



Distributed Preservation Services: Integrating Planning and Actions

Christoph Becker, Miguel Ferreira, Michael Kraxner,
Andreas Rauber, Ana Alice Baptista, José Carlos Ramalho

ECDL 2008
Aarhus, Denmark

The Longevity of Digital Objects

- ❑ Digital objects are the dominant way we exchange information
- ❑ Digital objects need technical environment to “function”
- ❑ Heterogeneity and complexity of file formats and speed of technological change make long-term access a challenge
- ❑ Digital preservation: Long-term access to digital objects
- ❑ Dominant types of preservation actions:
 - Migration
 - Emulation



Why do we need Digital Preservation?

- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...
- programs won't
- ...
- ...
- ...

.....

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.
-

Evaluating preservation strategies

- ❑ Variety of solutions and tools exist
- ❑ Each strategy has unique strengths and weaknesses
- ❑ Requirements vary across settings
- ❑ Decision on which solution to adopt is complex
- ❑ Documentation and accountability is essential

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



Preservation planning and services

- ❑ Planets preservation planning approach assists in decision making
- ❑ Evaluate potential actions and build plans for preserving digital objects
- ❑ Planning Tool supports the workflow

- ❑ Discovery of potential tools and actual application of them is still effort-intensive
- ❑ Automation and support needed



Outline

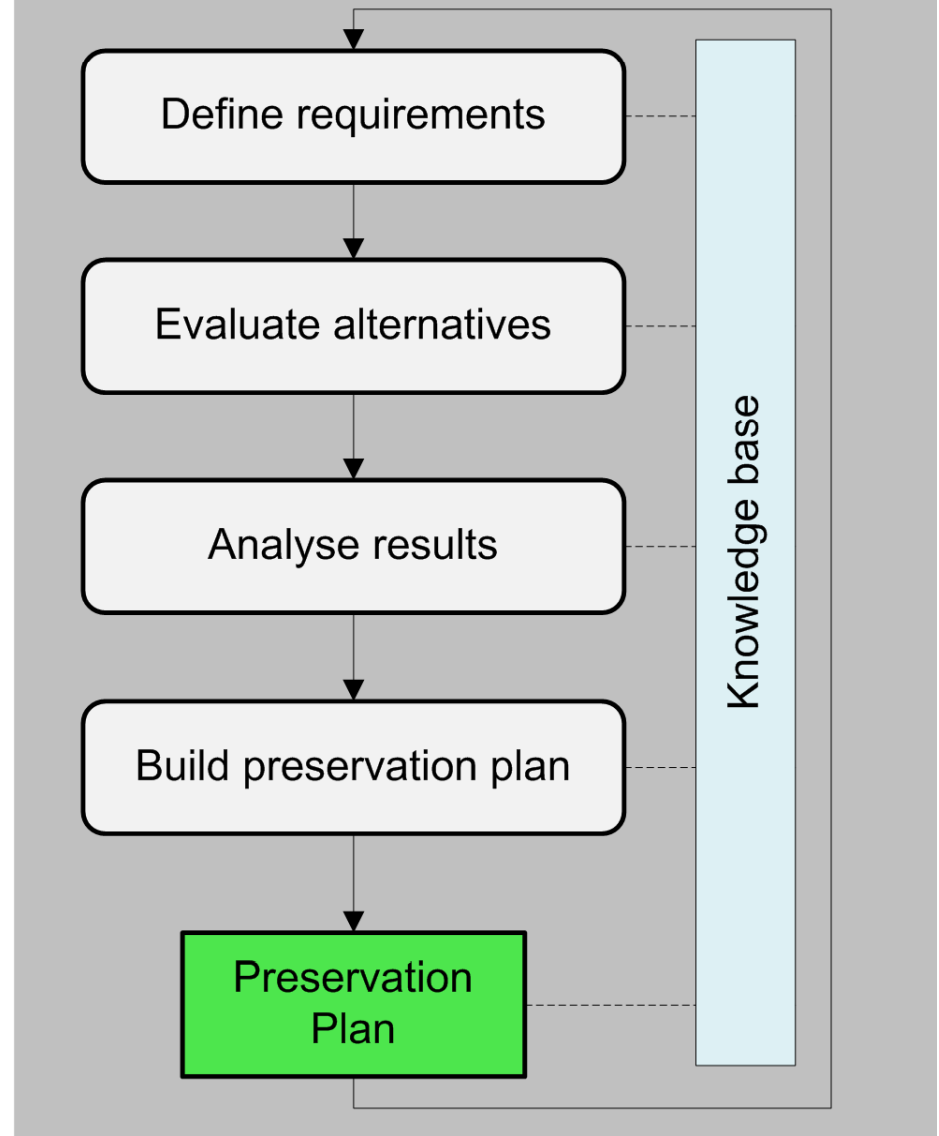
- Digital Preservation and Preservation Planning
- Preservation Planning in Plato
 - Empirical evaluation of potential actions
 - Plato: The Planets Preservation Planning Tool
- Distributed Preservation Services
 - CRiB
 - Planets
- Service discovery and integration
- Current and future work



