| | |
|---|---|
| Project Number | IST-2006-033789 |
| Project Title | Planets |
| Title of Deliverable | **Report on service integration in Plato 2** |
| Deliverable Number | **PP4/D3** |
| Contributing Sub-project and Work-package | SP/PP/4 |
| Deliverable Dissemination Level | External<br>Planets All |
| Deliverable Nature | Report |
| Contractual Delivery Date | 31st May 2008 |
| Actual Delivery Date | June 3, 2008 |
| Author(s) | TUWIEN |


**Abstract**

This report describes the PA/PC service integration in Plato v2 to automate decision support, and provides an outlook to the 3rd version of Plato. It also covers the risk assessment service developed in task PP/4.4



**Keyword list**

Digital Preservation, Preservation Planning, Tool support, Prototype, Distributed Preservation Services, Migration, Characterisation, Preservation Action, Registry, Design

### Contributors

| Person | Role | Partner | Contribution |
|---|---|---|---|
| Christoph Becker | Main author | TUWIEN | First draft |
| Hannes Kulovits | Reviewer | TUWIEN | Internal review and comments |
| Adrian Brown | Contributor | TNA | Section Risk assessment and various discussion points |

### Document Approval

| Person | Role | Partner |
|---|---|---|
| Andreas Rauber | WPL | TUWIEN |
| Hans Hofman | SPL | NANETH |
| Caroline van Wijk | Reviewer | KB-NL |
| Andrew Jackson | Reviewer | BL |

### Distribution

| Person | Role | Partner |
|---|---|---|
| Andreas Rauber | WPL | |
| Hans Hofman | SPL | NANETH |
| PP subproject mailing list | | |
| Deliverables mailing list | | |
| Caroline van Wijk | Reviewer | KB-NL |
| Andrew Jackson | Reviewer | BL |

### Revision History

| Issue | Author | Date | Description |
|---|---|---|---|
| 0.1 | Christoph Becker | April 28 | Initial draft |
| 0.2 | Christoph Becker | April 29 | Revised |
| 0.3 | Christoph Becker | April 30 | Revised after comments |
| 0.4 | Adrian Brown | May 2 | Added risk assessment and comments |
| 0.5 | Christoph Becker | May 2 | Incorporated comments |
| 0.9 | Christoph Becker | May 7 | Final draft for review |
| 1.0 | Christoph Becker | June 3 | Final version incorporating reviewers' comments and changes by Adrian Brown |

### References

| Ref. | Document | Date | Details and Version |
|---|---|---|---|
| 1 | Planets PP4/D1 Report on methodology for specifying preservation plans | 31st July 2007 | available at http://www.planets-project.eu/publications |
| 2 | Stephan Strodl, Christoph Becker, Robert Neumayer, Andreas Rauber: How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure | June 2007 | Proceedings of the ACM IEEE Joint Conference on Digital Libraries (JCDL'07), Vancouver, British Columbia, Canada, June 18-23, 2007. pp 29-38 |
| 3 | Christoph Becker, Stephan Strodl, Robert Neumayer, Andreas Rauber, Eleonora Nicchiarelli Bettelli, Max Kaiser: Long-Term Preservation of Electronic Theses and Dissertations: A Case Study in Preservation Planning | October 2007 | Proceedings of the Ninth Russian National Research Conference on Digital Libraries: Advanced Methods and Technologies, Digital Collections (RCDL'07), Pereslavl, Russia, October 15-18, 2007 |
| 4 | Christoph Becker, Günther Kolar, Josef | December | In: Proceedings of the Tenth Conference on Asian |

| | | | |
|---|---|---|---|
| | Küng, Andreas Rauber. Preserving Interactive Multimedia Art: A Case Study in Preservation Planning. | 2007 | Digital Libraries (ICADL'07). Hanoi, Vietnam, December 10-13, 2007. |
| 5 | Christoph Becker, Andreas Rauber, Volker Heydegger, Jan Schnasse, Manfred Thaller. | March 2008 | In: Proceedings of the ACM Symposium on Applied Computing (SAC'08), Track 'Document Engineering'. Fortaleza, Brazil, March 16-20, 2008. |
| 6 | Christoph Becker, Hannes Kulovits, Andreas Rauber, Hans Hofman. Plato: a service-oriented decision support system for preservation planning. | June 2008 | In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'08). Pittsburgh, Pennsylvania , June 16-20, 2008. (accepted for publication) |
| 7 | White Paper: Representation Information Registries. Planets deliverable PC3-D7 | | |
| 8 | Preservation plan definition version 0.83. Planets document | April 2008 | **http://www.planets-project.eu/private/pages/wiki/index.php/PP/Conceptual_issues** |
| 9 | Heydegger, V., Neumann, J., Schnasse, J., Thaller, M. Basic design for the extensible characterisation language. PC/2-D1, PC/2-D2 Planets internal deliverable. | October 2006 | |
| 10 | Ferreira, M., Baptista, A. A., and Ramalho, J. C. An intelligent decision support system for digital preservation | July 2007 | . Int. Journal Digital Libraries (IJDL) 6, 4 (July 2007), 295–304. |
| 11 | PP5-D1 Specification of basic metric and evaluation framework. Planets external deliverable. | May 2008 | |

# EXECUTIVE SUMMARY

This document describes the integration of services into the next two iterations of the Planets preservation planning tool, Plato.
We outline the underlying principles and the workflow that is implemented by the tool, and then provide an overview of integration points where services are being connected to the workflow.
We then discuss each of the following class of services and how it is going to be integrated in the planning workflow:

1. Preservation action services that perform primarily migration, but in principle also emulation, are discovered through registries;
2. Preservation characterisation services include format identification, validation and characterisation as well as collection profiling and risk assessment services. These are used for describing the set of objects that a preservation plan is created for, aid the selection of representative samples, and compare transformed representations to the original objects.

We describe where and how these services will be integrated, and provide a walk-through scenario to illustrate the concepts discussed.
A dependency diagram showing a roadmap for the future development of Plato concludes this report.

# TABLE OF CONTENTS

# 1. Introduction

This document describes the integration of Planets concepts and services into the upcoming releases of the Planets Preservation Planning Tool, Plato. Thus it provides a sort of functional outline for a central part of the deliverable PP4-D4, and an outlook to the third version of Plato.

We will first introduce and shortly describe the original workflow for evaluating preservation strategies as previously described in PP4-D1, "Report on methodology for specifying preservation plan", which forms one of the important precursors of this document, together with the first version of the Planning Tool, Plato 1, PP4-D2, which was released in November 2007.

We will then describe how Plato 2 takes forward this approach in two aspects:

1. It extends the workflow and takes the next step after evaluation to include the actual creation of preservation plans, and
2. In the Planets spirit of networked services, it integrates the distributed preservation services that have been and are being built and connects them to the preservation planning process.

In Section 2, we will describe the revised and extended preservation planning workflow and discuss the definition of a preservation plan as it has been agreed on in the preservation planning subproject.

Section 3 presents the risk assessment service developed within PP/4 and discusses its incorporation into the PC registry.

We will then discuss integration points for the various services and concepts developed in Planets in Section 4 and point to work outside of Planets that is relevant in this respect. Specifically, we will describe the integration of services as it is currently foreseen in Plato 2 and 3 step-by-step and discuss open issues that will be resolved during the next project period.

Section 5 provides an exemplary scenario walk-through to illustrate the concepts discussed before, and Section 6 provides a summary including a development and integration roadmap.

