The International Journal of Digital Curation
Issue 2, Volume 4 | 2009

Report on the “Digital Preservation - The Planets Way” Workshop

Gareth Knight

Digital Curation Specialist,

Centre for e-Research,

King’s College London

Summary

A report on the Planets a “Digital Preservation – The Planets Way” workshop, which took place on June 22-24, 2009 at the Royal Library, Copenhagen, Denmark. The workshop brought together representatives from archives, libraries, museums, academia, media and other institutions to consider the activities necessary to maintain content in the long-term and establish the methodologies and software tools developed by the EU-funded PLANETS project as a potential solution for preservation concerns. The event was the first of a series of three-day workshops that the Planets (Preservation and Long-term Access through NETworked Services) project is organizing across Europe during 2009-2010.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Planets (Preservation and Long-term Access via NETworked Services) is a four-year project co-funded by the European Commission Information Science and Technologies Framework Programme. It draws together a number of national libraries, archives, academic institutions and technology companies throughout Europe to create a framework of services that supports the long-term preservation of digital information. Key outputs of the project include services to: define, evaluate and execute preservation plans; characterise digital objects and establish their significant properties; and perform preservation action using a range of automated tools. Now in its penultimate year of funding, the project has initiated a series of three-day workshops across Europe to demonstrate the Planets approach to preservation and the set of tools and services that has been developed.

The aim of the workshop[1] was to present a practical approach to digital preservation, providing hands-on experience of the wide range of preservation tools and services being developed by the Planets Project. It also served as an opportunity for diverse institutions to come together to discuss common preservation issues.

Day One: June 22, 2009

Welcoming delegates to the workshop, **Clive Billenness,** PLANETS Programme Manager, outlined the aims of the day and introduced each of the speakers. The first day was intended to provide CEOs, IT managers and curatorial staff with an introduction to preservation issues and the approach that Planets has taken to manage the process. Delegates that were interested in gaining a high-level overview, but did not have time to attend all three days were able to attend the first day only. However, in practice the majority of delegates attended for the full workshop.

Introduction to Digital Preservation: Issues and Challenges

The proceedings commenced with a key note by **Dr. Ross King** of the Austrian Institute of Technology, who provided an introduction to digital preservation and the associated issues. Drawing upon the theme of a “digital universe”, Ross cited one estimate that there is approximately 700+ Exabytes of data in existence in 2009 (Gantz et al., 2008). Although some of it is transitory and redundant, much of it will require preservation. Examples of digital information that continue to be valuable include medical records, historical resources, patents and other sources of information. An institution that wishes to preserve its own data should consider a number of questions: “what information should be preserved?”, “what approach is necessary to preserve it?”, “how can you confirm that the information has been preserved correctly?” and “what incentives are there to preserve information?” Possible answers to these questions were considered by speakers throughout the three days of the workshop.

Ross went on to explore the different challenges posed when trying to preserve digital information. Several risks to the preservation of the bitstream were noted, including media deteriation and media obsolescence, as well as obsolescence to the operating system, software application and file format. The logical preservation strategies of migration and emulation were introduced as possible solutions to these

[1] Digital Preservation – The Planets Way, June 22-24, 2009, Royal Library, Copenhagen, Denmark http://www.planets-project.eu/events/copenhagen-2009/

issues and reference was made to presentations throughout the day for further information on these approaches.

To conclude, Ross rhetorically asked “why bother with preservation?” He immediately responded by suggesting that the potential cost of losing access to digital information may potentially be higher than the cost to perform digital preservation – a response that appealed to a number of managers in the audience.