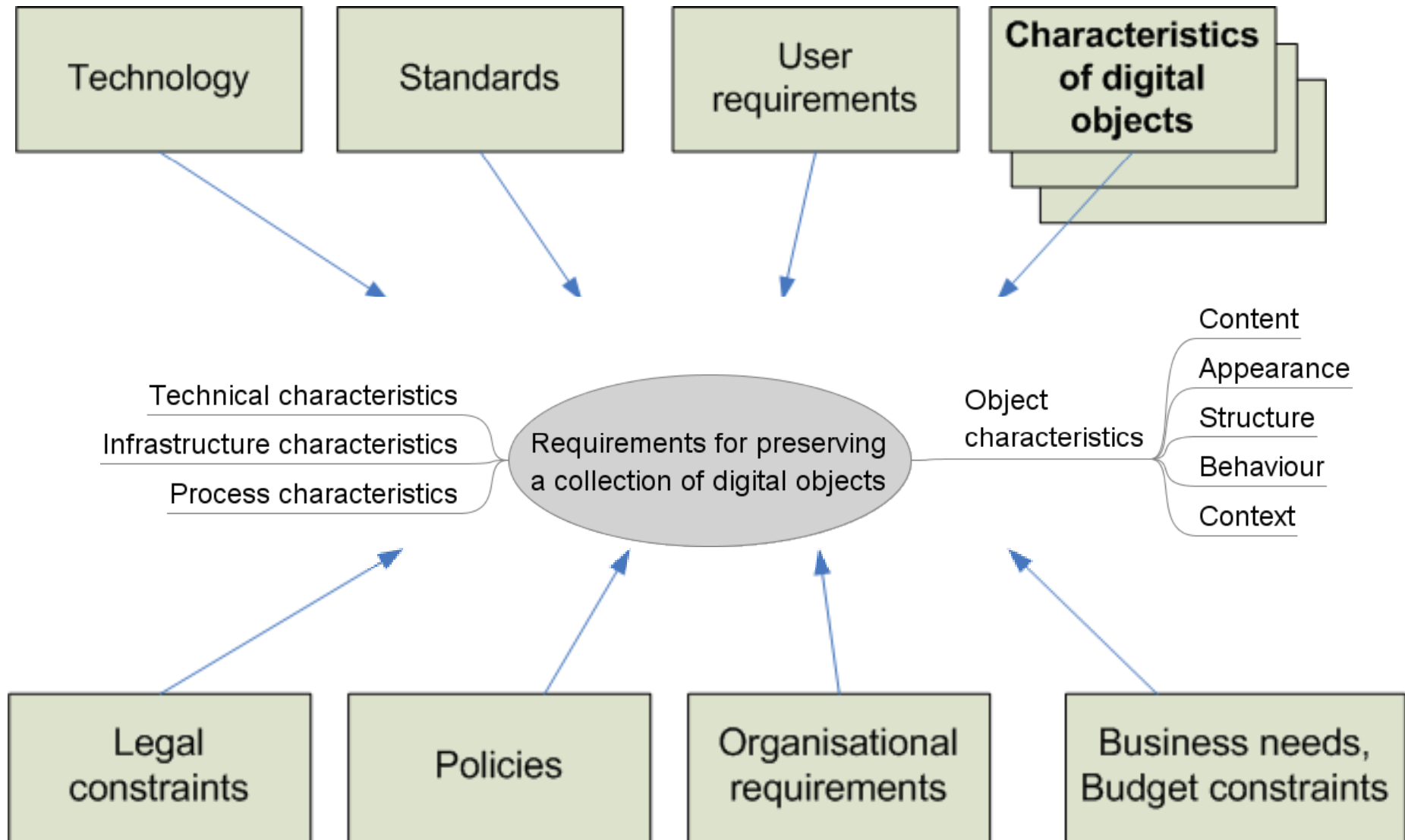
Planets Preservation Planning Workflow

- ❑ Define requirements
 - ❑ Context
 - ❑ Sample objects
 - ❑ Requirements
- ❑ Evaluate potential actions
- ❑ Analyse results
- ❑ Build a preservation plan