# 2. The preservation planning methodology

## Introduction

This section shortly recapitulates the preservation planning methodology and workflow. We will first describe the revised preservation planning workflow and the current state of the definition of a preservation plan as it has been agreed within the subproject.

## Preservation Planning Workflow

The deliverable PP4-D1 described in detail the methodology for evaluating potential preservation actions and strategies based on utility analysis. Figure 1 depicts this workflow as it has been specified in PP4-D1. It consists of three phases:

1. **Requirements definition** is the natural first step of the workflow, laying out the fundamental basis of the planning endeavour. The relevant context of the institution, the collection of objects in question, and the application of policies and constraints are defined. As the application and manual evaluation of preservation actions on a potentially large number of objects contained in a given collection is an infeasible task, the preservation planner selects a collection of sample objects that are representative of the total set, i.e. cover the essential properties and technical characteristics of the objects. Then the requirements are defined in a tree structure called *objective tree*, starting with high-level requirements such as object characteristics or process-related criteria and breaking them down to measurable requirements such as the fixity of image width or the time needed to transform a single object to a new representation. Examples of objective trees are presented in [1-4].
2. **Evaluation of potential strategies** first means discovering preservation actions that are applicable to the given set of objects. The actual evaluation is then done in an empirical

manner by applying the selected strategies to the samples defined in the first phase. The results are evaluated against the requirements specified in the objective tree.

3. **Analysis of results** needs to take the different importance factor of requirements into account. It thus involves a step of assigning relative weight factors to the requirements on each level in the tree hierarchy. A quantitative analysis of results is carried out by preservation planners, ranking the alternatives and giving a well-documented recommendation for a preservation action to apply on the collection of objects that shall be preserved.

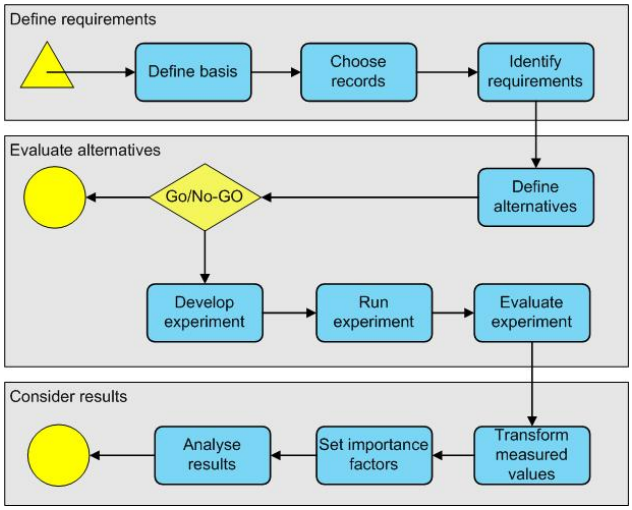


**Figure 1 PP workflow as specified in PP4-D1**

The first release of Plato implemented this workflow. Plato itself is a J2EE web application relying on open frameworks such as Java Server Faces and AJAX for the presentation layer and Enterprise Java Beans for the backend. It is integrated in the Planets Interoperability Framework that guarantees loose coupling of services and registries through standard interfaces and provides common services such as user management, security, and a common workspace. Based on this technical foundation, the aim is to create an interactive and highly supportive software environment that advances the insight of preservation planners and enables proactive preservation planning.

Based on the experience gained in case studies as well as feedback from partners, the workflow was extended by a fourth phase in which a preservation plan is created, based on the recommendation that results from the evaluation. The resulting four-phase workflow is shown in Figure 2.

4. This **fourth phase** takes this recommendation as the basis to **build a preservation plan**, i.e. *a definition of the steps of actions that shall be taken to preserve the given set of digital objects or records*. This plan thus includes a description of the planning context and environment, i.e. the institution's mission statement, the characteristics of the designated user community and applying policies; a definition of the collection which shall be preserved and its properties, such as number and type of objects, usage patterns and the chosen sample records; the objective tree specifying the institutions' requirements; the considered preservation actions, evaluation results and the resulting recommendation including a complete evidence base; and the **preservation action plan** as a core part, which can be an executable workflow defined in XML and accessing distributed services, provided that the chosen preservation action and its deployment support it.

While Plato 2 will implement the extended four-phase workflow, the creation of an executable workflow depends on a number of components that are going to be available at the release time of Plato 2, specifically the PA and PC service registry and the Planets data model. This will thus probably form part of the subsequent release of Plato.

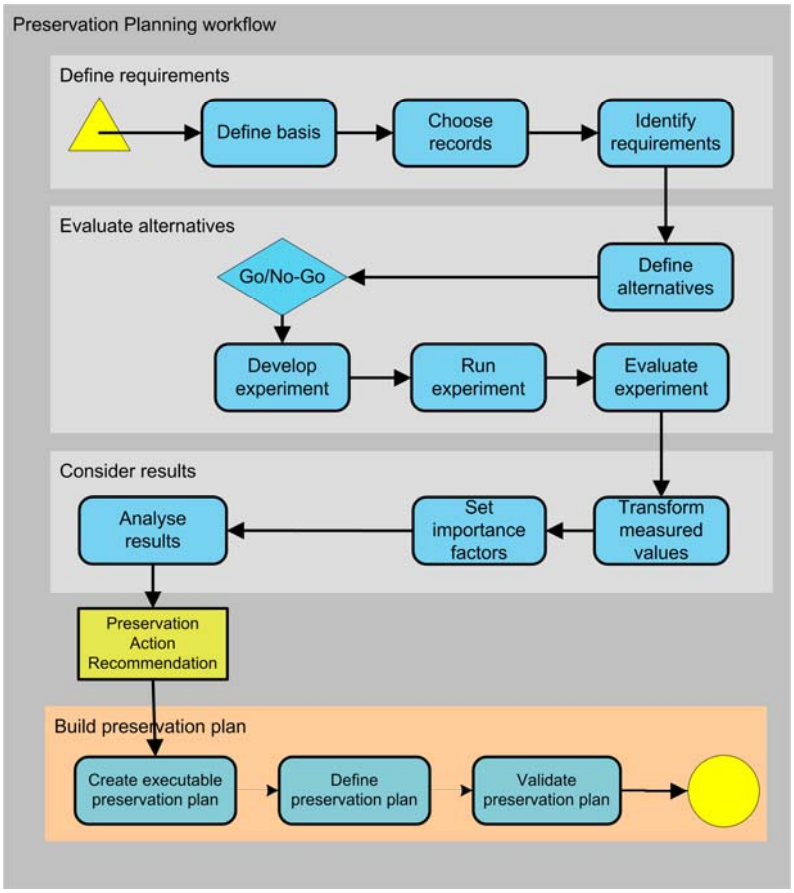


**Figure 2 Revised Planets Preservation Planning Workflow**

## Preservation Plan

According to the preservation plan template that has been specified within the PP subproject,

*A preservation plan defines a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records (called collection). The Preservation Plan takes into account the preservation policies, legal obligations, organisational and technical constraints, user requirements and preservation goals and describes the preservation context, the evaluated preservation strategies and the resulting decision for one strategy, including the reasoning for the decision.*

*It also specifies a series of steps or actions (called **preservation action plan**) along with responsibilities and rules and conditions for execution on the collection. Provided that the actions and their deployment as well as the technical environment allow it, this action plan is an executable workflow definition.[8]*

It contains the following sections:

1. Identification
2. Status
3. Description of the institutional setting
4. Description of the collection
5. Requirements for preservation
6. Evidence of decision for preservation strategy
7. Cost indications
8. Triggers for re-evaluation/update
9. Roles and responsibilities
10. Preservation Action Plan

# 3. Risk assessment service

The Preservation Characterisation Registry will provide a risk assessment service for digital objects, which will be used within the preservation planning workflow. This section provides an overview of the risk assessment service, and of its planned use by PLATO.