The Preservation Action Cycle: Introduction to Planets

The risk theme was continued by the first presenter of the day. **Clive Billenness** indicated that the inability to access digital information over time is frustrating for a user, but can be potentially costly for an institution. Many libraries and archives are legally bound to maintain access to digital information and so, must actively perform preservation actions. To ensure that management decisions are made with consideration for long-term preservation, he advocated the use of a risk management approach to identify and avert the risks that may limit the ability to access digital information over time. Clive went on to recommend that institutions adopt a risk management strategy that is built upon best practice, documented experience and corporate governance principles. Although, as he noted, there are relatively few standards or specifications that directly apply to the requirements of digital preservation, an institution may find it useful to consult the UK Office of Government Commerce “Management of Risk” and the British Standard 31100 Code of Practice, as well as the OAIS Reference Model to frame the decision-making and risk management process. In the final section of his talk, Clive examined the stages of the risk management cycle – from risk identification to the implementation of a risk management plan – and explained how the PLANETS services and tools (the PLATO planning tool, Testbed and Corpora, preservation characterisation services, comparator and others) could be used to address the requirements of each stage.

Preserving Digital Content: A Short Introduction to Digital Information in the Preservation Context

Volker Heydegger of the University at Cologne examined the question of how to distinguish information content from its encoding information, to consider what aspects of a digital object should be preserved. He began by considering the method by which information stored on an analogue and digital carrier is interpreted and rendered in a form that can be understood. Although the form in which information is stored differs, varying in the carrier type in use and the form of encoding, the process of information interpretation remains broadly similar. To illustrate, he compared the activities necessary for a reader to interpret and understand the content of the Rosetta Stone encoded in hieroglyphic, demotic, and ancient Greek on a granodiorite carrier, with the activities necessary to interpret and render the same information encoded as binary and stored on an optical disc or hard disk. In both instances, a decision has been made that the primary information to be communicated is the text and that aspects of the carrier are not required. Volker went on to explore the process by which digital information is interpreted and rendered in further detail. He indicated a combination of hardware and software was required to decode the binary data. At the storage level, digital information is encoded as binary data held on a carrier, such as an optical disc (CD, DVD), solid state (USB memory sticks, SD cards), or magnetic disc (hard disk, floppy disk). To access it, an operating system must interpret the encoding structure of a disc and enable its organization structure to be accessed and manipulated. To render binary data in a form understandable to the user, an appropriate software tool capable of interpreting and rendering the content is required. To illustrate the latter stage, he compared the information stored in an image format, as rendered by a hex viewer and image viewer. While the image viewer is able to interpret the content of the binary file and render the still image, the hex viewer displayed the content as a series of machine-interpretable hexadecimal values.

The latter half of the presentation considered the characteristics of a data object that should be considered essential to recreate the content. Volker drew upon the OAIS concepts of Data Objects, as well as work performed in the wider JISC community on significant properties to identify four types of information contained in a digital object that have different levels of significance:

1. “Basic information” that indicates the content type;
2. “rendering information” that indicates how the content should be recreated;
3. “Storage information” that indicates the method in which content should be deployed;
4. “data” that contains the information for rendering.

Of these, “basic information” and “data” were labelled as mandatory, “rendering information” was considered beneficial and “storage information” was considered useful for historical purposes.

To conclude, Volker considered how digital content can be maintained in the longterm in a way that preserves its characteristics. He indicated that an institution should combine preservation tasks in an orchestrated manner, by developing a preservation plan. It should evaluate the suitability of preservation action tools, extract and evaluate the characteristics of collections and objects; and ensure that preservation actions are performed in a controlled environment. Planets tools, such as the Planets Preservation Tool (Plato), eXtensible Characterisation Language (XCL), PRONOM and the Testbed may assist in the performance of these tasks.

Why Do We Have to Plan Digital Preservation?

Christoph Becker of the Vienna University of Technology gave a highly detailed talk on the steps that an institution should take to create a preservation policy and convert it into a preservation plan. He began by examining the reasons why an institution should consider the development of a policy, and identifying the need to develop a consistent, cost-effective management strategy that establishes trust from key stakeholders, as well as the potential risks associated, with technology obsolescence, data corruption and loss as key factors. This led to an exploration of the recommendations made by specifications, such as the RLG & NARA Trustworthy Repositories Audit & Certification (TRAC) criteria and checklist, NESTOR Catalogue of Criteria of Trusted Digital Repositories, or DRAMBORA Self-assessment toolkit. In each document, it was stated that a repository should have procedures and policies in place to preserve data as well as appropriate mechanisms to review, update and develop them to take account of evolving technology and community practice.