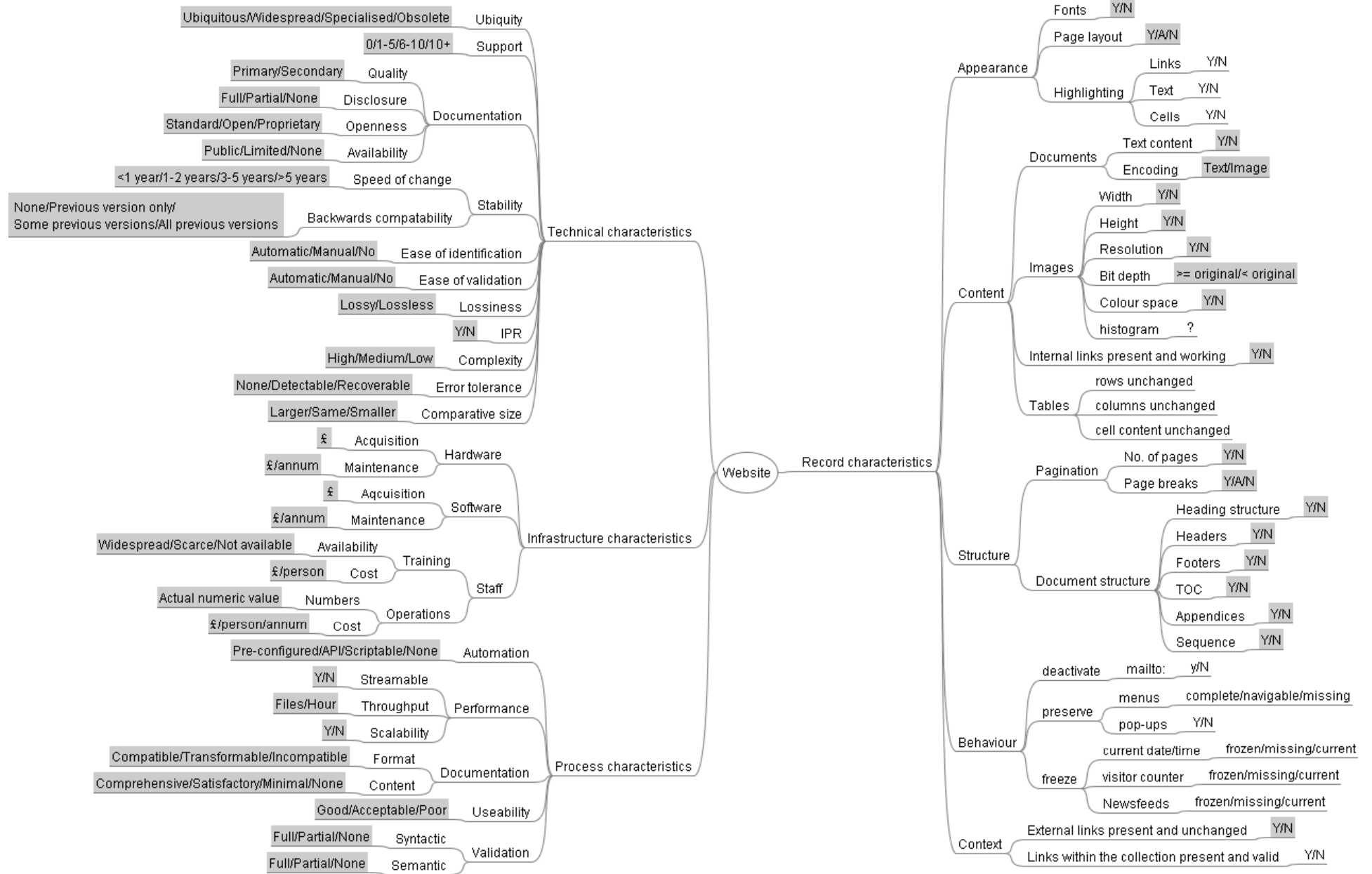
Preservation Planning in Plato



Requirements and Influence Factors



An Objective Tree



Preservation Planning in Plato

- ❑ Web based planning tool implementing the Planets preservation planning workflow

- ❑ Integration of registries and services for
 - File format identification
 - Preservation action
 - Characterisation and comparison

- ❑ A distributed architecture of preservation services



Objective Tree in Plato



Identify Requirements

[Expand All](#) | [Collapse All](#)

Website

Focus	Node	+	+	-	Single	Scale	Restriction	Unit
	Website	+	+					
X	Record characteristics	+	+					
X	Appearance	+	+					
X	Content	+	+					
X	Structure	+	+					
X	Behaviour	+	+					
X	deactivate	+	+					
X	mailto:				<input type="checkbox"/>	Boolean	Yes/No	
X	preserve	+	+					
X	menus				<input type="checkbox"/>	Ordinal	complete/navigable/missing	
X	pop-ups				<input type="checkbox"/>	Boolean	Yes/No	
X	freeze	+	+					
X	current date/time				<input type="checkbox"/>	Ordinal	frozen/missing/current	
X	visitor counter				<input type="checkbox"/>	Ordinal	frozen/missing/current	
X	Newsfeeds				<input type="checkbox"/>	Ordinal	frozen/missing/current	
X	Context	+	+					
X	Technical characteristics	+	+					
X	Ubiquity				<input type="checkbox"/>	Ordinal	Ubiquitous/Widespread/Specialised/Obs	
X	Tool Support				<input type="checkbox"/>	Positive Number		Number of tools
X	Documentation	+	+					
X	Stability	+	+					
X	Ease of identification				<input type="checkbox"/>	Ordinal	Automatic/Manual/No	
X	Ease of validation				<input type="checkbox"/>	Ordinal	Automatic/Manual/No	
						Ordinal	Lossy/Lossless	

Empirical evaluation of actions

- ❑ Define representative sample content
- ❑ Discover applicable actions in service registries
- ❑ Apply potential actions to sample content
- ❑ Evaluate outcomes
- ❑ Select most suitable action(s)
- ❑ Define an (executable) preservation plan



