The International Journal of Digital Curation
Issue 2, Volume 2 | 2007

Planets: Integrated Services for Digital Preservation

Adam Farquhar,
Head of Digital Library Technology,
The British Library

Helen Hockx-Yu,
Planets Project Manager,
The British Library

November 2007

Summary

The Planets Project[1] is developing services and technology to address core challenges in digital preservation. This article introduces the motivation for this work, describes the extensible technical architecture and places the Planets approach into the context of the Open Archival Information System (OAIS) Reference Model. It also provides a scenario demonstrating Planets' usefulness in solving real-life digital preservation problems and an overview of the project's progress to date.

[1] 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.

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 through Networked Services[2], is a four-year research and technology development project that commenced in June 2006. It is 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 services and tools to help ensure long-term access to digital cultural and scientific assets. The Planets consortium involves sixteen partners across Europe and brings together expertise from national libraries and archives, leading research universities and technology companies. Planets is coordinated by the British Library.[3]

Funded under the same Framework Programme are CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval)[4], coordinated by the UK Science and Technology Facilities Council, and DPE (DigitalPreservationEurope)[5] , coordinated 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 coordination, cooperation 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 organisational 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

[2] Planets: http://www.planets-project.eu/

[3] 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.

[4] CASPAR: http://www.casparpreserves.eu/

[5] DPE: http://www.digitalpreservationeurope.eu/

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 recognised 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 on-going 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 Planets aims to deliver is a range of tools and services, in the form of a downloadable "click-and-install" software package, that allows the administration, configuration, and deployment of preservation services and workflows. The Planets software supports a number of key preservation functions and will include the following components:

- Preservation Planning services that empower organisations to define, evaluate, and execute preservation plans
- Preservation Characterisation services that can automatically analyse digital objects to establish significant properties
- Preservation Action services for rendering digital objects and retaining the identified significant properties
- A Testbed providing evidence-base for the objective evaluation of different protocols, tools, services and preservation plans
- An Interoperability Framework that integrates seamlessly the tools and services to provide one easily managed preservation system

Planets Architecture

The Planets software architecture 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 organisations for ease of use, as well as those of software vendors, and third-party service providers.

The Planets software will be deployed in a wide range of environments. Some larger organisations, such as national libraries or archives, have strict requirements that they place on any software that they deploy. During the analysis phase, it became clear that Planets must support multiple application servers, databases, and authentication infrastructure. Furthermore, it needed to provide standard interfaces for monitoring, auditing, and logging. Some organisations have strict confidentiality needs that preclude any dependency on external services that might enable prying eyes to identify what sorts of content they are holding. Some smaller organisations need software that could be deployed on a single machine, whereas larger organisations have strict policies governing databases, computation, and network access to repositories.

These considerations drove towards a flexible service-oriented architecture based on enterprise-quality components that could be deployed in a wide range of configurations.

Planets is not a repository project. Although there is some provision for temporary workspace that is 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.

Preservation plans may comprise complex workflows 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. Thus, Planets needs to support the specification and execution of complex workflows.

The state-of-the-art ingest protocols, repository software, delivery environments, and preservation tools and services are all emerging. Substantial development in all of these areas is expected over the coming years. As a result, the Planets architecture has to be evolvable and support radical extensions.

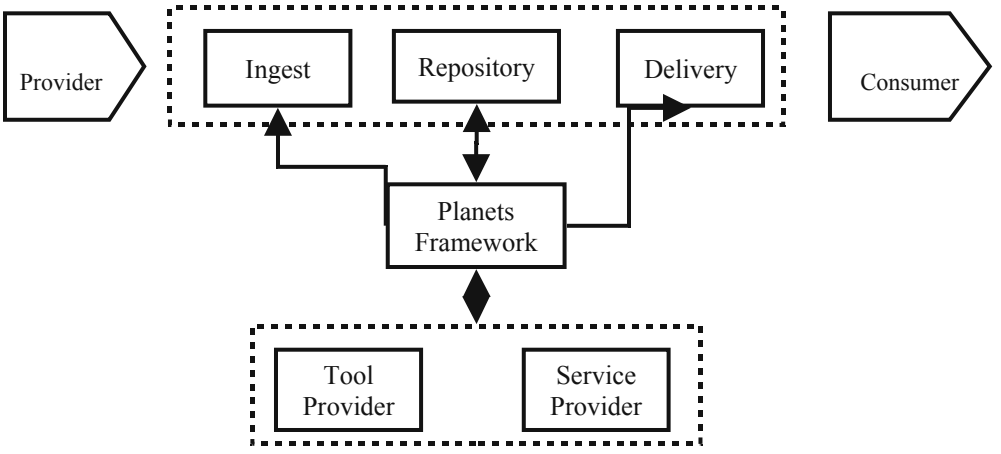

Figure 1: The Planets software framework interactions

Figure 1 provides an overview of the interactions which the Planets software framework has with an organisation'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 organisation'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 organisation's consumers are unable to access effectively; 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.

It is important to note that content objects may be compound and that migration may require a many-to-many mapping of files. It is also worth noting that emulation environments may be essential for both quality assurance for migrations and as an element of a migration process. Rothenberg has referred to this latter activity as "vernacular extraction" (2002).

Third-party service providers provide an important extension to the framework. The services can be very fine grained, such as characterising a specific type of file or performing a specific type of migration. They may also be coarser grained, such as executing an entire preservation plan. Services may be provided within an organisation, or be fully independently hosted and costed.

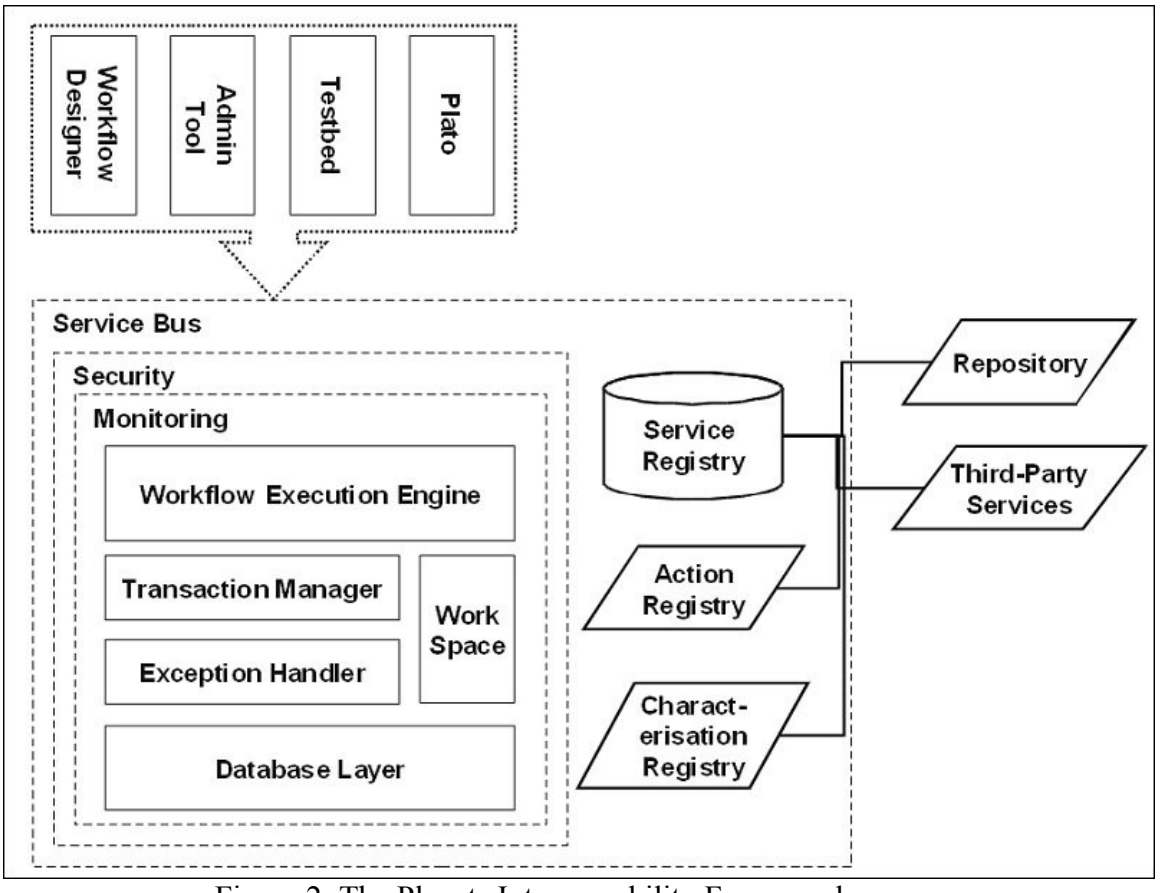

Figure 2: The Planets Interoperability Framework

Figure 2 illustrates the key components of the Planets Interoperability Framework and its relationship to Planets applications, repositories and third-party services. The Interoperability Framework establishes a service bus as well as essential shared services and components. These include the security component, which provides authentication and authorization services; the monitoring component, which provides flexible monitoring and logging services; the workflow execution engine, which takes workflows specified in the Business Process Execution Language (BPEL) and executes them in the context of the available Planets services; the transaction manager, which provides roll-back and compensations for complex transactions which may be

implemented by workflow elements; the exception handler, which provides a uniform set of services to register and handle exceptions that arise during service execution; a database or persistence layer; and workspace services so that workflows have appropriate levels of isolation. Services of every sort have basic definitions registered in the service registry. There are two important specialised registries – one for preservation actions, and one for content characterisation services. In addition, repository services and third-party services are also defined in the service registry.

The Interoperability Framework provides many of the key capabilities required to define and implement additional Planets application components. These include a workflow designer that provides preservation experts with the ability to define new preservation workflows in a graphical way and the administration tools that enable the framework administrator to configure components, define users, and allocate roles to them. The Plato preservation planning tool (Strodl, Becker, Neumayer, & Rauber, 2007) and the Planets Testbed application are also layered on top of the Interoperability Framework.

Preservation Planning in OAIS Reference Model and Planets

The Open Archival Information System Reference Model is a conceptual framework for a generic archival system which is committed to a dual role of preserving and providing access to information. It became an ISO standard in 2003 and has been widely adopted and used to inform the development of preservation tools and repositories. The reference model includes an OAIS Functional Model that describes six functional components which collectively fulfill the system's preservation and access responsibilities: *Ingest*, *Archival storage*, *Data management*, *Preservation planning*, *Access* and *Administration*. *Preservation planning*, the entity which designs preservation strategy based on evolving user and technology environment, is defined by OAIS as follows:

> **Preservation Planning**: This entity provides the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete. Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base. Preservation Planning also designs IP templates and provides design assistance and review to specialize these templates into SIPs and AIPs for specific submissions. . [sic] Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals. (Consultative Committee for Space Data Systems (CCSDS, 2002, p. 4-2)

A more detailed description of *Preservation Planning* includes a break-down into the following functions (CCSDS, 2002, pp. 4-13 - 4-14):

- Monitor Designated Community: responsible for tracking preservation requirements and available technologies through interactions with consumers and producers.
- Monitor Technology: responsible for tracking emerging technologies, information standards and computing platforms to identify risk of obsolescence.
- Develop Preservation Strategies and Standards: responsible for developing and recommending strategies and standards to enable better anticipation of changes in requirements or technology trends.
- Develop Packaging Designs and Migration Plans: responsible for developing information package (IP) designs, detailed migration plans, and their applications to specific holdings and submissions.

The above-mentioned comprises what is explicitly dedicated to *Preservation Planning* in the 148-page-long document describing the OAIS Reference Model. This leaves a lot to the imagination when it comes down to its practical implementation. There are references to additional requirements or functionalities scattered in different places in the document, but which are not categorised as parts of the six high-level functional components. An example is one of the mandatory responsibilities of an OAIS: "Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation" (CCSDS, 2002, p. 3-1). Many of these could be regarded as necessary requirements for *Preservation Planning* as they help define and implement the function. Sierman (2007) offers a detailed analysis of the additional functionalities and describes how they are treated within Planets.

Sierman (2007) also presents a Planets Functional Model which includes the following high-level functions:

- Preservation Watch: monitors content, users, organisation, producer and technological environment to provide preservation requirements, updates Representation Information and provides alerts as to when preservation action needs to be undertaken.
- Preservation Planning: evaluates requirements and selects the most appropriate preservation solution. It also requests development of new preservation actions in the absence of an appropriate choice.
- Preservation Action: Performs actions on digital objects to ensure continued accessibility. Also develops new tools (e.g. migration tools, emulators) upon requests from Preservation Planning.
- Preservation Characterisation: provides support to ingest and preservation action activities. It identifies file formats and extracts metadata and compares characteristics of digital objects before and after a preservation action.

Although comparison between the high-level functions in the Planets Model and the OAIS does not yield a one-to-one mapping, Planets in essence can be regarded as a practical implementation of the OAIS *Preservation Planning* function. Planets has brought some key processes, such as migration within the *Administration* function, into *Preservation Planning*. Relevant, but sometimes implicit requirements for

preservation planning in OAIS have been taken into account by Planets and brought together to form additional key processes.

Migration is the main digital preservation technique underpinning the OAIS reference model. Although hardware and software emulation have been discussed, they were at the time regarded as emerging but immature techniques worth significant comment (CCSDS, 2002, p. 5-12). Within Planets, the role of preservation actions that take into account the hardware and software environment required to interact with content is clearly identified. Emulation services may be used both in the end-user setting as well as during migration or performing quality control on preservation plans. Planets is progressing three distinct approaches to emulation. Firstly, it is essential to exploit the substantial cottage industry dedicated to developing emulators for specific machines. There is a wide range of such emulators, but making use of them can be complex. There are no established interfaces for the emulators themselves, so installing and invoking them can be time-consuming. Planets is developing a framework for describing, wrapping, installing, and invoking off-the-shelf emulation software, which includes commercial products and virtualisation tools. Secondly, many of the emulators have not been designed with their own longevity in mind. Planets is extending the Dioscuri modular emulation approach[6] to provide high-quality extensible emulation for specific families of hardware. Thirdly, high-quality emulators are software products themselves with reliance on specific hardware and software environments. Planets is extending the Universal Virtual Computer approach (Lorie, 2000) with models for peripheral devices to provide an even more durable basis for emulation.

It needs to be pointed out that at this stage of the project, Planets focuses on the preservation of digital content, rather than on the longevity of the digital repository holding the content, which has been addressed by the OAIS model. There are a number of areas within the OAIS *Preservation Planning* function which Planets does not yet fully address, such as packaging design and software prototyping. Comparison and alignment with OAIS is an ongoing area of work within Planets. It is hoped that the experience of Planets will lead to refinement or extension of the OAIS reference model.

A Scenario in 2010

Although Planets focuses on the needs of libraries and archives, it is hoped that the Planets preservation framework will become widely applicable and can be implemented by organisations other than libraries and archives which are responsible for safeguarding digital information. To demonstrate how Planets tools and services help solve real-life preservation problems, a delivery scenario for 2010 will illustrate what Planets could mean for a researcher looking for information on a specific topic online. This researcher, sitting at home behind her computer, browses e-journal articles published between 1999 and 2010. She finds a reference to an article stored at a European university. Firstly, the researcher requests the full-text version of the article, which was published in 1999 and was originally stored in PDF 1.1. The viewer for PDF 1.1 does not run on current computers and the document has effectively become obsolete.

