

The Planets Testbed: Science for Digital Preservation

The preservation of digital objects requires specific software tools or services. These can be characterisation tools that abstract the essential characteristics of a digital object from a file, migration tools that convert digital objects to different formats, or emulation tools that render digital objects in their original context on a new infrastructure. Until recently digital preservation has been characterised by practices and processes that could best be described as more art and craft than science. The Planets Testbed provides a controlled environment where preservation tools can be tested and evaluated, and where experiment results can be empirically compared. This paper presents an overview of the Testbed application, an analysis of the experiment methodology and a description of the Testbed's web service approach.

by Brian Aitken, Petra Helwig, Andrew Jackson, Andrew Lindley, Eleonora Nicchiarelli, Seamus Ross

Introduction and research framework

Planets, Preservation and Long-term Access through NETworked Services, is a four-year project which began in June 2006. It is co-funded by the Planets Consortium and the European Union under the Sixth Framework Programme to address the core digital preservation challenges that libraries, archives and the digital preservation community are facing. The primary goal for Planets is to build practical services and tools to help ensure long-term access to our digital cultural and scientific assets. Planets is pursuing this goal with both a practical and an intellectual approach, and is addressing such urgent research issues as the establishment of an automated characterisation and validation framework for digital objects, the preservation of databases and the role of emulation in digital preservation. For more details and a list of Planets publications see the Planets website [1] and in particular the publications page [2].

As the information society and its knowledge economy continues to expand, the underlying information assets of our society and the processes that we deploy are susceptible to contextual, syntactical and semantic obsolescence. Over the past 20 years there have been an increasing number of attempts to address these challenges. These have generally been haphazard and poorly co-ordinated, have relied on weak engineering methods and have proved difficult to replicate or deploy at any level of scale.

Until recently digital preservation has been characterised by practices and processes that could best be described as more art and craft than science; studies such as the DELOS/NSF Research Agenda on Digital Preservation [3, 4] and the DigitalPreservationEurope Research Agenda [5] note that frameworks for *experimentation* must be central to the design and practice of digital preservation research. Researchers such as Gladney [6], Ross [7], Thibodeau [8], and most recently Watry [9] have all argued that digital preservation investigations need to be more engineering focused. Following on from work done by the National Archives of the Netherlands [10] and the Digital Preservation Cluster of the DELOS Network of Excellence in Digital Libraries [11], Planets identified that a key step in taking this agenda forward should be the development of a testbed framework to support the assessment of preservation approaches and tools.

In order to perform research that has a scientific grounding, the digital preservation community needs to evaluate preservation approaches within a controlled environment. This environment must be able to simulate diverse real-life settings whilst avoiding duplication of work and maximising the use of invested resources. These are some of the issues that the Planets Testbed addresses. The Testbed provides a dedicated research environment that allows the systematic execution of experiments by distributed actors, enabling the automated evaluation of experiment results, the reproducibility of experiments, the long-term availability of structured experiment documentation and shared access to the experiments themselves.

The preservation of digital objects requires specific software tools or services. These can be characterisation tools that abstract the essential characteristics of a digital object from a file, migration tools that convert digital objects to different formats, or emulation tools that render digital objects in their original context on a new infrastructure. If we can gain an understanding about which tools best serve our needs we can compile preservation plans that indicate which of these preservation tools should be applied to a collection of digital objects under which circumstances.

The goal of the Planets Testbed (consisting of a software application and a research framework) is to extend the already existing information on preservation tools with new information based on the outcomes of practical experimentation. This extended knowledge will enable digital preservation practitioners to make sensible, informed decisions regarding the applicability of the tools in various preservation settings. Experiments can be made visible to other Testbed users and be exported to publicly available registries, enabling the results from previous experiments to shape a common understanding of the best preservation approaches for specific types of data in specific situations. The Testbed also provides the mechanisms

to enable users to repeat existing experiments in order to validate the available results.

The testbed concept is not a new one in the field of digital libraries; the development of testbeds was already a key component of the US Digital Library Initiative (DLI) which led to the development of the D-Lib Test Suite [12]. Other, more recent digital library initiatives include the Open Video Digital Library [13]; see the DELOS Framework for Testbed for Digital Preservation Experiments [11] for a detailed introduction to this topic.

In a report prepared at the request of the Dutch Government in 1999, the digital preservation expert Jeff Rothenberg recommended the establishment of a testbed project as part of a broader series of initiatives to address digital preservation ([10], [14]). As a result, the Dutch digital preservation Testbed was founded in October 2000. This research project tested the practical applicability of approaches to the preservation of governmental and other digital information. The National Archives of the Netherlands have utilised the Dutch Digital Preservation Testbed since 2003 to ensure that they select and implement the most appropriate and effective tools to preserve their electronic records over time. [10].

Following on from this the DELOS Digital Preservation Testbed [11] built upon the experiences of the Dutch digital preservation Testbed. The DELOS approach was to propose the abstraction of an operational retrieval environment that provides the means for researchers to explore the relative benefits of different preservation strategies in a laboratory setting. While at least one of the D-Lib Test Suite collections was used to test the transformation of varied SGML formats into well-formed XML [15], the Dutch and DELOS testbeds were the first serious attempts to develop a generic framework for the evaluation of digital preservation strategies.

The Planets Testbed was in turn inspired by the work undertaken by the Dutch and DELOS testbeds. However, the scope of the Planets Testbed is considerably broader, both in terms of the available services and the intended user community. The Planets Testbed's web-service approach enables a large number of users to perform experiments on a wide variety of services with a focus on the workflow-based design of the experiments. Consultations with the eight national memory institutions that are partners in Planets has led to a number of refinements to the experiment methodology of the Dutch and DELOS testbeds, resulting in a simplification of the experiment process. Testbed users can execute single migration and characterisation experiments, but also load pre-existing or create new service workflows. A typical Testbed workflow-based experiment could involve the following sequence of steps:

1. invoking a characterisation service on the input data to determine their significant properties and appropriate migration tools;
2. subsequently invoking a migration service for the execution of data migration;
3. finally, invoking a second characterisation service to automatically assess the results of the migration.