A preservation policy was defined as a high-level document that specifies the institution’s commitment to maintaining digital resources. It will indicate the requirements that must be met at different levels of the organization and in relation to various types of digital content. The policy may provide an explicit statement on the strategy that will be adopted to preserve content (e.g., migration) and the criteria that will be used to evaluate formats, or may refer to alternative documents. To prepare for the creation of a preservation policy, Christoph advised institution staff to examine the context in which the institution operates, the policies published by other institutions, and recommendations made in appropriate management specifications. For the former, he suggested that research should consider four broad categories:

1. the organizational context, as expressed in mandates, legislation, organizational policies and user community;
2. technological infrastructure and staff expertise that are available;
3. the type and characteristics of the digital resources that are being maintained; and
4. the options that are available to manage the data.

The second section of his talk examined the process that institutional staff may use to convert abstract policy into working practice. A preservation plan should provide a documented, practical set of actions to be performed to address one or more preservation risks. The plan should enable staff to make an informed decision on the action to be performed to handle particular types of digital information.

Finally, Christoph outlined how the preservation planning workflow that is a key component of Plato may be used to record the decision-making process (Becker, Brown & Kulovits, 2008; King, Lindley & Schmidt, 2009). The workflow is composed of three high-level stages - define requirements, evaluate other options and consider results – and a series of individual steps that should be performed to develop a preservation policy and plan that meets the requirements of the institution and its data. He also referred to outputs by other Planets partners to develop a technology watch service and collection profiling and risk assessment system, which has some relation to The National Archives PRONOM characterisation service.

How to Preserve?

The final presentation of the morning was given by **Sara van Bussel** of The National Library of The Netherlands, who spoke of ongoing work to integrate migration and emulation tools into the Planets technical architecture. Migration and emulation are both considered to be effective strategies to support “logical preservation”: migration refers to the process of converting content from one encoding format to another, to ensure it remains accessible and usable in contemporary operating environments and software applications; while emulation refers to the process of recreating an original technical system – hardware, operating system, software tool, or a combination of all three – in contemporary operating environments. An emulator may be used to render a website as it was intended for viewing, use obsolete software to perform formation migration and recreate computer games. Although both strategies offer a practical method of maintaining access to content, they are not without risk. A migration activity may introduce unexpected content change or result in functionality loss. Similarly, the emulator may not be sufficiently accurate to recreate the original environment used to access content, resulting in unexpected differences in the user experience.

Sara outlined the gap analysis work with which she had been involved. The survey was performed to identify the formats that should receive priority support in the Planets toolset. To establish the range and type of formats that required preservation, the project team surveyed 65 organisations across Europe, including archives, libraries, academic universities and other institutions. The survey identified a total of 107 distinct formats (excluding version differences). The most common formats being stored by institutions were TIFF, JPEG, PDF, XML, MS Word document, MP3 and HTML, followed by content stored in Microsoft Excel, Microsoft Access, Microsoft PowerPoint, BMP and PDF/A. A small number of institutions also stored discipline-specific data, such as sheet music stored in a number of proprietary formats and audio book formats for blind users.

In the latter part of her presentation, Sara spoke about the preservation action services being created by the project. To provide a consistent interface for interacting with characterisation, migration and other software, the tool is "wrapped" into a web service. More than 20 tools have been integrated into the testbed, so far. They include SOX (audio format converter), SIARD (for converting relational databases to XML), JHOVE (for format characterisation), XENA (a format normalization tool developed by the National Archives of Australia) and others. Preservation action web services may be accessed through an API or controlled via a web interface.

Tools: How to Understand Files

Jan Schnasse of the University at Cologne opened the afternoon session of day one with a presentation on the content characterisation work being performed by the project. His presentation considered the process by which content is interpreted by a user and introduced the Extensible Characterisation Language (XCL)[2] as a method of evaluating the success of format action.

