Planets: integrated services for digital preservation

Planets (Preservation and Long-term Access through NETworked Services) is a four-year research and technology development project co-funded by the European Union under the Sixth Framework Programme to address core digital preservation challenges. The primary goal for Planets is to build practical tools and services that will help ensure long-term access to large scale digital collections. This article describes Planets' approach to digital preservation and explains how it interacts with an organization's ingest, repository and delivery services. It also presents the deployment options of the Planets digital preservations software framework, and includes an overview of the project's progress to date.

ADAM FARQUHAR
Head of Digital Library Technology
British Library

HELEN HOCKX-YU
Web Archiving Programme Manager
(formerly Planets Project Manager)
British Library

Introduction

Planets (Preservation and Long-term Access through NETworked Services)[1], is a four-year research and technology development project that commenced in June 2006. It is co-funded by the European Union[2] under the Sixth Framework Programme[3] to address core digital preservation challenges. The primary goal for Planets is to build practical services and tools to help ensure long-term access to large scale digital collections. The Planets consortium involves 16 partners across Europe and brings together expertise from national libraries and archives, leading research universities and technology companies. Planets is co-ordinated by the British Library.[4]

Funded under the same Framework Programme are CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval)[5], co-ordinated by the UK Science and Technology Facilities Council and DPE (DigitalPreservationEurope)[6], co-ordinated by the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow. CASPAR develops solutions to the long-term preservation of scientific, cultural and artistic data. DPE fosters collaboration and synergies between existing national initiatives and aims to improve co-ordination, co-operation and consistency in current digital preservation activities. All three projects commenced in the first half of 2006 and are part of an ambitious European initiative aimed at keeping today's digital content alive in the future.

Plugging the gap

The impetus for Planets comes from national libraries and archives across Europe which have the legal responsibility as well as the legislative framework to safeguard digital information and provide sustained access to digital cultural and scientific knowledge. While much progress has been made in digital preservation research in recent years, the current state of the art has fallen short of implementing integrated solutions to the preservation of large-scale real life digital collections. Preservation planning, the process of selecting the appropriate preservation strategy against organizational requirements and collection characteristics,

for example, remains a manual one in practice. In addition, tools available for identifying file formats and extracting metadata only cover a limited range of specific formats and are often not specifically developed for preservation purposes. A fundamental problem with the current tools, including specific preservation tools which migrate digital objects to newer formats or emulate their original environments, is that they usually exist as stand-alone applications and are not geared to preserve a *collection* of digital objects, which can include embedded, complex objects in multiple formats. The tools cannot be easily combined to perform chained actions and there is little or no support for handling dynamic datasets or compound content. There is also a lack of methodology or testbed for comparing tools and accessing their effectiveness, making it difficult to plan, select and evaluate them.

Being most directly confronted with the challenges of digital preservation, the national libraries and archives have recognized the value of integrating the fragmented state of the art to introduce innovations that fill the gap in current understanding and practice. The intention is to provide an environment to encourage ongoing development of tools and services and a framework that software vendors and commercial service providers can implement and augment. Achieving these ambitious goals goes beyond the capabilities of any single institution. Planets benefits from the complementary expertise of the partners and will consolidate existing and emerging technologies. The end product that Planets aims to deliver is a range of tools and services supporting key preservation functions, in the form of a downloadable 'click-and-install' software package, which allows the administration, configuration and deployment of preservation services and work flows.

The Planets approach

The Planets software supports a number of key digital preservation functions: preservation planning, characterisation, preservation action, the testbed and the interoperability framework. Figure 1 shows their interrelationships, as well as their interactions with the external environment.

Preservation planning is a comprehensive set of services that enables an organization to identify risks to its digital collection, come up with a set of alternative treatment plans to address the problem, evaluate the plans to determine the most effective one, and then execute the plan with as much automation as possible.

In order to identify risks to a digital collection, the preservation planning services need to know about: the digital content held in the collection; the

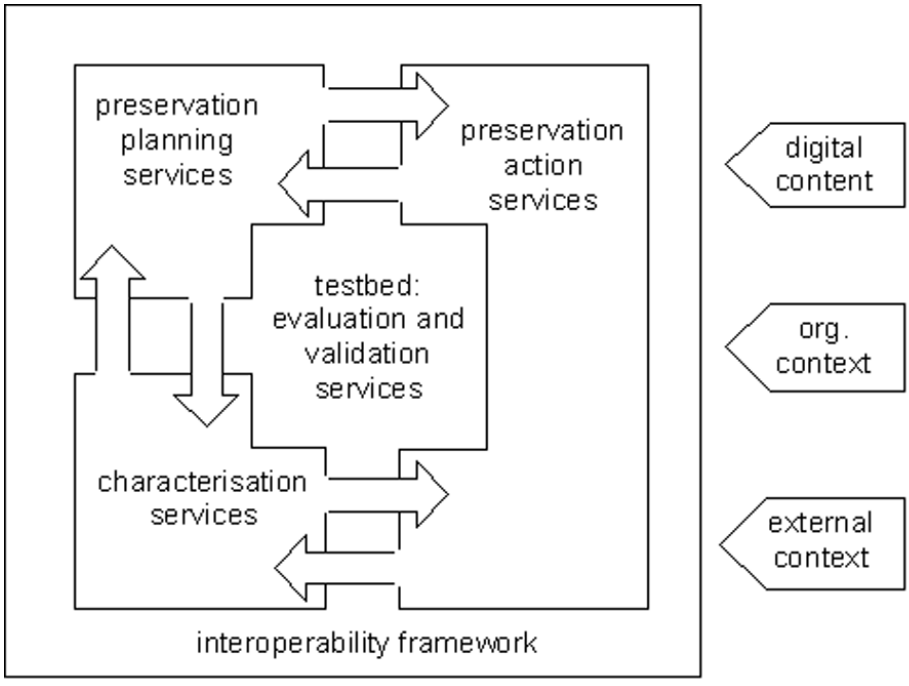

Figure 1. Planets architecture

organization's policies, goals, and constraints; the environments in which end-users interact with the content, and the goals that the end-users have. The preservation planner uses two key sets of Planets services: content characterisation and preservation action.

Content characterisation services identify and extract features of the content in the collection that are relevant to the planning process. Examples include basic technical features such as the pixel-depth, height, width of images, rendered features such as the number of pages in a document, and other aspects such as whether a database uses stored procedures. In addition, content characterisation services include comparison services that can be used to determine the extent to which objects share the same significant properties.

Preservation action services are the essential components in any preservation plan. They determine what can be done. For example, one preservation action might take a specific type of digital object and convert it to a new format; another might produce an emulation environment that enables users to interact with the original digital object using the original software application; another might provide a viewing tool that allows limited interaction with a digital object, but can be readily executed on a user's mobile phone.

Although some preservation plans will be simple, others will comprise complex multi-step work flows that involve extracting content from a repository, characterising it, using the results to select one or more services to treat, transform, or encapsulate the content, and then either returning the result to the repository with a detailed record of treatment, or providing a capability that can be used in a delivery environment so that end-users can get appropriate access.

Critically, content characterisation and preservation actions capabilities will come from a wide range of sources including open-source projects, commercial vendors, and third-party service providers. Thus, it is essential to enable these tools, applications and services to be incorporated.

The Planets ***interoperability framework*** enables the intimate relationship between preservation planning, preservation actions and content characterisation as well as extensibility with a plug-in approach. This is why the interoperability framework is depicted behind all of the other components. It provides essential shared services such as auditing and security in addition to an extensibility mechanism which allows plug-in of third-party tools and services.

Finally, the ***testbed*** provides a controlled environment that enables researchers and practitioners to conduct experiments and gather data on a scientific basis. This data can be used during preservation planning to select the most appropriate actions and characterisation methods.

Planets and the repository

Planets is not a repository project. Although there is some provision for temporary workspace required for analysing and manipulating digital material, Planets expects an institution to hold its content in a repository or archive system. Therefore, the Planets software components have been designed to work with a wide range of repository or archive systems. The most common usage pattern requires an adaptor to be defined for a specific product so that the Planets software is able to extract content and metadata, manipulate them, and possibly provide new content and metadata to the repository.