Another typical scenario is that of large-scale experiments (thousands of files or files of extremely large size) which evaluate the non-functional characteristics of tools and services. This will be made possible by the establishment of the Testbed corpora of test objects, as described below.

The Current Functionality of the Testbed

The core functionality of the Testbed enables a user to design, execute and evaluate an experiment on certain data with certain services using the web-based front-end of the Testbed application. In general, every experiment follows a fixed experimentation workflow that consists of six steps, as Figure 1 below demonstrates.

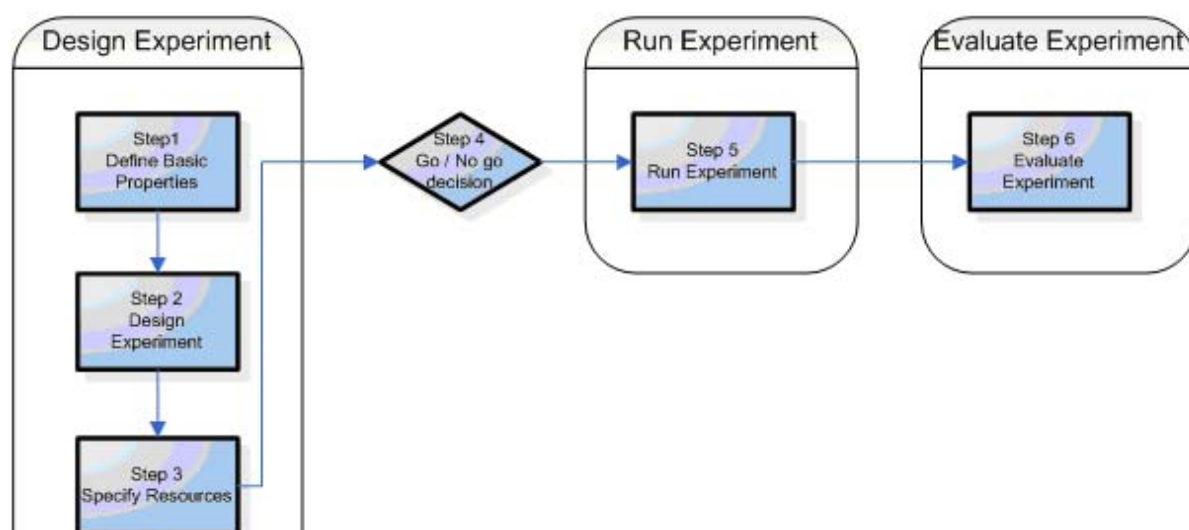


Figure 1: The Testbed Experiment Workflow

In the first three steps, the Testbed user designs the experiment, specifying the focus and intended goals of the experiment and selecting the services and data on which the experiment will execute. The user also specifies how the outcomes of the experiment will be evaluated. Following on from this a decision is made about whether experiment execution should proceed or not. The experiment is then carried out in step 5 and is then evaluated in step 6. For each of these steps there is a corresponding webpage in the Testbed application that enables the experimenter to supply the required information. The user's current position in the workflow and access to all previous stages is available in a panel on the left of the page, which allows the user to return and edit existing information if necessary. In the following section each of the experiment steps are described in more detail.

Step 1 - Define Basic Properties

In the first step, the basic properties for the experiment are defined. In addition to supplying general information such as an experiment name and a brief description, this stage enables the user to identify the purpose of the experiment, the pertinent research questions and references to any relevant research and literature. The experimenter must also indicate the type of experiment and the types of digital object the experiment will focus upon.

PLANETS Testbed - New Experiment - 1. Define Basic Properties - Mozilla Firefox

http://testbed.hatit.arts.gla.ac.uk:8080/testbed/exp_stage1.faces

eleonora Edit Profile Logout

Home New Experiment Import Experiment My Experiments Browse Experiments Browse Services Browse Data Help

PLANETS Testbed - New Experiment - 1. Define Basic Properties

EXPERIMENT PROGRESS

- 1. Define Basic Properties
- 2. Design Experiment
- 3. Specify Outcomes
- 4. Experiment Approval
- 5. Run Experiment
- 6. Evaluate Experiment

Create a new experiment using the form below. Once you have supplied the required information and have submitted the form the experiment will be added to your list of experiments as found in the My Experiments section.

General Information

Experiment Name: Migration Experiment #1

Summary: This experiment will test the migration of Doc to PDF files.

Participants: experimenter

Main Experiment: ☒

Contact Information

Contact Name: A Testbed Experimenter

Contact Email: testbed-experimenter@planets-project.eu

Contact Tel:

Contact Address:

References

External Reference ID:

Experiment Reference:

Fertig

Figure 2: The Testbed Application showing Step 1 of the Experiment Process

Step 2 - Design Experiment

For step 2 the specific services on which the experiment will focus are selected and the input files that will be accessed by these services can be uploaded or selected. The list of services will include any characterisation, migration and emulation tools that have been wrapped as web services and have been made available within the Testbed, as discussed in a following section. The input files for the experiment may consist of new files that the experimenter uploads, existing files that are stored in the Testbed's data registry or files taken from the digital preservation corpora.

PLANETS Testbed - Migration Experiment #1 : Stage 2: Design Experiment - Mozilla Firefox

http://localhost:8080/testbed/exp_stage2.faces

eleonora Edit Profile Logout

Home New Experiment Import Experiment My Experiments Browse Experiments Browse Services Browse Data Help

PLANETS Testbed - Migration Experiment #1 : Stage 2: Design Experiment

EXPERIMENT PROGRESS

- 1. Define Basic Properties
- 2. Design Experiment
- 3. Specify Outcomes
- 4. Experiment Approval
- 5. Run Experiment
- 6. Evaluate Experiment

Selected Experiment Time

Step 2: Specify Service and Operation

Please select from the list of available service that have been registered by the Testbed Administrator. You can restrict the query by selecting predefined tags and values.

Available: simple migration

Services for type:

Selected Service Name: Tiff2JpegActionService

Selected Operation Name: convertFile

display all services OR [restrict query by tags](#)

- Service Description:
- Operation Description:
- max. supported files: 1

Select **Proceed**

Figure 3: Step 2 of the Experiment Process

Step 3 - Specify Outcomes

In the third step, a list of possible evaluation criteria is presented to the experimenter. The contents of this list are dependant on the types of object and experiment that have previously been chosen and the user can make a selection by ticking the corresponding criterion. The selections made here will define the evaluation options that are available in the final step of the experiment process.

Step 3: Specify Outcomes

Supply information about the estimated resources your experiment will require and the objectives you expect your experiment to meet.

Select Benchmark Goals

Here you can specify the Benchmark Goals for your experiment. These will be used during the evaluation of your experiment to provide a record of the goals of your experiment and whether these goals were achieved.

Criteria	Explanation
<input type="checkbox"/> Footer font colour	Identifies the font colour of footers. The font colour of footers. If there are several instances of footers with different values, choose one to focus on. (for first version) "appearance"
<input type="checkbox"/> Header line spacing	Identifies the line spacing of headers. The spacing between two lines of headers. If there are several instances of headers with different values, choose one to focus on. (for first version) "appearance"
<input type="checkbox"/> Machine readability	Indicates which percentage of the characters of the text are machine readable. Define this value with a reference application, e.g. an OCR application. 0% means nothing readable, 100% means every character readable.
<input type="checkbox"/> Page height	Total height of the page. "appearance"
<input type="checkbox"/> Footer spacing-left	Identifies the space from left of Footer to element at the left. Space from left of footer to element at the left. If there are several instances of footers with different values, choose one to focus on. (for first version) "appearance"
<input type="checkbox"/> Stereo	Describes whether files are opened with one or two sound channels yes for stereo, no for mono.
<input type="checkbox"/> Paragraph spacing-top	Identifies the space from top of paragraph to element above. Space from top of paragraph to element above. If there are several instances of "plain text" with different values, choose one to focus on. (for first version) "appearance"
<input type="checkbox"/> Human readability	Indicates how well the text is readable. Grade how well the text is readable. 1 is very bad, 10 is excellent.
<input type="checkbox"/> Paragraph font colour	Identifies the font colour of plain text. The font colour of plain text. If there are several instances of "plain text" with different values, choose one to focus on. (for first version) "appearance"
<input type="checkbox"/> Header spacing-bottom	Identifies the space from bottom of header to element below. Space from right of header to element below. If there are several instances of headers with different values, choose one to focus on. (for first version) "appearance"
<input type="checkbox"/> Page width	Total width of the page. "appearance"
<input type="checkbox"/> Footer spacing-top	Identifies the space from top of footer to element above. Space from top of footer to element above. If there are several instances of footers with

Select **Proceed**

Figure 4: Step 3 of the Experiment Process

Step 4 - Experiment Approval

After the first three steps a decision whether or not to proceed with the execution of the experiment is taken. This decision is initially made automatically by the Testbed application; if the input data consists of a handful of files, the estimated intensity of the experiment is low or there are no currently executing experiments the application may automatically approve the experiment, in which case the user is informed that s/he may proceed to step 5. If this is not the case then the Testbed administrator will need to manually approve and schedule the execution of the experiment. If this manual decision results in

approval the user is then informed and may proceed with the next step.

Step 5 - Run Experiment

In step 5 the experiment is executed. An experiment may take a considerable amount of time to execute depending on the number and size of the input files and the service that is called. During execution the webpage for this step will display run-time statistics about the experiment.

In the current version of the Testbed experiments are automatically executed following approval. However, as concurrent execution of multiple large-scale experiments will influence run-time and performance related benchmarks, additional scheduling and load balancing facilities will be added to subsequent Testbed versions.

Step 6 - Experiment Evaluation

After execution has completed the experimenter is presented with the results and may then evaluate the experiment. The evaluation criteria that were selected in step 3 are presented again, enabling the user to manually evaluate each criterion. For each criterion the user can enter values for the input and output files and may then assign a value for the evaluation based on the similarity of the input and output values. For example, if the value of "header spacing-left" was 1.5 cm in an input document stored as a Microsoft Word file and, after migration to PDF, the value of "header spacing-left" in the output file is 1.4 cm, the user can enter these values in the corresponding fields and rate the evaluation of this criterion. For more information on experiment evaluation please see below.

PLANETS Testbed - Migration Experiment #1: Stage 6: Evaluate Experiment

View the experiment's input and output data, evaluate the outcomes of your experiment against the benchmark goals you specified during Stage 3 of the experiment and supply a report about your experiment. You may save and re-edit your evaluation as often as required.

Experiment Data

Experiment Input Data	Experiment Output Data
19081231_0001.tif	http://localhost:8080/planets-testbed/outputdata/daec946-8e4d-4c11-b7f9-d24f1597d074.jpg

Benchmark Goals

Below are the benchmark goals you specified during stage three of the experiment design. Evaluate the success of your experiment by selecting the level of success of each goal.

Criteria	Explanation	Evaluation	Source Value	Target Value
Resolution	The granularity of the picture <i>p.p.</i>	bad		
Size of smallest detail	Defines the smallest detail that can be distinguished in an image. <i>Can be resolution in a jpeg or .bmp image, but different for a vector image.</i>	bad	Float	Float
Page height	Total height of the page. <i>"appearance"</i>	bad	Float	Float
Page width	Total width of the page. <i>"appearance"</i>	bad	Float	Float

Experimental Report

Title:

Report:

Fertig

Figure 5: Step 6 of the Experiment Process

Once the experimenter has completed step 6 the experiment process draws to a close and the results can be considered by other users, who are given the option of posting comments about the experiment. The results from the experiments are shared with other Planets applications, enabling the output from Testbed experiments to be used by decision support tools for the compilation of preservation plans. Based on the Testbed results, recommendations for refinements and enhancements for future experiments can be made and further experiments can be proposed.

Evaluation

A digital object has syntactic and semantic properties or characteristics that can relate to content, context, appearance, structure or behaviour. A specific digital object type, such as Text, Image, Sound or Video, can be stored in many different file formats. The question when migrating a digital object from one file format to another is to what extent the intellectual characteristics of the object remain untouched. It should be possible to ask questions such as: How well is the font size preserved if we use tool X to migrate a text from a Microsoft Word format to a PDF format?