He began by examining the process by which encoded data is interpreted and rendered in a form that can be understood by the user, expanding upon the theme outlined in Volker Heydegger's earlier presentation. Jan considered the process by which a digital object is converted into a form that can be perceived by a user, drawing upon the three-tier Performance Model developed by the National Archives of Australia (Heslop, Davis & Wilson, 2002). Although preservation actions, such as migration and emulation, enable content to be recreated in a form that can be interpreted by a user, it can be difficult to establish if all of the significant properties necessary to communicate the meaning have been successful transferral. Until recently, the only effective approach to address these concerns has been to perform a manual inspection of the object – a prohibitively time-consuming task. A more effective approach is to automate the process of content validation when converting between file formats. The characterisation results created by current tools, such as JHOVE, ImageMagick and TIFFInfo are unsuitable for comparison, due to the differences in the amount and type of information extracted for each format.

Jan introduced the Extensible Characterisation Language (XCL) as a solution for measuring the success of preservation actions. XCL enables the content of a digital object to be described using a common structure. It may be used to compare and contract the properties stored in Format A to those stored in Format B, subsequent to format conversion being performed. XCL is composed of two components – an Extensible Characterisation Description Language (XCDL) that is used to describe the properties associated with the content of a digital object; and the Extensible Characterisation Extraction Language (XCEL) which describes the structures of a file format.

[2] Extensible Characterisation Language (XCL) http://planetarium.hki.uni-koeln.de/planets_cms/

In closing, Jan spoke about the work that had been performed to create XCELs to describe the structure of several common raster image formats. He also described their current efforts to develop XCELs that will support the comparison of document formats, such as Microsoft Word and Adobe PDF.

Digital Preservation: How to Verify

Petra Helwig of The National Archives of the Netherlands spoke on the need to take a preservation-led approach to preservation that learns from experience gained in the user community. Digital objects often have specific properties that must be maintained during preservation action. However, it is difficult for a single institution to develop sufficient experience to find the most effective solution. To illustrate, she demonstrated the conversion of a Dutch-language word-processing document using XENA. Although the document contained a small amount of text only, the conversion action had resulted in the umlauts[3] [4] being removed from several letters. By using the PLANETS Testbed as an environment to perform and share preservation experiments, she indicated a large body of information will be created on different preservation tools that may be used by preservation staff to evaluate and select the best and most appropriate approach to preserving a digital object.

On conclusion of her presentation, she invited questions from the audience. Several delegates representing libraries, museums and academic institutions expressed an interest in using the testbed to experiment with different preservation actions, but were concerned that the testing of high-quality digital masters would result in the public availability of data that they do not have the legal right to publish on a third-party site. In its current version, the testbed publishes a copy of the data on which an experiment was performed, for re-testing by others. Petra indicated it was not currently possible to restrict the test files, but the institution could install a local copy of the testbed software. The only limitation is that the local instance would not contain the experiments performed by others. Another delegate asked if it was possible to submit new tools for integration. Petra responded that it was not possible to upload tools themselves, due to the complexity involved with integrating each tool into a service wrapper. However, the PLANETS Project will issues guidelines for integration of new tools for use by third parties in November 2009.

Digital Preservation: How to Plan

For his second presentation of the day, **Christoph Becker** demonstrated the use of the Plato tool to implement the preservation plan introduced in the morning session. It is beneficial to define the requirements that must be met prior to performing preservation action. They may be divided into three categories – Object Characteristics, Record Characteristics and Process Characteristics – and assigned some form of measurement. Each factor is assigned a quantifiable value (a Boolean or numeric value) that may be measured and compared. By defining a set of quality controls for the preservation process, the staff members may evaluate different preservation approaches available and determine the file format or software tool that best meets their needs.

