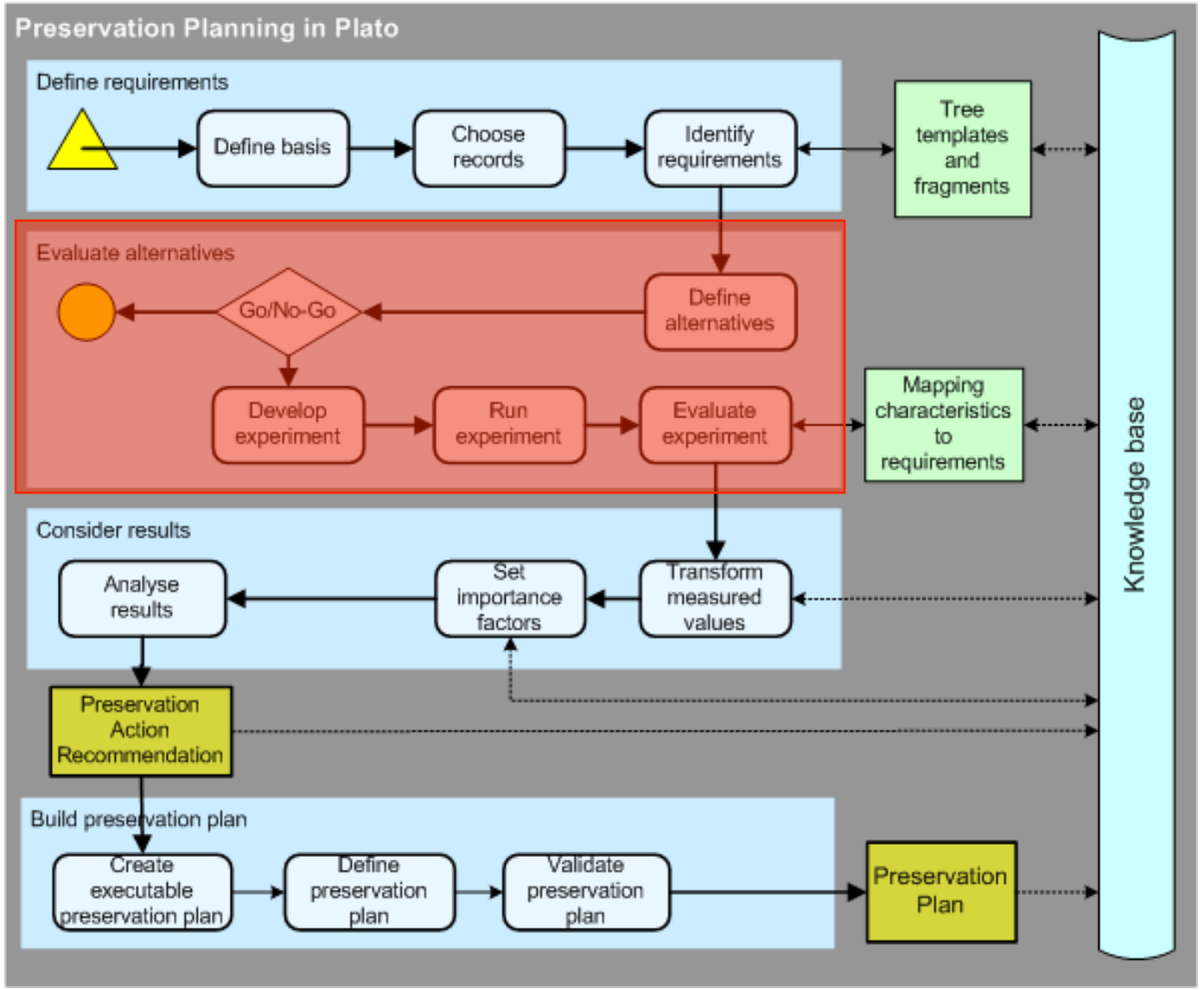


- Visitor counter and similar functionalities can be
  - Frozen at harvesting time
  - Omitted
  - Remain operational, i.e. the counter will be increased upon archival calls
    (is this desired? count? demonstrate functionality?)

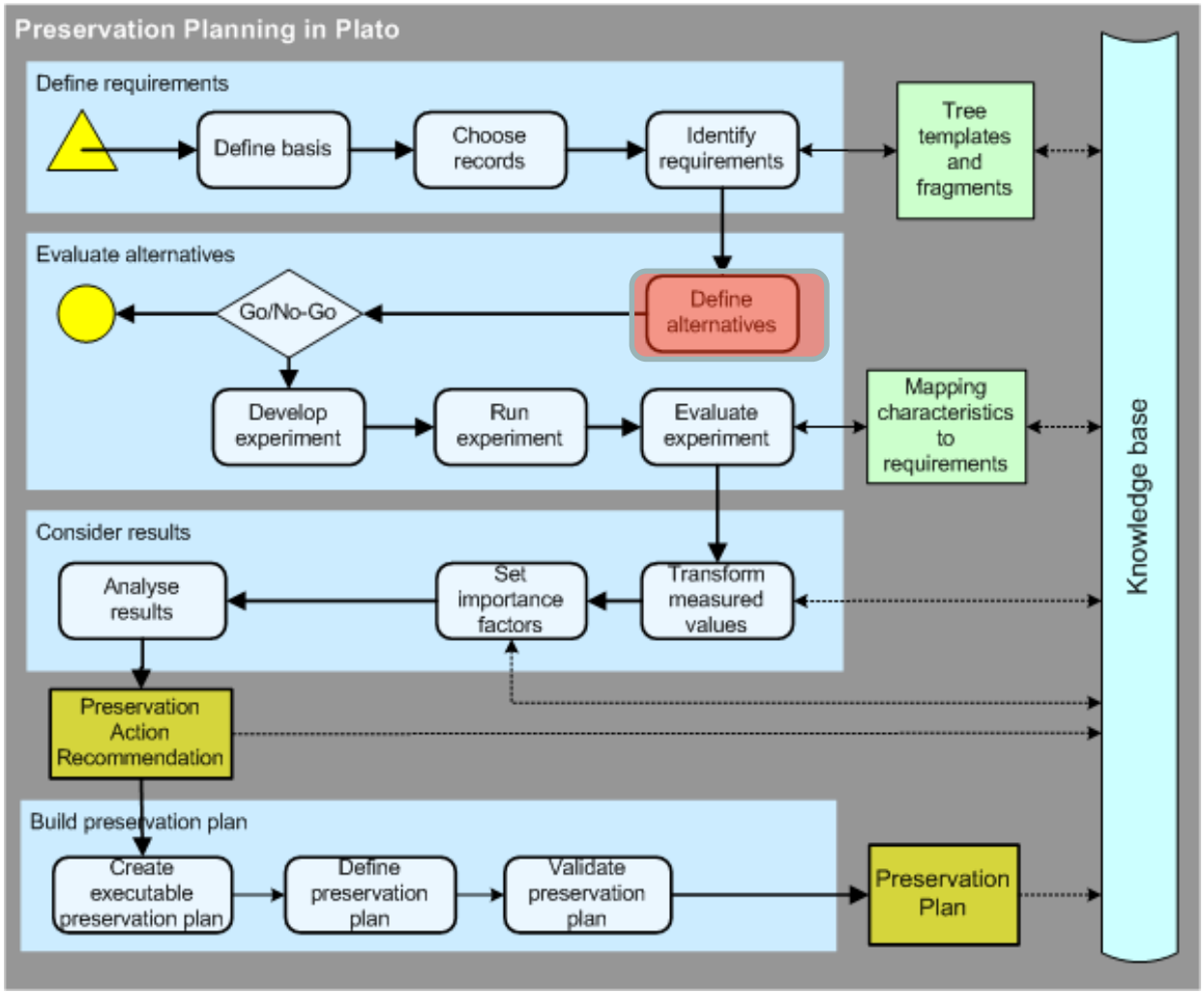
TU WIEN planets UNIVERSITY OF Southampton eprints

Practise time!

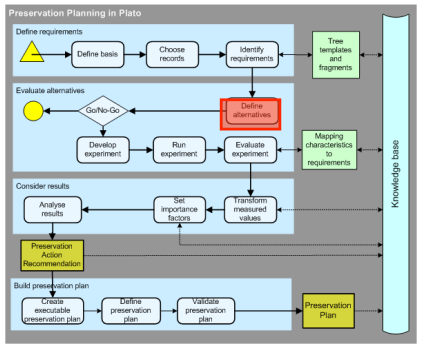
PP Workflow


Preservation Planning in Plato
Define requirements
Define basis
Choose records
Identify requirements
Tree templates and fragments
Evaluate alternatives
Go/No-Go
Define alternatives
Develop experiment
Run experiment
Evaluate experiment
Mapping characteristics to requirements
Consider 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
Knowledge base

Orientation

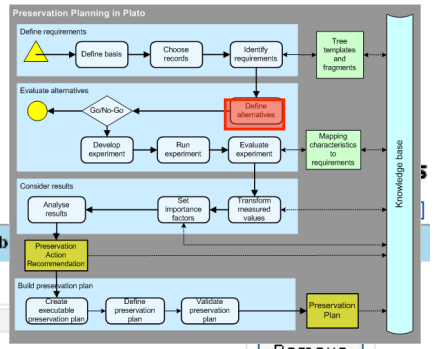


Define Alternatives



- 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)
- Define manually or use registries via web services

Define Alternatives



PLANETS Preservation Planning Tool *(Plato)*

Project | Define Requirements ::: Evaluate Requirements ::: Consider Results | Polar b

Define the alternatives of the Project

ID	Name	Description	
196616	TIFF (tool A)	Convert to TIFF using the well-tested and expensive tool 'A'	Remove
196613	TIFF (tool B)	Convert to TIFF/4 using this new tool named 'B'	Remove
196614	GIF (tool C)	Convert to GIF using the well-tested tool 'C'	Remove
196615	PNG (tool D)	Convert to PNG using the well-tested tool 'D'	Remove

Add new Alternative

Save | Discard changes | Proceed

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

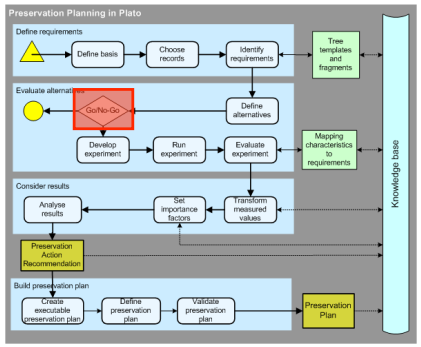
CRiB Service Registry

Show Preservation Services

	Preservation Action	Target Format	Info
☐	JPG > BMP	Windows Bitmap, version 3.0	JPG>BMP
☑	JPG > TIF	Tagged Image File Format, version 3	JPG>BMP>TIF
☐	JPG > TIF #2	Tagged Image File Format, version 3	JPG>TIF
☑	JPG > TIF_2	Tagged Image File Format, version 3	JPG>TIF_2
☐	JPG > PNG	Portable Network Graphics, version 1.0	JPG>PNG
☐	JPG > JP2	JPEG 2000	JPG>JP2

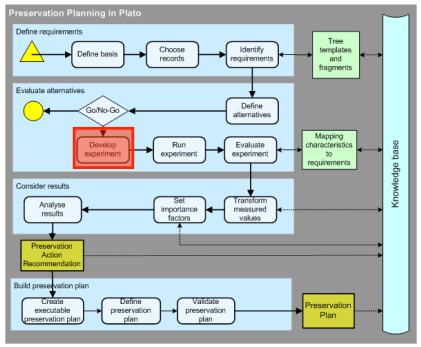
Create alternatives for selected services

Go/No-Go



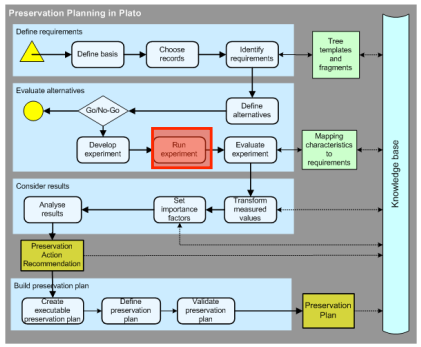
- Deliberate step for taking a decision if it will be useful and cost-effective to continue the procedure, given
  - The resources to be spent (people, money)
  - The availability of tools and solutions,
  - The expected result(s).
- Review of the experiment/ evaluation process design so far
  - Is the 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...)
- Standardized Testbed-environment simplifies this step (PLANETS Testbed)
- Ideally directly accessible Preservation Action Services
- Ensures that results are comparable and repeatable

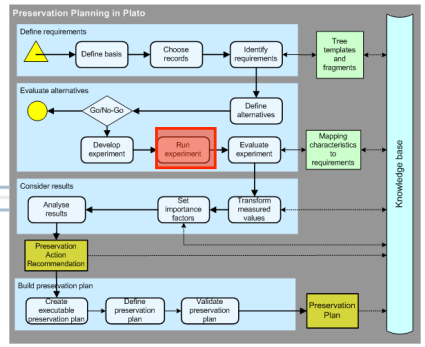
Run experiment



- Before running experiments: Test
- Call migration / emulation tools
- Local or service-based
- Capture process measurements
  (Start-up time, time per object, throughput, ...)
- Capture resulting objects, system logs, error messages,…

TU WIEN planets UNIVERSITY OF Southampton eprints

Develop and Run Experiment



JPG > JP2

Save | Discard changes | Proceed

Result Files

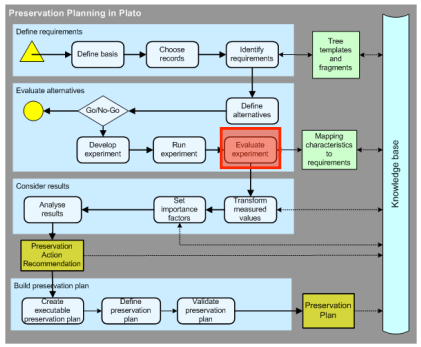
Results	Sample Records		
TIFF (tool A) ?	Polar bear 1	Upload File	
	Polar bear 2	Upload File	
TIFF (tool B) ?	Polar bear 1	Upload File	
	Polar bear 2	Upload File	
GIF (tool C) ?	Polar bear 1	Upload File	
	Polar bear 2	Upload File	
PNG (tool D) ?	Polar bear 1	Upload File	
	Polar bear 2	Upload File	
JPG > BMP ?	Polar bear 1	Upload File	Polar bear 1.bmp Download File
	Polar bear 2	Upload File	Polar bear 2.bmp Download File
JPG > TIF ?	Polar bear 1	Upload File	
	Polar bear 2	Upload File	

Release 1.2 - Institute of Software Technology and Interactive Systems: « off-ice bears »

TU WIEN planets UNIVERSITY OF Southampton eprints

Demo!

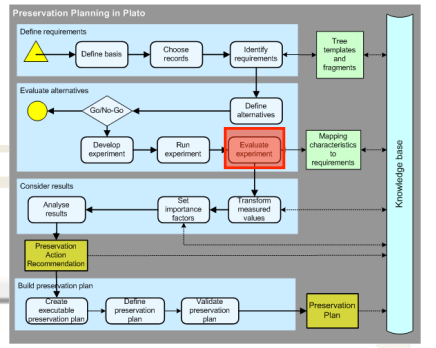
Evaluate experiment



- Analyse the results according to the criteria specified in the Objective Tree
- Preservation Characterization: *Characterization Services*
- Evaluation analyses
  - Experiment measurements, results
  - Necessity to repeat an experiment
  - Undesired / unexpected results
- Technical and intellectual aspects

TU WIEN planets UNIVERSITY OF Southampton eprints

Evaluate Experiment



Designated community:

Durchsuchen...

Applying policies:

Durchsuchen...

Relevant organisational procedures

Release 1.2 - Institute of Software Technology and Interactive Systems: « off-ice bears »

Figure 8 Define

In the first step of the workflow (Figure 8), the number environment are described in the text boxes. The proje and organisation) can be altered on this page. Throug

Evaluate Experiment

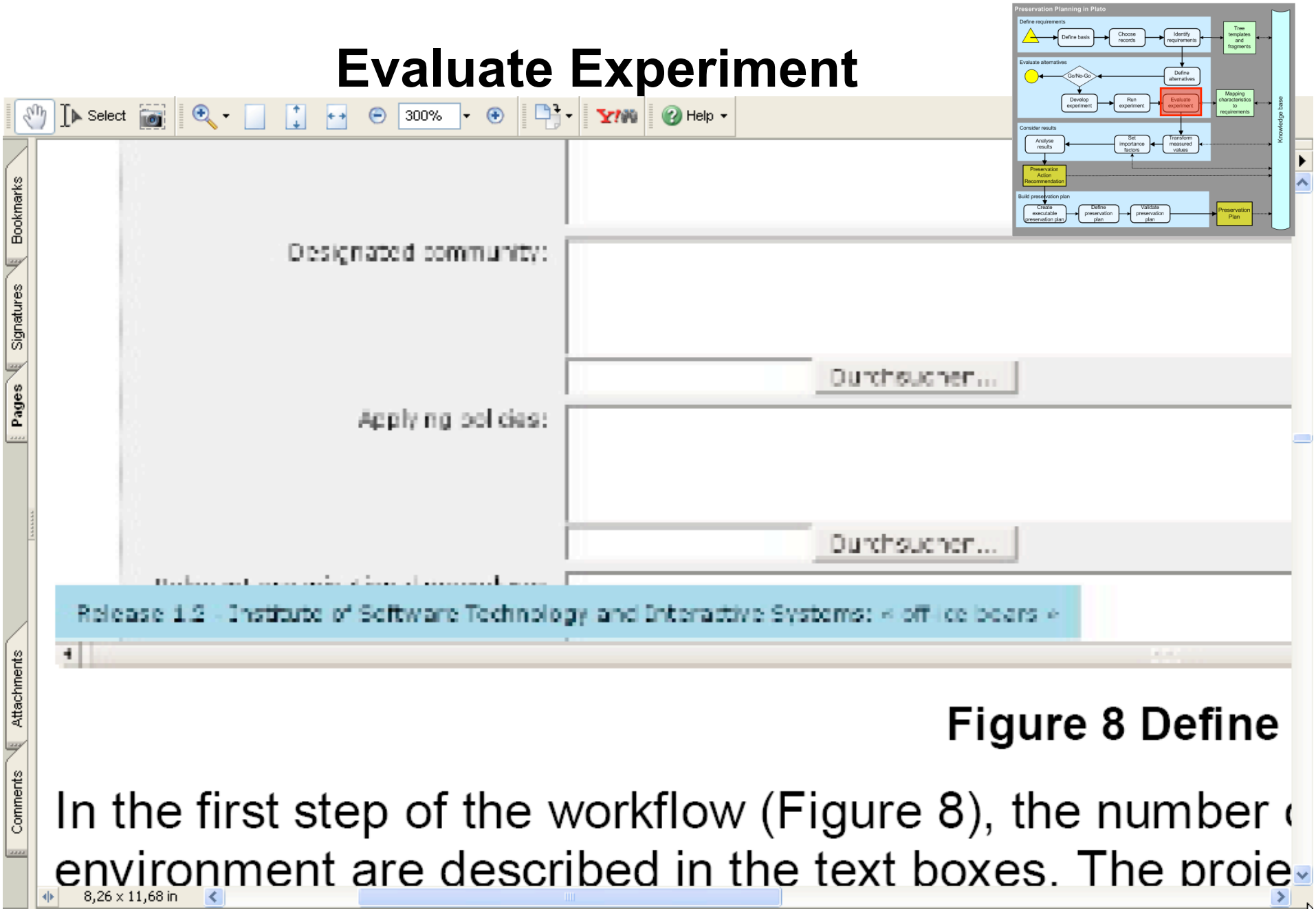
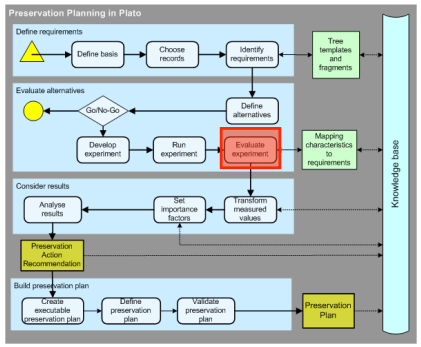


Figure 8 Define

In the first step of the workflow (Figure 8), the number
environment are described in the text boxes. The proje

Evaluate Experiment



PLANETS Preservation Planning Tool *(Plato)*

Project | | Define Requirements ::: Evaluate Requirements ::: Consider Results | |

Evaluate Experiment

Expand All | Collapse All

Polar bear image preservation

Focus	Node
	▼Polar bear image preservation
X	▼Process
X	▼Complexity
X	▼Cost
X	▼Image properties
X	▼Bits of colour depth
X	▼Technical characteristics
X	▼Official standard
X	▼Filesize (in Relation to Original)

Comments:

Process > Complexity

Alternative	Single result		Comments
TIFF (tool A)	Simple	*	
TIFF (tool B)	Simple	*	
GIF (tool C)	Complex	*	
PNG (tool D)	Medium	*	

Process > Cost

Alternative			Comments
TIFF (tool A)	173	*	
	152	*	
TIFF (tool B)	100	*	
	88	*	
GIF (tool C)	140	*	
	128	*	
PNG (tool D)	79	*	
	80	*	

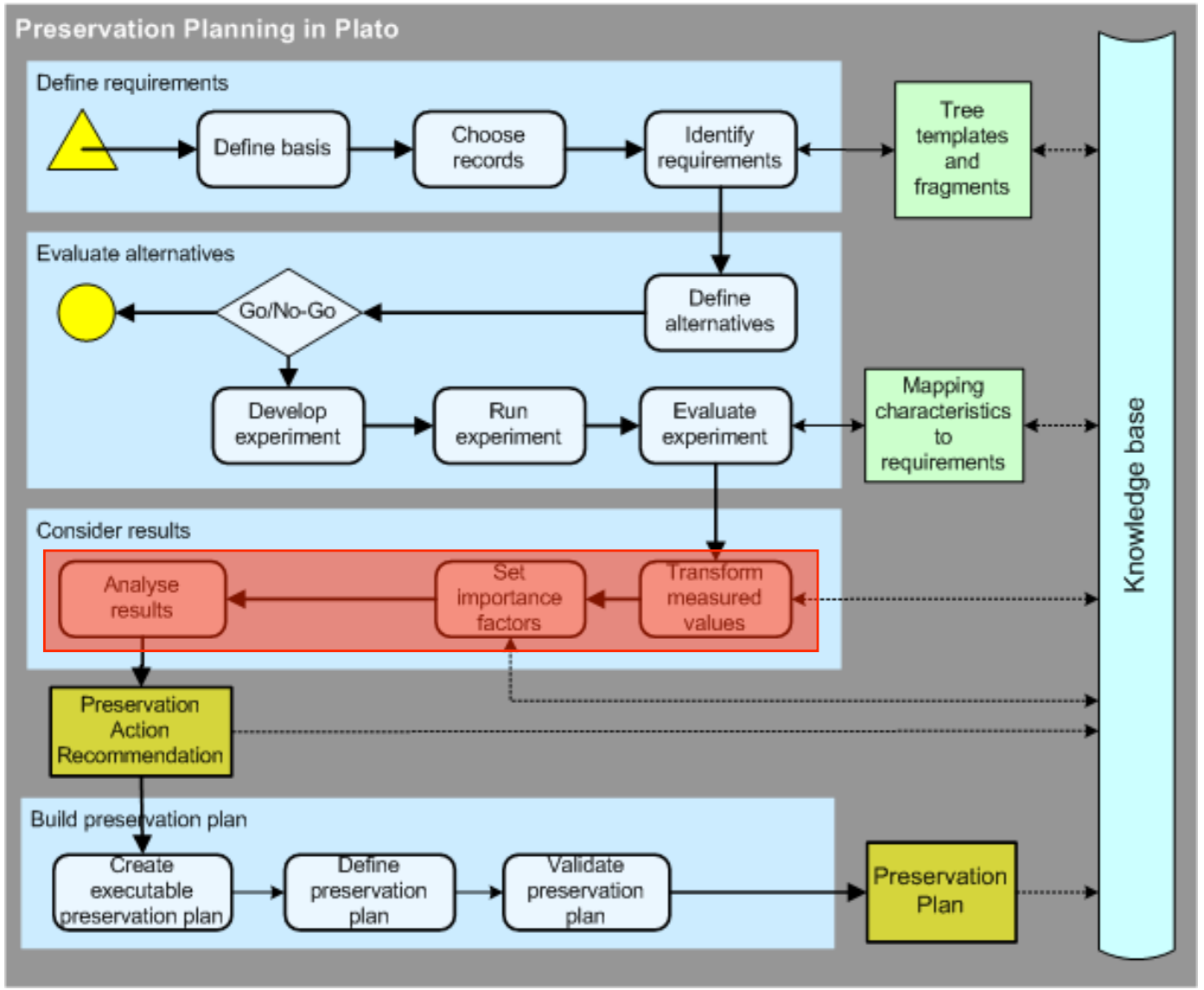
Image properties > Bits of colour depth

Alternative	Single result		Comments
TIFF (tool A)	32	bit *	
TIFF (tool B)	32	bit *	
GIF (tool C)	32	bit *	
PNG (tool D)	32	bit *	

TU WIEN planets UNIVERSITY OF Southampton eprints

Practise time!

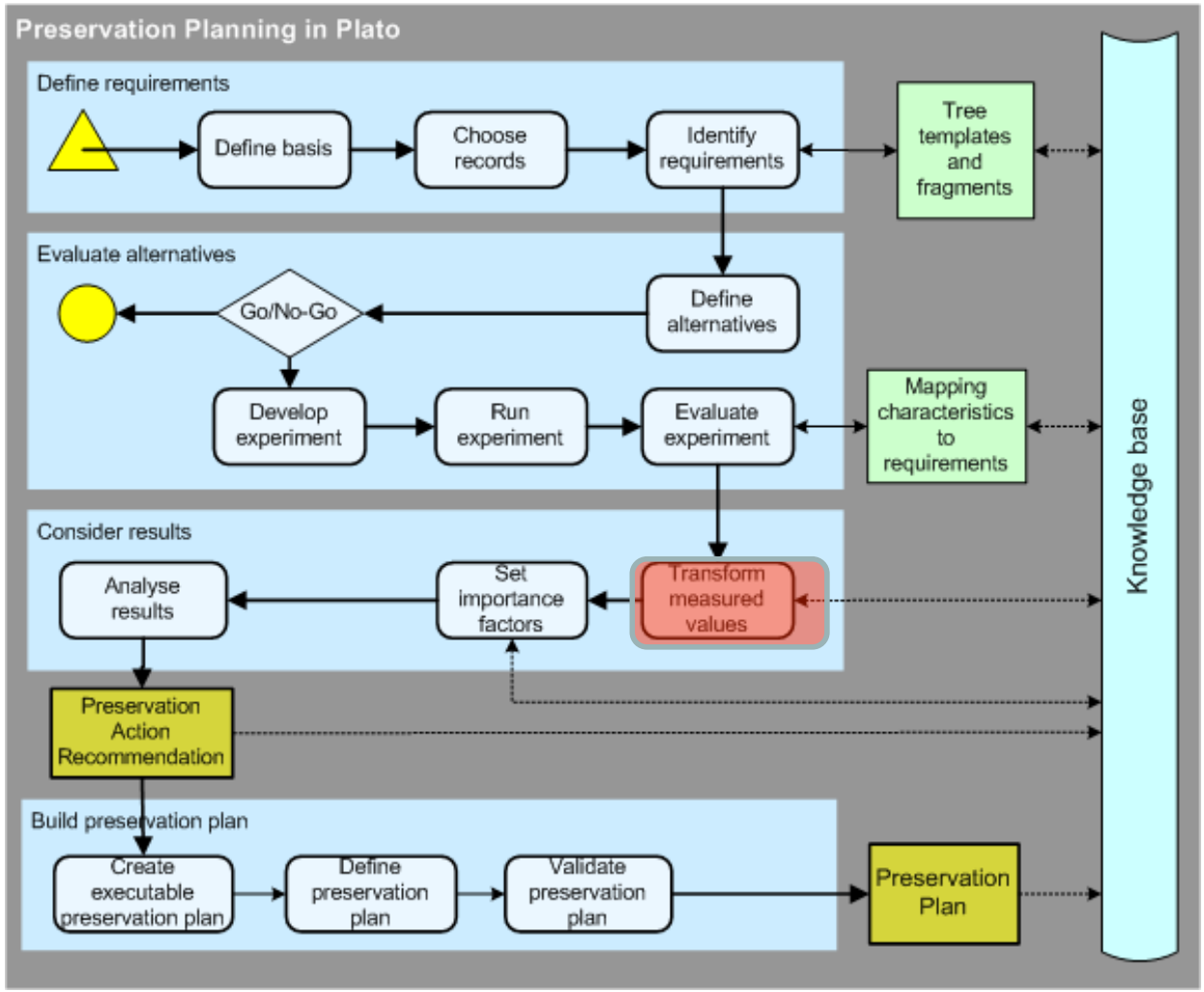
PP Workflow



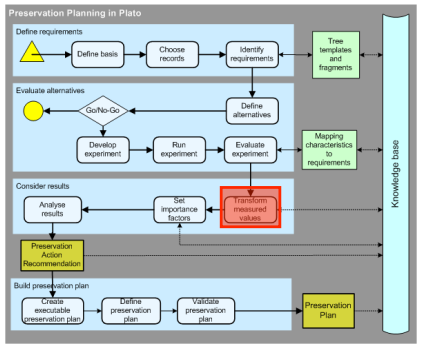
TU WIEN planets

UNIVERSITY OF Southampton eprints

Orientation



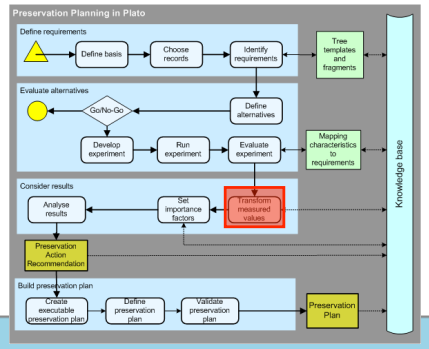
Transform measured values



- 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"

TU WIEN planets UNIVERSITY OF Southampton eprints

Transform Measured Values



PLANETS Preservation Planning Tool *(Plato)*

Project | | **Define Requirements** ::: **Evaluate Requirements** ::: **Consider Results** | |

Transform Measured Values

Expand All | Collapse All

Polar bear image preservation

Focus	Node
	▼Polar bear image preservation
X	▼Process
X	▼Complexity
X	▼Cost
X	▼Image properties
X	▼Bits of colour depth
X	▼Technical characteristics
X	▼Official standard
X	▼Filesize (in Relation to Original)

Comments: The limit for the process cost was decided to be 150¢ per picture in the last polar-bear-enthusiasts-association-

Technical characteristics > Official standard

Ordinal Value		Target Value
Yes	->	5.0 *
No	->	0.0 *

Results	Single
TIFF (tool A)	Yes
TIFF (tool B)	Yes
GIF (tool C)	No
PNG (tool D)	Yes

Aggregation mode:

(•) Worst result ( ) Arithmetic mean

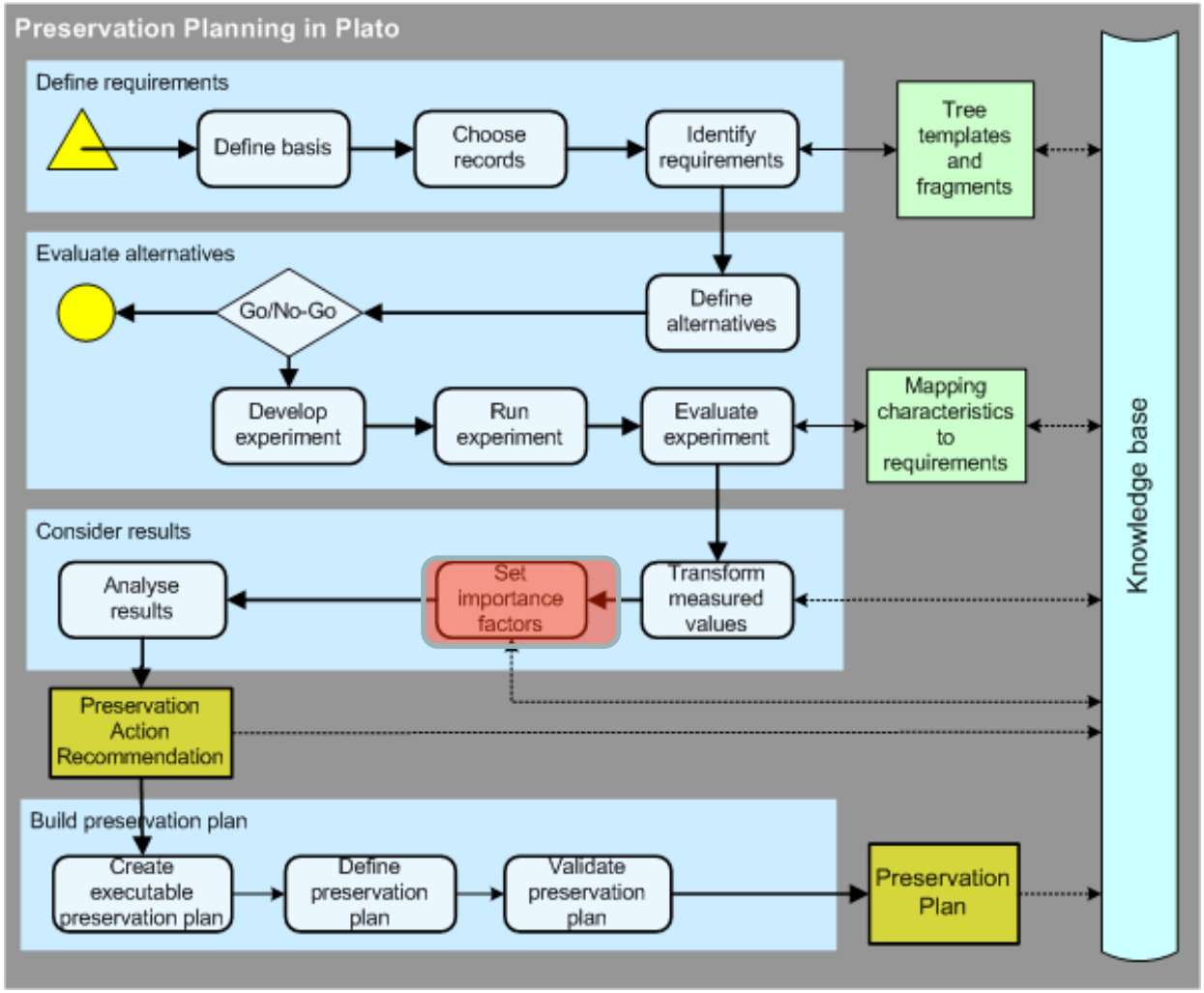
Technical characteristics > Filesize (in Relation to Original)

Threshold	Target value
10.0 *	x -> 1
5.0 *	x -> 2
2.0 *	x -> 3
1.2 *	x -> 4
0.8 *	x -> 5

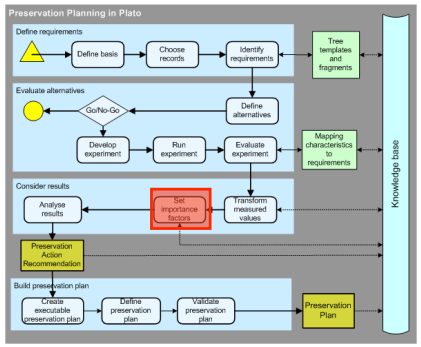
Results	1	2
TIFF (tool A)	2.73	2.6
TIFF (tool B)	2.3	3.1
GIF (tool C)	1.1	0.9
PNG (tool D)	1.3	1.5

TU WIEN planets UNIVERSITY OF Southampton eprints

Orientation

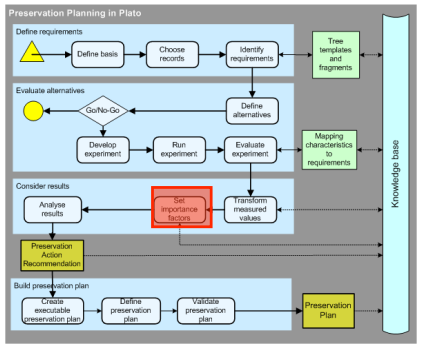


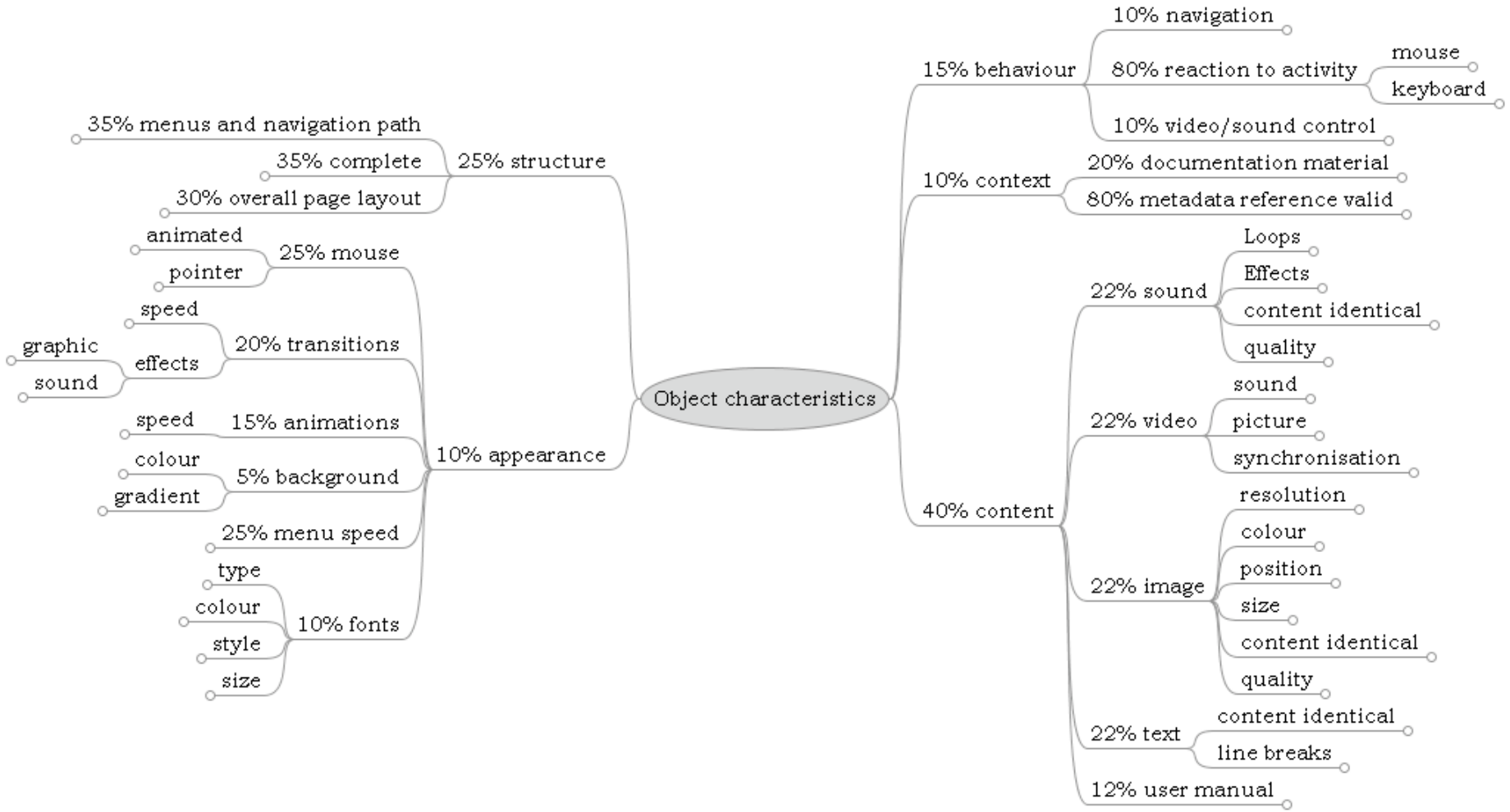
Set Importance Factors



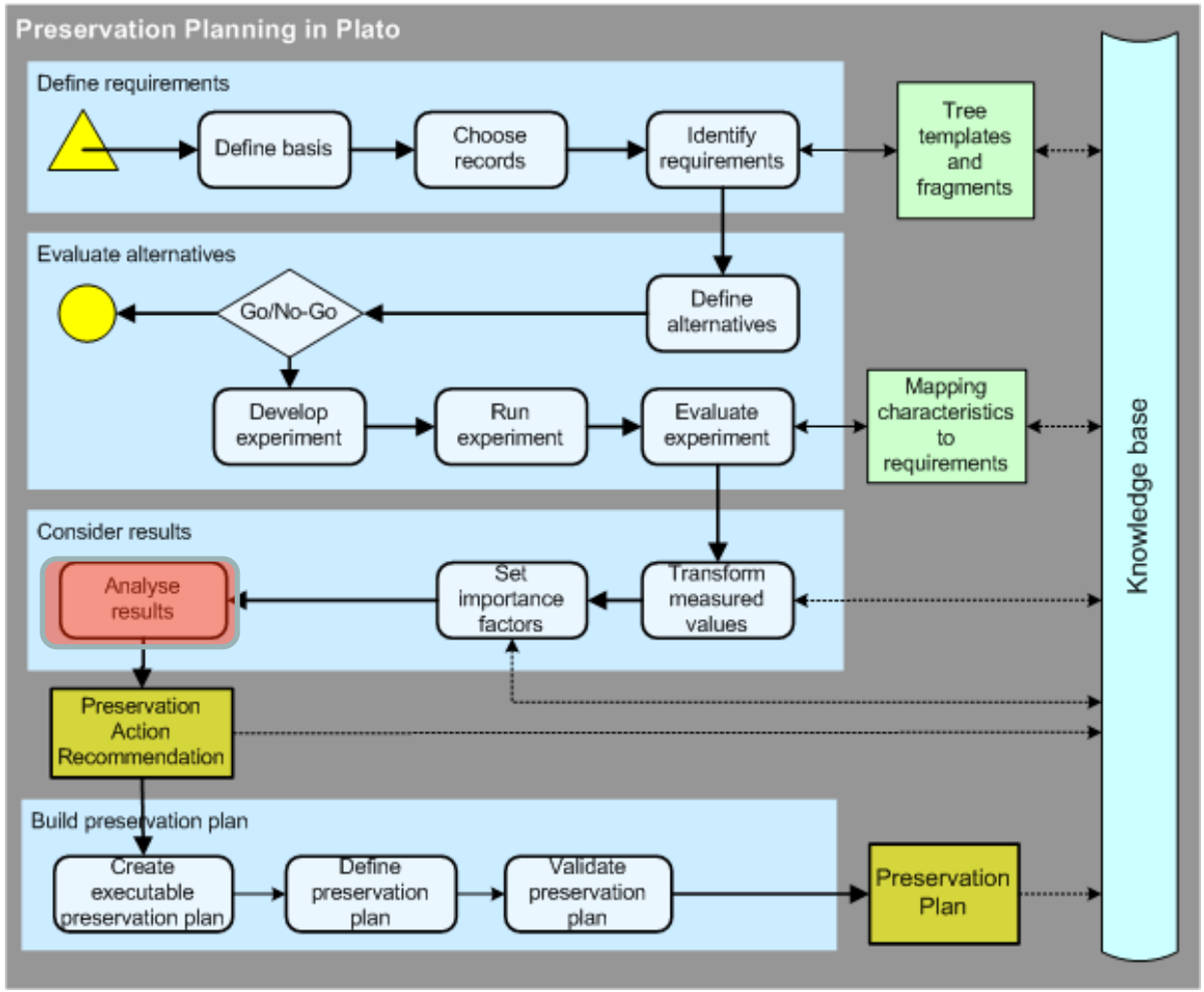
- Not all leaf criteria are equally important
- By default, weights are distributed equally
- Adjust relative importance of all siblings in a branch
- Weights are propagated down the tree to the leaves

Set Importance Factors

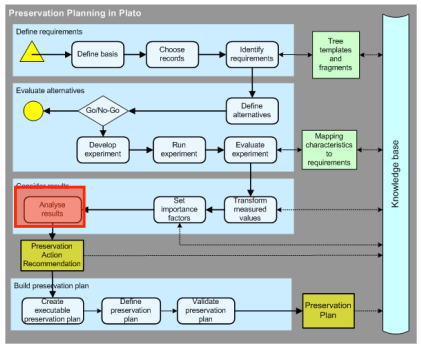




Orientation



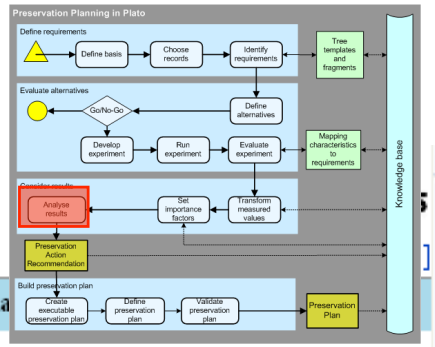
Analyse results



- Aggregate values in Objective Tree
  - Multiply transformed measurements in leaves with weights
  - Sum up across tree
- Results in accumulated performance value per alternative at root level
  → ranking of alternatives
- Also results in performance value for each alternative in each sub-branch of the tree
  → combination of alternatives
- Basis for well-informed and accountable decisions
- Different aggregation methods, e.g. sum and multiplication

TU WIEN planets UNIVERSITY OF Southampton eprints

Analyse Results



PLANETS Preservation Planning Tool *(Plato)*

Project | Define Requirements ::: Evaluate Requirements ::: Consider Results | Minimalist test project in sta

Analyse Results

Aggregation method: Sum

Select	Alternative
☑	PDF/A ToolA
☑	PDF/A ToolB

Show

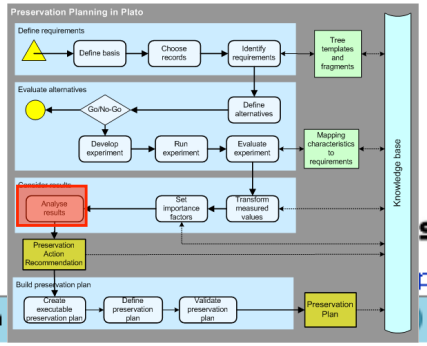
Expand All | Collapse All

Minimalist root node

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	▶IntRange 0-10	PDF/A ToolA: 0,70 PDF/A ToolB: 0,60

TU WIEN planets UNIVERSITY OF Southampton eprints

Analyse Results



PLANETS Preservation Planning Tool *(Plato)*

Project | Define Requirements ::: Evaluate Requirements ::: Consider Results | Minimalist test project in

Analyse Results

How do the aggregation mechanisms work?

Aggregation method: Multiplication

Select	Alternative
☑	PDF/A ToolA
☑	PDF/A ToolB

Show

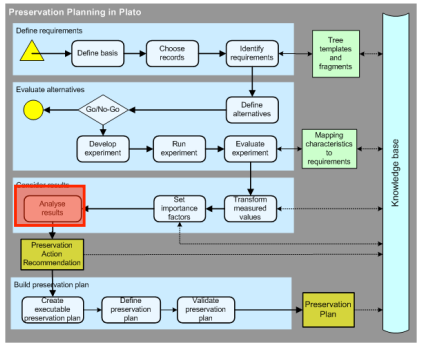
Expand All | Collapse All

Minimalist root node

Focus	Name	Result
	▼Minimalist root node	PDF/A ToolA:2,86 PDF/A ToolB:0,00
X	▼Image properties	PDF/A ToolA:1,28 PDF/A ToolB:1,32
X	▼Amount of Pixel	PDF/A ToolA:3,50 PDF/A ToolB:4,00
X	▼Karma	PDF/A ToolA:1,15 PDF/A ToolB:0,00
X	▼Filesize (in Relation to Original)	PDF/A ToolA:1,31 PDF/A ToolB:1,38
X	▼A Single-Leaf	PDF/A ToolA:1,15 PDF/A ToolB:1,32
X	▼IntRange 0-10	PDF/A ToolA:1,28 PDF/A ToolB:1,25

TU WIEN planets UNIVERSITY OF Southampton eprints

Analyse results



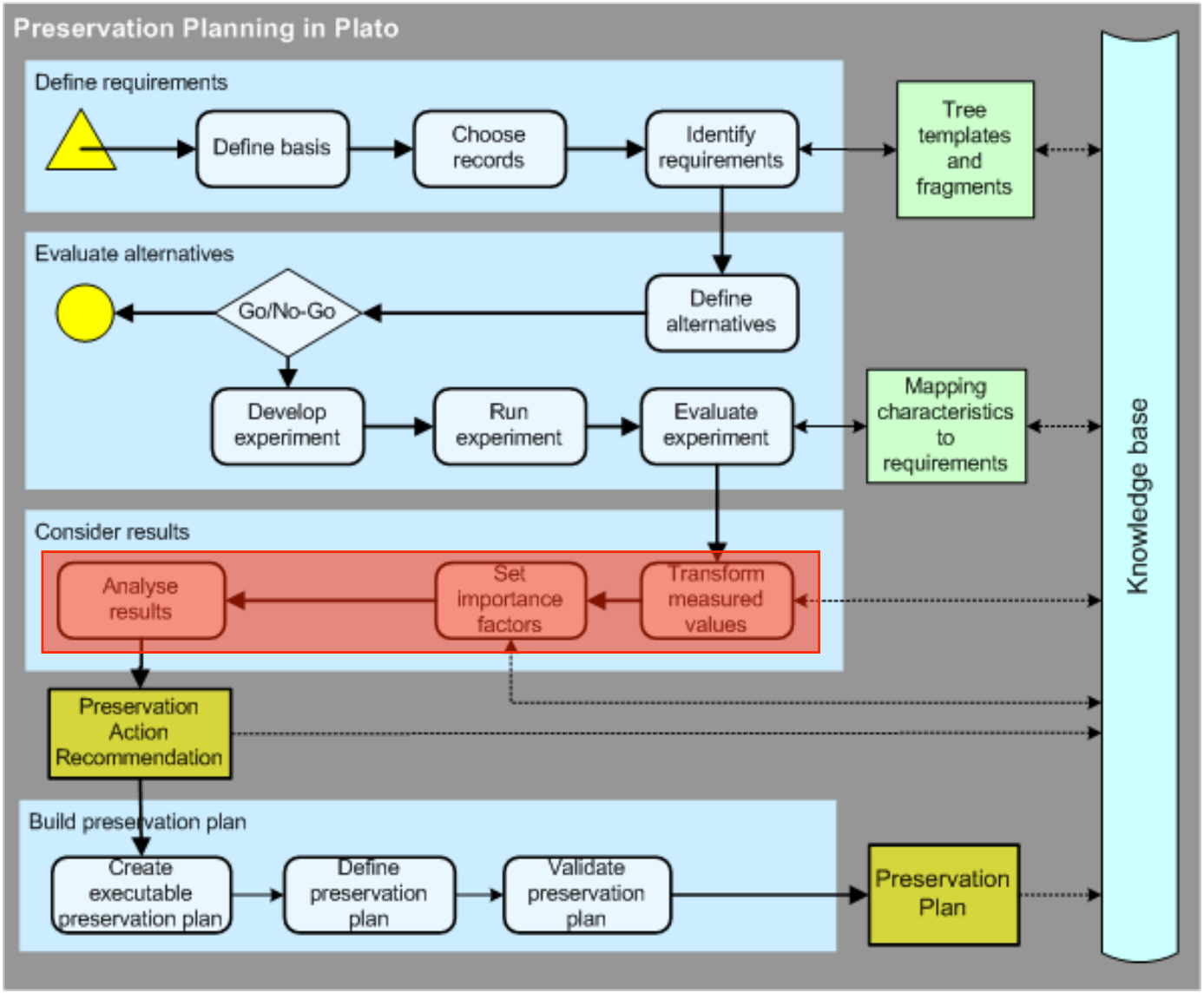
Example: Electronic documents

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

TU WIEN planets UNIVERSITY OF Southampton eprints

PP Workflow


Preservation Planning in Plato
Define requirements
Define basis
Choose records
Identify requirements
Tree templates and fragments
Evaluate alternatives
Go/No-Go
Define alternatives
Develop experiment
Run experiment
Evaluate experiment
Mapping characteristics to requirements
Consider 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
Knowledge base

Schedule

(1) Introduction
- What is Digital Preservation?
- EPrints
- Preservation Planning and Plato

(2) Preservation in EPrints

(3) Preservation Planning with Plato

(4) Bringing it all together and Closing

TU WIEN planets UNIVERSITY OF Southampton eprints

Questions

Overview

Part 4: Bringing it all together

- Why are we doing this: Trust and Authenticity
- Recap and closure

Compliance

- Trustworthy repositories
- Compliance to best practices, standards
- 3 core initiatives, of which 2 prescriptive
  - RLG- National Archives and Records Administration Digital Repository Certification Task Force:
    Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)
  - NESTOR:
    Catalogue of Criteria of Trusted Digital Repositories
  - DCC/DPE:
    DRAMBORA: Digital Repository Audit Method Based on Risk Assessment
- Embedding into OAIS model

Compliance

TRAC:

- Three sections
  - A. Organisational Infrastructure
  - B. Digital Object Management
  - C. Technologies, Technical Infrastructure & Security

Compliance

TRAC and Preservation Planning 1:

- **A3.2** 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
  - Watch Services, triggers
  - Verification against changes in the environment
  - Update of preservation plans

- **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
  - History of preservation plans (created, reviewed and updated)
  - Plato: Automated documentation of planning activities

Compliance

TRAC and Preservation Planning 2:

- **A3.7** Repository commits to transparency and accountability in all actions supporting the operation and management of the repository, especially those that affect the preservation of digital content over time
  - Solid workflow in consist manner enables informed and well-documented decisions
  - Explicit definition of objectives and measurement units
- **B1.1** Repository identifies properties it will preserve for digital objects
  - Objective Tree

Compliance

TRAC and Preservation Planning 3:

- **B3.1** Repository has documented preservation strategies
  - Preservation Plan
- **B3.3** Repository has mechanisms to change its preservation plans as a result of its monitoring activities.
  - Watch Services, triggers
  - Verification against changes in the environment
  - Update of preservation plans

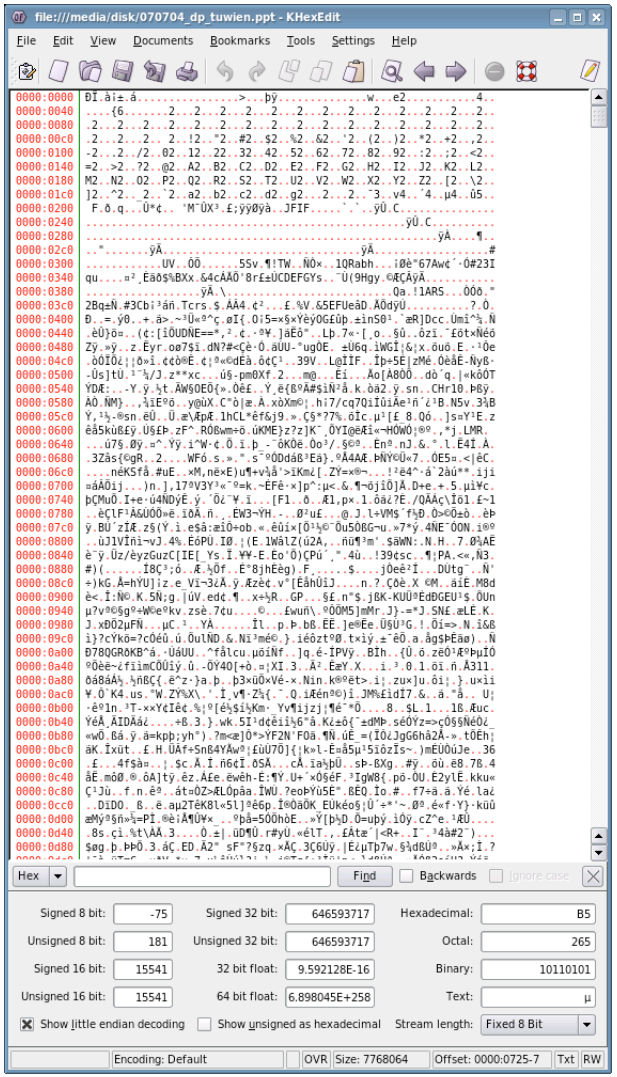
Overview

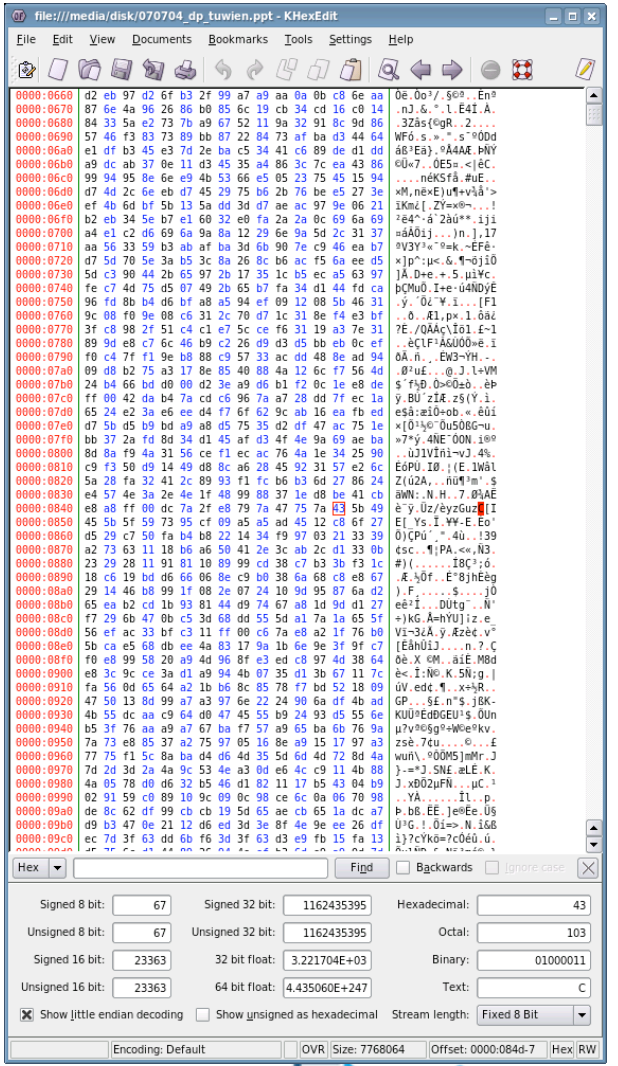
Part 4: Bringing it all together

- Questions and answers
- Why are we doing this: Trust and Authenticity
- Recap and closure

Why do we need Digital Preservation?

Why do we need Digital Preservation?





TU WIEN planets

UNIVERSITY OF Southampton eprints

Why do we need Digital Preservation?

- Digital Objects require specific environment to be accessible :
  - Files need specific programs
  - Programs need specific operating systems (-versions)
  - Operating systems need specific hardware components
- SW/HW environment is not stable:
  - Files cannot be opened anymore
  - Embedded objects are no longer accessible/linked
  - Programs won‘t run
  - Information in digital form is lost
    (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.

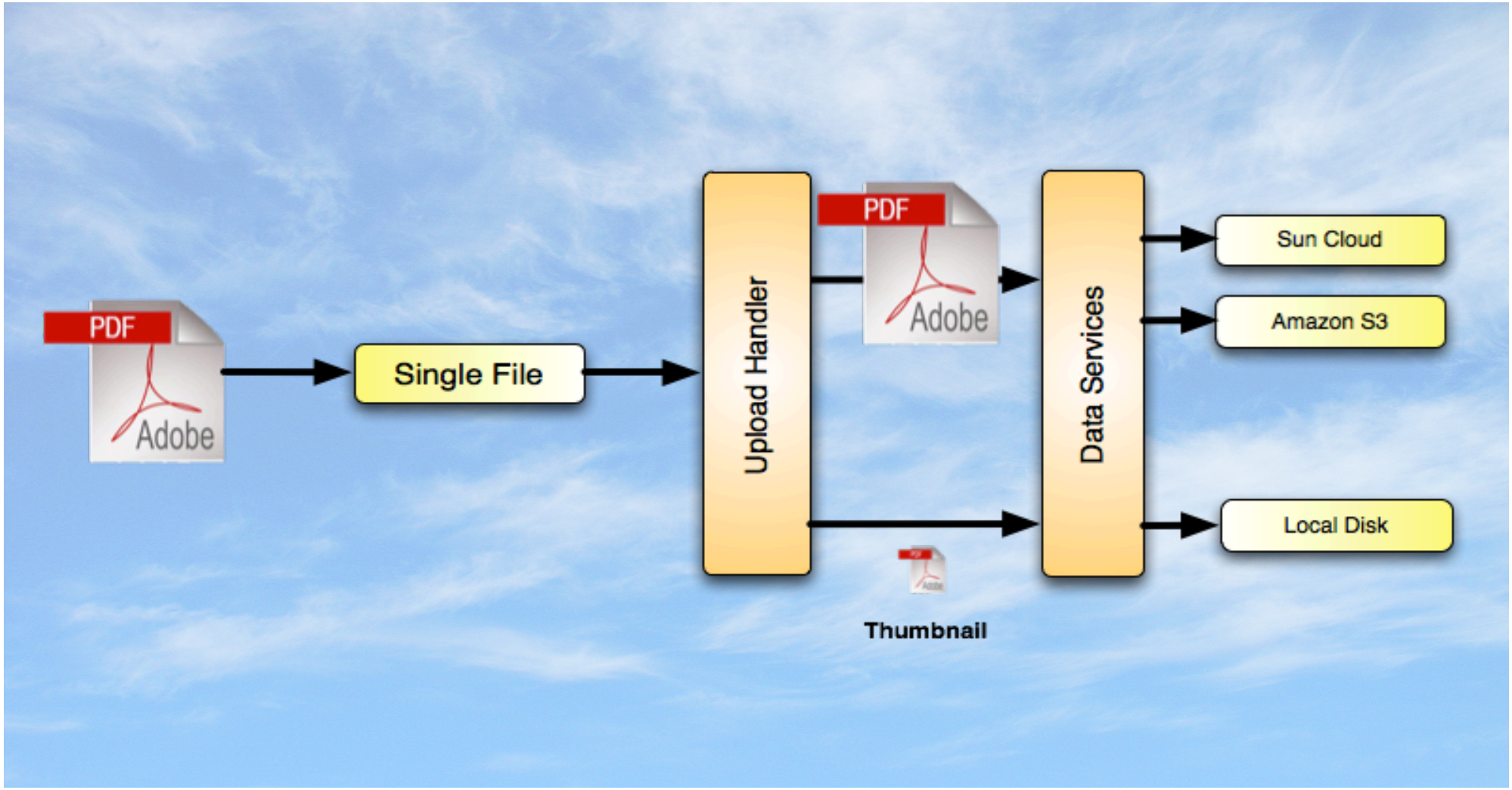
TU WIEN planets UNIVERSITY OF Southampton eprints

Why do we need Digital Preservation?

- Essential for all digital objects
  - Office documents, accounting, emails, …
  - Scientific datasets, sensor data, metadata, …
  - Applications, simulations,…

- All application domains
  - Cultural heritage data
  - eGovernment, public administration
  - Science / Research
  - Industry
  - Health, pharmaceutical industry
  - Aviation, control systems, construction, …
  - Private data
  - …

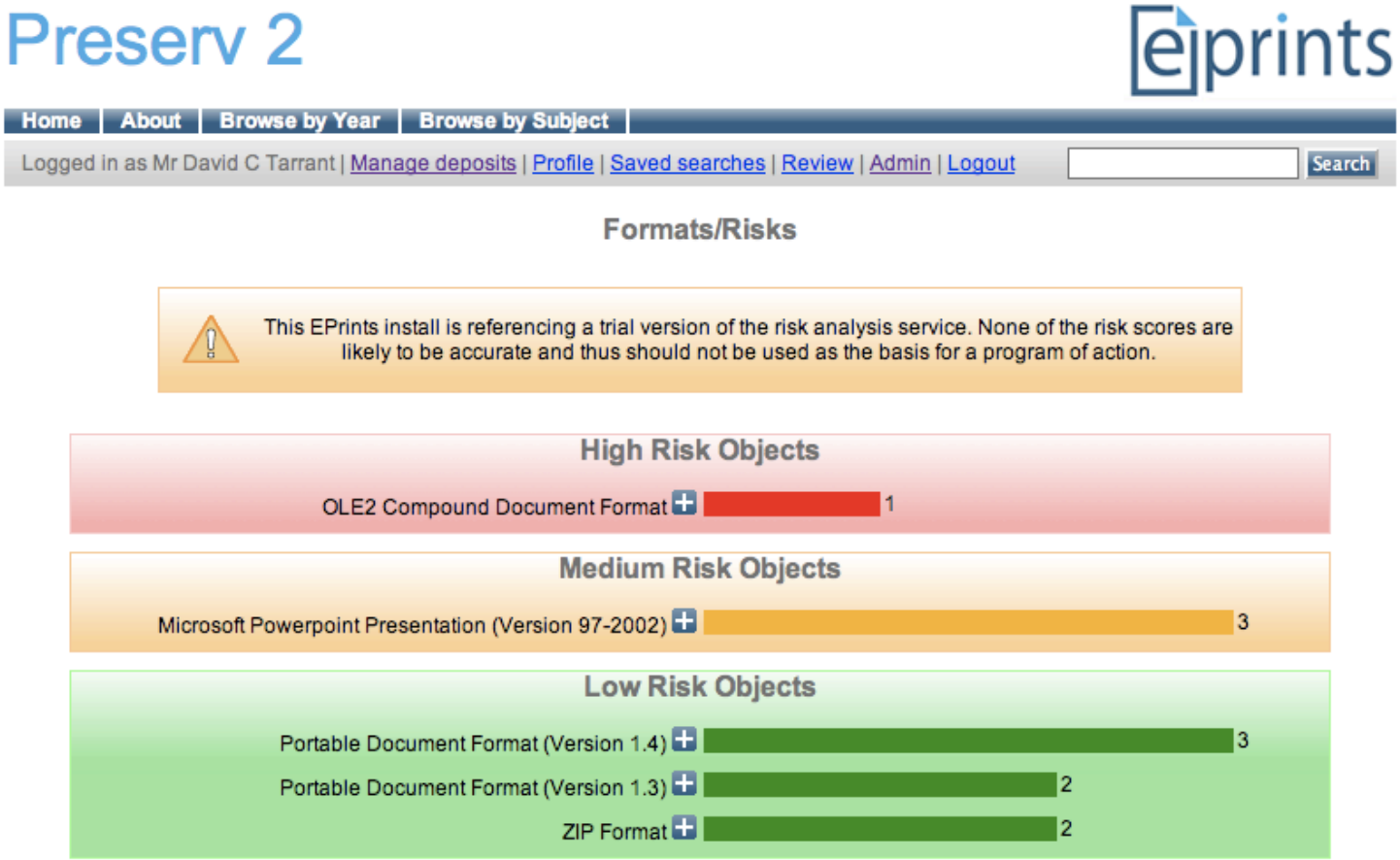
Hybrid Storage



Risk Analysis In EPrints

Preservation - Analyse

EPrints File Classification + Risk Analysis



TU WIEN planets

UNIVERSITY OF Southampton eprints

Preservation Planning

Why Preservation Planning?

- 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?

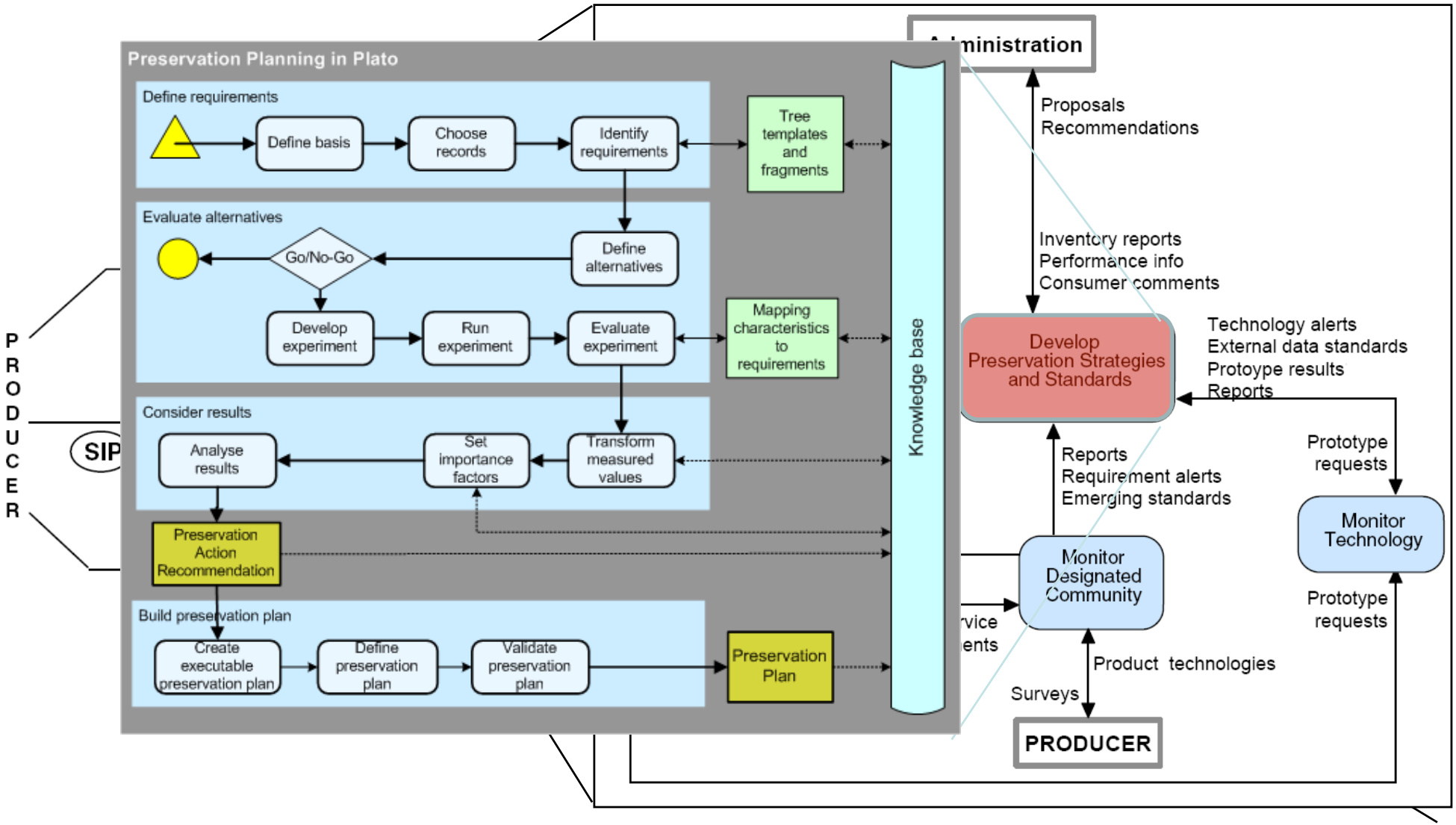
Digital Preservation

What is a preservation plan?

- 10 Sections
  - Identification
  - Status
  - Description of Institutional Setting
  - Description of Collection
  - Requirements for Preservation
  - Evidence for Preservation Strategy
  - Cost
  - Trigger for Re-evaluation
  - Roles and Responsibilities
  - Preservation Action Plan

Preservation Plan Template

Preservation Planning



TU WIEN planets

UNIVERSITY OF Southampton eprints

Preservation Planning with Plato

What we have now:

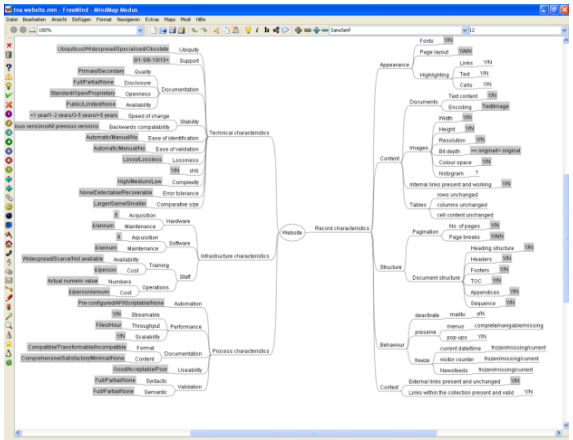
- Basic Preservation Plan:
  - *PDF: Preservation Plan.pdf*
  - *XML: Preservation Plan.xml*
- That was developed in a solid, repeatable and documented process
- That is optimal for the needs of a given institution and for the data at hand

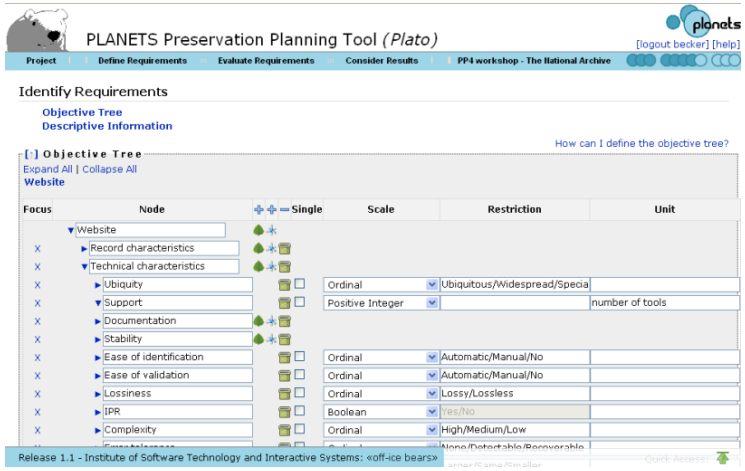
TU WIEN planets UNIVERSITY OF Southampton eprints

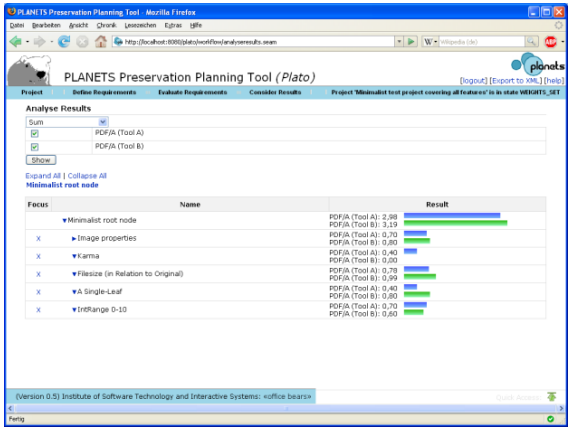
Preservation Planning

Plato

- Preservation Planning Tool
- Reference implementation of planning workflow
- Documents the process and ensures all steps are considered
- Creates a preservation plan







Conclusions

- Physical preservation ensures longevity of resources.
- Simple risk analysis reporting
- Preservation Planning to ensure “optimal” preservation
- A simple, methodologically sound model to specify and document requirements
- Repeatable and documented evaluation
- Basis for well-informed, accountable decisions
- Follows recommendations of TRAC and nestor
- **Plato**:
  - Tool support to perform solid, well-documented analyses
  - Creates core preservation plan
- **EPrints:**
  - Software to manage the institutional repository.
  - Accounting, reporting and preservation.

TU WIEN planets UNIVERSITY OF Southampton eprints

Thank you!

http://www.ifs.tuwien.ac.at/dp

http://www.eprints.org/