Figure 2 provides an overview of the interactions which the Planets software framework has with an organization's ingest, repository and delivery services, as well as with third-party tool and service providers. Typically, the content producers or consumers do not interact directly with the Planets software. Consider three scenarios:

1. In a content migration scenario, the Planets software would extract content from the repository, characterise it, select an appropriate migration service, apply the service to derive new content, validate the result, and then ingest the content back into the repository along with appropriate information about the process. The organization's delivery component is then able to provide access to the derived content to consumers as appropriate.
2. In a plug-in scenario, the Planets software identifies content that the organization's consumers are unable to effectively access; identifies a software tool, such as a browser plug-in, which would enable effective access; and packages the plug-in for the delivery component which provides it to consumers.
3. In an emulation scenario, the Planets software packages an emulator along with the required

software so that the delivery component is able to provide consumers with an environment which enables interaction with the content.

Deployment options

There are four very different ways in which the Planets software may be deployed.

1. A content holder may use the software directly. For example, an archive might deploy the Planets software locally and install an adaptor for their repository. This will enable them to build and execute preservation plans. Within this model, there are two important subsets. The content holder may install the software in a test or pilot environment on a single machine. In this case, ease of set-up is essential. Alternatively, the content holder may require an enterprise class installation. In this case, flexibility, scalability, distribution across multiple machines and network zones, security, reliability and integration are essential.
2. A service provider may deliver a suite of preservation services. For example, an integrated service provider might implement a suite of services to transform complex content, such as relational databases. In order to accomplish, they would need to implement Planets service interfaces. This task could be accelerated by using substantial pieces of the Planets software package.
3. A tool provider might wrap their preservation-relevant tools so that content holders could use them together with the Planets software. For example, a rich image conversion utility could be wrapped to provide a local Planets service that could be used in a wide range of preservation work flows. This requires them to use the wrapper framework.
4. A product developer might integrate preservation services into an application. For example, a repository vendor might 'preservation enable' its product by integrating it into the Planets preservation framework. This requires them to build various hooks into the application's existing work flows. Their customers could then either use an all-you-could-eat preservation service from a provider, a set of à la carte services, or install the software directly.

Progress to date

Almost two years into the project, Planets has now completed initial releases of each of the major software components in the Planets software framework, including:

- The Planets preservation planning tool (Plato), which enables organizations to make informed selection of the most appropriate preservation plan against requirements[7]
- The modular emulator[8], which allows 1980s software applications to run in a simulated computer environment, and the Universal Virtual Computer, which provides an alternative approach to emulation that allows interaction with software long into the future.
- The preservation characterisation registry, a core component that supports automatic identification of the essential characteristics of digital materials we wish to preserve.

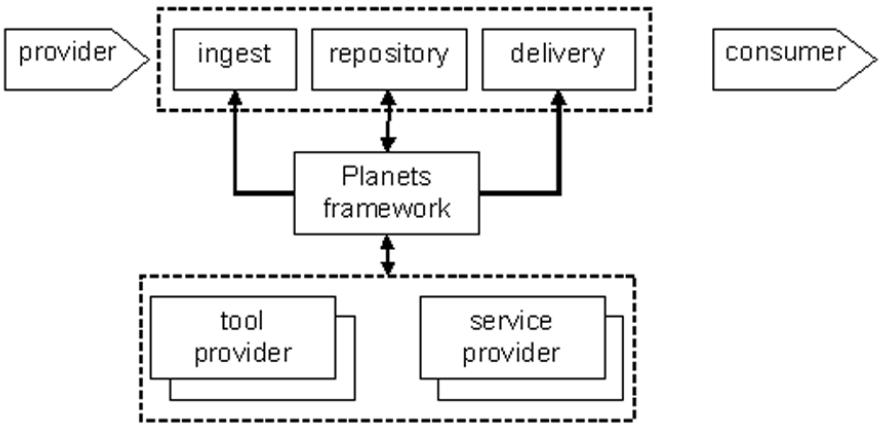

Figure 2. The Planets software framework interactions

- Extension to the XCEL-based significant property extraction tool, which now supports textual formats in addition to image, sound and other non-textual formats.[9]
- The Planets testbed, a controlled software environment that allows researchers and practitioners to conduct digital preservation experiments and gather data on a scientific basis.
- The Planets interoperability framework that provides shared functions and integrates the Planets tools and services into an easily managed preservation system. Its extensibility allows plug-in of third party tools and services.
- These initial releases will enable Planets partners to conduct case-studies and refine support for key preservation work flows. They will lead to public releases that members of the wider digital preservation community will be able to use.

Conclusion

The Planets software design is rooted in the vision of a click-and-install framework that meets the strict demands of larger enterprises for secure scaleable deployment, the demands of smaller organizations for ease of use, as well as of software vendors, and third-party service providers.

Planets is making excellent progress towards its goals of advancing, integrating and automating key digital preservation processes. The Planets digital preservation framework will enable organizations to improve decision-making about long-term preservation, ensure long-term access to their valued digital content and control the costs of preservation actions through increased automation and scaleable infrastructure. The extensible architecture will enable commercial tool and service providers to compete in a new market place for differentiated preservation services and tools.

References and notes

1. Planets: http://www.planets-project.eu/ (Accessed 3 June 2008)
2. Work presented in this paper is partially supported by European Community under the Information Society Technologies (IST) Programme of the 6th FP for RTD - Project IST-033789. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.
3. http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-preservation_en.html . (Accessed 3 June 2008)
4. The Planets partners include: The British Library, The National Library of the Netherlands, Austrian National Library, The Royal Library of Denmark, State and University Library, Denmark, The National Archives of the Netherlands, The National Archives of England, Wales and the United Kingdom, Swiss Federal Archives, University of Cologne, University of Freiburg, Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, Vienna University of Technology, Austrian Research Centers GmbH, IBM Netherlands, Microsoft Research Limited and Tessella Support Services Plc.
5. CASPAR: http://www.casparpreserves.eu/ (Accessed 3 June 2008)
6. DPE: http://www.digitalpreservationeurope.eu/ (Accessed 3 June 2008)
7. Strodl, S, Becker, C, Neumayer, R, & Rauber, A, How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 2007 Conference on Digital Libraries*, Vancouver, BC, Canada, June 18–23, 2007. JCDL '07. ACM: New York, 29–38. DOI http://doi.acm.org/10.1145/1255175.1255181. (Accessed 7 December 2007)
8. Dioscuri - the modular emulator: http://dioscuri.sourceforge.net/ (Accessed 3 June 2008)
9. Becker, C, Rauber, A, Heydegger V, Schnasse, J, Thaller, M, A Generic XML Language for Characterising Objects to Support Digital Preservation. In *Proceedings of the ACM Symposium on Applied Computing (SAC'08)*, Track 'Document Engineering'. Fortaleza, Brazil, March 16–20, 2008. http://www.ifs.tuwien.ac.at/~becker/pubs/becker_sac08.pdf. (Accessed 17 April 2008)

Acknowledgements

Planets is co-funded by the European Union under the Sixth Framework Programme under contract IST-033789. It is a substantial collaborative project that builds on and brings together the work of many talented individuals contributing from a consortium of committed organizations.

A different version of this article was published in the International Journal of Digital Curation, 2008, 2(2).*

■ **Adam Farquhar**
Head of Digital Library Technology
e-Strategy and Information Systems
The British Library
Wetherby
West Yorkshire LS23 7BQ, UK
Tel: T:+44 (0)1937 546515
E-mail: Adam.Farquhar@bl.uk

■ **Helen Hockx-Yu**
Web Archiving Programme Manager
(formerly Planets Project Manager)
British Library
E-mail: Helen.Hockx-Yu@bl.uk

* Farquhar, A, Hockx-Yu, H, 'Planets: Integrated Services for Digital Preservation', *International Journal of Digital Curation*, 2007, 2(2), 88-98, http://www.ijdc.net/./ijdc/article/view/30. (Accessed 17 April 2008)