[3] a diacritical mark ¨ placed over a vowel to indicate a more central or front articulation (Merriam-Webster Online), for example, in German, “Universität zu Köln”. [editor]

[4] However, in Dutch such a mark represents a trema or diaeresis which serves to distinguish the separate pronunciation usually of two succeeding vowels (e.g., in Dutch, “zoölogie”, and French and English, “Noël” [editor].

How to Integrate the Components of Digital Preservation

Ross King presented a case study describing the use of the Planets services to automate the processing of a raster image collection stored by the British Library. The collection consists of 80TB of TIFF-encoded newspaper scans that require characterizing, rotating and cropping to correct for scanning errors and conversion into an alternative preservation format. He began by outlining a set of terms that the project uses to describe components of a workflow. In the context of the PLANETS toolset, a “Workflow” consists of a set of web services that each performs a specific action, such as characterize, modify, or migrate. The sequence in which the services appear in the workflow and the interface to be used is defined in a “Workflow Template”. Each workflow may be expressed as a “Workflow Description” XML serialization that contains information on the workflow template, parameters associated with each of the Planets and other technical information. By refining an existing Workflow Template or developing a new one that fits local requirements, an institution may automate several activities performed by an OAIS-compliant system. To prepare the image collection for ingest into the British Library archive storage system, a Workflow Template was created that consists of five stages – Validate, Identify, Characterize, Modify and Normalize – and one or more tools were selected to perform each action. In this instance, they chose: JHOVE to validate that the TIFF parameters were correct; DROID to identify the image and obtain a PRONOM ID; JHOVE to characterize each image; ImageMagick to rotate and crop the images; and OpenJPEG to normalize the set of TIFF to JPEG2000. To conclude, he spoke about the work performed to produce Workflow Descriptions that can automate a standard set of preservation activities using common preservation action tools. The latter may be adopted and used by others, or expanded to fix bespoke workflow activities.

Planets at the National Library of the Netherlands

Barbara Sierman of The Koninklijke Bibliotheek (KB)/ National Library of The Netherlands presented a case study of the e-Depot and the library’s contribution to the PLANETS Project. The e-Depot was established in 2003 and has since collected over 12 million objects. The majority of objects are PDF documents, though staff expect to receive an increasing number of websites and digitized data in the next five years. At its current stage of development, the library performs preservation watch, planning and action for various content types. However, there is concern that the growing number of objects will overwhelm their current, primarily manual process. To ensure that the Library can continue to preserve digital objects that are being captured and deposited, it requires a scalable management system that has integrated support for preservation actions. Specifically, the system should automate large sections of the ingest process, performing format identification, characterisation, normalization and metadata extraction and other appropriate activities. It should also support additional functionality, such as data versioning and management activity monitoring.

Barbara went on to describe the reasons why the National Library had sought to participate in the PLANETS Project and the work it was were performing. The Library views the project as an opportunity to work with a community of experts and develop a preservation solution that meets a wide range of preservation needs. It would be time-consuming and expensive to produce similar tools in isolation. In return, the Library is managing the Preservation Action sub-project, developing the migration and emulation tools registry and creating an inventory of preservation policies.

Day Two: June 23, 2009

The second day of the workshop was opened by **Clive Billenness**, who welcomed delegates who had not enrolled for the first day of the event. The second and third day were more practical in focus, providing a detailed description of many of the technologies in use and providing delegates with training in the use of the various tools and services in development. Clive concluded his introduction by inviting delegates to submit feedback on the use of the tools, to help the project to address technical issues and expand their functionality in the final year of the project.

The Digital Preservation Scenario

Following the welcome and introductions, the first presentation of the day was given by **Vittore Casarosa** from HATII at the University of Glasgow, who spoke about the growing need for digital preservation solutions expressed by institutions operating in many different fields of expertise.

In the first part of his talk, Vittore examined the different management approaches available to curate and preserve data. The list of preservation strategies used terminology outlined by Wilson (2007), describing four approaches to preservation:

1. Techno-centric in which the original hardware and software is maintained;
2. Data-centric approach in which the object is maintained in current formats using migration or emulated in a virtual environment;
3. Process-centric in which the digital content is migrated;
4. Post-hoc in which digital archaeology and forensics are performed to rescue the data.