For each digital object type the Testbed contains a list of intellectual characteristics, or properties, which may be of interest to

the experimenter. For example, properties for “text” include the number of pages, font size and background colour. These properties can then be used to assess the functional performance of the services.

In characterisation experiments, experimenters may test whether the characterisation tool *characterises* these properties correctly: to what extent are the values of these properties in test objects, as characterised by the characterisation tool, the same as the “actual” values.

In migration experiments, experimenters may test if these properties remain untouched when using the migration tool: to what extent are the properties of the migrated digital objects the same as the values of the original objects when using the migration tool.

In emulation experiments, experimenters may test to what extent the values of the content, context, appearance, structure and behaviour of digital objects in the environment emulated by the emulation tool are the same as the values of the objects in their original environment.

Using the intellectual characteristics as evaluation criteria in this way it is possible to build up an overall picture of a tool's performance over the course of several individual experiments. These results can be aggregated to give average information about the performance of tools on various types of digital objects and the more experiments that are carried out, the more trustworthy this kind of experimentally gained information will prove to be.

Within the Testbed it is possible to both manually and automatically evaluate an experiment based on its experiment criteria. At the time of writing the automatic evaluation of experiments is currently being researched. In order to enable the automated extraction and evaluation of digital object characteristics the Testbed will make use of the eXtensible Characterisation Definition Language (XCDL) and eXtensible Characterisation Extraction Language (XCEL) [16], which are being developed by another part of the Planets project.

Experiment Search and Retrieval

A major goal of the Testbed and its research framework is to ensure that the results of the experiments are comparable, retrievable and documented so that they can be repeated. The Testbed's database of existing experiments is a research tool of immense value to preservation officers and it is imperative that users can effectively search the experiments database in order to gain answers to their research questions. The Testbed's search facilities provide the necessary depth to gain answers to such questions as:

- Which image migration tools are available that are known to preserve the colour correctly?
- Which image migration tool is considered the best at preserving sharpness when migrating from TIFF to JPEG2000?
- Is there a characterisation tool that can automatically extract the values for font family and variant from Word 2003 files?
- Which tools are available for extracting textual information from an image representing a page of a printed book in uncial script?
- Is there any tool that saves correctly undisclosed-format information (as maker notes) contained in JPEG EXIF tags? On which examples has it been benchmarked?

Testbed users may explore the database of completed experiments, comment on the design, methodology and outcomes of experiments and generate or browse knowledge trees based on experiment references.

The Architecture of the Testbed

In designing the Testbed some key principles shaped the development of the software. Firstly, the Testbed application was designed to be platform independent, robust and scalable and for this reason development was undertaken using Java Enterprise Edition. Secondly, at the outset it was agreed that the Testbed had to be able to execute experiments on as wide an array of preservation tools and services as possible, including legacy tools and tools that may not run natively on the platform on which the Testbed itself is installed. A web service approach was considered to be the most flexible way of meeting this challenge; tools are wrapped as web services which can then be accessed by the Testbed application by means of a platform-neutral URI. Thirdly, the Testbed application was designed to be a part of the overall Planets software suite rather than as a standalone entity. A different Planets sub-project, the Interoperability Framework, was tasked with providing many of the registries and components, such as the application server, storage systems and user management, which would be required by many of the various Planets applications. This approach enabled the Testbed developers to focus primarily on the components that were unique to the Testbed. This sharing of common components across the entire project also makes it easier for the various Planets applications to communicate with one another, allowing (for example) the results from Testbed experiments on preservation tools to be aggregated within the appropriate registry and then fed into the preservation planning components.

The design and development of the Testbed began in September 2006 and since then the developers have followed a traditional structured and iterative methodology for defining the system's functional range and the implementation of the software. This methodology, which is based on the Rational Unified Process, included interviews, use case descriptions, a workflow checklist for functional and non-functional requirements, a detailed architectural system design document and several phases of development during which the design was refined and aspects of the system were enhanced.

The Planets Testbed application was built around the JBoss Java Enterprise Edition 1.4 certified application server [17], and the Testbed utilises and extends various features that JBoss offers, such as messaging and transaction support. The application was designed and developed using the Model-View-Controller (MVC) architectural pattern, an approach which ensures that the business logic of an application is isolated from its user interface. By adopting this approach the developers were able to successfully divide the Testbed development, enabling certain developers to focus on the complex business processes that the system required whilst other developers focused on the ways these processes could be best represented to users in the web-based front-end.

Within the Testbed architecture the 'View' layer is realised as a web-based user interface, using the Java Server Faces (JSF) standard [18] as the core technology and outputting valid XHTML / CSS web pages. The 'Controller' layer, which processes all of the business logic, was developed using Java Enterprise Edition (EE), and many of the enhancements offered through EE, such as Enterprise Java Beans and Session Beans, have been utilised. The 'Model' layer for the Testbed has primarily been provided by the Planets Interoperability Framework and consists of a number of registries shared by the Planets applications where data and services are stored. Figure 6 below demonstrates how the Testbed MVC approach interoperates.

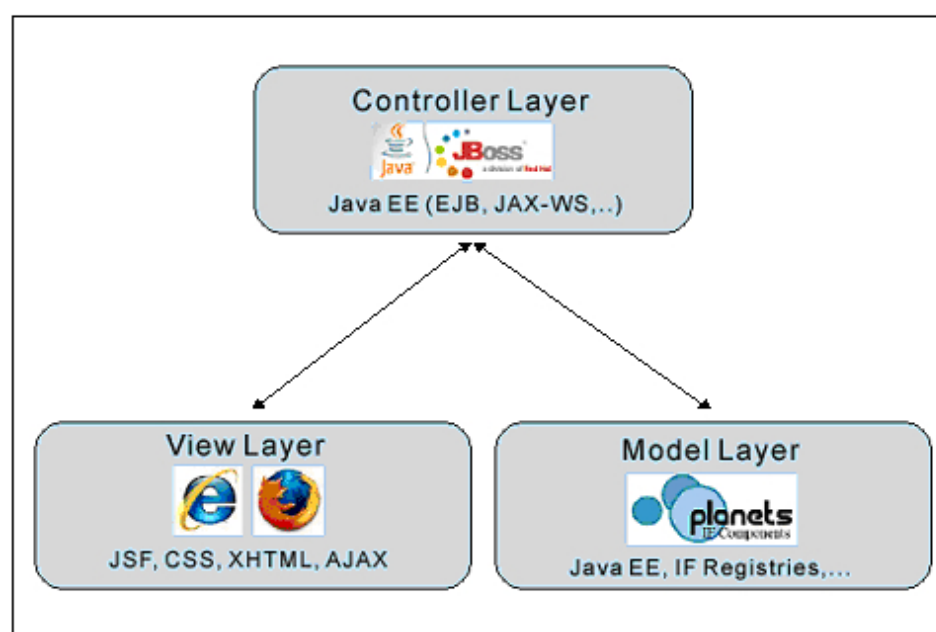


Figure 6: The Testbed Architecture

Built on Planets

Overall, Planets aims to build software components for digital preservation following a service-oriented architecture, as illustrated in Figure 7 below. This approach allows complex preservation workflows to be built up as chains of simpler preservation services. For example, a migration workflow might start by retrieving a digital object from a repository via a data-access service. This object may then be passed through one or more services in order to characterise it, e.g. to determine the file format. Based on this characterisation, the object may then be migrated to a new format and after this the new object may also be characterised to see how closely it retains the character of the initial form. If all is well, the object can then be saved back to the repository in the new format. Such processes are easy to build using the Planets infrastructure, and many different types of workflow can be created in order to meet any specific needs, such as those dictated by institutional policy.

During the development of these preservation services and workflows, it must be possible to test them against any digital objects to which the institution has access. This is the purpose of the Testbed, which leverages the Planets infrastructure to allow preservation services and workflows to be discovered, modified, executed and evaluated.

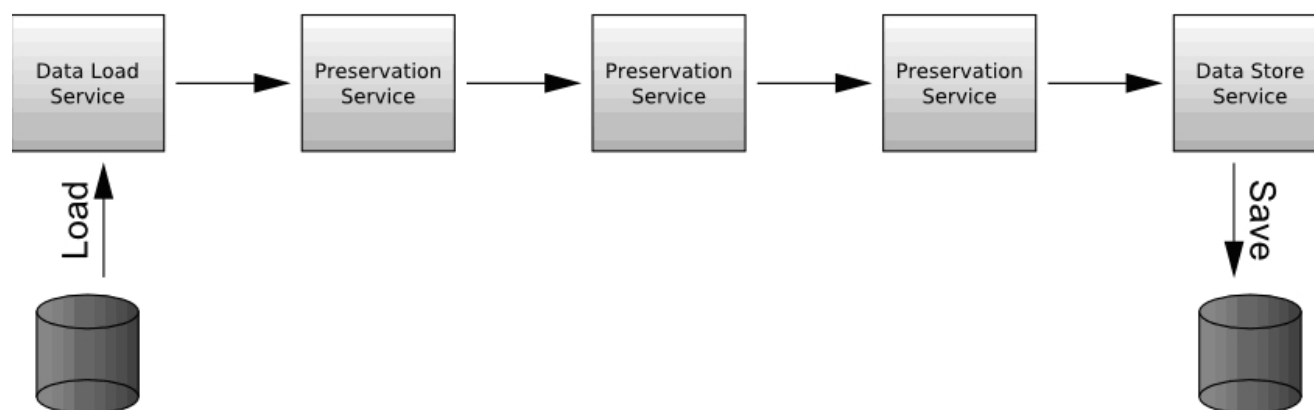


Figure 7: Simple schematic overview of the Planets workflow model.

Working within JBoss, each of the Planets components provide classes, Enterprise Java Beans, web services and web pages that can be used to build up a preservation system (see Figure 8 below). The Testbed is designed to be deployed as an Enterprise Archive within the Planets Application Server, and as it leverages the same technologies as the wider framework, the shared components and code can be reused easily. The Planets infrastructure provides user accounts, security management systems and administration pages, each of which the Testbed utilises. User authentication is achieved by configuring the application server, so no explicit authentication code is required. For authorisation, the Testbed defines the roles of Experimenter, Reader and Administrator, but the management and assignment of these roles is handled externally to the Testbed. The Planets application server also provides a set of standard Enterprise Java Beans which the Testbed can use to access (for example) the user account information.

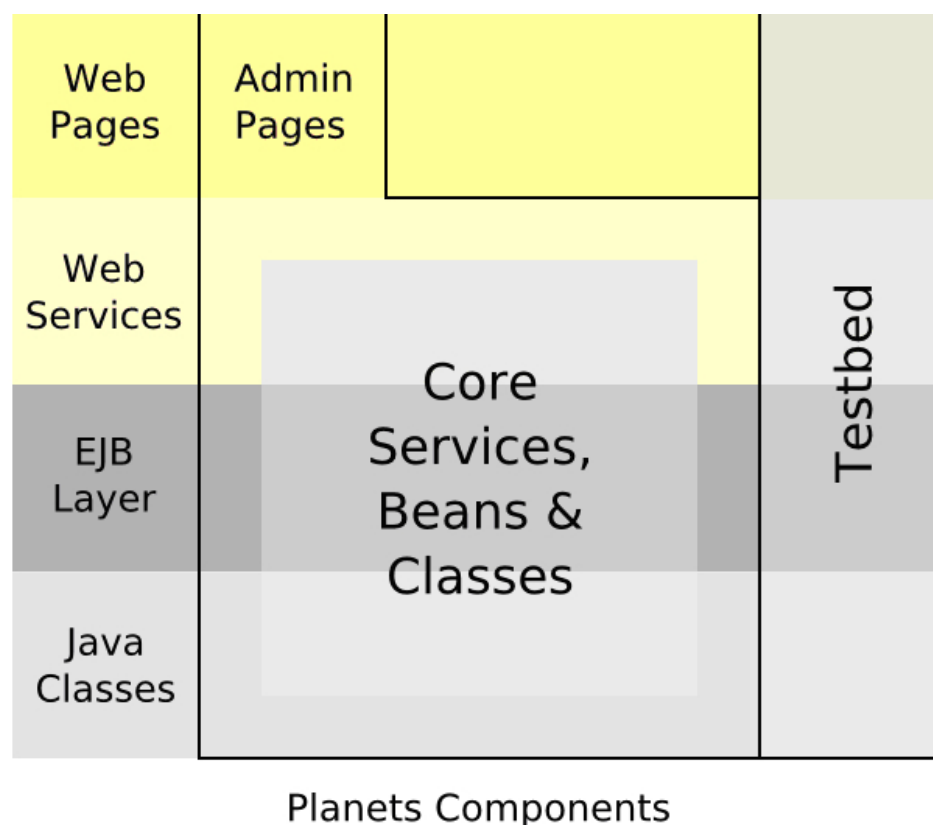


Figure 8: The different code layers of the Planets infrastructure, with the two 'Code' levels at the bottom and the two 'User' levels at the top.

As indicated above, the Planets project is attempting to standardise access to digital preservation data (via Data Registries), and to develop a common set of data types and preservation services (discoverable via Service Registries). The Testbed is designed to interoperate with the Planets infrastructure to access data on which experiments may be executed, and will be able to test and explore the preservation services that conform to the Planets specifications. Users of the Testbed can discover services via the Planets Service Registry and in addition to this may explore third-party services beyond this registry.

The Web Service Approach

Given the centrality of a service-oriented architecture approach to the Planets project, it was fundamental to ensure that the preservation tools required for Testbed experiments could be deployed and accessed as web services. The web service solution adopted by the project is built upon SOAP with Attachments API for Java (SAAJ) [19]. This API handles the creation, population and transmission of SOAP messages, and it was adapted by Planets to execute generic WSI-compliant web services [20].

The main challenge in adopting a web service approach for the Testbed was to ensure that a generic and extensible web service execution approach could be successfully integrated with the Testbed's workflow and user interface, could be easily understood by the Testbed users and could enable access not only to Planets services but also to third party services.

In order to perform experiments on preservation tools these tools must first be wrapped as WSI-compliant web services. These services can then be registered within the Testbed and service templates can be created. Experimenters can then access these templates to simulate the specific usage of a tool.

The steps involved in registering and configuring a service are handled by the Testbed administrator. This enables experimenter users such as librarians and archivists to access and use the available services without requiring any knowledge of invocation details at the SOAP message level or complex service configuration options. Different sets of service configuration options are currently presented to users as individual templates. However, user feedback has suggested that this simplified approach is insufficiently flexible; the fine-tuning of tools by end-users is one of the main scenarios of Testbed use and the desire for a more flexible approach has been noted in the Planets service requirements specification and will be incorporated into the Testbed's future releases.

The service registration process consists of a five-step wizard based interface which guides the administrator through the tasks required to make a service available. The following screenshot demonstrates step one of the process.

The screenshot shows a web browser window titled "Testbed: Register Services - Mozilla Firefox". The address bar shows "http://localhost:8080/testbed/admin/register_TBServices.faces". The page features a navigation bar with links: Home, Browse Experiments, Browse Services, Browse Data, Experiment Scheduler, Upload Data, Manage Users, Register Services, and Help. The main content area is titled "Testbed: Register Services" and contains a "REGISTRATION WIZARD PROGRESS" section with five steps: 1. select operation (selected), 2. add template, 3. invoke sample, 4. mapping of results, and 5. service metadata. Below this is a "NAVIGATION OPTIONS" section with links: register new service operation, browse registered services, and remove registered services. The main form area is titled "This is the Registration Wizard for deploying executable services within the Testbed. Every service used within the Testbed must follow this five-step registration process. Once the process is complete the service will be available for experimenters to select during the creation of new experiments." It includes a "For Stage 1 you will need to find the required service endpoint on the server, either by entering a known URL by hand or by browsing the list of service endpoints that are currently registered with JBoss. You can view a list of services registered with JBoss here." and a "Once you have entered a service endpoint you will then be able to select the appropriate Service Name and Selected Operation Name." section. The form has two main sections: "Analyze Service Endpoint" with a text input for "Please enter any locally available Service WSDL URL by hand or browse the underlying JBoss server instance." and a dropdown for "On JBoss deployed service endpoints:" showing "http://miesnb:8080/sample-ift-sample-ejb/SimpleCharacterisationService?wsdl". The second section is "Select Service and Operation" with a dropdown for "Selected Service Name:" showing "SimpleCharacterisationService" and a dropdown for "Selected Operation Name:" showing "characteriseFile". A "continue" button is at the bottom. The footer text reads "The Testbed is part of The Planets Suite. | PTB Version 0.5".

Figure 9: Step 1 of the service registration wizard

During step 1 of the process the administrator must locate and analyse a web service endpoint in order to select the actual service operation that is to be registered with the Testbed. Step 2 involves creating a service request template with placeholders for supported service input. For step 3 the administrator must upload sample data in order to test that the service can be invoked successfully. In step 4 the administrator has to note where the service's output (such as a migrated file) can be located and extracted from the XML service response message. In the final step the administrator may supply additional descriptive metadata for the service and apply restrictions on the minimum and maximum number of files the service can accept. The administrator may also supply some semantic metadata in the form of annotation tags which will help the experimenters to locate the service.

Once a service has been registered it can be experimented upon within the Testbed, and all the information required by the Testbed to execute an experiment will be available. This includes the information needed to re-factor the SOAP service

request message, construct the actual service execution proxy, invoke the service upon the given experiment data and finally extract the actual service's response (e.g. the migrated file) in order to store it within the application.

The first internal release of the Testbed incorporated three Planets preservation services: OpenOfficeXMLMigration, ImageMagicTiff2Jpeg and a simple characterisation action. However, the web service approach enables any service that adheres to the following constraints to be registered within the Testbed's infrastructure and used within an executable experiment:

1. The web-service must comply with the WSI basic profile v 1.1
2. The service's WSDL endpoint description must be located and accessible at the URI: endpointaddress?WSDL
3. Service endpoints must identify a service uniquely
4. Services and tools currently must be installed locally on one and the same machine
5. The service input parameters may only include a single file reference or an array of file references. Other service configuration options are not currently supported and are assumed to be implicitly hidden behind a given service endpoint.

Next steps of development

The first version of the Planets Testbed was released for internal use by the Planets partners in March 2008. A new cycle of requirements gathering, design and implementation will integrate partners' feedback and intermediate releases will follow, culminating in the release of the Testbed to external institutions which is expected to take place in Spring 2009.

The public version of the Testbed will be hosted in a central instance by HATII at the University of Glasgow. Planets partners and other users will be encouraged to use this central instance to ensure the seamless aggregation of experiment results. This will lead to a higher statistical significance and will make it possible for the community to create an experimental dataset in which they can look for patterns and processes that might otherwise be overlooked. A downloadable version of the Testbed will also be made available. The central instance of the Testbed will include a number of features that will make it a unique and invaluable reference for the evaluation of digital preservation tools and strategies:

Integration of digital preservation corpora

In digital preservation research, a corpus can be defined as an annotated collection of digital objects, where the annotations should contain the criteria against which given algorithms will be evaluated. [21]. In the Planets context, the algorithms will be those implemented by the tools that perform characterisation and preservation actions on the digital test objects.

The Planets Testbed corpora will ensure that a sufficient knowledgebase is available for each experiment, which will avoid duplication of effort for experimenters who would otherwise have to find and upload their own data. The corpora will also facilitate the statistical analysis and aggregation of data. Each corpus will contain individual "difficult" objects for the worst-case analysis of a given tool, and large sets of objects of a given type, or very large files, for average-case analysis.

An exhaustive list of base types of digital objects used in Testbed experiments will be established. The experience gained through the Dutch testbed project [10] suggests that it is advisable to aim to collect simple objects with few properties each, and to keep the corpora as uniform as possible with respect to variation of significant properties.

The management and maintenance of the Testbed corpora will be carried out according to procedures established with the other Planets partners; these procedures will formalise the coordination of additions of new objects and object types, of requests to incorporate pre-existing material and the management of access to the corpora.

Deployment of digital preservation tools to the Testbed; certification of third-party tools

One of the main results of the Planets project will be the development of new digital preservation tools. These tools will be deployed within the Testbed and will be evaluated using data from the Testbed corpora or data supplied by Testbed users.

The availability of the corpora of digital objects is also a necessary basis for the certification of third-party tools, an additional feature that the Planets Testbed will offer in a future release. The Testbed will not only make available existing, well-known digital preservation tools, it will also present a unique opportunity to software developers, enabling them to certify their tools against a collection of digital objects unsurpassable for its breadth, following a strict experiment process, and benefiting from previous experiment results collected in the Testbed database.

Preservation Plan Assessment Services

The Testbed corpora, together with the experiments database, will also be the basis on which preservation plans created by other Planets components will be assessed. In the Planets general workflow a preservation plan is devised by executing tool

evaluation (following a model inspired by utility analysis) on a small number of files that should represent a sample of the collection that a preservation officer wants to preserve. This plan will consist of a general section that will take into consideration the institution's policies and usage profiles, and an executable section that will detail the services needed for the actual execution of the plan itself. The Planets Testbed will allow the assessment of the executable plan by testing the services, or chains of services, on its corpora and comparing the results with those that can be found in the experiments database.

Evaluation of emulation approaches for digital preservation

Analysis of digital preservation strategies should not be restricted to migration [22]. The Planets project is also conducting research on two different emulation approaches: the Dioscuri modular emulator [23] and the IBM Universal Virtual Computer [24].

Dioscuri is an x86 computer hardware emulator written in Java. It is being jointly developed by Tessella Support Services and the National Library of the Netherlands and is managed by the National Archives of the Netherlands (the three partners are all members of the Planets consortium). Dioscuri is completely component-based; each hardware component is emulated by a software surrogate called a module. Combining several modules allows the user to replicate the configuration of any computer system, as long as these modules are compatible. New or upgraded modules can be added to the software library, thus expanding the emulator's capabilities.

The IBM Universal Virtual Computer (UVC) was initially conceived by Raymond Lorie of IBM Research in Almaden, California as a combined emulation/migration approach. IBM (also a Planets partner) conducted proofs of concept with the National Library of the Netherlands to test the UVC approach in a library environment. Within Planets, work on the UVC is focussed on improving the performance of the UVC runtime model and expanding the development environment.

With emulation, the Planets Testbed web-service architecture faces the challenge of evaluating tools and services that do not change the object, but rather the representation environment of the object, while leaving the object itself untouched. How to perform and evaluate emulation experiments conducted with Dioscuri and UVC as web services is currently being investigated within Planets by Albert-Ludwigs of the University of Freiburg using the GRATE tool [25]; the results of this research will inform the next steps of the Testbed development in this domain.

Conclusion

It is widely acknowledged [4, 5] that the management of the long-term accessibility of digital materials depends upon the automation of the processes involved. Automation of processes requires a rich understanding of them: their aims and objectives, the mechanisms needed to abstract and represent them, and often new technologies. In developing the Testbed, Planets has begun a first step in digital preservation to tackle this challenge both by providing tools to manage experimentation and also by creating an experimental evidence base to enable the analysis of the effectiveness of preservation services and strategies. This experimentation framework will foster the development of digital preservation solutions that will significantly increase automation in the daily practice of librarians and archivists, will make software developers more aware of the issues connected to obsolescence, and will ultimately change the way we all produce our digital content. It also ensures that digital preservation research is tackled as an engineering problem rather than as an art and craft.

Notes

[1] Preservation and Long-term Access through NETworked Services (Planets), viewed 23 May 2008 <<http://www.planets-project.eu>>

[2] Planets Publications, viewed 23 May 2008 <<http://www.planets-project.eu/publications/?l=1>>

[3] Ross, Seamus, and Hedstrom, Margaret (2005), 'Preservation Research and Sustainable Digital Libraries', *International Journal on Digital Libraries*, Vol. 5/4, 317-324

[4] M. Hedstrom and S. Ross (eds.) (2003), *Invest to Save: Report and Recommendations of the NSF and DELOS Working Group on Digital Archiving and Preservation*, (National Science Foundation's (NSF) Digital Library Initiative & The European Union under the Fifth Framework Programme by the Network of Excellence for Digital Libraries (DELOS))

<<http://eprints.erpanet.org/48/>>

[5] DigitalPreservationEurope, *DPE Digital Preservation Research Roadmap* (2007), viewed 23 May 2008 <http://www.digitalpreservationeurope.eu/publications/dpe_research_roadmap_D72.pdf>

[6] Gladney, Henry (2007), *Preserving Digital Information*, Springer Verlag

- [7] Ross, Seamus (2007), *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries*, Keynote Address at the 11th European Conference on Digital Libraries (ECDL), Budapest, viewed 23 May 2008 <http://www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf>
- [8] Duranti, L and Thibodeau, K (2006), 'The Concept of Record in Interactive, Experiential and Dynamic Environments: the View of InterPARES,' *Archival Science*, vol. 6, no. 1, 13-68
- [9] Watry, Paul (2007). Digital Preservation Theory and Application: Transcontinental Persistent Archives Testbed Activity, *International Journal of Digital Curation*, Vol 2, No 2, viewed 23 May 2008 <<http://www.ijdc.net/ijdc/article/view/43/50>>
- [10] Potter, M (2002), 'Researching Long Term Digital Preservation Approaches in the Dutch Digital Preservation Testbed (Testbed Digitale Bewaring)', *RLG DigiNews*, Vol 6 No 3, viewed 23 May 2008 <<http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070519:000006287741&reqid=3550#feature2>>
- [11] DELOS deliverable WP6, D6.1.1 (2004), *Framework for Testbed for digital preservation experiments*, November, viewed 23 May 2008 <[http://www.dpc.delos.info/private/output/DELOS_WP6_D611_finalv2\(0\)_denhaag.pdf](http://www.dpc.delos.info/private/output/DELOS_WP6_D611_finalv2(0)_denhaag.pdf)>
- [12] D-Lib Test Suite, viewed 23 May 2008 <<http://www.dlib.org/test-suite/research.html>>
- [13] Open Video Digital Library, viewed 23 May 2008 <http://www.open-video.org/project_info.php>
- [14] Rothenberg, J & Bikson, T (1999), *Carrying Authentic, Understandable and Usable Digital Records Through Time, Report To the Dutch National Archives And Ministry of the Interior*, viewed 23 May 2008 <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf>
- [15] Cole, T. W., Mischo, W. H., Ferrer, R., & Habing, T. G. (2001). "Using XML and XSLT to process and render online journals." *Library Hi Tech*, 19(3), 210-222.
- [16] Thaller, M (2007), *Characterizing with a Goal in Mind: The XCL approach*, viewed 23 May 2008 <<http://www.kb.nl/hrd/congressen/toolstrends/presentations/Thaller.pdf>>
- [17] JBoss, viewed 23 May 2008 <<http://www.jboss.org/>>
- [18] Java Server Faces, viewed 23 May 2008 <<http://java.sun.com/javaee/javaserverfaces/>>
- [19] The SOAP with Attachments API for Java, viewed 23 May 2008 <<https://saaj.dev.java.net/>>
- [20] Web Services Interoperability, viewed 23 May 2008 <<http://www.ws-i.org/>>
- [21] Neumayer, R, Kulovits, H, Rauber, A, Thaller, M, Nicchiarelli, E, Day, M, Hofman, H & Ross, S, *On the Need for Benchmark Corpora in Digital Preservation*, viewed 23 May 2008 <http://www.ifs.tuwien.ac.at/~neumayer/pubs/NEU07_delos.pdf>
- [22] Emulation Expert Meeting Statement, viewed 23 May 2008 <http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-eemstatement-en.html>
- [23] Dioscuri, viewed 23 May 2008 <<http://dioscuri.sourceforge.net/>>
- [24] Van Diessen, R (2006), *IBM's UVC Approach*, viewed 23 May 2008 <http://www.kb.nl/hrd/dd/dd_projecten/slides/eem_ibm_rvdiessen.pdf>
- [25] Hoeven, J, Houtman, F, Suchodoletz, D, Lohman, B, Schroder J (2007), *Next steps integration of PA5 tools (emulation)*, viewed 23 May 2008 <http://gforge.planets-project.eu/gf/download/docmanfileversion/49/2190/PA5_next_steps_integration_v3.doc>

About the Authors

Brian Aitken is a designer and developer of Web based information management systems. He has been employed both as sole developer and as leader of a development team on a wide range of successful web based projects, most recently the Digital Curation Centre, Digital Preservation Europe and Planets.

Petra Helwig worked for several years for the Dutch Ministry of the Interior where she specialised in software engineering and information management. She joined the Dutch National Archives in 2006 where she works as a Digital Longevity advisor.

Andy Jackson worked for a number of years as an applications consultant at the Edinburgh Parallel Computing Centre. After a few years spent as a post-doctoral researcher in Edinburgh and Wellington (New Zealand), Andy joined the British Library to work on the Planets digital preservation project as a senior software developer.

Andrew Lindley is currently working as a junior researcher (PhD student) and software engineer for the Austrian Research Centers GmbH - ARC in Vienna. He is part of the group's digital preservation team and the main contact at ARC for the EU FP6 IP Planets Testbed sub-project.

Eleonora Nicchiarelli has been a researcher since June 2006 at the Austrian National Library in Vienna, where she is responsible for the Planets Testbed subproject.

Seamus Ross is Professor of Humanities Informatics and Digital Curation and Founding Director of HATII (Humanities Advanced Technology and Information Institute) (<http://www.hatii.arts.gla.ac.uk>) at the University of Glasgow. He is an Associate Director of the Digital Curation Centre in the UK (<http://www.dcc.ac.uk>), a co-principal investigator in the DELOS Digital Libraries Network of Excellence (<http://www.dpc.delos.ac.uk>), and Principal Director of Digital Preservation Europe (DPE) (<http://www.digitalpreservationeurope.eu>). He was Principal Director of ERPANET, a European Commission activity to enhance the preservation of cultural heritage and scientific digital objects (<http://www.erpanet.org>), and a key player in The Digital Culture Forum (DigiCULT Forum) which worked to improve the take-up of cutting edge research and technology by the cultural heritage sector (<http://www.digicult.info>).

Subscribe to comments: [For this article](#) | [For all articles](#)

This work is licensed under a Creative Commons Attribution 3.0 United States License.