[6] Dioscuri: modular emulator for digital preservation: http://dioscuri.sourceforge.net/ .

The university, like many others, has only a modest budget and cannot afford to employ digital preservation experts or develop its own preservation solutions. Maintaining a digital repository where the university's research outputs can be stored keeps the two full-time repository staff busy. Fortunately they have implemented the Planets preservation framework which has allowed the university easy access to a range of preservation tools and services produced elsewhere in Europe and around the world.

A choice is offered to the user as to how she wants to render the article using Planets technology. She can:

- Look at the article with the original viewer, running under emulation.
- Launch a viewing tool that renders the article on screen, but provides no other functionality.
- Read and print a migrated version of the article in current software.

A simple decision tool assists the researcher to make an appropriate choice based on her needs. The third option provides access to the significant properties required by the user, in a software environment with which she is familiar. The researcher views the migrated document, noting a useful diagram which illustrates some particular aspects of the work.

Each of the three solutions was created by a different developer. Each conformed to the Planets framework enabling the busy repository managers at the university to integrate them easily into their preservation and access workflows. The emulator was produced by one of partners in the Planets consortium, and was made freely available to other academic institutions. The viewer was written by an enthusiast and released as open source. The migration tool was written by a commercial company which had identified the potential market offered by Planets-compliant tools. The university employs this commercial tool under a national academic licence purchased by the country's higher education funding body.

The researcher would not have to go to the library where the University repository is based, because emulation-based viewing is offered through browser functionality online. She can also be sure that the migrated version of the article and the representation provided by the viewer is of maximum quality as the output of the tools has been independently evaluated and validated in the Planets Testbed.

Migration of the article was possible because a decision model identified the right tool for migrating PDF 1.1 to PDF 4.5 (the current version of PDF in 2010). The PDF contained an embedded diagram in a vector graphics format. This may not have migrated successfully in some tools. Fortunately, Planets automated characterisation tools ensured that this possible danger was identified, providing the decision model with the necessary data to select an effective migration solution.

In the article the researcher finds a reference to the underlying research data that were made available at the time of publication. In 1999 these data were deposited in a data archive run by the organisation that funded the original research.

Since the time of deposit, the archive has used the Planets Preservation Planning tools to characterise its collections, formulate appropriate preservation solutions and automatically record preservation metadata.

The researcher completes a retrieval request. In this request she has to fill in what properties of the data are essential for her and in what way she wants to re-use the data. Ideally the researcher would like to be able to interrogate the database as it was used way back in 1999. The request is sent to the data archive.

The researcher's request is analysed and an appropriate rendering solution is selected. In this particular case rendering the database including the original database management system (DBMS) and front-end interface requires specific software and operating system components that are no longer available. Fortunately Planets Preservation Planning tools have guided the archive's preservation staff to migrate the proprietary database to a common format, ideal for preservation. Viewing and interrogating the database on an emulated platform developed by Planets for database rendering would then be possible. The platform was none-trivial to develop but it provides support for a range of migrated database formats. By combining migration and emulation technologies, Planets Preservation Action technology made this difficult preservation process cost-effective, and straightforward for an existing digital repository to implement.

The data are retrieved and delivered to the researcher in the requested format and the emulation process with the database is launched. The researcher is able to query the database and examine in detail the research and context behind the e-journal article she first discovered.

Progress to Date

Work within Planets broadly follows an iterative and incremental process which begins with requirements specification and requirements analysis, followed by design specification, implementation and testing. In order not to reinvent the wheel and to take advantage of the current state of the art in digital preservation, an element of requirement specification and analysis is to survey the current landscape and carry out gap analysis. The intention is to identify the tools currently available which can be leveraged within the Planets Interoperability Framework. Each iteration over the project lifecycle is expected to result in a certain piece of software, ranging from a primitive prototype to the final software release, which is evaluated, consequently leading to a more elaborated and better tailored prototype. Early iterations of the software will not implement the complete functionality but include functions that are essential and thus achieve first useful results quickly.

Planets commenced in June 2006 and will complete in May 2010. Eighteen months into the project, Planets has completed the first iterative cycle and is in the process of refining requirements and specification, which will lead to more mature prototypes of the Planets software. Planets has made significant advances in developing the key components underpinning the Planets architecture, including registries for storing and accessing information about file formats, implementation of file migration tools as web services so that they can be included in distributed preservation workflows, and prototypes of the Plato preservation planning tool, the Testbed and the Planets Interoperability Framework.

Among the early results of Planets, it is worth highlighting two new XML-based languages for extracting and describing the essential properties of digital objects: the Extensible Characterisation Description Language (XCDL) and the Extensible Characterisation Extraction Language (XCEL).

Preservation planning relies on the characteristics of digital content held within a repository. Planets leverages excellent current tools, such as JHOVE for extracting file properties, and DROID for identifying file types. JHOVE takes a procedural approach which requires new code modules to be developed for each new format and existing modules to be extended or rewritten to extract additional properties. DROID uses a simple language that enables the identifying properties of many formats to be described. It does not extend to enabling properties to be extracted from specific digital objects. It is designed to answer the question "what is this?", not the questions "what properties does it have?". Preservation planning however requires the identification of many features of digital objects. For example, a page-oriented file format such as PDF might provide an excellent migration target format for office documents – but only if they do not use interactive features such as macros, complex formulae, or animated slide transitions. The Planets characterisation tools must be able to identify these and many other subtle properties.

This requirement has led to the development of XCDL and XCEL. XCEL moves well beyond the expressive power of the languages found in DROID and similar tools and aims at providing a general purpose capability for a very wide range of modern digital object structures. In addition to specifying the language, an execution engine has been developed that is capable of applying them efficiently to extract XCDL properties from specific objects. It is expected that these languages will support the next generation of DROID services (Heydegger, Neumann, Schnasse, & Thaller, 2006).

Conclusion

Planets is making good progress towards its goals of advancing, integrating and automating key digital preservation processes. The Planets preservation framework will enable organisations 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.

Acknowledgements

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

References

CCSDS - Consultative Committee for Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1: Blue Book. Retrieved December 7, 2007, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Heydegger, V., Neumann, J., Schnasse, J., & Thaller, M. (2006). *Basic design for the extensible characterisation languages*. Retrieved December 7, 2007, from http://www.planets-project.eu/docs/reports/Planets_PC2-D1D2_BasicDesign-Final.pdf

Lorie, R. (2000). *Long-Term Archiving of Digital Information*. Retrieved December 7, 2007, from http://domino.watson.ibm.com/library/CyberDig.nsf/7d11afdf5c7cda94852566de006b4127/be2a2b188544df2c8525690d00517082

Rothenberg, J. (2002, March/April). Preservation of the Times. *The Information Management Journal*, pp. 39-43. Retrieved December 7, 2007, from http://www.panix.com/~jeffr/Prof/Pubs/DigitalLongevity/arma.paper.from-journal.pdf

Sierman, B. (2007). *Report on Comparison of Planets with OAIS*. Retrieved December 7, 2007, from http://www.Planets-project.eu/docs/reports/Planets_PP7-D1_ReportOnComparisonOfPlanetsWithOais.pdf

Strodl, S., Becker, C., Neumayer, R., & Rauber, A. (2007). 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. Retrieved December 7, 2007, from DOI http://doi.acm.org/10.1145/1255175.1255181.