He went on to introduce the Open Archival Information System (OAIS) as a reference model for organizing the structure of a data management system to consider preservation and dissemination requirements that met the requirements of the institution's Designated Community.

For the second part of his talk, Vittore spoke of the broad interest in digital preservation issues from a wide range of institutions. He drew upon information provided by delegates when registering, to illustrate the growing need for digital preservation solutions. The majority of delegates were employed by National Libraries (38 per cent), academic institutions (28 per cent) and archives (21 per cent) in different countries. A comparatively smaller number of representatives was representing government (7 per cent), media (3 per cent) and national museums (3 per cent). These institutions currently store a range of object types, including documents and images, audio, websites and databases. A small number of institutions also stored software, GIS, scientific data and disc images. Finally, Vittore provided statistics on the number of institutions that possessed, were developing, or were implementing a digital curation solution. In total, 18 delegates stated their respective institutions had or were in the process of developing a curation system; two institutions were assessing their needs, one institution was in the process of tendering for a management solution and five had no current plans to introduce a digital preservation system

To conclude, Vittore reflected on the interpretation that may be derived from the set of statistics. The diverse range of institutions, content types and management approaches made it difficult, if not impossible to advocate a "one-size-fits-all" solution. However, the adoption of a flexible approach should enable the Planets project to create services that can be used to fulfil a wide range of needs.

Preservation Planning with Planets

The first practical exercise of the workshop challenged delegates to develop a set of criteria against which a preservation action may be evaluated. **Christoph Becker** supplied a walk-through of the initial steps of the preservation planning workflow provided in the Plato preservation tool. Delegates were separated into three groups and given the task of defining the technical characteristics of the encoding format, infrastructure characteristics of the institution, and record characteristics of the information content that should be maintained.

Characterisation of Digital Documents

Volker Heydegger and **Jan Schnasse** provided an enlightening presentation that explored the constriction of the eXtensible Characterisation Language (XCL) in further detail and provided examples of its application to a set of objects[5].

Volker introduced the eXtensible Characterisation Language (XCL) for delegates who had not attended the first day of the workshop. XCL is composed of two components, both of which conform to an XCL ontology:

1. An eXtensible Characterisation Extraction Language(XCEL) that contains a formal, abstract description of each file format.
2. An eXtensible Characterisation Definition Language (XCDL) that is used to document the contents of a specific digital object.

An XCDL XML-encoded file contains a series of embedded statements that describe the contents of an analysed object. Each element provides a distinct value measurement and the measurement type (e.g., an integer value; UTF8 text, or hexadecimal value for binary). The explanation was followed by a practical demonstration of XCL to evaluate the success of converting a set of TIFF images to PNG. The demonstration comprised four stages:

1. An analyzer tool is executed and configured to examine digital information stored in Format A. A set of properties is extracted and recorded as an XCDL output file.
2. A preservation action tool is used to convert content stored in Format A to Format B.
3. The analyzer tool is configured to examine digital information stored in Format B and a set of properties is extracted and stored as an XCDL file.
4. Finally, a comparator tool is executed to compare the two XCDL files and differences are noted.

In the question-and-answer session, Volker stated that the project team had created XCELs to describe several raster image formats and had recently moved on to examining the PDF and Microsoft OfficeOpen formats. However, he indicated that the current version of the software imposes a small number of limitations. It is not currently possible to perform one-to-many comparisons, in the event that a curator wishes to compare a source object stored in one file to a normalized object that has been exported to two or more distinct files. He also noted that, due to the nature of the work, the comparator is unable to compare format-specific properties.

Preservation Actions

Sara van Bussel explored the preservation requirements of relational databases,

[5] Extensible Characterisation Language (XCL) http://planetarium.hki.uni-koeln.de/planets_cms/

examined the progress that had been made to develop an emulation web service and described the function performed by the Planets Core Registry.