Action vs. characterisation

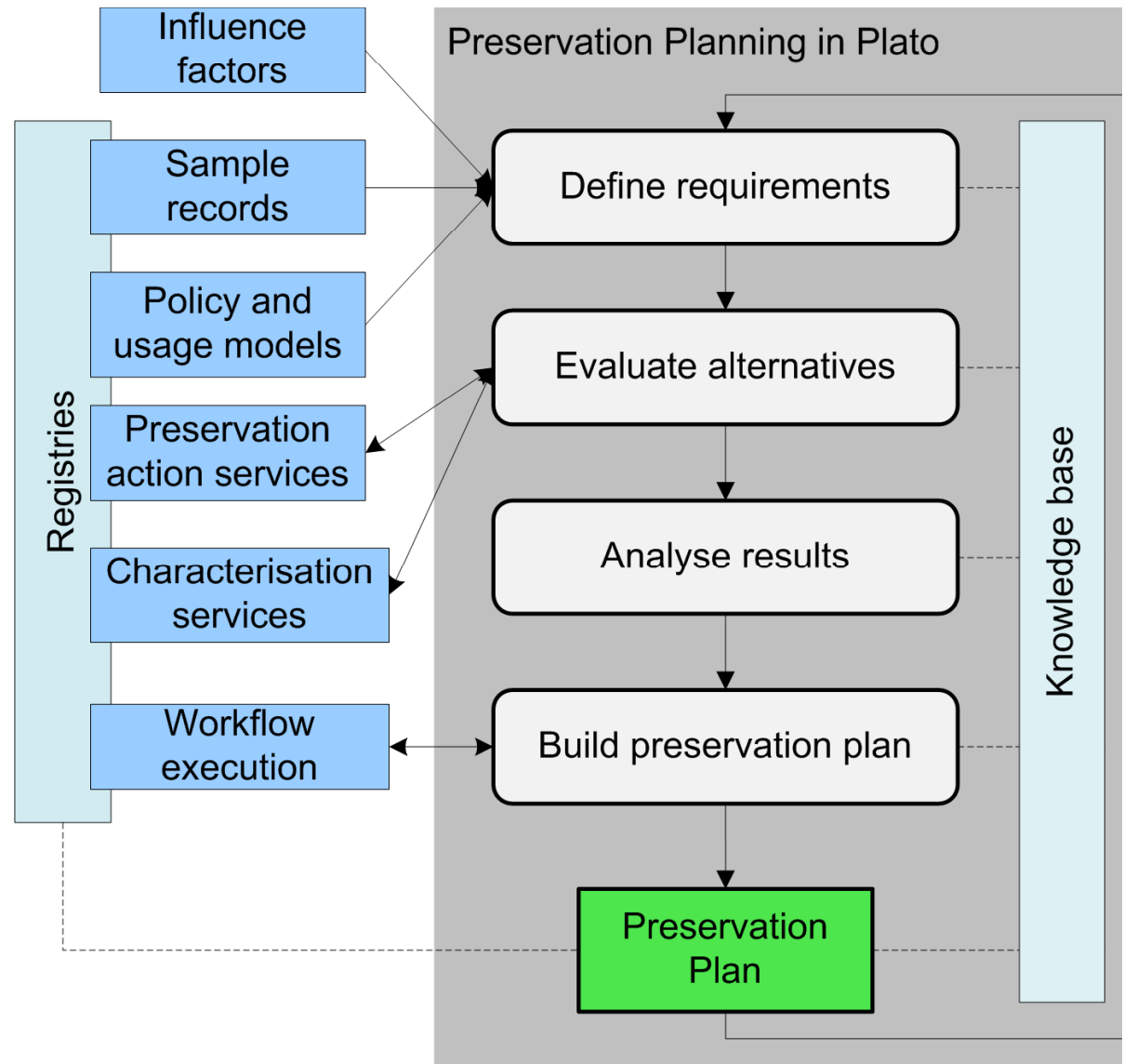
- ❑ Preservation action
 - ❑ Tools are applied to sample content
 - ❑ Discovered in registries
 - ❑ Invoked online

- ❑ Preservation characterisation
 - ❑ Identify format
 - ❑ Describe and assess objects
 - ❑ Compare transformed content

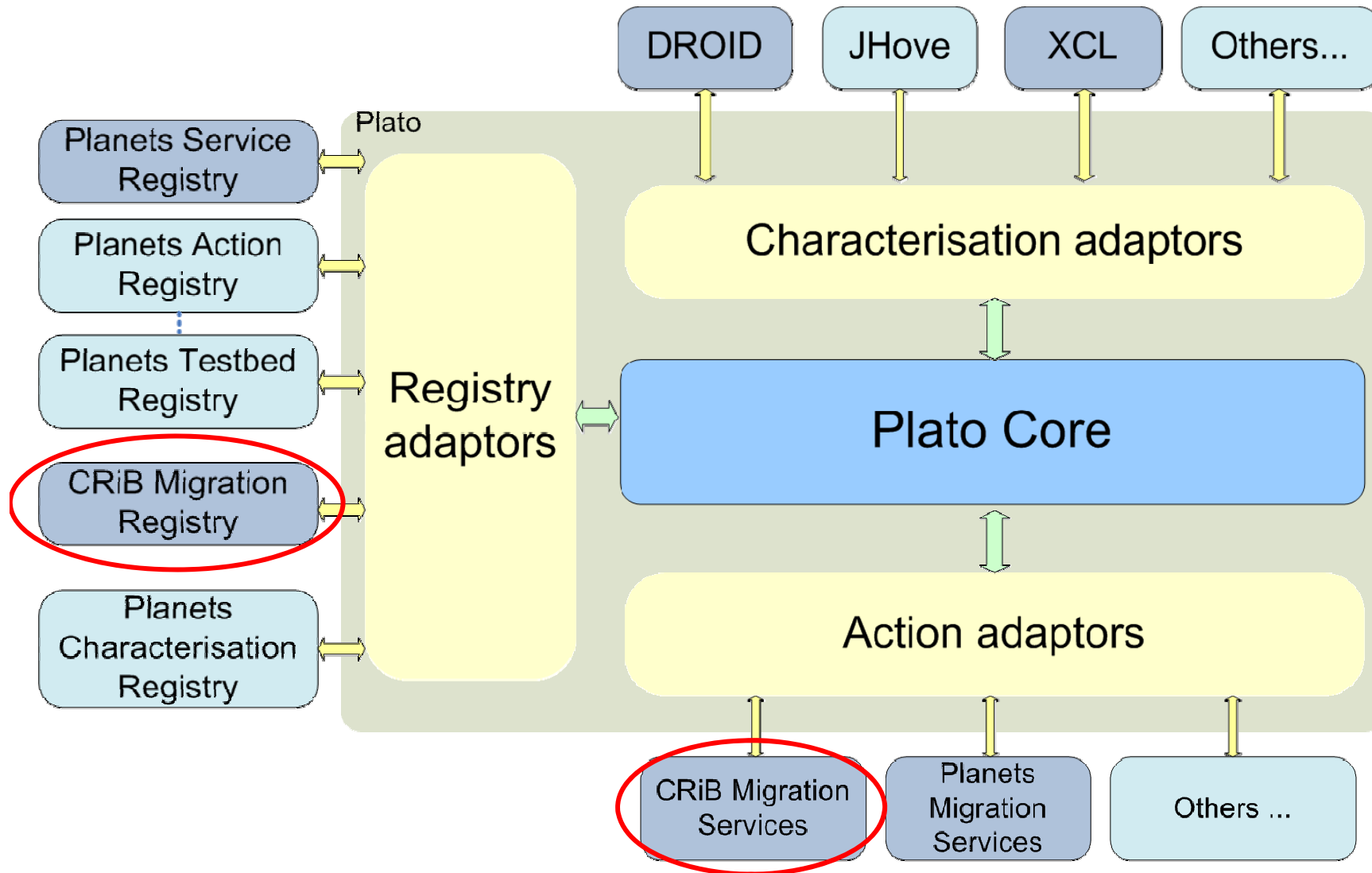


Collecting information: Registries

- ❑ Format registries
- ❑ Preservation action registries
- ❑ Preservation characterisation registries
- ❑ Data registries



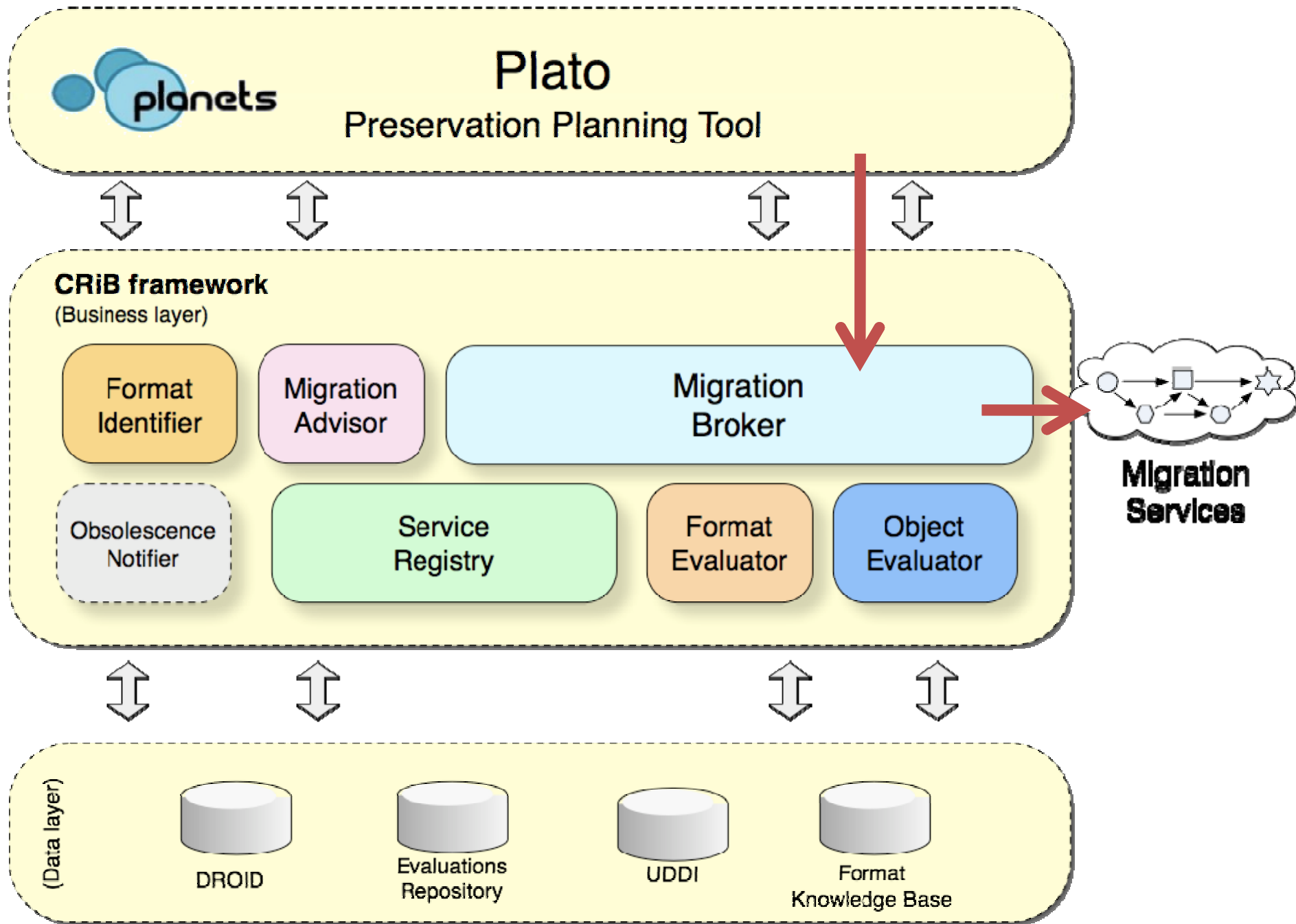
Integration architecture



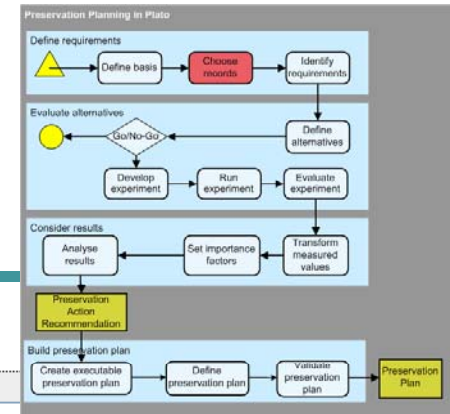
CRiB

- ❑ Conversion and Recommendation of Digital Object Formats
- ❑ Service Oriented Architecture assists institutions in migration-based preservation activities
- ❑ Central components
 - ❑ Format identifier, format evaluator
 - ❑ Migration advisor, object evaluator
 - ❑ Migration broker, Service registry
- ❑ Currently 39 atomic services for standard object types
 - ❑ CRiB itself is distributed
 - ❑ Unix: ImageMagick, sam2p
 - ❑ Windows services for Office documents
 - ❑ Migration paths are chained automatically







Format identification

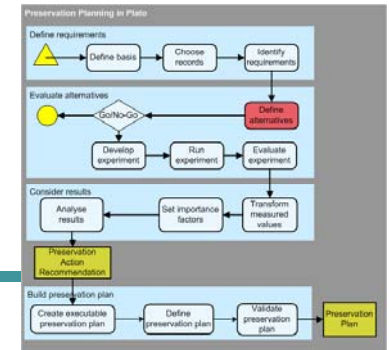


Sample Records

Description of sample records: some test samples

Sample Record	Object Format
Full name: <input type="text" value="sample 1"/> * ?	PUID: <input type="text" value="fmt/11"/> ?
Short name: <input type="text" value="eins.png"/> * ?	Name: <input type="text" value="Portable Network Graphics"/> ?
Has data: <input checked="" type="checkbox"/> <input type="button" value="download"/>	Version: <input type="text" value="1.0"/> ?
Original technical environment: <input type="text"/> ?	Mime-type: <input type="text" value="image/png"/> ?
Description: <input type="text"/> ?	<input type="button" value="Identify format"/> 
<input type="button" value="Remove record"/>	
Full name: <input type="text" value="sample number two"/> * ?	PUID: <input type="text" value="fmt/11"/> ?
Short name: <input type="text" value="zwo.png"/> * ?	Name: <input type="text" value="Portable Network Graphics"/> ?
Has data: <input checked="" type="checkbox"/> <input type="button" value="download"/>	Version: <input type="text" value="1.0"/> ?
Original technical environment: <input type="text"/> ?	Mime-type: <input type="text" value="image/png"/> ?
Description: <input type="text"/> ?	<input type="button" value="Identify format"/> 
<input type="button" value="Remove record"/>	

Discovering possible actions



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
<input type="checkbox"/>	JPG > BMP	Windows Bitmap, version 3.0	JPG>BMP
<input checked="" type="checkbox"/>	JPG > TIF	Tagged Image File Format, version 3	JPG>BMP>TIF
<input type="checkbox"/>	JPG > TIF #2	Tagged Image File Format, version 3	JPG>TIF
<input checked="" type="checkbox"/>	JPG > TIF_2	Tagged Image File Format, version 3	JPG>TIF_2
<input type="checkbox"/>	JPG > PNG	Portable Network Graphics, version 1.0	JPG>PNG
<input type="checkbox"/>	JPG > JP2	JPEG 2000	JPG>JP2

Create alternatives for selected services

Calling migration services

Run Experiments

Alternative	Description
Adobe Acrobat->DOC	Document is loaded in Adobe Acrobat and saved as a Microsoft Word .Doc File.
Convert Doc->DOC	Convert Doc is configured to transform the PDF Document to a Microsoft Work Document and the process of conversion is started.
Adobe Acrobat->HTML	Document is loaded in Adobe Acrobat and saved as an HTML File.
PDF > TIF #7 	
PDF > MultipageTIF # 	
PDF > TextLayout #21 	

Run all experiments



Save



Discard changes



Save and proceed

Analysing results

Results: Weighted multiplication

Result-Tree with all Alternatives, Aggregation method: Weighted multiplication

[Expand All](#) | [Collapse All](#)

National Library Publications

Focus	Name	Result
	▼ National Library Publications	Adobe Acrobat->DOC: 0,00 Convert Doc->DOC: 3,44 Adobe Acrobat->HTML:3,18
X	▶ Object characteristics	Adobe Acrobat->DOC: 1,55 Convert Doc->DOC: 1,63 Adobe Acrobat->HTML:1,52
X	▶ Technical characteristics	Adobe Acrobat->DOC: 1,14 Convert Doc->DOC: 1,14 Adobe Acrobat->HTML:1,16
X	▼ Process Characteristics	Adobe Acrobat->DOC: 0,00 Convert Doc->DOC: 1,14 Adobe Acrobat->HTML:1,08
X	▶ Duration	Adobe Acrobat->DOC: 0,00 Convert Doc->DOC: 1,23 Adobe Acrobat->HTML:1,06
X	▶ Automation of the process	Adobe Acrobat->DOC: 1,55 Convert Doc->DOC: 1,90 Adobe Acrobat->HTML:1,55
X	▶ Integrity	Adobe Acrobat->DOC: 1,00 Convert Doc->DOC: 1,00 Adobe Acrobat->HTML:1,00
X	▶ Costs	Adobe Acrobat->DOC: 1,67 Convert Doc->DOC: 1,63 Adobe Acrobat->HTML:1,67

Summary

- ❑ Digital preservation and preservation planning
- ❑ Distributed planning environment
- ❑ Service discovery and invocation

- ❑ CRiB migration services
- ❑ Integration in the planning tool Plato

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:

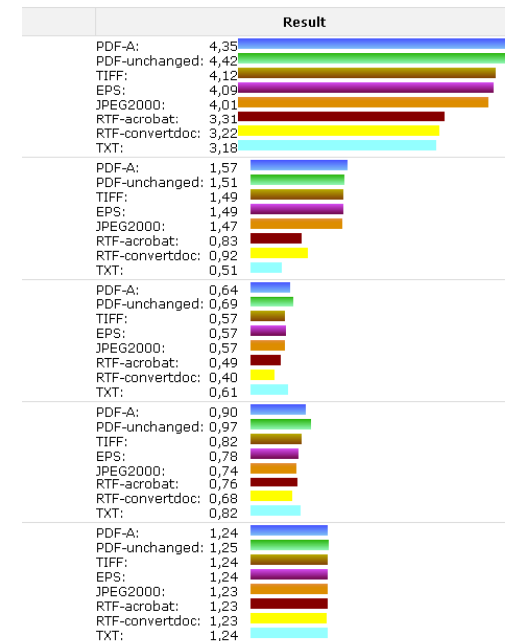
Preservation Action	Target Format	Info
<input type="checkbox"/> JPG > BMP	Windows Bitmap, version 3.0	JPG>BMP
<input checked="" type="checkbox"/> JPG > TIF	Tagged Image File Format, version 3	JPG>BMP>TIF
<input type="checkbox"/> JPG > TIF #2	Tagged Image File Format, version 3	JPG>TIF
<input checked="" type="checkbox"/> JPG > TIF_2	Tagged Image File Format, version 3	JPG>TIF_2
<input type="checkbox"/> JPG > PNG	Portable Network Graphics, version 1.0	JPG>PNG
<input type="checkbox"/> JPG > JP2	JPEG 2000	JPG>JP2

Planets Preservation Action Tool registry

Planets Service Registry

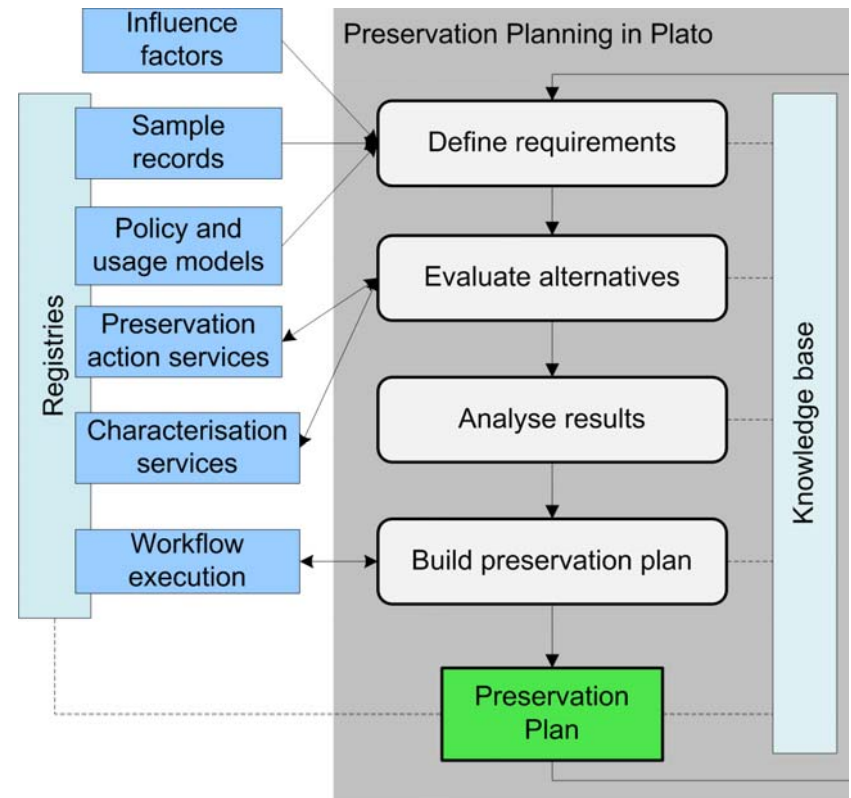
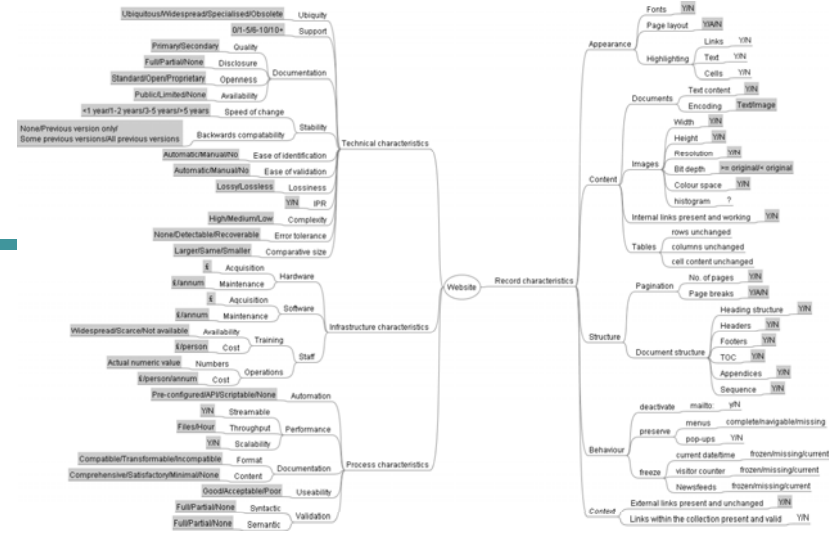
CRiB Service Registry

Create alternatives for selected services



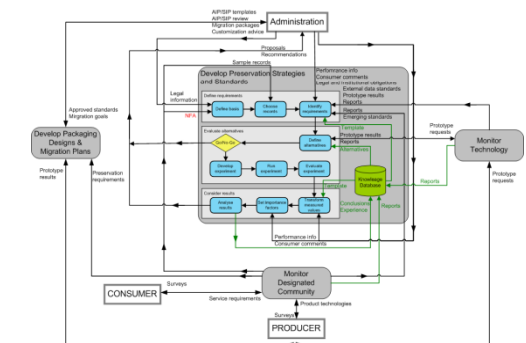
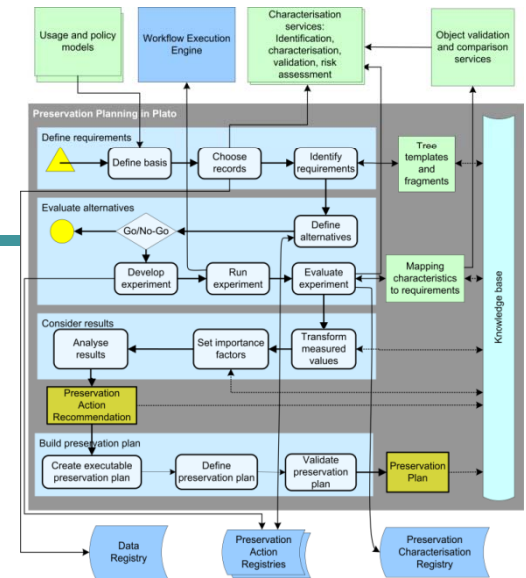
Plato

- ❑ Plato 2.0 scheduled for October 2008
 - ❑ Service integration
 - ❑ Preservation plan
- ❑ Plato is available at www.ifs.tuwien.ac.at/dp/plato
- ❑ Plato Demonstration here at ECDL

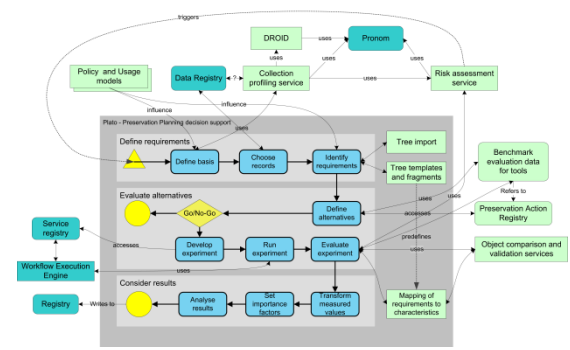


Current and future work

- ❑ Improved integration of emulation
- ❑ Integration of characterisation and comparison services
- ❑ Pluggable infrastructure for the automated evaluation of preservation actions
- ❑ Integrated knowledge base and recommender systems
- ❑ Case studies evaluating strategies and building preservation plans



Integrating Preservation Planning Decision support with Planets components
November 2007



Thank you very much for your attention.

Questions?

becker@ifs.tuwien.ac.at

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

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

www.planets-project.eu