## PC Risk Assessment

The PC registry incorporates a risk assessment service developed to support the Planets preservation planning service (PLATO). Risk assessments may be calculated at two levels. Generic format risks are calculated using a set of standard criteria, based on format properties recorded on the registry (e.g. use of open standards, level of tool support, age). These risks can be modified for a given object based on specific properties of the format: for example, PDF files have an associated property which reflects their ability to support encryption. For objects where this property is true, their *instance risk* is increased above the generic risk which applies to PDF, to reflect the additional preservation risks which encryption entails. The registry will expose a web service to return current risk scores, which may be invoked either for a generic format (using the PUID) or for a specific instance (by supplying additional parameters to describe applicable property values). These risk scores and property values are mapped onto nodes in the requirements trees created within Plato.

Each format recorded in the registry may be associated with a standard set of generic risk factors. An authority-controlled value can be assigned to each risk factor based on information recorded in the registry. For example, the extent to which the format uses open standards may be assessed by reviewing the status of the format documentation. The assignment of risk values will be manual, to allow for a degree of subjectivity and the inevitably incomplete status of information in the registry. However, in the future it may be possible to automate at least part of this process. The allowed values will also need to be backed-up by clear definitions, to reduce the degree of subjectivity to a minimum.

Examples of generic risk factors which are currently envisaged include:

| Risk factor | Definition | Allowed values | Score |
|---|---|---|---|
| Ubiquity | The degree of adoption of the format | Format is most widely adopted of type | 10 |
| | | Format is one of most widely adopted of type | 20 |
| | | Format is in occasional/specialised use | 30 |
| | | Format is no longer in current use | 40 |
| Support | The number of access tools currently available | Format is supported by 10+ tools | 10 |
| | | Format is supported by 6-10 tools | 20 |
| | | Format is supported by 1-5 tools | 30 |
| | | Format is supported by no tools | 40 |
| Disclosure | The extent to which the format documentation is publicly disclosed | Documentation is freely available online | 10 |
| | | Documentation is not available online | 20 |
| | | Documentation is available through NDA | 30 |
| | | Documentation is not available | 40 |
| Document quality | The accuracy and completeness of the available documentation | Documentation is complete and of a high standard | 10 |
| | | Documentation is complete | 20 |
| | | Documentation is poor | 30 |
| | | Documentation is not usable | 40 |
| Stability | The speed and backward-compatibility of format changes | Format is very stable and is backwards compatible | 10 |
| | | Format changes but is backwards compatible | 20 |

| | | | |
|---|---|---|---|
| | | Format is not backwards compatible , but versions change infrequently | 30 |
| | | Format changes frequently and is not backwards compatible | 40 |
| Ease of identification | The ease with which the format can be automatically identified | Can be positively identified as a specific format version. | 10 |
| | | Can be positively identified as a generic format type. | 20 |
| | | Can be tentatively identified | 30 |
| | | Cannot be identified | 40 |
| Ease of validation | The ease with which the format can be automatically validated | Validation tools available | 10 |
| | | No Validation tools available | 40 |
| Use of compression | The nature of any compression used | No Compression | 10 |
| | | Lossless Compression | 30 |
| | | Lossy Compression | 40 |
| IPR | The extent to which the format is encumbered by IPR issues | Format is not encumbered by IPR | 10 |
| | | Format is encumbered by IPR | 40 |
| Complexity | The degree of content and behavioural complexity supported | Low complexity format | 10 |
| | | Medium complexity format | 30 |
| | | High complexity format | 40 |

Risk factors and allowed values may both be extended and amended through the registry. Processes for adding and amending risk information will be defined as part of the procedures for maintaining the Planets registries.

Each risk factor may also be assigned a weighting, which determines its influence in calculating the overall risk pertaining to a particular format. Using the risk values and weightings, an overall risk score for a format can then be automatically calculated and stored in the registry.

The registry will expose generic risk assessment information via web services. Organisations may wish to apply different weightings to particular risk factors, depending on their preservation policies; indeed, they may wish to calculate risks in a fundamentally different manner. The registry will therefore provide access both to the individual risk factor values for a format, and to the overall risk score.

The risk assessment service may also be used in conjunction with the PC framework to determine the instance risks which apply to individual digital objects. The PC framework provides an interface for invoking Planets-enabled characterisation tools. The framework can therefore be used to measure those properties of a digital object which may affect its instance risk. Thus, for example, JHOVE might be used to determine that a particular PDF document is encrypted. The measured instance risk factors mapped onto the objective tree nodes in PLATO result in tailored risk values and may thus result in a different risk assessment then the generic assessment provided by the PC registry. It should, however, be noted that the registry cannot determine instance risks itself, since these are object-dependent.

### Use within PLATO

A digital repository might begin by using Planets characterisation services across its content. The Planets characterisation framework will allow them to deploy a range of characterisation tools appropriate to their content. Typically, this would begin with format identification, using a tool such as DROID, which draws its signature file from the PC registry. Next, the characterisation framework will automatically determine which characterisation tools are available to validate and measure properties for the formats identified, and automatically deploy these tools. Again, the PC registry provides tool information to allow the framework to make these decisions. Depending on the format, one of the characterisation tools which might be deployed would be the Planets XCEL tool, which utilises XCDL descriptions stored in the PC registry. The output of the characterisation process will be a set of metadata documenting both the process, and the characterisation properties determined for each digital object processed.

Secondly, the repository might use PLATO to undertake preservation planning for the characterised objects. PLATO will request risk factor values and risk scores from the PC registry

risk assessment service for formats of interest, i.e. the formats of sample objects and the formats that result from applying preservation action services on the sample objects, by citing their PUIDs. Plato will moreover characterise both sample objects and results from preservation actions and request the according risk assessment values from the registry.

# 4. Service integration in Plato

This section describes how services are going to be integrated in Plato 2 and 3. We are first giving an introductory overview, pointing out at which points in the workflow new developments are of interest that are going to be integrated, and then go through several of these points in more detail.

## Introduction and Overview

Figure 3 illustrates the preservation planning environment, putting the described workflow in the working context of services and registries as they are currently being implemented.

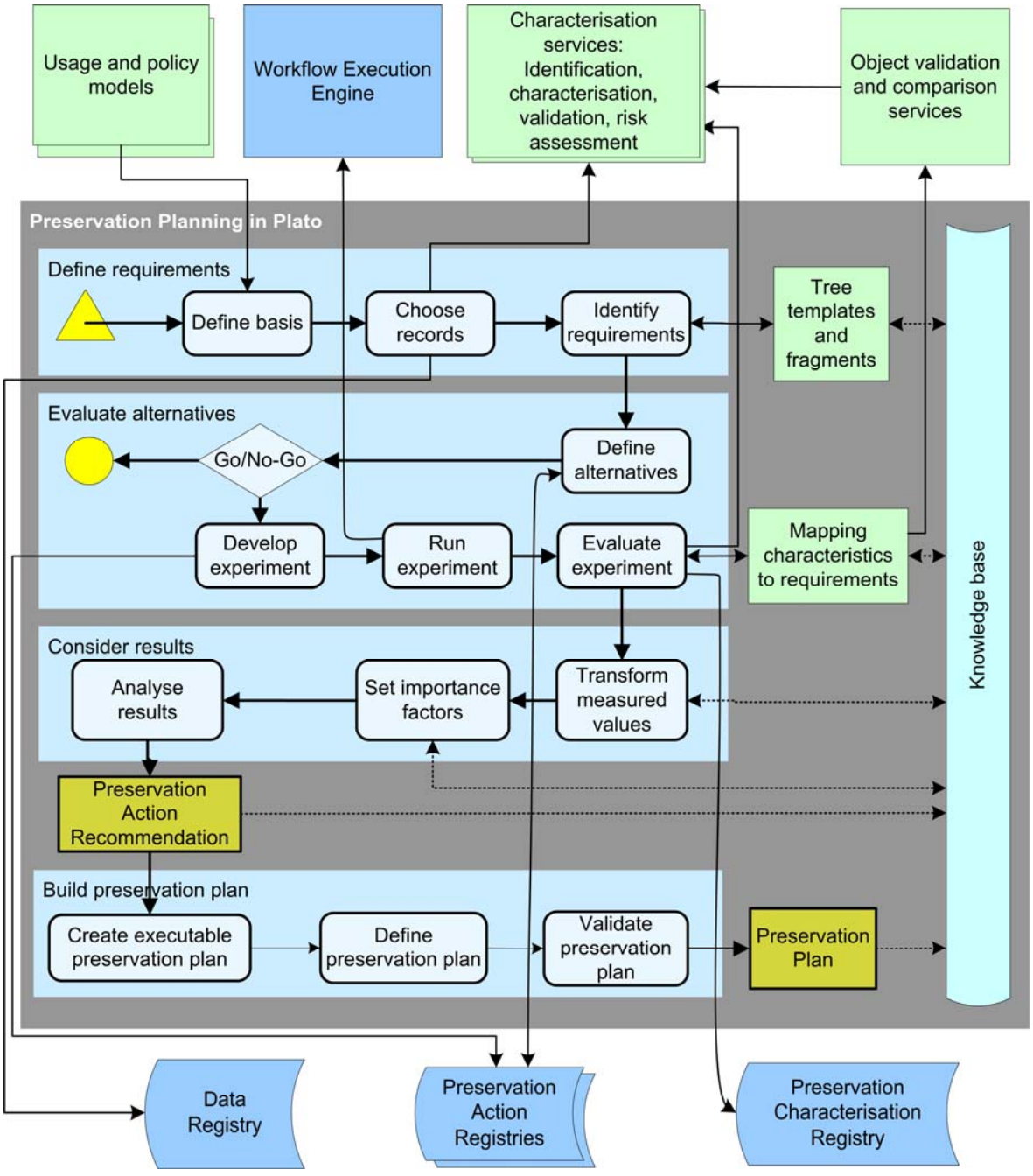


**Figure 3 Preservation Planning Environment**

In principle, there are two aspects to consider:

1. Integrating services for preservation action and characterisation of objects, and
2. Integrating registries for discovery of information and services.

### Define sample records

**Defining sample records** can profit considerably by integrating services for identification and validation of object formats as well as more extensive characterisation and risk assessment services that aid in the selection of suitable sample records.

### Define alternatives

**Discovery of applicable preservation action services** is the prime issue during the definition of alternatives to consider for evaluation. Starting from the sample objects and their formats, the system queries available registries of preservation actions and looks up applicable tools such as emulators of the original environment of migration tools that can handle the provided input format. The Planets registry moreover holds information on benchmark evaluation results produced by experiments carried out in the Planets Testbed, which provides a controlled environment for preservation experiments.

### Develop and run experiment

Preservation action tools that are accessible through a web service are directly invoked during the execution of experiments on the sample objects; other tools such as emulators have to be executed externally.

### Evaluate experiment

The evaluation of experiments is probably the most complex and, so far, least automated step in preservation planning. Until now, most of the judgment, e.g. if a migration tool accurately preserves the colour model of an image or the line breaks in a document, has to be carried out manually by looking at the rendered objects. However, **characterisation services** are available that can measure some of the essential characteristics of objects such as the dimensions of images. In contrast to characterisation tools like JHove, the extensible characterisation languages (XCL) [9,5] do not attempt to extract a set of characteristics from a file, but instead are able to express the complete informational content of a file in a format independent model.

Comparison services specify measurable properties as well as property-specific metrics and their implementation as algorithms in order to identify degrees of equality between two objects. This is in principle independent of the applied strategy, i.e. migration or emulation. The compared objects can be both the original and a migrated object, or the original object in two different environments. To allow comparison and evaluation, a mapping is created between the requirements specified in the objective tree and the characteristics that can be measured and compared automatically by the available characterisation tools. This mapping mechanism is part of the evaluation framework that is being developed in PP5. Where the planner has been using criteria stemming from the template library, the mapping will be already selected. It can however also be adapted and extended by the user.

As described in Section 3, the risk assessment service in the Planets characterisation framework addresses two categories of risks:

1. General risks of formats, such as complexity or lack of documentation, and
2. *Instance risks* that can apply to objects of a certain kind.

For example, Word documents with more than 1000 pages are much more difficult to preserve than documents that contain just a few pages. Similarly, compound documents with spreadsheets embedded in text documents present a higher challenge to migration tools.
The risk scores obtained by the services support the correct selection of sample objects and ensure the stratification of samples over the given set of objects to be preserved. Additionally, assessing the risks of original and migrated objects allows the comparison of risk scores and therefore assists in the evaluation of potential preservation actions.

Figure 4 shows a possible deployment of the distributed infrastructure, which can be changed and extended dynamically. The shown deployment consists of seven server instances; additional registries and services can be dynamically added and registered in the planning tool. The Planets server instance on the top-left side corresponds to the application server running the main deployment of Plato. The interoperability framework provides features such as a workflow

execution engine, a data registry based on a Java Content Repository (JCR) implementation, and services such as user management, Single-Sign-On, persistence, and logging.

Plato builds on these common services to provide a proactive decision support environment for preservation planning, integrating registries and services at the corresponding points in the preservation planning workflow.
The knowledge base holds reusable patterns and templates for requirements recurring in different planning situations, while the tree importer allows the planner to use free mind-mapping software for the requirements definition and import the results directly into the planning tool.

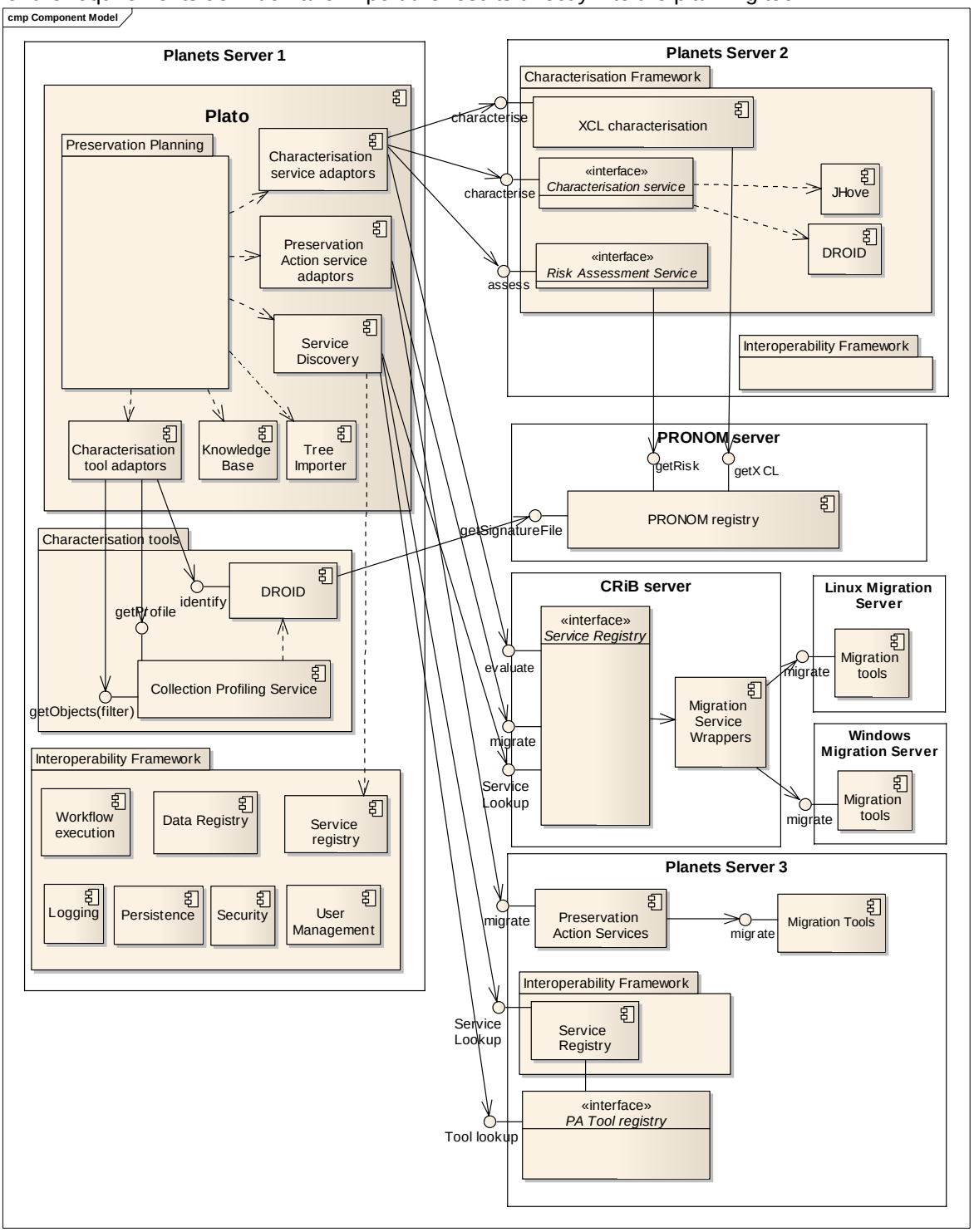


**Figure 4 A distributed preservation planning environment**

Characterisation tools deployed on the same server as the planning tool can be directly accessed locally through corresponding adaptors. Currently, DROID is being used for identifying the file

format of sample objects. The tool regularly downloads the current database of file signatures from the PRONOM server. Furthermore, the collection profile created by the corresponding service documents the characteristics of the set of objects for which preservation planning takes place and supports the selection of representative sample records.
Characterisation services can also be accessed remotely. The Planets Characterisation framework will include DROID together with other tools such as JHove to deliver detailed characterisation of digital objects. These detailed characteristics are then assessed by the risk assessment service.

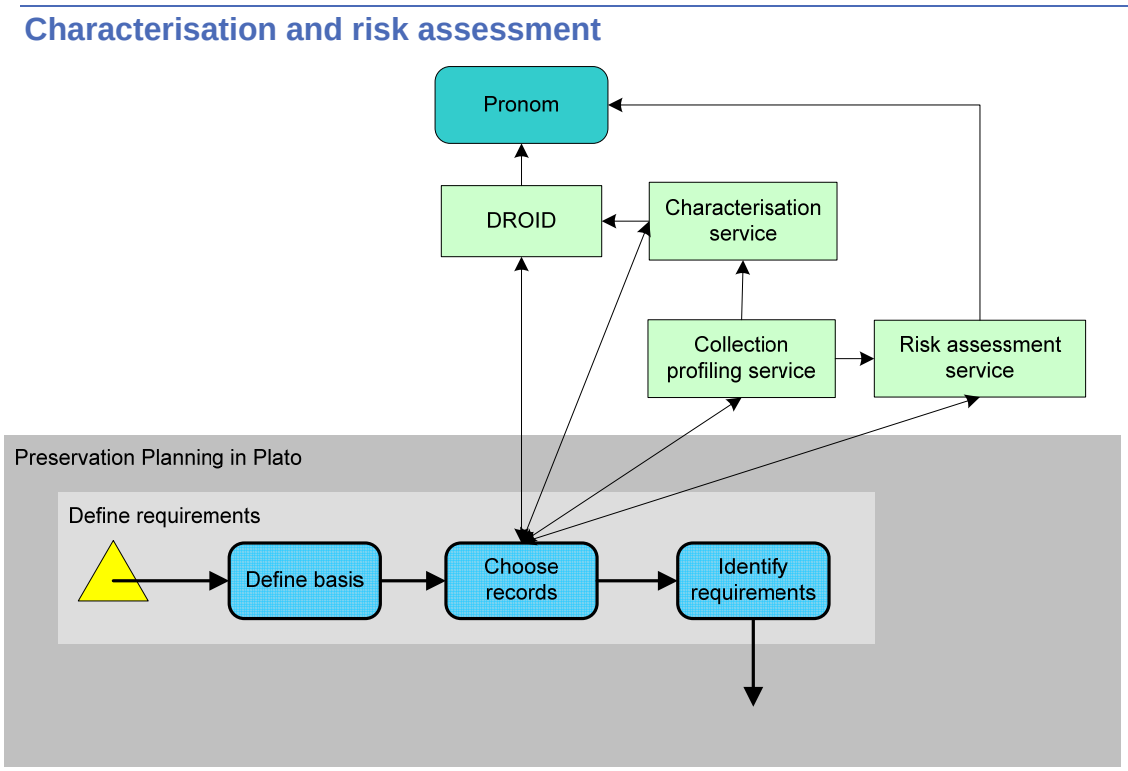


**Figure 5 Integrating registries and services for the specification of sample records**

During the first phase of requirements specification, sample objects are defined by the planner that cover the essential characteristics of the objects in question and are used for experimentation in the subsequent stages. DROID is used to identify the format of these objects; the Planets characterisation services can be used to further describe them and to assess risk scores for sample objects.
Furthermore, the XCL engine can hierarchically decompose sample objects and produce a representation in an abstract XML language, the eXtensible characterisation description language (XCDL). This allows the later comparison of migrated representations to the originals.

The output of the characterisation services is saved as an important evidence for the planning procedure. It can moreover inform the requirements definition process about significant properties of objects and potential risk factors that need to be addressed. The characterisation services may also be used to produce preservation metadata, and can ensure authenticity by highlighting which object properties have been preserved adequately and which not.

For large or very diverse collections, the selection of sample records that are truly representative for a set of objects, i.e. cover the range of features and essential properties present in the whole set, can be quite difficult.
Collection profiling services that automatically characterise and describe a set of objects can be of great value and both inform and in the future automate the process of selection.

Plato 3 will in the future be able to make use of the upcoming collection profiling service to analyse the collection profile and recommend suitable sample objects that are representative for the set of objects for which a preservation plan is being created.
It is envisioned that this takes place in two steps.

1. Plato calls the collection profiling service, which delivers a profile in XML format detailing the distribution of formats and characteristics.
2. Plato can then request the actual files that correspond to a set of filters, for example
   a. all files in PNG1.1 that contain transparency,
   b. all files with a risk score > 900,
   c. all documents in Word95 with more than 200 pages.

The collection profile then delivers the actual objects, including the bitstream.

## Preservation Action Discovery

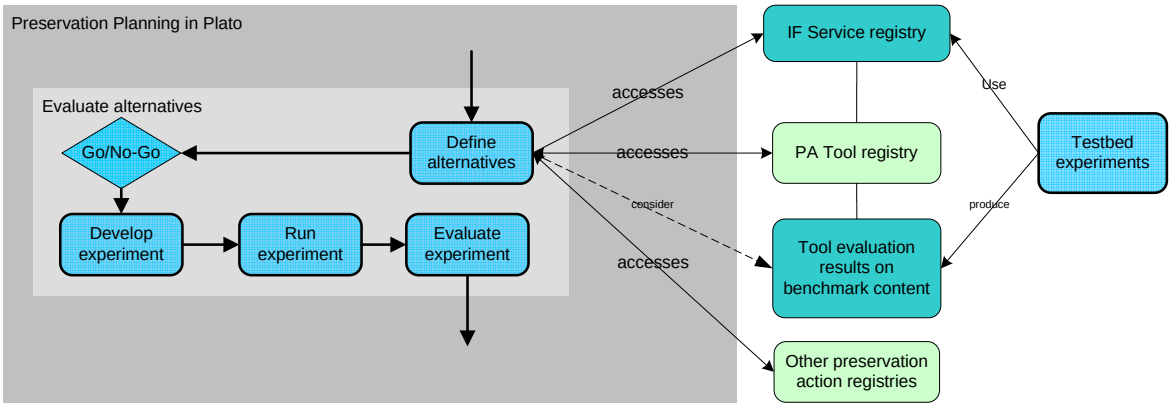


**Figure 6 Preservation Action Service discovery**

During the second phase of the planning process, preservation actions are evaluated in experiments. Tools that are accessible through Web services are executed within Plato; other tools such as emulators are executed externally by the planner. This phase starts by looking up potentially applicable actions in service registries. A query to a preservation action service registry such as the CRiB registry described in [10], using the PRONOM unique identifier obtained from the Format Identification Service as input parameter, yields a list of both atomic and chained migration services that can convert files from the input format to other more desirable preservation formats.

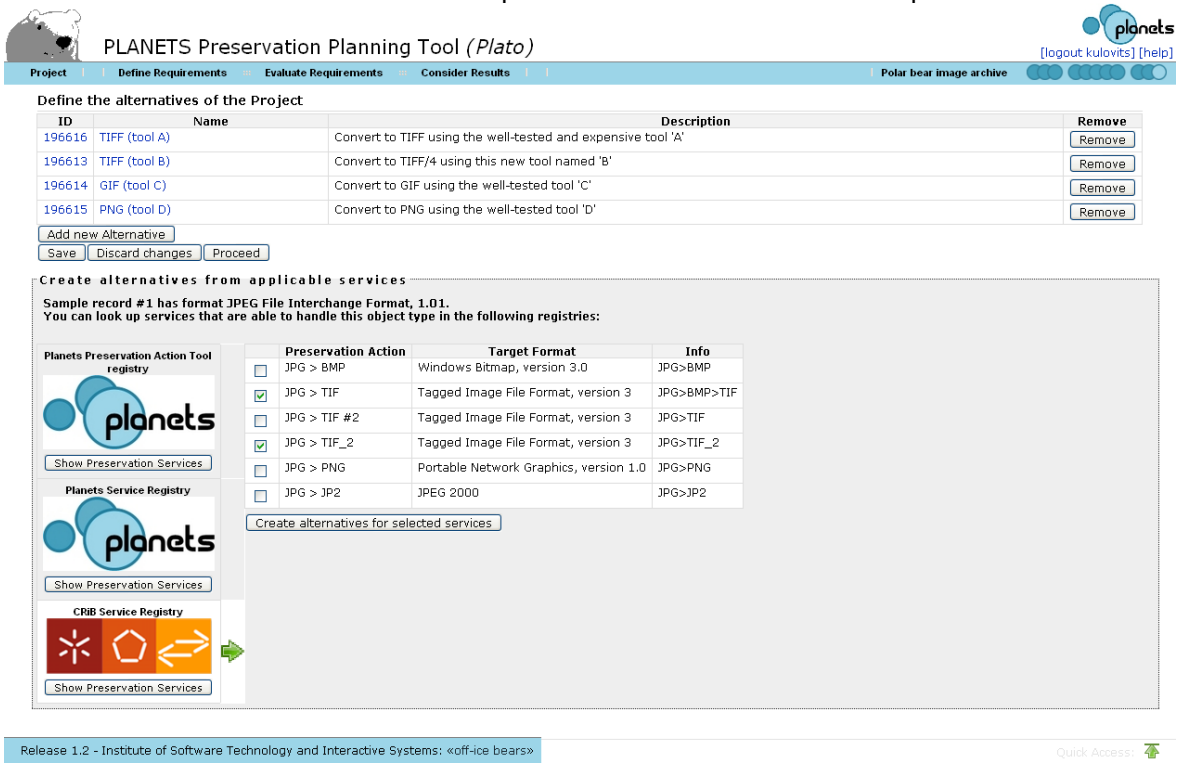


**Figure 7 Plato 1.2 querying applicable migration services for JPEG images**

Figure 7 shows Plato 1.2 querying the CRiB registry for migration services that can operate on JPEG images. Based on the returned list of applicable services, the preservation planner can select PA services of choice to include in the evaluation procedure.

Depending on the registry, different levels of information are present here that can aid this selection process.

The IF service registry contains services of different classes such as migration and characterisation services. It provides a taxonomy that allows filtering services according to their category – for example, *migration.* The Planets PA tool registry will hold information such as empirical evidence referencing experiments conducted on benchmark content within the Testbed environment and thus provide a rich foundation for suggesting suitable preservation services to the planner. How to use this information is being investigated within Planets PP/6 and described in deliverable PP6-D3.

In contrast to the IF service registry, the PA tool registry moreover contains not only information about tools accessible as web services, but also applications and tools such as emulators. The CRiB registry is a pure web service registry containing information about available migration pathways and migration services.

## Evaluation of preservation actions using characterisation services

After applying the selected preservation actions to the sample objects, the outcome is evaluated against the requirements defined in the objective tree. This can be assisted by characterisation and comparison services using the Planets characterisation framework, the XCL languages and the comparison services. To automate this process, a mapping is introduced between the requirements defined in the tree and the characteristics that can be extracted automatically from the objects.

Figure 8 shows a scenario for applying XCL in the context of format migration. After converting a document from ODF to PDF/A, the XCDL documents of the original and the transformed object can be compared using an interpretation software. A comparison tool ('Comparator') for XCDL documents is currently under development. Key objectives are the property-specific definition of metrics and their implementations as algorithms in order to identify degrees of equality between two XCDL documents. In its core functionality the comparator loads two XCDL documents, extracts the property sequences and compares them according to comparison metrics which are defined with respect to the types of the values in the value sets.

While the preservation planning approach implemented in Plato defines requirements on the significant properties of objects to preserved starting from high-level characteristics and breaking them down to measurable properties, the XCL language delivers a bottom-up characterisation of technical characteristics. The validation framework that is being developed in workpackage PP5 connects these two and thus allows a quantified evaluation of the quality of preservation actions with respect to object properties.

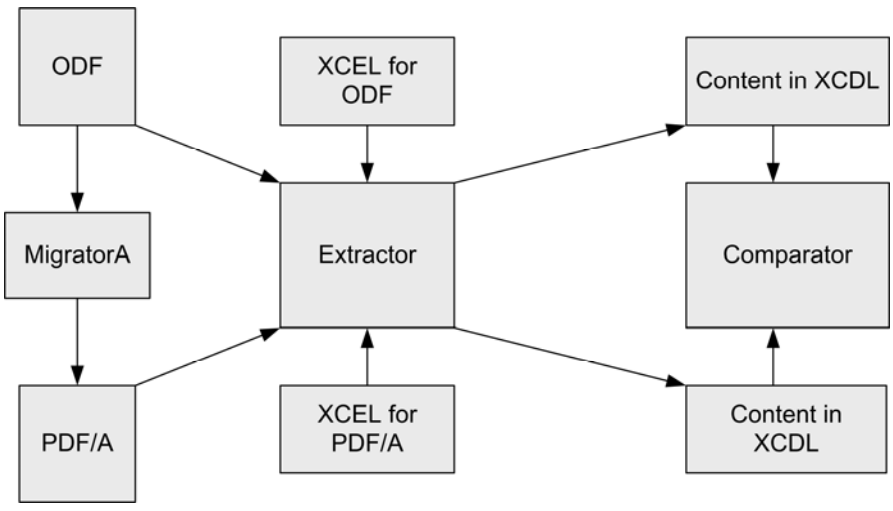


**Figure 8 Using XCL to compare migrated documents**

This means that within Plato, a mapping is being created between the requirements and the properties that can be measured by XCL and compared by the comparison service. Figure 9 depicts the current prototype of Plato 1.3 where the user can manually create this mapping. Future work in PP6 will investigate approaches to automatically select and recommend suitable mappings taking into account the XCL ontologies and suggesting appropriate measurement scales.

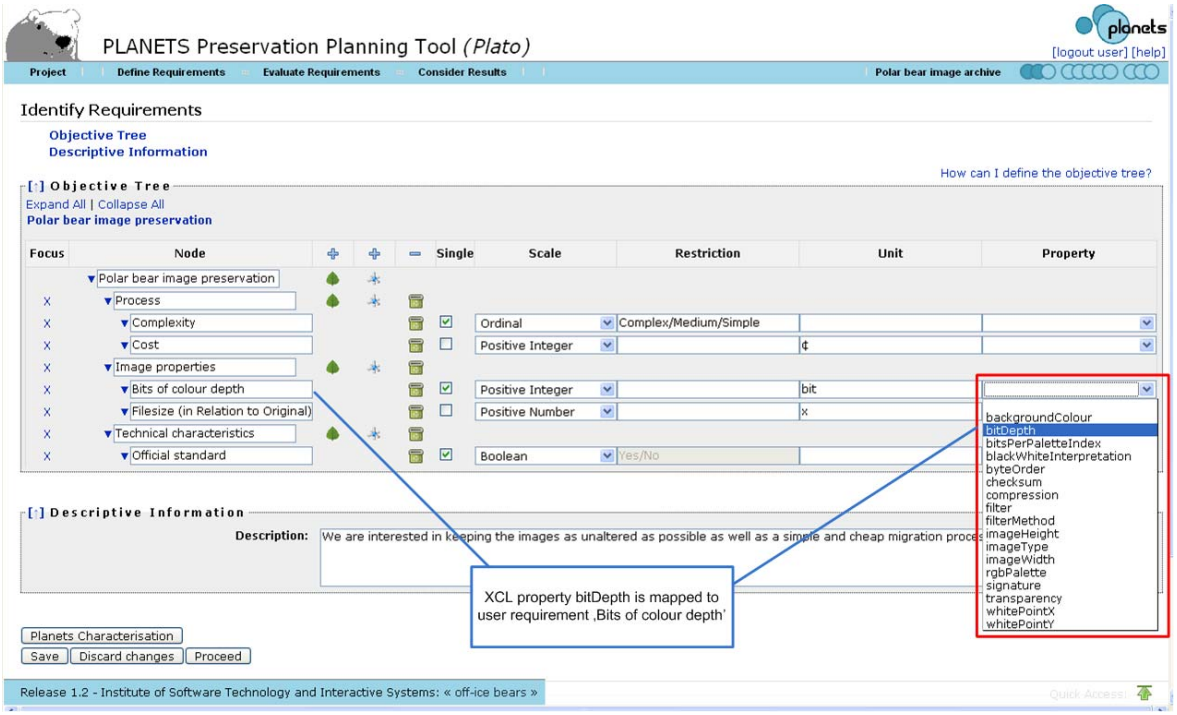


**Figure 9 Plato 1.3 mapping properties to requirements**

## 5. An example scenario

Consider a curator who is aiming to create a plan for preserving a collection of image files in various versions of the PNG format using a future version of the Planets system. We will accompany her on the various steps of preservation planning that she is going through, supported by the Planets system and using Plato as a preservation planning tool.

Assuming that she has already documented her institution's policy and usage using the models developed in the Planets work packages PP2 and PP3, she references these models during the definition of the basis that forms the first step of the preservation planning workflow.

In workflow step 2, she is able to run the collection profiling service and get detailed statistics on the distribution of images, the generic format risk of PNG 1.0, 1.1, and 1.2, potential risks that can be present in each of the versions, and actually occurring properties that

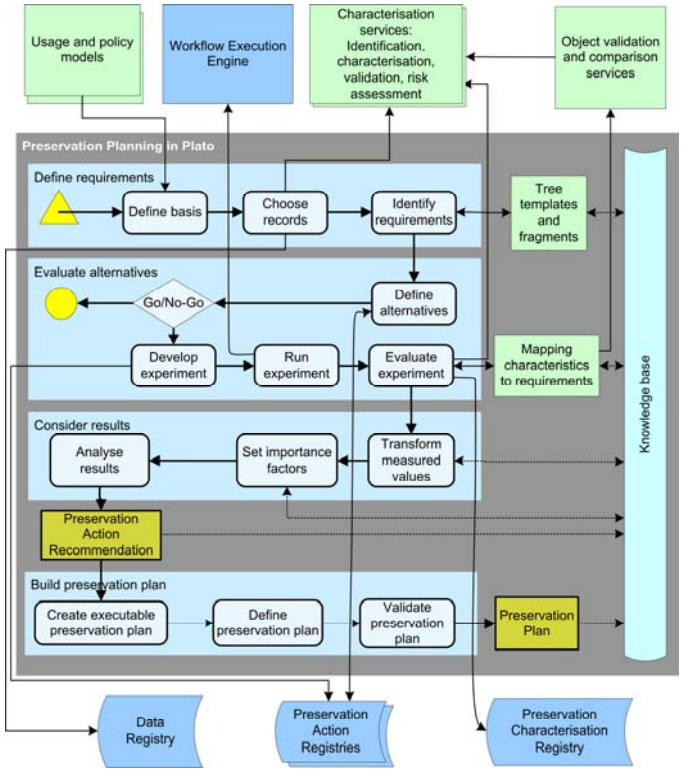


**Figure 10 PP environment**

present a risk – e.g., transparency could be present in several of the images.

She would then select sample objects that are representative of the collection at hand, based on this collection profile.

Following that, the curator would characterise the chosen sample objects using the PC framework and the XCL engine, documenting the characterisation results.

During requirements definition, for which she will strive to seek a broad participation from all stakeholders to have a complete coverage of influence factors and viewpoints, she maps requirements dealing with object properties to characteristics that can be extracted and compared automatically according to specified metrics.

Querying preservation action registries for applicably migration services for images as well as emulators that are able to simulate the original technical environment in which the images were created, she retrieves a list of potential preservation actions. She is able to narrow this list down to 5 alternatives that are, in her opinion, worth closer consideration. She chooses in this scenario to not further evaluate the emulator option because the experiment data available about the migration of PNG images, coming from large-scale experiments in the Planets Testbed, look very promising, and because the possibilities of automated QA for migration of images mean that in this case, the transformation seems to incur a low risk to authenticity.

During the second phase, evaluation of alternatives, she can automatically execute the preservation actions that are accessible through a web service. This is done within Plato by sending the sample objects to the service and retrieving the resulting files. There is one migration tool that she chooses to evaluate apart from the migration services offered online. She applies this tool manually and uploads the resulting objects.

Evaluation of the results is to a large extent carried out by automated validation and comparison through the XCL comparison service. Moreover, the target file format risks are retrieved from the PC registry.

She analyses the quality of alternatives supported by a visualisation as shown in Figure 11, being able to directly compare specific strengths and weaknesses of each alternative action.

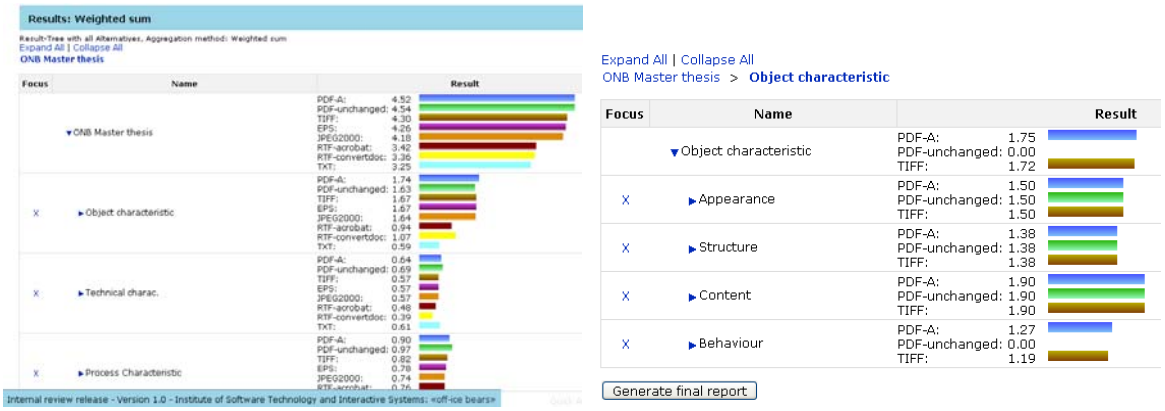


**Figure 11 Analysing results**

Based on this analysis, she opts for one migration service to be applied to all the objects at question. She documents this recommendation as the core decision of the preservation plan.

She is then in the fourth phase aided in creating an executable preservation plan that applies this preservation action to a set of objects, and is able to select QA services, i.e. comparison services, that are to be executed automatically on each migration step that is carried out on an object. These QA services conform to the subset of the requirements tree that can be measured automatically.

This executable preservation action plan is the core part of the preservation plan. Apart from that, she also identifies roles and responsibilities as well as specific events that would trigger a re-evaluation and possibly update of the plan.

# 6. Summary and Outlook

This document described the level of service integration into Plato 2 and 3. We shortly introduced the underlying workflow to provide the necessary background of information and then outlined the integration points we identified in the workflow where preservation action and characterisation services can be integrated.

The dependency roadmap shown in Figure 12 provides an overview of the ongoing and future development of the preservation planning methodology and workflow and its implementation in Plato, and relates this to other work going on within Planets. With respect to these dependencies, it is essential that the work team of Plato and the teams working on results that need to be integrated work together in ensuring both timely completion of products and seamless communication about intermediary results that can aid prototypical integration during the work on these products.

For example, the team working on the usage model which will be integrated in the planning workflow in the first step and subsequently influence aspects such as the requirements specification in step 3 is going to produce a preliminary model in time to incorporate this into Plato 2. Subsequent feedback will inform the specification of the final usage model that will be integrated in Plato 3. For the integration of services, a proper interface specification as provided for instance in [11] is necessary.

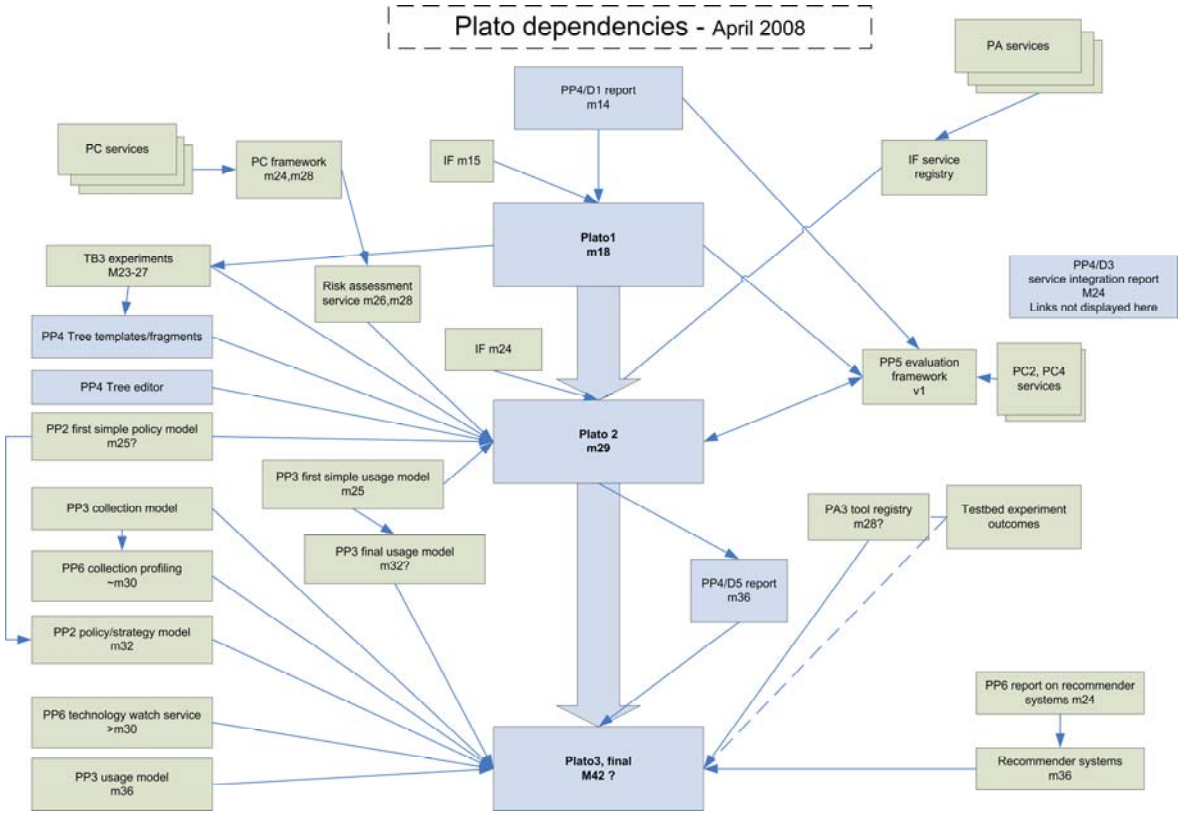


**Figure 12 PP4 dependencies**