The presentation began with an overview of the SIARD (Software-Independent Archiving of Relational Databases) prototype that is being developed by the Swiss Federal Archives. SIARD provides a set of tools that may be used to convert a relational database – consisting of a set of inter-connected tables which are joined by primary and foreign keys – stored in Microsoft Access, Microsoft SQL Server, Oracle, or other formats and resave in a non-proprietary format. The SIARD format is a zip-compressed archive that stores database content as a set of XML files, one for each table, accompanied by a metadata record (based upon SQL: 1999). A SIARD software tool, executable in any JAVA 1.5 environment, may be used to manipulate and access the relational database.

Sara moved on to describe the emulation development work taking place in the project. The National Library of the Netherlands has created a Java-based application called Dioscuri (van der Hoeven, 2006; van der Hoeven, Lohman & Verdegem, 2007) that is able to emulate a 16-bit Intel 8086 PC and its various components. Although the processor capabilities of the emulation environment are relatively limited in technology terms, screenshots of the emulator running MS-DOS, FreeDOS and Microsoft Windows 3.0, as well as a DOS-based web browser and computer games were shown. Interesting features of value to the preservation community include an XML-based module configuration and the ability to copy text from the emulated environment into the clipboard of the host operating system. She also demonstrated GRATE (Global Remote Access to Emulation), an interesting project that provides users with access to an emulated environment via their web browser (Welte, 2009). GRATE consists of a client-side application, requiring JRE 1.5 that connects to a remote server. The server offers several Linux-based emulators that the user can select and use to access or manipulate objects. GRATE was shown running Microsoft Windows 98 in QEMU to access the contents of a document. Interesting features that were mentioned during her talk include the ability to upload local objects to the emulated environment, copy and paste between the local and remote system, mount virtual disk images and make network connections using FTP, Samba and others. However, many of these features were dependent upon the emulator software in use.

Finally, Sara spoke about the PLANETS Core Registry, an online tool that contains information on file formats, software applications, hardware and media carriers. The registry builds upon functionality offered by PRONOM, while addressing specific issues such as the vague format definitions, for example, identification of a Microsoft Word document as OLE2 rather than MS Word 97. She indicated that the Testbed also contained a feature called "pathways" that provides users with a list of step-by-step actions that may be performed to extract content and convert it into a form suitable for preservation or distribution.

Benchmarking Preservation Tools: The Testbed Environment

For the final session of the day, **Brian Aitken** demonstrated the latest version of the Planets Testbed, which was currently undergoing testing and would be made publicly available at a later date. He demonstrated the use of the preservation planning workflow that underpins the PLANETS methodology and the use of tool-based services to perform format conversion actions.

Day Three: June 24, 2009

The final day of the workshop focused upon practical use of the Planets services and tools that had been mentioned and demonstrated during the previous two days. In the welcome session, Clive Billenness explained that the tools had undergone considerable development and invited delegates to provide bug reports, suggestions and feedback.

Practical Session 1

Hannes Kulovits of the Vienna University of Technology introduced the first session of the day, explaining the practical task to be performed by each group and providing a refresher on the use of the PLATO tool. Delegates were asked to prepare a preservation plan for a test collection – a set of GIF- and TIFF-encoded images – by identifying the Object, Record and Process characteristics that should be considered. To perform the task, delegates were asked to create an Object Tree – the set of targets that should be met – using Freemind, a Java-based mind mapping software tool and subsequently import it into PLATO. The preparation activity proved to be a slow and slightly confusing process for many, due to unfamiliarity with Freemind and the PLATO tool, but was considered to be a worthwhile exercise.

