

Distributed Preservation Services: Integrating Planning and Actions

Christoph Becker¹, Miguel Ferreira², Michael Kraxner¹, Andreas Rauber¹,
Ana Alice Baptista², and José Carlos Ramalho²

¹ Vienna University of Technology, Vienna, Austria

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

² University of Minho, Braga/Guimarães, Portugal

<http://www.dsi.uminho.pt>

Abstract. Digital preservation has turned into an active field of research. The most prominent approaches today are migration and emulation; especially considering migration, a range of working tools is available, each with specific strengths and weaknesses. The decision process on which actions to take to preserve a given set of digital objects for future access, i.e., preservation planning, is usually an ad-hoc procedure with little tool support and even less support for automation.

This paper presents the integration of tools and services for object migration and characterization through a service oriented architecture into a planning tool called Plato, thus creating a distributed and highly automated preservation planning environment.

1 Introduction

The longevity of digital objects used to be something taken for granted by many, until in the last decade several instances of spectacular data loss drew the public's attention to the fact that digital objects do not last forever. Last year, a survey among professional archivists underlined the growing awareness of the urgency of digital preservation [15].

This awareness has led to the development of various approaches that deal with the question of preserving digital objects over long periods of time. The number of tools that are available for preserving standard types of objects such as images or electronic documents is steadily increasing. However, choosing the right treatment for a given set of objects is a crucial decision that needs to be taken based on a profound and well-documented analysis of the requirements and the performance of the tools considered. Until now, this *preservation planning* process was a rather time-consuming ad-hoc procedure in which most steps had to be carried out manually.

In this paper we present a web-based planning tool called *Plato* that supports and automates the planning process underlying digital preservation endeavors. It implements an existent solid methodology for preservation planning[14] and integrates distributed services through an interoperability framework to provide

a proactive planning platform for distributed preservation activities. The conceptual structure and final set of services that Plato will provide are outlined in [2]. In this paper we demonstrate the existing flexible integration of services for preservation action and characterisation through web service registries. Examples are the existing migration services delivered by CRiB (‘Conversion and Recommendation of Digital Object Formats’) and the preservation action services deployed within the EU project Planets¹ (‘Preservation and Long-Term Access via Networked Services’).

We describe the workflow underlying the planning tool, depict the basic architecture of the systems integrated, and describe the technical approaches accomplishing the integration. We further provide a working example of format identification, discovery and invocation of preservation action services.

The remainder of this paper is organised as follows. Section 2 provides an overview of previous work in the area of digital preservation, object migration, and distributed preservation services, while Section 3 presents the architecture and components of the CRiB system. Section 4 introduces the Planets approach to preservation planning and outlines the architecture and features of the planning component Plato, which forms the basis of the work presented here. We then describe the integration of file format identification and object migration services through service discovery and invocation in Section 5. In Section 6 we draw conclusions and give a short outlook on future work.

2 Related Work

Two principal approaches to preservation actions have been pursued particularly intensively during the last years: migration and emulation. While emulation operates on the environment of the objects to make sure that everything that is needed for rendering (or performing) those objects is in place, migration transforms them to new representations that are considered to be safer or better suited for long-term preservation. In contrast to emulation, it is widely used in practice [12].

There is a range of tools available for converting standard content such as images and documents to target formats that are considered more stable. An important part of ongoing efforts in digital preservation is the advocacy of technologies for sustainable documents. The effects can be seen in standards such as PDF/A [10] or the Open Document Format (ODF) [11]. Still, there are risks that need to be considered. Transforming the digital representation of an intellectual object always risks damaging its content. Yet, the preservation of digital objects also requires that one is able to prove authenticity by documenting in which ways the preserved representation differs from the original one. Lawrence investigated risks in migrating objects to different file formats [13].

Various migration tools are available for standard file formats such as office documents; the selection is less wide for more exotic and complex compound objects. However, even within migration tools for e.g. office documents, the quality

¹ <http://www.planets-project.eu>

of conversion varies a lot. Some tools for example fail to preserve the structure of documents, while others might miss essential layout characteristics or discard important metadata. A thorough comparison of objects before and after transformation is necessary to ensure that the essential properties of the objects have been preserved correctly.

Clausen presents a mechanism for semi-automatic quality assurance to reduce this uncertainty in [5]. Within the Planets project, a family of generic XML languages for characterising digital objects to support digital preservation is being developed. The idea is to hierarchically decompose an object and thus represent objects from different format sources in an abstract XML language. This allows the automatic comparison and evaluation of preservation actions and thus supports automated quality assurance [3].

Several approaches to distributed service infrastructures for digital preservation have been developed. Hunter presents PANIC, a prototypical system for semantic web service composition based on ontologies and open web technologies [9]. PANIC captures and monitors preservation metadata to provide obsolescence notification and integrates existing software for emulation and migration.

The Planets project is creating a distributed service oriented architecture for digital preservation. Farquhar presents an overview of the distributed service infrastructure and the main components that form the Planets system [6].

Distributed preservation infrastructures such as these need registries that hold up-to date information on aspects such as object formats, available characterisation services, and applicable preservation action services. Format registries such as PRONOM[4] or the Global Digital Format Registry (GDFR) partly cover this need.

3 CRiB: Conversion and Recommendation of Digital Object Formats

CRiB is a Service Oriented Architecture (SOA) designed to assist institutions in the implementation of migration-based preservation interventions. It is publicly available² and described in detail in [7] and [8].

The general architecture of the CRiB system is depicted in Figure 1. The top layer illustrates Plato, an example of a client application that accesses the services provided by the CRiB. Other example of client applications might be: digital repository systems (e.g. DSpace, Fedora or Eprints) or custom applications developed by individuals. The following layers illustrate the whole set of components that constitute the CRiB system.

The **Format Identifier** is used to determine the underlying encoding of a digital object by using the PRONOM registry described above through the DROID tool for format identification.³

The **Service Registry** is responsible for managing information about existing conversion services. It stores metadata about its producer/developer

² <http://crib.dsi.uminho.pt>

³ <http://droid.sourceforge.net>

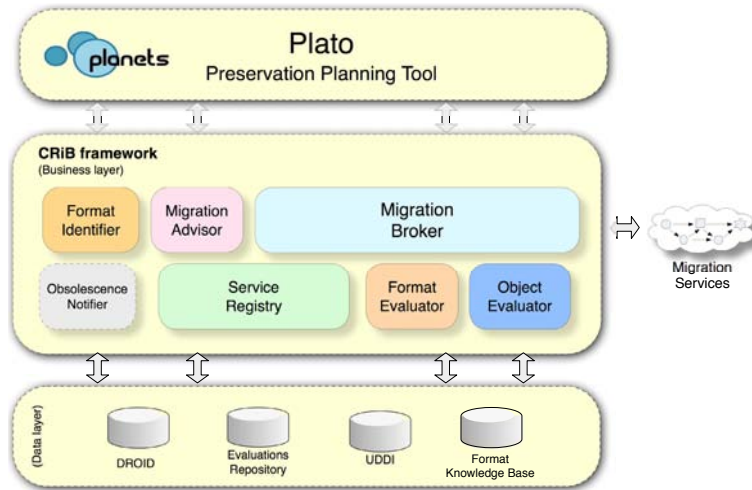


Fig. 1. The general architecture of the CRiB system with Plato as an exemplary client

(e.g. name, description and contact), about the service itself (e.g. name, description, the source/target formats, cost of invocation, etc.) and information on how the service should be invoked by a client application (i.e. its access point).

The **Migration Broker** is responsible for carrying out object migrations. In practice, this component takes care of all aspects of a composite conversion and makes sure that it is performed atomically from CRiB's point of view. Additionally, this component is responsible for recording the performance of each migration service. The results of these measurements are stored in the Evaluations Repository, a knowledge base that supports the recommendation system.

The **Format Evaluator** provides information about the current status of file formats. This enables the Migration Advisor to determine which formats are better candidates for preserving the original objects by looking at the characteristics of each pair of formats. The service is supported by a data store containing facts about formats (i.e. Format Knowledge Base), but could also exploit other external sources of information, such as PRONOM or Google Trends, to automatically determine a format's ubiquity or intrinsic characteristics.

The **Object Evaluator** is in charge of judging the quality of the migration outcome. It accomplishes this by comparing objects submitted to migration with its converted counterparts. Again, these evaluations will be performed according to multiple criteria. These criteria, also known as significant properties, constitute the set of attributes of an object that should be maintained intact during a preservation intervention. This is a fundamental piece in digital preservation as it brings quality control to migration-oriented preservation strategies.

The **Migration Advisor** produces recommendations of migration actions. In reality this component acts as a decision support centre for client institutions which helps them to determine the best possible migration option for their specific preservation problem. It accomplishes this by confronting the preservation

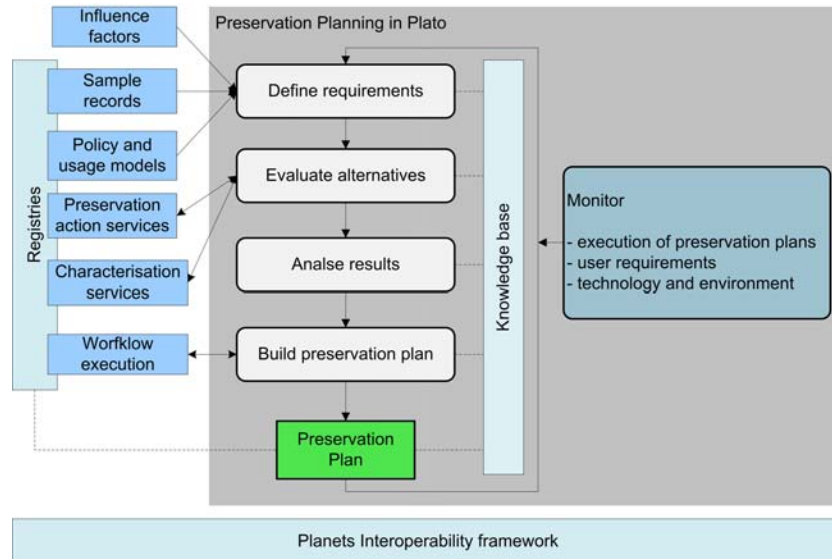


Fig. 2. Preservation planning environment

requirements outlined by client institutions with the accumulated knowledge about the performance of each accessible migration service (composite or not).

4 Plato: Preservation Planning

The Planets preservation planning workflow provides a solid methodology for building preservation plans by defining institution-specific requirements in a formal way and evaluating alternative strategies against the specified criteria. This is done by applying preservation actions to a representative set of digital objects in an experimental setting and evaluating the outcomes. The workflow as described in [14] consists of three phases: (1) Define requirements, (2) Evaluate potential preservation actions, and (3) Analyse results to arrive at a recommendation for a preservation action. Case studies validating the methodology have been described in [1,14]. The insights gained during these and other case studies have led to a recent refinement and extension of the process with a fourth phase in which an executable preservation plan is created, based on the well-documented recommendation that comes from the evaluation procedure.

Figure 2 illustrates the resulting four-phase workflow in the context of the preservation planning environment.

1. **Requirements definition** is the 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

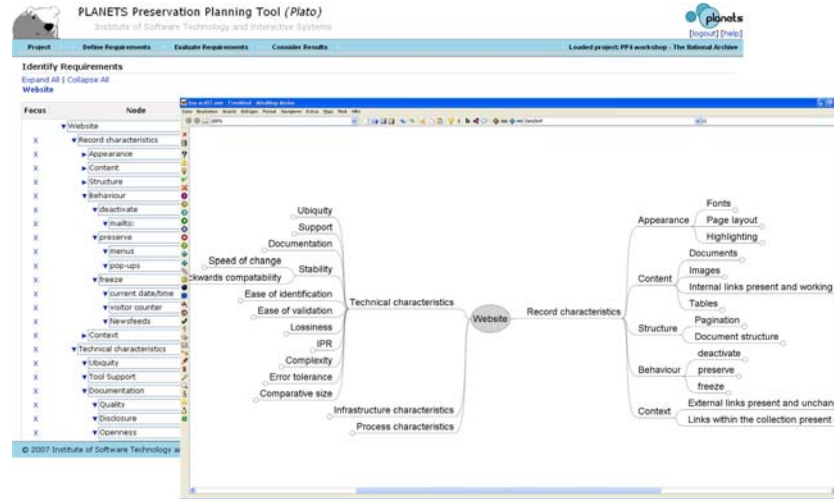


Fig. 3. Requirements definition in Freemind and Plato

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 [1,14]. As the trees are often created in workshop settings where mind-mapping can be very helpful, the software can directly import trees created in mind-mapping software⁴. Figure 3 shows the tree in both the mind-mapping software and imported into Plato. The objective tree forms the basis for evaluation in the subsequent stages.

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 visualisation of results assists preservation planners in ranking the alternatives and giving a well-documented recommendation for a preservation action to apply on the collection of objects that shall be preserved.

⁴ <http://freemind.sourceforge.net>

4. The 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 BPEL workflow accessing distributed services provided that the chosen preservation action and its deployment support it.

The **planning tool Plato** currently implements the three-stage workflow and includes services for preservation action and characterisation to automate the planning process. Plato is an enterprise Java web application relying on open frameworks such as Java Server Faces and AJAX for the user interface and Enterprise Java Beans for the business logic. It is integrated in the Planets Interoperability Framework that supports loose coupling of services and registries through standardised interfaces and provides common services such as user management, single-sign-on, security, and logging. Based on this technical foundation, Plato provides a highly interactive software environment that supports preservation planners and enables proactive preservation planning.

Service discovery and integration is a prime issue throughout the workflow, be it discovery of preservation actions that are applicable to the objects to be preserved, identification of object formats or the characterisation of objects before and after migration (or during emulation) to evaluate the effects of applying a specific preservation action to the sample objects. To this end, we are currently integrating a range of registries covering different needs.

The next section will describe the distributed planning environment that results from integrating both Planets components such as preservation characterisation and preservation action services, the technical registry PRONOM, and CRiB as an advanced and stable system performing format migrations.

5 A Distributed Preservation Planning System

Figure 4 shows a possible deployment of the distributed infrastructure. 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 Plato⁵. The interoperability framework provides features such as a workflow execution engine, a data registry based on

⁵ <http://www.ifs.tuwien.ac.at/dp/plato>

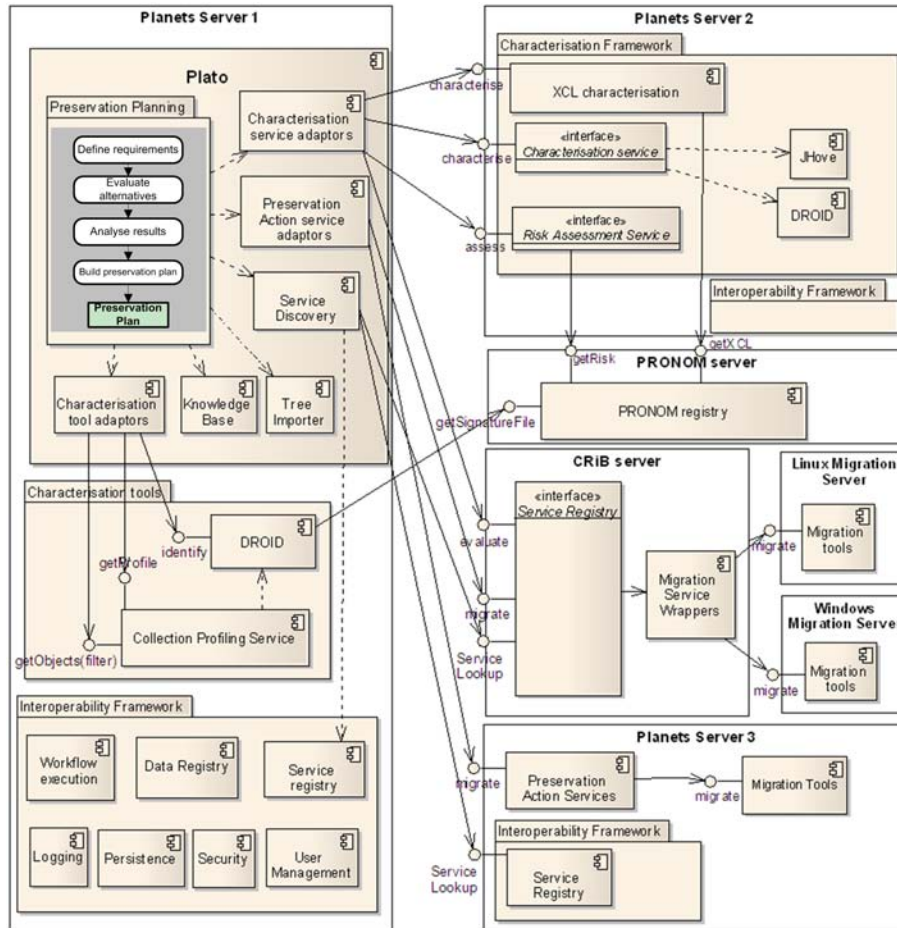


Fig. 4. A distributed preservation planning environment

a Java Content Repository (JCR)⁶ 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, as pictured in Figure 3.

Characterisation tools can be directly accessed 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

⁶ <http://jcp.org/aboutJava/communityprocess/final/jsr170/index.html>

from the PRONOM server. The Planets Characterisation framework will include DROID together with other tools such as JHove⁷ to deliver detailed characterisation and risk assessment of digital objects.

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. Alternatively, the identification service in CRiB can be used for this purpose, which is based on DROID as well. Furthermore, the XCL engine can hierarchically decompose sample objects and produce a representation in an abstract XML language, the eXtensible characterisation description language (XCDL) [3]. 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.

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) 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 short documents. 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.

During the second phase of the planning process, preservation actions are experimented by executing tools that are accessible through Web services; other tools such as emulators are executed externally. This phase starts by looking up potentially applicable actions in service registries. A query to CRiB using the PRONOM unique identifier obtained from the Format Identification Service yields a list of both atomic and chained migration services that can convert files from the input format to other more desirable preservation formats. Table 1 shows the 39 atomic services which are currently deployed. By composing these migration services, we get an overwhelming number of 7345 possible migration paths. For example, 147 atomic and chained migration services are available for migrating JPG images to 8 different file formats, as shown in Figure 5.

CRiB offers migration services for standard object types such as images and documents. To this end, it relies on open software such as ImageMagick and sam2p running on Unix, but also offers Windows-based migration services for office documents. The CRiB system itself is distributed across multiple servers

⁷ <http://hul.harvard.edu/jhove/>

Table 1. CRiB's list of atomic migration services

GIF2JP2	PNG2TIF	BMP2JP2	PDF2Text	PDF2TextLayout
MultipageTIF2JP2	DOC2PDF	BMP2JPG	TIF2GIF	TIF2PDF3
TIF2PDF2	TIF2PDF	TIF2JP2	ODT2DOC97	JPG2BMP
JP22TIF	PDF2MultipageTIF	PNG2JP2	PDF2JPG	JPG2PDF
ODT2DOC	JPG2MultipageTIF	JPG2TIF2	RTF2ODT	TIF2JPG
ODT2PDF	TIF2PNG	ODT2RTF	GIF2TIF	BMP2TIF
JPG2TIF	PDF2JP2	DOC2ODT	ODT2TEXT	JPG2PNG
TIF2BMP	PNG2JPG	JPG2JP2	PDF2TIF	

running the respective platforms. The service facade provides a unified interface to these services through the tool wrappers.

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 comparison services as well as the evaluation services offered by CRiB. 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 [3].

Web service integration is technically not always straightforward. Due to the inherent incompatibilities between different frameworks and environments, web

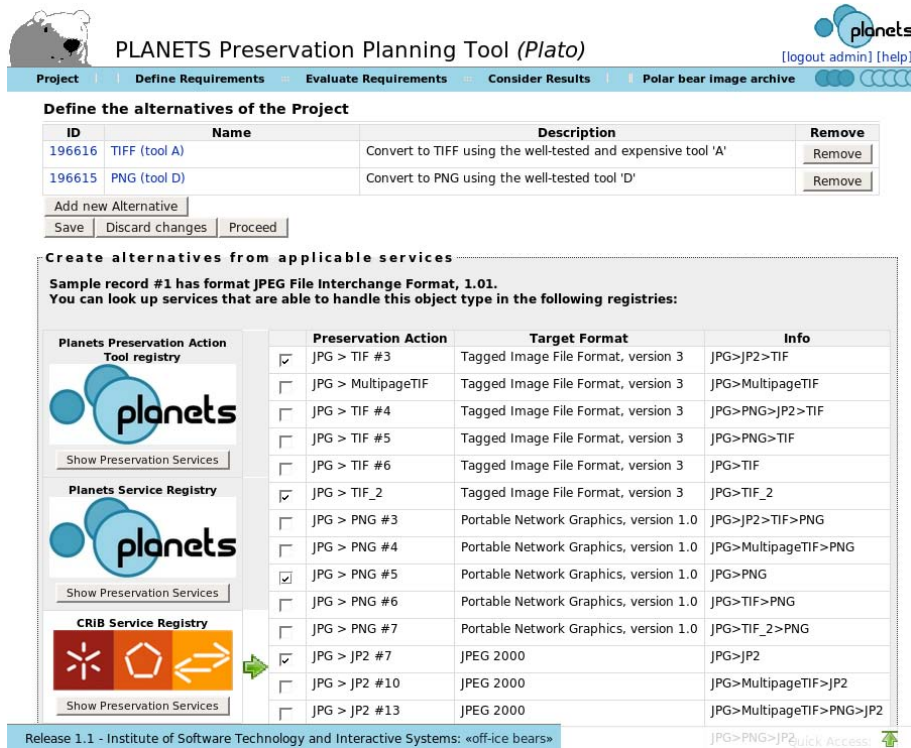


Fig. 5. Plato listing CRiB migration services for JPEG images

service adaptors might need to use specific implementations of the web service stack to access various services. For example, CRiB is using version 1.4 of the Apache Axis implementation of the web service stack, which uses *RPC/encoded* request transmission and does not properly support the currently recommended mode *document/literal*. The newer framework of Planets is based on the JBoss Application Server and thus uses the JBoss-WS web service implementation, which does not fully support the older *RPC/encoded* request transmission.⁸ A service adaptor is therefore needed that makes use of the Axis client to generate an *RPC/encoded* SOAP request compatible to CRiB.

6 Discussion and Outlook

This paper presented a distributed preservation planning environment that results from integrating existing registries and services with the decision support tool Plato that is being developed in the Planets project.

It also depicts the CRiB system, a Service Oriented Architecture that delivers preservation services such as format identification, quality assessment, automatic chaining of migration services, and a recommendation system that learns from every executed migration [7,8]. These systems collaborate to provide flexible service discovery and invocation. The resulting distributed planning environment allows the planner to query registries for preservation actions that are applicable to the given set of objects to be preserved, and directly invoke migration services through web services. It furthermore integrates advanced characterisation services for file format identification, validation, and risk assessment.

The first version of Plato is publicly accessible.⁹ It implements the workflow described in [14] and provides file format identification as well as service discovery and invocation of CRiB services as described above. An upcoming public release will include the fourth stage where an executable preservation plan is defined [2].

Ongoing efforts are dedicated towards integrating preservation action, characterisation and risk assessment services, building a validation framework for object migration based upon the eXtensible characterisation languages, and developing proactive decision support and technology watch functions.

While the system presentation in this paper focuses on migration, emulation is of course a valid alternative that can be evaluated with Plato. A case study comparing migration and emulation approaches for the preservation of computer games is being finalised. So far, emulators are executed manually in an external environment; service-based integration is currently under investigation.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789. CRiB has been

⁸ http://labs.jboss.com/jbossws/docs/jaxws_userguide-2.0/index.html

⁹ <http://www.ifs.tuwien.ac.at/dp/plato>

funded by the FCT (Fundação para a Ciência e a Tecnologia, Portugal) under the grant SFRH/BD/17334/2004.

References

1. Becker, C., Kolar, G., Kueng, J., Rauber, A.: Preserving interactive multimedia art: A case study in preservation planning. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 257–266. Springer, Heidelberg (2007)
2. Becker, C., Kulovits, H., Rauber, A., Hofman, H.: Plato: A service oriented decision support system for preservation planning. In: Proc. Joint Conf. Digital Libraries (JCDL 2008) (2008)
3. Becker, C., Rauber, A., Heydegger, V., Schnasse, J., Thaller, M.: A generic XML language for characterising objects to support digital preservation. In: Proc. 23rd Annual ACM Symposium on Applied Computing (SAC 2008), Fortaleza, vol. 1, pp. 402–406. ACM Press, New York (2008)
4. Brody, T., Carr, L., Hey, J.M.N., Brown, A., Hitchcock, S.: PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *Int. Journal of Digital Curation* 2(2), 3–19 (2007)
5. Clausen, L.: Opening schrödingers library: Semi-automatic QA reduces uncertainty in object transformation. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 186–197. Springer, Heidelberg (2007)
6. Farquhar, A., Hockx-Yu, H.: Planets: Integrated services for digital preservation. *Int. Journal of Digital Curation* 2(2), 88–99 (2007)
7. Ferreira, M., Baptista, A.A., Ramalho, J.C.: A foundation for automatic digital preservation. *Ariadne* 48 (July 2006)
8. Ferreira, M., Baptista, A.A., Ramalho, J.C.: An intelligent decision support system for digital preservation. *Int. Journal on Digital Libraries* 6(4), 295–304 (2007)
9. Hunter, J., Choudhury, S.: PANIC - an integrated approach to the preservation of complex digital objects using semantic web services. *Int. Journal on Digital Libraries: Special Issue on Complex Digital Objects* 6(2), 174–183 (2006)
10. ISO. Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A) ISO/CD 19005-1. International Standards Organization (2004)
11. ISO. Information technology - Open Document Format for Office Applications. International Standards Organization (2006)
12. Lee, K.H., Slattery, O., Lu, R., Tang, X., McCrary, V.: The state of the art and research in digital preservation. *Journal of Research of the National Institute of Standards and Technology* 107(1), 93–106 (2002)
13. Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., Kenney, A.R.: Risk management of digital information: A file format investigation. *CLIR Report* 93 (June 2000)
14. Strodl, S., Becker, C., Neumayer, R., Rauber, A.: How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In: Proc. Joint Conf. Digital Libraries (JCDL 2007), pp. 29–38 (June 2007)
15. The 100 Year Archive Task Force. The 100 year archive requirements survey (2007), http://www.snia.org/forums/dmf/programs/ltacsi/100_year/