Practical Session 2

The second practical exercise of the day was set by **Brian Aitken**, who gave the groups the task of characterizing and converting data to other formats using the Planets Testbed. In the reporting session, each group explained the activity it had performed and its results. Experiments performed by the various groups, included conversion from HTML to XHTML, MSWord to PDF, SVG to PDF and JPEG, and animated GIF to PNG and JPEG. The session provided delegates with hands-on experience of the experiment workflow which helped to familiarize them with the approach that it takes. A few issues were noted during testing, most notably the time overhead required to perform characterization and conversion tasks in comparison to the use of the same tools in a native environment and confusion over the interpretation of status messages that indicate the success of the action (in its current iteration, Testbed indicates that the service process has completed successfully, but does not indicate if the activity has been performed correctly). These issues were noted by the development team, who indicated that they would be addressed in later versions.

Pulling It All Together - Implementing Digital Preservation Using the Planets Interoperability Framework

The final session on the workshop programme was presented by **Clive Billenness**, who provided a high-level view of the PLATO, Core Registry and Testbed technical requirements and outlined the actions necessary to install and test the PLANETS software on an institution’s own servers. The core PLANETS toolset[6] is written in Java and has been tested in Unix- and Windows-compatible environment. The current iteration of the workflow engine uses a customized version of JBOSS and cannot be used with the default JBOSS setup, although Clive observed it was possible that this may change at a later date. Separately available components, such as preservation action services, may have specific operating system-specific functionality, or licence restrictions, that limit their use to certain environments. To conclude, he invited

[6] PLANETS Interoperability Framework: PLANETS IF Sub-Project gForge http://gforge.planets-project.eu/gf/project/if_sp/ifrs/?action=index

delegates to download the software and evaluate the software for themselves.

Conclusions

The workshop generated a high level of interest among the delegates, reflected by the discussion that took place throughout the workshop, via twitter and blog posts. It successfully demonstrated the diverse (and prior to the event, slightly confusing) range of tools and services that have been developed by the partners, and raised some interesting issues that may be addressed in the final year of Planets or as part of a follow-on project. The key message communicated by the workshop was the value in taking a joined-up approach to preservation. The various services are not currently as seamlessly integrated as the project clearly intends, and further work is necessary to integrate the outputs with other software operated by institutions. However, the high level of integration is encouraging and should provide a solid framework for supporting preservation activities in the future.

References

Becker, C. Brown, A., & Kulovits, H.(2008). *Report on service integration in Plato 2*. Retrieved August 6, 2009, from http://www.planets-project.eu/docs/reports/Planets_PP4-D3_ReportonServiceIntegrationInPlato-final.pdf

Gantz, J. F., Chute, C., Minton, S., Reinsel, D., Schlichting, W. & Toncheva, A. (2008). *The diverse and exploding digital universe - An updated forecast of worldwide information growth through 2011*. IDC White Paper. Retrieved September 30, 2009, from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Heslop, H., Davis, S., & Wilson, A. (2002). *An approach to the preservation of digital records*. Retrieved August 6, 2009, from http://www.naa.gov.au/Images/An-approach- Green-Paper_tcm2-888.pdf

King, R., Lindley, A., & Schmidt, R. (2009). *Guidelines for creating and installing IF preservation workflows and templates*. Retrieved August 6, 2009, from http://www.planets-project.eu/docs/reports/Planets_IF5-D1_Creating&Install_IF_Pres_Workflows.pdf

van der Hoeven, J. (2006). S*econd version of Dioscuri*. Retrieved August 6, 2009, from Planets website: http://www.planets-project.eu/docs/reports/Planets_PA5-D6-Second_version_of_Dioscuri_final.pdf

van der Hoeven, J., Lohman, B., & Verdegem, R. (2007). Emulation for digital preservation in practice: The results. *International Journal of Digital Curation, 2(2)*, pp. 123-132. Retrieved September 30, 2009, from http://www.ijdc.net/index.php/ijdc/article/viewFile/50/35

Welte. R. (2009). *First version of GRATE*. Retrieved August 6, 2009, from Planets website: http://www.planets-project.eu/docs/reports/Planets_PA5-D7_GRATE.pdf

Wilson, A. (2007). *Significant properties report*. InSPECT Work Package 2.2. Retrieved September 30, 2009, from http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf