



Project Number	IST-2006-033789
Project Title	PLANETS
Title of Deliverable	White Paper: Representation Information Registries
Deliverable Number	D7
Contributing Sub-project and Work-package	PC/3
Deliverable Dissemination Level	External Restricted to other programme participants (including the Commission Services)
Deliverable Nature	Report
Contractual Delivery Date	31 st January 2008
Actual Delivery Date	29 th January 2008
Author(s)	Adrian Brown, The National Archives (UK)

Contributors

Person	Role	Partner	Contribution
Adrian Brown	Author	TNA	

EXECUTIVE SUMMARY

This document is a report on the state-of-the-art in the field of Representation Information Registries (RIRs). RIRs are widely recognised as a critical component of digital preservation architecture in general, and a number of such registries are being developed as part of the Planets architecture in particular. This document discusses the development of the concept of representation information, and of the use of registries as a means of exposing that information for use by digital preservation services; it describes the RIR implementations which currently exist or are under development globally; it assesses planned and potential future developments in this area; it discusses the role which RIRs play within the Planets project, and concludes with recommendations for future areas of research within Planets and beyond.

TABLE OF CONTENTS

1	Introduction	4
2	Representation information: the concept	4
	2.1 OAIS and representation information	4
	2.2 Formats and representation information.....	8
	2.3 Representation information and significant properties	8
3	The role of Representation Information Registries	9
	3.1 Drivers.....	10
	3.2 Use cases	10
4	Representation Information Registries: the current state-of-the-art	11
	4.1 Cedars Demonstrator.....	11
	4.2 PRONOM.....	12
	4.3 Library of Congress.....	14
	4.4 KB Preservation Manager.....	14
	4.5 FOCUS.....	16
	4.6 Representation Information Registry Repository	17
	4.7 Swedish File Format Registry	17
	4.8 National Geospatial Data Archive Format Registry	17
	4.9 Global Digital Format Registry (GDFR)	18
5	Planets Representation Information Registries.....	19
	5.1 The Characterisation Registry	19
	5.1.1 Introduction.....	19
	5.1.2 Content.....	19
	5.1.3 Functionality	20
	5.1.4 Future development	20
	5.2 The Preservation Action Registry	21
	5.2.1 Introduction.....	21
	5.2.2 Content.....	21
	5.2.3 Functionality	21
	5.2.4 Future development	22
	5.3 Providing integrated registry services.....	22
6	Conclusions and recommendations.....	23
7	References.....	25

Acknowledgements

Figures 1, 2 and 3: © Consultative Committee on Space Data Systems
 Figures 5 and 6: Based on illustrations © Koninklijke Bibliotheek

1 Introduction

This document is a report on the state-of-the-art in the field of Representation Information Registries (RIRs). RIRs are widely recognised as a critical component of digital preservation architecture in general, and a number of such registries are being developed as part of the Planets architecture in particular. This document discusses the development of the concept of representation information, and of the use of registries as a means of exposing that information for use by digital preservation services; it describes the RIR implementations which currently exist or are under development globally; it assesses planned and potential future developments in this area; it discusses the role which RIRs play within the Planets project, and concludes with recommendations for future areas of research within Planets and beyond.

2 Representation information: the concept

Any meaningful digital preservation activity requires some form of knowledge base regarding the technical environments necessary to support access to digital objects. As a concept, this has been recognised since the first significant analyses of the digital preservation problem¹.

However, the first comprehensive articulation of this concept arose from the work of the Consultative Committee on Space Data Systems to develop the Open Archival Information Systems (OAIS) Reference Model². The following discussion is therefore based upon OAIS concepts.

2.1 OAIS and representation information

Digital data has no inherent meaning – information can only be extracted from it through the correct interpretation of that data in accordance with some predefined knowledge base. As a simple example, a digital image in TIFF format can only be rendered as an image using software which has been designed to interpret the bitstream in accordance with the TIFF format specification. In other words, the logical *Information Object* (the image) can only be derived from the physical *Data Object* (the bitstream) via a process of interpretation. OAIS uses the term *Representation Information* to describe the knowledge base required for this interpretation. These relationships are illustrated in Figure 1:

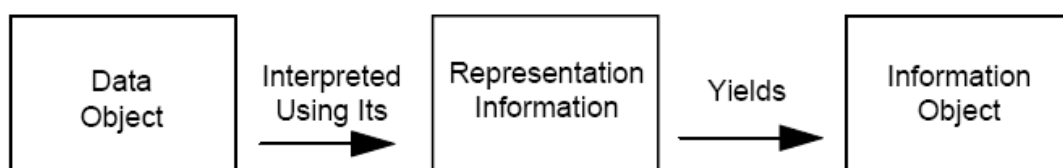


Figure 1: Obtaining information from data

It should be noted that this corresponds closely with the National Archives of Australia's Performance Model³, wherein the *Performance* (Information Object) is produced through the interpretation of a *Source* (Data Object) by a *Process* (Representation Information).

In OAIS terms, an Information Object therefore comprises both the Data Object, and all of the representation information required to interpret the data object in the context of the designated user community's knowledge base. The data object may be a physical or digital object, the latter ultimately decomposing to a series of bits. It is important to note that representation information is

¹ See, for example, Garrett & Waters (1996) p. 18

² ISO 14721:2003: Space data and information transfer systems -- Open archival information system -- Reference model

³ See Heslop, Davis & Wilson (2002)

itself a type of information object, which may exist in physical or digital form. Figure 2 illustrates these concepts:

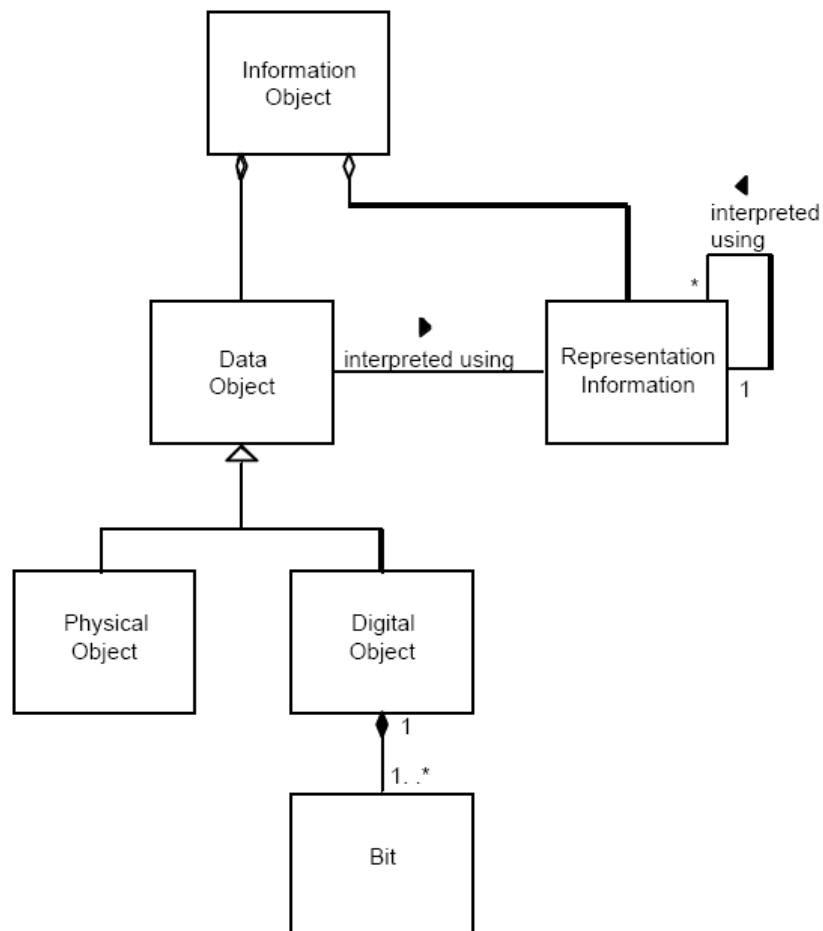


Figure 2: OAIS Information Object

Figure 3 expands on the concept of representation information:

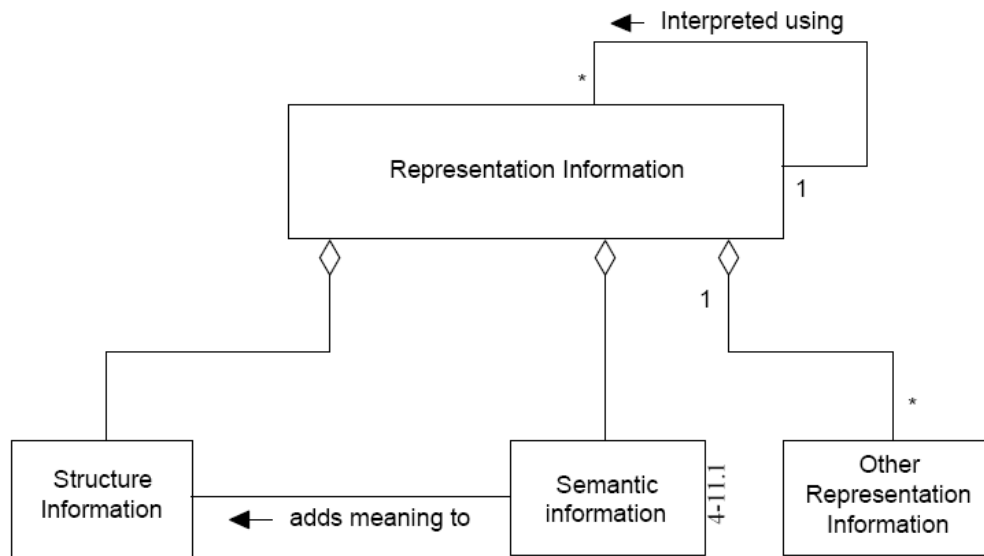


Figure 3: OAIS Representation Information Object

Representation information comprises structure information and semantic information. Structure information describes data types and structures which map the bit sequences of a data object to meaningful values. Semantic information describes how those values are to be interpreted. As a simple example, documentation of data in comma-separated values format requires not only the number, order and type of the values to be described, but also the interpretations which should be applied to those values.

As an information object, representation information may itself require further representation information for interpretation. Representation information therefore exists within recursive networks. For example, an image stored in TIFF format could be interpreted by reference to representation information comprising the TIFF specification. This specification could be represented in either physical (hard copy) form, or as a digital object, such as a PDF document. The PDF document in turn requires its own representation information – the specification for PDF. The TIFF image may implement JPEG compression, requiring additional representation information defining the JPEG standard. In addition to their dependent standards, each of these documents can only be interpreted correctly through knowledge of the character set and language used, and descriptions of these also therefore form part of the representation network.

In practice, a representation network must be terminated, and the point of termination is determined by the knowledge base of the designated user community. For example, it can reasonably be assumed that the user community of this document will have sufficient understanding of the Latin alphabet and English language to not require explicit documentation of these in its representation network.

OAIS also allows for two special types of representation information: *Representation Information Software* provides a means to interpret representation information (e.g. a PDF viewer could be used to access the PDF version of the TIFF specification), and *Access Software* provides a means to interpret a Data Object. The software therefore acts as a substitute for part of the representation information network – a PDF viewer embodies knowledge of the PDF specification, and may be used to directly access a data object in PDF format. The distinction which OAIS makes here seems somewhat arbitrary: since representation information is itself simply another Information Object, it is unnecessary to distinguish between the two types of software – it is enough to say that representation information may take the form of software, which may be used to interpret a data object.

The OAIS model primarily discusses representation information as a purely descriptive entity; although the inclusion of two specific categories of software implies that it can also encompass technologies which instantiate these descriptions, OAIS does not discuss this important distinction

directly. However, for the purposes of this paper, it is proposed that representation information be explicitly defined as encompassing either information which describes how to interpret a data object (such as a format specification), or a component of a technical environment which supports interpretation of that object (such as a software tool or hardware platform). These two classes can be termed *Descriptive Representation Information* and *Instantiated Representation Information* respectively.

OAIS also omits any discussion of elements of the technical environment, other than software, which are equally essential to the interpretation of a data object, such as hardware platforms, and media types. A more useful distinction than that provided in OAIS might be to classify representation information by type. For software, a basic division can be made between application software and operating systems. From a preservation perspective, it is possible to further distinguish between application software which offers different types of preservation service. A possible extension of the OAIS model along these lines is illustrated in Figure 4:

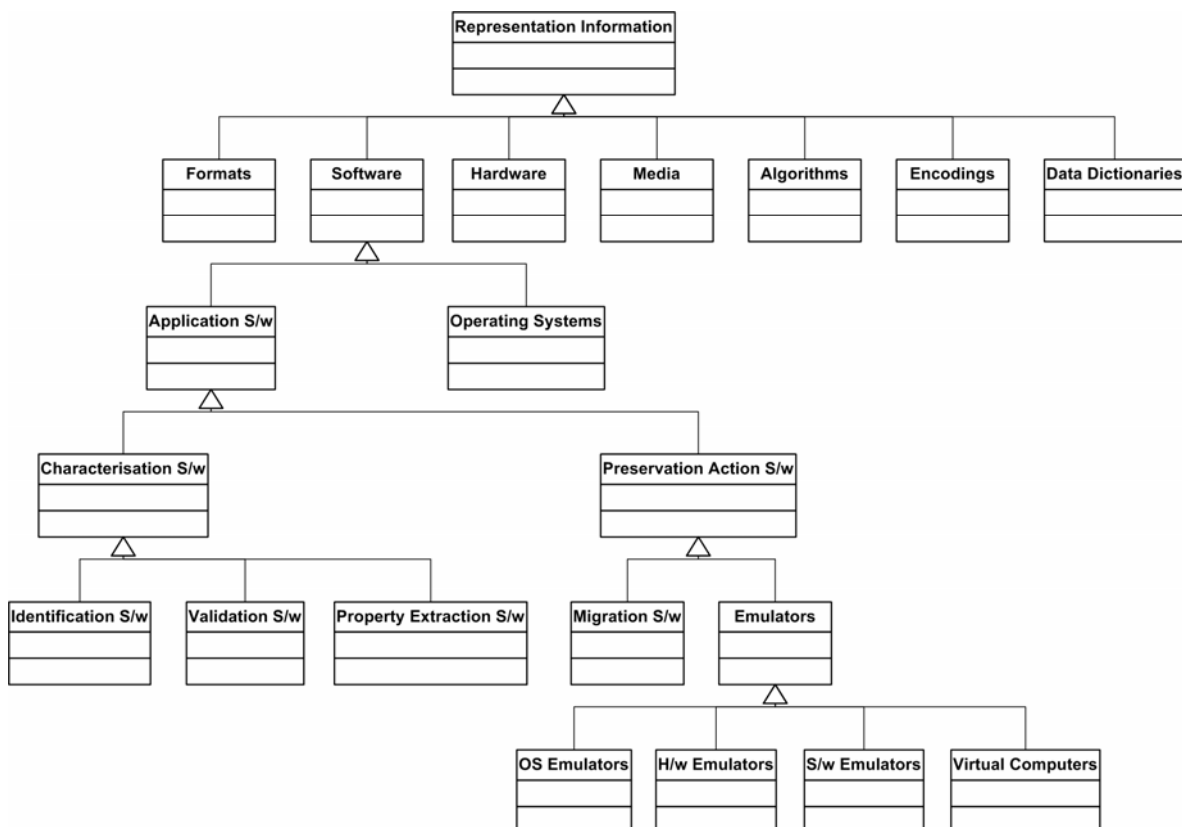


Figure 4: Example extended categories of representation information

This is not intended to be exhaustive, merely indicative of how representation information may be categorised to greater levels of granularity. This provides an orthogonal view to the OAIS classification of *structure*, *semantic*, and *other*. The last of these is not a very helpful category, and the first two are largely encompassed by the *formats*, *algorithms*, *encodings*, and *data dictionaries* categories. The Cedars Project proposed an extension to OAIS along similar lines in 2000⁴, which distinguishes between *data format definitions*, *render/analyse engines* and *platforms* (see 4.1). These map well to Figure 4, which particularly illustrates how the categorisation of preservation tools employed by Planets may be used to further decompose these render/analyse engines. Similar extensions are also discussed in Giarretta, et al (2005), and in research by the Koninklijke Bibliotheek on Preservation Layer Models (see 4.3). Further research is required to develop a more comprehensive model for representation information networks. Such a model is particularly required to underpin the metadata structures which repositories will use to reference RIRs.

⁴ See Holdsworth & Sergeant (2000)

2.2 Formats and representation information

There has been considerable debate regarding the importance of the format of an object in relation to other types of RI. For the purposes of this paper, the definition of a format proposed by the Global Digital Format Registry⁵ will be used:

“A byte-wise serialization of an abstract information model”.

The GDFR format model extends this definition more rigorously, using the following conceptual entities:

- Information Model (*IM*) – a class of exchangeable knowledge.
- Semantic Model (*SM*) – a set of semantic information structures capable of realizing the meaning of the *IM*.
- Syntactic Model (*CM*) – a set of syntactic data units capable of expressing the *SM*.
- Serialized Byte Stream (*SB*) – a sequence of bytes capable of manifesting the *CM*.

This equates very closely with the OAIS model, as follows:

- Information Model (*IM*) = OAIS Information Object
- Semantic Model (*SM*) = OAIS Semantic representation information
- Syntactic Model (*CM*) = OAIS Syntactic representation information
- Serialized Byte Stream (*SB*) = OAIS Data Object

Registries of technical information about formats are probably the earliest and most widely recognised incarnations of RIR. It is essential to recognise that the concept of representation information extends far beyond the format of a data object, and that knowledge of format alone is frequently insufficient to interpret a data object. This is particularly true with regard to structured data, such as scientific datasets. However, the format of a data object is perhaps the most universal type of representation information: without an understanding of this, further levels of representation information will be redundant.

The importance of format varies according to the type of data object. This arises from the extent to which all necessary structural and semantic information is integral to the format specification itself, or separate from it. For example, the structure and semantics of a TIFF image are determined entirely by the format, and are constant for all objects in that format. However, for data stored in a FITS file, understanding the format is an essential but preliminary step towards interpretation of the underlying information object, since the FITS format is essentially a container: interpretation of the content requires additional representation information, such as data dictionaries.

For the purposes of this report, format registries are therefore considered a subtype of RIRs.

2.3 Representation information and significant properties

The concept of significant properties must be considered in relation to representation information. The term was coined by the Cedars project in 2000⁶, although the concept has been considered very widely within the digital preservation community, principally from an archival perspective (e.g.

⁵ See Abrams (2007a)

⁶ Cedars Project (2002), Section 13

InterPares, National Archives of Australia⁷). Cedars introduced the concept of an *Underlying Abstract Form*, which appears to align closely with the OAIS *Information Object*, and encapsulates all of the significant properties of that object. The InSPECT project in the UK is currently developing a formal methodology for describing and measuring significant properties⁸, and its definition of the term will be adopted in this document:

“The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects.”⁹

InSPECT further categorises significant properties as applying to the following aspects of a digital object:

- **Content:** That which actually conveys information, not necessarily human readable, e.g. text, image, slides, programming code, etc.
- **Context:** Background information that enhances understanding of technical and business environments to which the digital objects relate, and the provenance (creator and subsequent changes in custody and ownership) of the object, e.g. who, when, why.
- **Appearance:** How the content of the object appears to an agent interacting with it, e.g. font and size, colour, layout, etc.
- **Structure:** The arrangement of component parts of the content of the object and how they relate to each other, e.g. embedded files, pagination, headings, etc.
- **Behaviour:** Functionality that is intrinsic to an object, e.g. hypertext links, updating calculations, active links, etc.

In OAIS terms, significant properties may therefore be defined as attributes of the information object, and are therefore quite distinct from representation information, which applies to the data object. Although this may be considered to place them outside the scope of any discussion of RIRs, it must be noted that some registries model both significant properties and representation information: PRONOM and the Planets Characterisation Registry are notable examples. This approach is founded on the principle that successful preservation requires access to both types of information. Although representation information is crucial to both preservation planning and accessibility, it is not sufficient to validate the success of a preservation action: this can only be achieved through an understanding of the significant properties of the preserved object. Registries which are intended to support the complete preservation process must therefore encompass both. The role of significant properties in such registries is discussed further below.

3 The role of Representation Information Registries

A Representation Information Registry (RIR) may be defined as a systematic collection of representation information objects or locatable references to objects held elsewhere. The RIR exposes these objects for discovery and processing by human or automated systems. RIRs may be designed to describe any class of representation information, or may specialise in a particular class, such as file formats.

This section identifies the major identified drivers and use cases for RIRs.

⁷ The NAA use the term *essence* in this context.

⁸ See www.significantproperties.org.uk/

⁹ Wilson (2007) p.7

3.1 Drivers

Key drivers for the establishment of RIRs include:

- **Efficiency of description:** Representation information forms a substantial element of the technical metadata required to describe digital objects in a repository. It would be redundant and inefficient for a repository to duplicate the complete representation information network for every object in its custody: instead, it can simply store pointers to the appropriate records in an RIR. The Planets PUID scheme is an example of how this can be achieved in practice¹⁰. A repository can simply store the PUID for the representation information which applies to an object in its local metadata, and use this to point to comprehensive technical information about that representation information in the Planets Characterisation Registry. An equivalent function is provided by the DCC's concept of *information labels*, which attach to a digital object, and contain pointers to the appropriate information in a RIR (see 4.5).
- **Knowledge sharing:** It is generally acknowledged that no single organisation has the resources or expertise to create and maintain information about every conceivable representation information network. RIRs have the potential to allow many organisations to collaborate in this activity. This possibility is being explored within a number of projects, most notably Planets and GDFR (see 5 and 4.8 respectively).
- **Sustainability and redundancy:** The creation of dedicated repositories for representation information can enhance the sustainability of that knowledge base, by decoupling it from the content repositories which it supports. Distributed registry structures can also provide additional redundancy, and increase confidence within the user community that it is secure to rely upon external RIRs to support local preservation infrastructures.

3.2 Use cases

RIRs support a number of use cases, which range from that of a passive repository of reference information, to the active enabling of a number of preservation processes. The latter potential has been most actively explored by The National Archives (UK) through its PRONOM service (see 4.1) and subsequently by Planets. The use cases which RIRs may support include:

- **Reference:** An RIR may serve as a passive repository of representation information, available for interrogation by human users or automated processes. Whether explicitly or by inference, an RIR also defines a boundary between the termination of the representation information network, and the knowledge base of the designated user communities. Irrespective of whether this boundary is correctly defined for a given RIR, this at least gives clarity to the assumptions about those communities contained within the RIR.
- **Characterisation:** Characterisation is one method by which the appropriate representation information for a digital object is determined - recording of the results of characterisation processes is therefore a key example of the reference use case. However, characterisation tools can themselves be viewed as elements of the representation information network: for example, a metadata extraction tool must be capable of correctly interpreting a data object, and extracts information from it which is independent of the data encoding. A registry can allow appropriate characterisation tools to be identified and executed. An example of this is provided by the Preserv project (see 4.1). Characterisation also supports the validation of preservation actions, by allowing the comparison of the significant properties of a digital object before and after the action has been undertaken.
- **Preservation planning:** Preservation planning encompasses all activities which identify the need to perform preservation actions, and the most appropriate actions to perform in order to meet specified objectives. RIRs can potentially support these activities in a

¹⁰ See Brown (2007b)

number of ways. Analysis of representation information can be used to trigger preservation actions, by identifying when elements of a representation information network are under threat. This may, for example, use some form of risk assessment methodology. Different types of representation information analysis, for example using objective trees, may then be used to evaluate alternative preservation actions.

- **Preservation action:** A preservation action tool, such as a format migration tool or an emulator, may be described as representation information. The justification for this is most obvious in the case of an emulator, which provides an alternative access environment for a data object. However, a migration tool can also be considered as such: it must be capable of interpreting a given data object correctly, and of transforming it such that the new data object will still yield the same information object. Registries which describe preservation action tools may therefore be identified as a discrete subtype of RIR; one such preservation action registry is being developed by Planets. These registries can also automate the identification and execution of tools.
- **Ingest:** RIRs may be used as part of the ingest of Submission Information Packages (SIPs) into an archive, most typically to perform initial characterisation of the SIPs.
- **Dissemination:** An RIR may support the dissemination of Dissemination Information Packages (DIPs) from an archive to users. This may take a number of forms. Firstly, it may enable the automated generation of a DIP, either by generating a DIP which has been transformed to a form accessible within the user's current technology environment, or by deploying an appropriate environment to access the object. This is essentially a specialisation of preservation action for dissemination purposes. Alternatively, it may provide information to the user about the required environment. Both options are potentially complex: preservation action for dissemination requires access to knowledge about the technological capabilities of the designated user community, which in turn has implications for preservation planning. The LOCKSS migration-on-demand demonstrator¹¹ illustrates one possible method, but much analysis remains to be done in this area. The alternative approach of specifying an appropriate access environment is also far from trivial: the generation of either a human-friendly or machine-interpretable description of the appropriate representation information network currently remains an open research topic.

4 Representation Information Registries: the current state-of-the-art

This section describes RIRs which have either already been implemented, or are at an advanced stage of development, in chronological order. It includes some prototypes and demonstrator projects which may no longer be actively maintained, but which have advanced the state of the art.

4.1 Cedars Demonstrator

The Cedars project developed a Distributed Archive Prototype demonstrator in 2000, which implemented an example representation information network¹². Cedars was probably the first project to examine many of the practical implications of the OAIS representation information model.

The Cedars demonstrator used a scheme of Cedars Reference Identifiers (CRIDS) as persistent unique identifiers for each element in the representation information network; these were initially implemented using a simple name resolving service, but were ultimately envisaged as being realised as Uniform Resource Names (URNs). The demonstrator implemented three main types of node, as follows:

¹¹ See <http://lockss.stanford.edu/index.html>

¹² See www.leeds.ac.uk/cedars/index.html

- **Data Format Definitions (DFD):** These might define a data format, a physical media type, or a software interface for a human user.
- **Render/Analyse Engines (RAE):** These transform data from one format to another, where a format is anything which can be defined by any DFD.
- **Platforms:** These are systems which provide the environments necessary for the execution of RAEs.

Each DFD node includes a list of RAEs capable of either receiving that format as an input, or delivering it as an output. Each node is further defined in terms of structure and semantic information. This model offers a simple but powerful approach to modelling representation networks¹³.

The demonstrator was intended to support distributed repository environments, and was tested within a number of organisations holding significant digital content, including the British Library, Birmingham University, Exeter University, University College London, and Manchester University.

At the present time, the demonstrator appears to be no longer available.

4.2 PRONOM

The PRONOM Technical Registry¹⁴, developed by The National Archives in the UK (TNA), can reasonably claim to be the first full implementation of an RIR to be deployed in an operational environment. First developed in 2002, and made freely available on the Web in 2004, PRONOM not only provides a registry of many classes of representation information, but also demonstrates the integration of an RIR within a wider digital preservation infrastructure: it forms the central component of the end-to-end digital preservation service developed by TNA as part of its Seamless Flow programme¹⁵. Within the Planets project, it forms the basis for both the Characterisation and Preservation Action registries (see 5).

The stated design objectives for PRONOM are as follows (ref):

- Providing an unambiguous and persistent means of describing the representation information required to access digital objects.
- Supporting the automatic characterisation of digital objects.
- Automatically identifying the formats of digital objects with sufficient precision and granularity to support digital repository management and preservation planning.
- Providing a rigorous basis for making preservation planning decisions.
- Supporting the identification of appropriate migration pathways for digital objects.

The PRONOM information model encompasses a wide range of classes of representation information, including formats, character encoding schemes, compression algorithms, application software, operating systems, and hardware. It enables direct description of a range of attributes for each type of representation information, but also allows description of, and linking to, other sources of representation information held externally, such as documentation, or records in other RIRs. Formats do play a central role in the PRONOM model, reflecting the nature of the digital objects which predominate within The National Archives' collections. PRONOM is consequently sometimes

¹³ See Holdsworth & Sergeant (2000)

¹⁴ See www.nationalarchives.gov.uk/pronom/

¹⁵ See www.nationalarchives.gov.uk/electronicrecords/seamless_flow/default.htm

characterised as purely a format registry; however, it can reasonably be categorised as a full RIR, albeit one which focuses on unstructured and semi structured data.

PRONOM also supports the modelling of significant properties, which can be described in terms of their expression within specific formats, and in terms of the capabilities of software tools (e.g. to characterise or migrate those properties in relation to specified formats). These capabilities are at a formative stage, but will be enhanced in response to the demands of projects such as Planets and InSPECT.

PRONOM is designed to actively support the full range of preservation processes, i.e. the use cases described in 3.2 above, and offers a range of services accordingly. The services which it provides have been abstracted from the underlying registry architecture via interfaces, to support both sustainability within TNA's preservation infrastructure, and interoperability with arbitrary third-party users, such as external repositories. At present, these interfaces have been implemented using web services and REST, although these could be translated to new technologies in future¹⁶.

TNA's DROID (Digital Record Object IDentification) format identification tool provides an example of the integration of a preservation tool with a registry service¹⁷. Developed in 2005, DROID uses internal and external signatures to identify and report the specific file format versions of digital files. Internal signatures are specific patterns of bytes which can be used to identify a format. These signatures are expressed as sequences of hexadecimal values, and can also incorporate wildcard operators. As a secondary method, DROID also attempts to match against external signatures, such as file extensions, although any identification based purely on these is accorded a low priority. The signatures are recorded in PRONOM, and exported as an XML file which is used by DROID. As new and updated signatures are added to PRONOM, users can configure DROID to automatically download updated signature files via a web service. DROID supports batch processing of large numbers of files. It is freely available to download under an Open Source license and, is written in platform-independent Java. It provides a graphical user interface, a command-line interface, and a public API for integration with other systems via web services.

PRONOM also implements an extensible scheme of PRONOM Unique Identifiers (PUIs), which provide persistent and unambiguous identifiers for records in PRONOM, such as file formats¹⁸. PUIs have been adopted as a recommended encoding scheme in the latest version of the UK e-Government Metadata Standard. PRONOM also provides a persistent resolution service for PUIs, which enables them to be resolved to the relevant PRONOM record in either human-readable XHTML or machine-readable XML. The PRONOM scheme forms the basis for the Planets PUID scheme.

PRONOM and DROID have now gained a level of international adoption, both within research projects and operational environments. Examples include:

- The National Archives of Scotland are using PRONOM and DROID within their production Digital Archive.
- As part of the JISC-funded PRESERV project¹⁹, TNA worked with Southampton University and partners to integrate DROID and PRONOM with the Eprints digital repository software. This has enabled the automatic identification of file formats on ingest, and also provided a format profiling component to the Registry of Open Access Repositories (ROAR)²⁰.
- Los Alamos National Laboratory used PRONOM to provide a format registry service as part of the NSF-funded Pathways project with Cornell University²¹.

¹⁶ See Brown (2007a)

¹⁷ See <http://droid.sourceforge.net/>

¹⁸ See Brown (2006)

¹⁹ See <http://preserv.eprints.org/>

²⁰ See Brody et al (2007)

²¹ See www.dlib.org/dlib/october06/vandesompel/10vandesompel.html

- The Los Angeles Times' Editorial Art Database uses PRONOM content to describe the format of digital objects in its collection.
- The Dspace digital repository software²² now provides integration with both PRONOM and DROID. MIT have also been active in supplying new and updated registry content.

TNA is a member of the Technical Working Group for the Global Digital Format Registry project (see 4.8 below). If the project bears fruit, it is anticipated that PRONOM will participate as one of the first nodes in the GDFR network.

4.3 Library of Congress

The Library of Congress maintains web-based, structured information about a wide range of formats²³. Although essentially a set of web pages designed for human access, this can be considered an RIR according to the definition provided in Section 3. The recent adoption of an XML-based approach will support automated metadata harvesting, and thus a degree of M2M interoperability. The registry was initiated in 2004, and is one of the most actively maintained registries. It currently contains information on c. 200 formats, although this number includes multiple versions and profiles of formats (for example, it lists 33 variants of MPEG4), and also includes codecs.

The LoC registry provides considerable depth of content in many cases, with a particular emphasis on assessing the suitability of formats for local use within the Library. Each format record is identified with a unique identifier of the form 'fdd' followed by a six digit number. It is understood that there have been some discussions between LoC and GDFR regarding content sharing, however the LoC registry does not currently reference PRONOM content or PUIDs.

4.4 KB Preservation Manager

The Koninklijke Bibliotheek has developed a facility for storing digital objects, the e-Depot, using IBM's DIAS system. In 2003, the KB started developing a preservation system for the e-Depot, based on a joint study by the KB and IBM²⁴. This system consists of a Preservation Manager, a Preservation Processor and tools for permanent access. The Preservation Manager uses technical metadata to manage and control the long-term preservation of the digital objects stored in the e-Depot, and is a form of RIR.

The Preservation Manager describes the technical environments required to support access to digital objects stored in the e-Depot, using two key concepts: *Preservation Layer Models* (PLM) and *View Paths*. A Preservation Layer Model describes the different layers of technology required to access a digital object. A typical PLM might consist of a data format layer, a viewer application layer, an operating system layer, and a reference platform layer. Each layer can be described using attributes such as "operating system name", "operating system version" and "operating system patch level". An example PLM is illustrated in Figure 5:

²² See www.dspace.org/

²³ See www.digitalpreservation.gov/formats/index.shtml

²⁴ See van Diessen (2002)

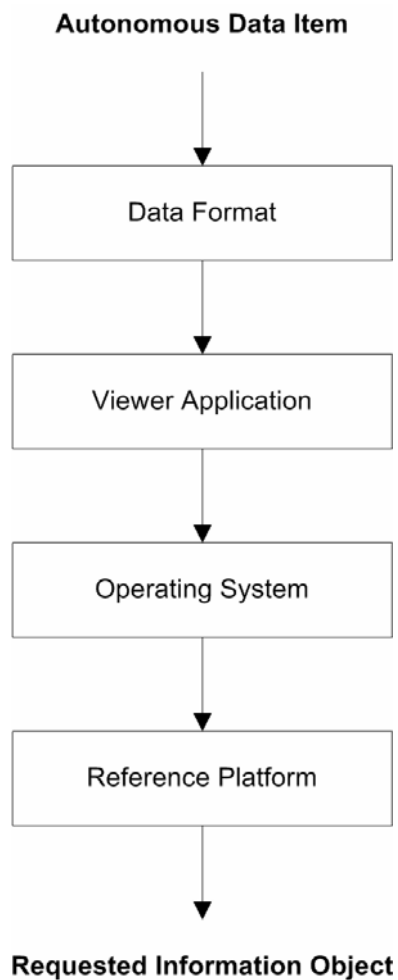


Figure 5: Example PLM

A PLM therefore represents a form of classification model for representation information networks, which extends the OAIS model as discussed in 2.1. A View Path describes the PLM required for a specific object. An example View Path for an object in PDF 1.2 format is illustrated in Figure 6:

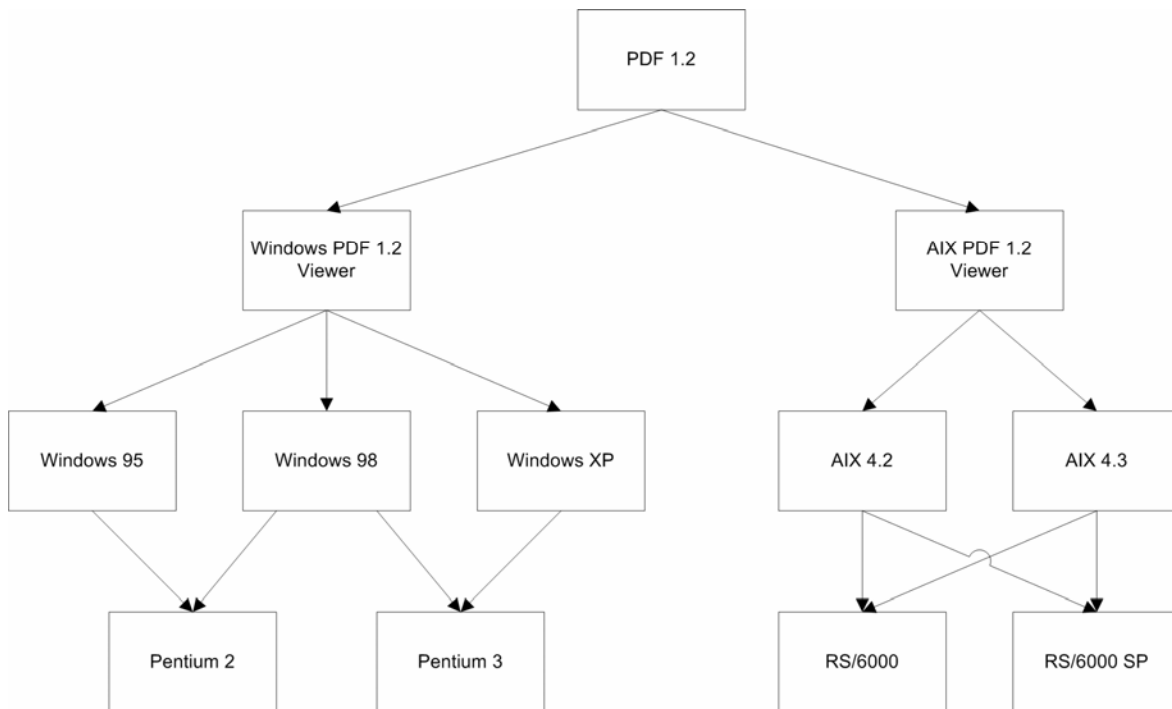


Figure 6: Example View Path

A View Path therefore describes the representation information network for a specific digital object.

Although it shares many similarities with other RIR implementations, the work of the KB has several unique elements. Firstly, although few would dispute that emulation tools should be considered as elements of representation information networks, the Preservation Manager is probably the only RIR to address the modelling of emulation-based networks in any depth. Secondly, the KB plan to use analysis of view paths as the basis for assessing risks to digital objects. This risk assessment is based on the number of view paths currently available, and is comparable with the PRONOM risk assessment service, although the later considers a wider range of variables.

A stand-alone version of the system was developed in 2004, and is currently being integrated with the main DIAS system.

4.5 FOCUS

FOCUS, the Format Curation Service²⁵, was a demonstrator project developed in 2004 by the Institute for Advanced Computer Study, University of Maryland (UMIACS), and funded by NDIIPP under the DIGARCH program. The objective was to demonstrate a scalable and secure environment for a global digital format registry, based on extensible and proven web technologies, such as LDAP and web services. The registry was designed to support a range of preservation services, including:

- Identification of file formats
- Verification of formats
- Delivery
- Transformation
- Risk assessment
- Characterisation

²⁵ See www.umiacs.umd.edu/research/adapt/focus/

Despite some differences in terminology, these correspond almost exactly with the use cases identified in 3.2. Interestingly, it would appear that the basis for defining the services in FOCUS was provided by the GDFR data model²⁶, which in turn was derived from the PRONOM information model²⁷.

Unfortunately, as of January 2008, the demonstrator service is no longer accessible, although some documentation is maintained on the website.

4.6 Representation Information Registry Repository

The Digital Curation Centre began developing its Representation Information Registry Repository (RIRR) in 2005, which is intended to fully implement the OAIS model. Its classification scheme extends the OAIS model along similar lines to those discussed in 2.1 (see Figure 4). The RIRR model is based on two key concepts: *Curation Persistent Identifiers* (CPIDs) and descriptive *labels*. A CPID is a unique identifier for an information object: this may be an item of representation information in the registry, or an object being preserved in a repository. In the former case, it may be considered equivalent to a PRONOM PUID. A label provides a means of describing and structuring multiple elements of representation information which relate to a digital object, by referencing their CPIDs. Thus, a digital object in a repository can be associated with a label, which points to items of representation information in the RIRR. Those items of RI may in turn point to further items, in a recursive structure. The recursion is terminated when the item of representation information refers to an assumption about a defined knowledge base. The representation information network for the original object is therefore expressed as the set of representation information items yielded by recursively following the pointers from its label.

It is envisaged that a network of interoperating RIRRs could be established, in a similar manner to the GDFR network. A prototype of the registry is now available²⁸, which has been implemented using ebXML, and specifically the freebXML Registry Reference Implementation Project (ebxmlrr)²⁹. Further development is expected to take place as part of the Caspar project³⁰.

4.7 Swedish File Format Registry

In Sweden, the Electronic Publishing Centre at the University of Uppsala has released a prototype format registry³¹. This was developed in 2005 within the preservation strand of the SVEP project, which is aimed at promoting, coordinating and supporting electronic publishing at Swedish universities and university colleges. The registry is still under development, and was partially populated using an export from PRONOM.

4.8 National Geospatial Data Archive Format Registry

In 2006, the National Geospatial Digital Archive in the US established a prototype registry for geospatial formats³². Aimed at the expert community in geospatial data formats, it uses a moderated, wiki-based approach to creating format records, using a standard template. Once approved, each record is exported in XML format for storage in the archive. This approach may offer a useful exemplar when considering governance models for community-based registries, such as GDFR.

²⁶ See Abrams (2007b) p.39

²⁷ See Brown (2005)

²⁸ See <http://registry.dcc.ac.uk/omar/>

²⁹ See <http://ebxmlrr.sourceforge.net/>

³⁰ See www.casparpreserves.eu/

³¹ See <http://svep.epc.uu.se/ffr/>

³² See http://ngda.library.ucsb.edu/format/index.php/Main_Page

4.9 Global Digital Format Registry (GDFR)

The Global Digital Format Registry project has undergone a lengthy genesis. An initial meeting of invited international participants was held in Philadelphia in 2003, at the instigation of Harvard University Library. Although a limited demonstrator (FRED)³³, based on the existing TOM service, was developed in 2004 to showcase some of the principles under discussion, the funding necessary to begin serious development work was only secured in 2005, via a grant from the Andrew W. Mellon Foundation. Since that time, work on the project has progressed to the stage whereby prototype software should soon be available.

The scope of GDFR was initially limited to the description of formats. However, the current data model does extend into other categories of representation information. GDFR has very much been designed to meet the reference use case for a registry, although it does have the potential to offer more dynamic support of preservation activities. The distinguishing feature of GDFR is that it is intended to provide an interoperable network of registries, whereby individual registry nodes can submit and share content. If this vision can be successfully realised, it offers huge advantages for knowledge sharing: the concept of a distributed network of specialist registry nodes, each contributing representation information expertise from their own domain, is a very attractive one. However, the practical and political difficulties of establishing and maintaining a sustainable governance structure for such a network are formidable – it remains to be seen whether GDFR can resolve these.

A number of innovative ideas have already emerged from the GDFR project. The work on defining a rigorous format model has already been discussed in 2.2. GDFR has also introduced a faceted classification scheme, which allows for sophisticated typing of formats³⁴. Each format registered in the GDFR may be associated with a set of classification entries. Each entry is specified in terms of a facet type and value. The currently defined facets are:

- **Genre:** This indicates the broad type of content associated with a format. This facet could be equated to an ontology of information object classes
- **Role:** This indicates the ontological role of the classified format
- **Composition:** This indicates the compositional nature of the classified format, e.g. whether it is an atomic entity or a container
- **Form:** This indicates the nature of the format encoding, e.g. whether text or binary encoded
- **Constraint:** This indicates the organizational nature of the classified format, e.g. whether structured or unstructured
- **Basis:** this indicates the nature of the representation of content used by the format, e.g. whether information is represented by sample or notational values
- **Domain:** This indicates an intellectual domain in which the format is commonly used, which equates to an OAIS designated community
- **Transform:** This indicates a format that defines a transformation on the form of content, such as compression or encryption

This scheme is likely to be further refined as a result of operational use.

GDFR and PRONOM share many common features, and each has built on concepts introduced by the other. For example, the GDFR data model is based on the information model for PRONOM³⁵.

³³ See <http://tom.library.upenn.edu/fred/>

³⁴ See Abrams (2007c)

³⁵ See Brown (2005)

Conversely, PRONOM is adopting the faceted classification scheme introduced by GDFR. GDFR is introducing its own unique identifier scheme, the current draft of which is based on the PRONOM PUID scheme.

The designers of GDFR have indicated³⁶ that it would be comparatively simple to implement a GDFR interface on existing registries such as PRONOM. The progress of GDFR is being closely monitored by many institutions, some of whom are also contributing directly through membership of its Technical Working Group. Planets is represented on this through the British Library and TNA.

5 Planets Representation Information Registries

The Planets project is currently implementing two distinct RIRs: a Characterisation Registry and a Preservation Action Registry. In both cases, these are being developed as enhancements of the PRONOM registry. The role of each registry is discussed in the following sections:

5.1 The Characterisation Registry

5.1.1 Introduction

The Planets Characterisation Registry implements an RIR designed specifically to support the reference, characterisation, and preservation planning use cases identified in 3.2. It also indirectly supports the ingest and dissemination use cases.

5.1.2 Content

For the reference use case, the registry provides a passive knowledge base of broad scope. It can describe the following classes of representation information:

- Formats
- Encoding schemes
- Algorithms
- Software (applications and operating systems)
- Hardware platforms
- Storage media

In each case, the registry contains a core record, but also allows linking to external resources, such as externally held documentation or entries in another RIR. Each record is identified by a typed PUID.

The PC registry also acts as a tool registry, describing the characterisation tools which may be used to perform specific characterisation functions for given formats. Each tool recorded in the registry is categorised according to the functions which it can perform for a specific format. The characterisation functions currently supported in this way are identification, validation, metadata extraction, and embedded object extraction. The registry also supports the automated deployment of characterisation tools which are available as services within the Planets interoperability framework.

The PC registry is intended to support specific characterisation tools, by providing data required for the execution of those tools. The following examples illustrate this:

- **DROID signatures:** The PC registry stores the internal and external signatures used by DROID to identify formats, and makes these available via automatically generated XML files which may be downloaded using web services.

³⁶ See GDFR (2007) pp. 4-5

- **XC*L:** The PC registry will store the XCEL descriptions which are used by the Planets XCEL extraction tool, and provide them for download in a similar manner to DROID signatures.
- **XML schemas:** The PC registry can store arbitrary XML schema documents associated with a format. These are used by the XML validation component of the Planets characterisation framework.

The registry has the capability to describe significant properties, although this aspect of the data model has potential for further refinement in the light of new research within Planets, and external projects such as InSPECT. At present, significant properties are associated with particular formats, and can also be related to tools. For example, it is possible to model the fact that TIFF images have a significant property of 'Image Width', and that this property can be measured using the JHOVE TIFF module. In future, it would be desirable to introduce entities to model abstract classes of information object, such as 'Still Image', and attach the significant properties to these, rather than to formats; formats would then be associated with the appropriate object classes. Properties are also currently modelled as simple name/value pairs, but this could potentially be extended to allow the description of additional attributes, such as constraints and measurement tolerances.

5.1.3 Functionality

The other use cases supported by the registry can be summarised as follows:

- **Characterisation:** The registry exposes an interface to allow the automated selection and execution of appropriate characterisation tools. In a typical usage scenario, this would begin with a request for available identification tools (of which DROID is the only example currently supported). Once the identification tool has executed, the registry can be queried to return a list of other available tools, based on the identified formats (expressed as PUIDs).
- **Preservation planning:** The registry incorporates a risk assessment service developed to support the Planets preservation planning service (PLATO). Risk assessments are calculated at two levels. Generic format risks are calculated using a set of standard criteria, based on format properties recorded on the registry (e.g. use of open standards, level of tool support, age). These risks can be modified for a given object based on specific properties of the format: for example, PDF files have an associated property which reflects their ability to support encryption. For objects where this property is true, their *instance risk* is increased above the generic risk which applies to PDF, to reflect the additional preservation risks which encryption entails. The registry exposes a web service to return current risk scores, which may be invoked either for a generic format (using the PUID) or for a specific instance (by supplying additional parameters to describe applicable property values). These risk scores can be then be used as parameters for specific nodes in decision trees constructed using PLATO.
- **Preservation action validation:** The registry exposes an interface to allow the automated selection and execution of appropriate characterisation tools to support validation of preservation actions. In future, this could be extended to provide other information required by validation tools, such as property measurement constraints.

5.1.4 Future development

The PC registry is being developed as an enhancement of PRONOM. The first iteration was released in 2007, with a second iteration due in 2008. It is likely that at least one further iteration will be developed before the end of the project.

5.2 The Preservation Action Registry

5.2.1 Introduction

The Planets Preservation Action Registry will implement an RIR devoted specifically to the description of tools for performing preservation actions. This will meet the reference and preservation action use cases defined in 3.2 and, indirectly, the dissemination use case.

5.2.2 Content

In the context of Planets, a preservation action tool is a software program that performs a specific action on a digital object to ensure the continued accessibility of the object. Planets has defined a distinction between preservation actions which transform objects (e.g. migration tools), and those which transform the environments which support those objects (e.g. emulators). A pragmatic decision has also been taken by the sub-project to include access tools, such as viewers, in the latter category: although these do not strictly perform transformations, they occupy a position close to emulators in the spectrum of preservation actions, i.e. they provide an alternative environment for accessing the original data object.

The PA registry will support both categories of tools. It makes a further distinction between the description of software which could potentially be used for preservation action, and of tools which are available for deployment as services within the Planets infrastructure. In the former case, software descriptions are intended to support assessment of their suitability for use; in the latter case, this is extended to support automated deployment of the appropriate tool itself. This mirrors the existing model of the underlying PRONOM registry, which distinguishes between 'tools' (software which has been wrapped for deployment within the preservation action framework) and 'software' (which has not).

The preservation action registry will also serve as a source of information on preservation action tools for general users such as employees from institutions that are concerned with digital preservation. This serves the more generic reference registry use case.

The PA registry encompasses both representation information and significant properties: preservation action tools form a specific subtype of representation information, but the selection of appropriate tools depends in part upon knowledge of their capabilities with respect to maintaining the significant properties of objects. It is anticipated that the significant properties of specific content types will be described in the characterisation registry; the PA registry will then reference these property descriptions in relation to specific preservation action tools.

The PA registry will include the following categories of information:

- Information about tools (information about the creator of the tool, operating specifics, licensing information etc.)
- Information about pathways (e.g. specific input file format and specific output file format, required technical environments etc.)
- Information on how to invoke tools (for tools which are available as Planets services)
- Links to experiments and evaluations in the Planets Testbed

5.2.3 Functionality

The Planets preservation planning tool (PLATO) will make use of the preservation action registry for the planning and execution of preservation action plans. Registry information on the application of preservation action tools is therefore designed to meet the information requirements of the PLATO component that will enable preservation plans to be designed or executed. The precise

nature of this interaction requires further definition. However, the capabilities of preservation action tools form key nodes in PLATO decision trees, and it is anticipated that these nodes will be populated by reference to the PA registry. The registry will then support the testing of alternative plans through automated deployment of the appropriate tools within the Testbed environment and, finally, the execution of the agreed plan in a similar manner.

5.2.4 Future development

The first iteration of the PA registry is being developed as an enhancement to PRONOM, and is due for release in 2008. It is likely that at least one further iteration will be developed before the end of the project.

5.3 Providing integrated registry services

Although the development of Planets registry services, and indeed the other Planets services which will interoperate with them, is still at a comparatively early stage of development, it is possible to outline some scenarios in which these integrated services might operate. Figure 7 illustrates some of the potential interactions between a repository and Planets registry services:

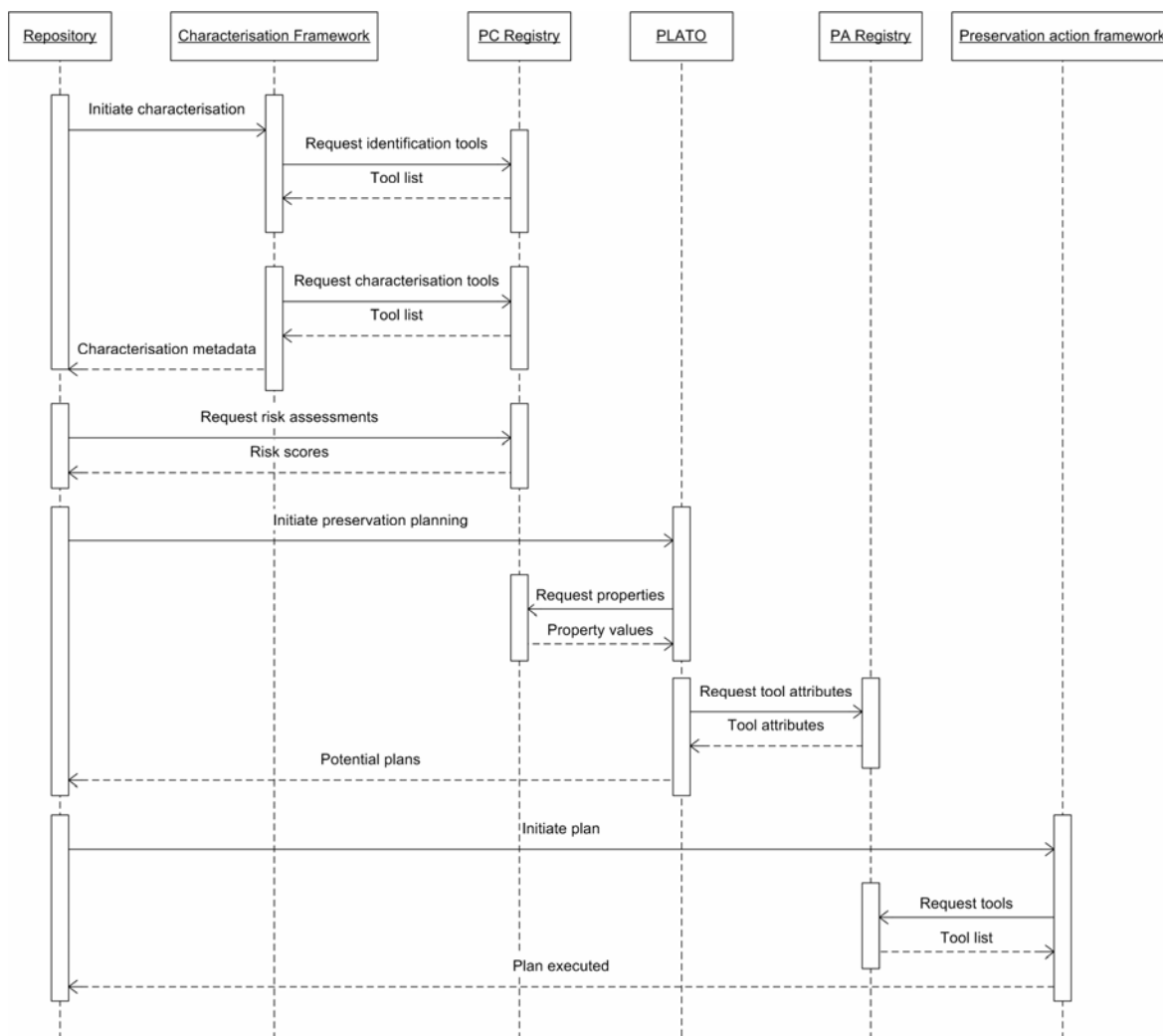


Figure 7: Example sequence diagram

A digital repository might begin by using Planets characterisation services across its content. The Planets characterisation framework will allow them to deploy a range of characterisation tools appropriate to their content. Typically, this would begin with format identification, using a tool such as DROID, which draws its signature file from the PC registry. Next, the characterisation framework will automatically determine which characterisation tools are available to validate and measure properties for the formats identified, and automatically deploy these tools. Again, the PC registry provides tool information to allow the framework to make these decisions. Depending on the format, one of the characterisation tools which might be deployed would be the Planets XCEL tool, which utilises XCDL descriptions stored in the PC registry. The output of the characterisation process will be a set of metadata documenting both the process, and the characterisation properties determined for each digital object processed.

Secondly, the repository might use PLATO to undertake preservation planning for the characterised objects. Although the precise interaction here has not yet been fully defined, this is likely to begin by using the PC registry risk assessment service to identify objects at risk. The repository will then use PLATO to develop decision trees for each object type. Some nodes of these trees will describe significant properties to be preserved, which will be populated by reference to the PC registry. Others will describe the required capabilities and behaviours of the preservation tools to be used, which will be populated from the PA registry. Once the decision trees have been analysed, candidate preservation plans will be evaluated in the Testbed environment, supported by automated tool deployment via the PA registry service. Once an approved plan has been determined, it will finally be enacted in a similar manner.

To validate the success of a given preservation plan, the repository could apply the Planets characterisation services to the resultant data objects, and then use the Planets validation service to compare the significant properties of the source and target. The PC registry supports these in exactly the same manner as for the original characterisation.

6 Conclusions and recommendations

This section draws together conclusions from the previous discussion, and makes a number of recommendations to be considered within future stages of Planets.

Although there is a clear distinction between representation information and Significant Properties, knowledge bases describing both are required to support preservation activities. Some RIRs, including the Planets PC registry, encompass both and offer a vision of a class of registry which extends beyond an RIR. The relationship between representation information and significant properties, and the role which registries can play in supporting these, are areas which warrant further research.

- **Recommendation 1:** Full account should be taken of the results emerging from the InSPECT project: specifically, the Planets conceptual data model, the characterisation registry, and the XC*L methodology should be revised to incorporate a more rigorous approach to modelling significant properties. This work should be coordinated by the PC sub-project. Planets should also feed its results into InSPECT, through partner representation on both projects.

There is clear scope for further development of the OAIS model of representation information networks, and of ontologies for classifying classes of representation information, and there are a number of current research activities in this area.

- **Recommendation 2:** Planets should develop a comprehensive classification scheme for representation information networks, building on existing work by TNA in PRONOM, and the KB on Preservation Layer Models. This work should be undertaken within the PC/3 workpackage, liaising closely with PP/7. The potential for collaboration with Caspar in this area should also be explored. This scheme could potentially be developed further to provide a standard syntax for associating digital objects with their representation information networks. Planets should develop ontologies for classifying categories of

representation information: one strand of this work could build on the GDFR faceted classification scheme for formats.

It is essential that the digital preservation community develops a practical and sustainable model for providing operational services which utilise current research into RIRs.

- **Recommendation 3:** The potential for developing full interoperability between the Planets registries and other emerging registries should be explored: TNA is already considering the possible development of an interface to enable PRONOM to participate as a node in the GDFR network, and this work could be taken forward within Planets to enable all Planets registries to participate. Such an interface could also provide the most effective means to enable multiple instances of PRONOM to interoperate, irrespective of future take-up of GDFR. Planets should also consider the potential for interoperability with the RIRR, which is being developed further as part of the Caspar project. This work should be taken forward by the PC/3 workpackage.
- **Recommendation 4:** As part of ongoing work to develop an appropriate sustainability model for Planets services beyond the life of the project, particular thought should be given to how the registry services will be maintained. TNA is committed to providing continued public access to its PRONOM service, which will include all Planets enhancements, but Planets should consider the potential to establish a distributed registry infrastructure, whether through existing project partners or other bodies.

7 References

ISO 14721:2003: Space data and information transfer systems -- Open archival information system -- Reference model

Abrams, S. (2007a) *Global Digital Format Registry (GDFR): Format Model and Relationships, Version 1.0.10*

https://collaborate.oclc.org/wiki/gdfr/images/e/e6/GDFR-Format-Model-and-Relationship-1_0_10.rtf

Abrams, S. (2007b) *Global Digital Format Registry (GDFR): Data Model, Version 5.0.10*

https://collaborate.oclc.org/wiki/gdfr/images/8/82/GDFR-data-model-5_0_10.rtf

Abrams, S. (2007c) *Global Digital Format Registry (GDFR): Classification, Version 1.0.5*

https://collaborate.oclc.org/wiki/gdfr/images/4/4f/GDFR-Classification-1_0_5.rtf

Brody, T., Brown, A., Carr, L., Hey, J. M. N. and Hitchcock, S. (2007) PRONOM-ROAR: Adding Format Profiles to a Repository Registry to Inform Preservation Services, *International Journal of Digital Curation*, **2** (2), 3-19

www.ijdc.net/ijdc/article/view/53

Brown, A. (2005) PRONOM 4 information model, The National Archives

www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_info_model.pdf

Brown, A. (2006) The PRONOM PUID Scheme: a scheme of persistent unique identifiers for representation information, version 2, *Digital Preservation Technical Paper*, **2**, The National Archives: London

www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf

Brown, A. (2007a) Developing practical approaches to active preservation, *International Journal of Digital Curation*, **1** (2), 3-11

<http://www.ijdc.net/ijdc/article/view/37/42>

Brown, A. (2007b) A scheme of persistent unique identifiers for formats, Planets project

www.planets-project.eu/private/pages/wiki/images/4/4e/PC3_D2_v2.doc

Cedars Project (2002) *Cedars Guide To Digital Preservation Strategies*

www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html

Garrett, J. and Waters, D. (eds) (1996) *Preserving digital information: Report of the Task Force on Archiving of Digital Information*, The Commission on Preservation and Access and The Research Libraries Group

www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf

Giaretta, D. et al (2005) *Supporting e-research using representation information*, Poster: Fourth E-Science All Hands Meeting, 19-22nd September 2005, Nottingham, UK

www.allhands.org.uk/2005/proceedings/papers/447.pdf

Global Digital Format Registry (2007) *Global Digital Format Registry: Technology Platform, version 0.2*

<https://collaborate.oclc.org/wiki/gdfr/images/5/55/GDFR-technology-platform.pdf>

Heslop, H., Davis, S. and Wilson, A. (2002) *An approach to the preservation of digital records*, National Archives of Australia

www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf

Holdsworth, D. & Sergeant, D. M. (2000) *A blueprint for Representation Information in the OAIS model*, Cedars Project

<http://esdis-it.gsfc.nasa.gov/MSST/conf2000/PAPERS/D02PA.PDF>

Wilson, A. (2007) *Significant properties report*, InSPECT Project
www.significantproperties.org.uk/documents/wp22_significant_properties.pdf

van Diessen, R. J. (2002) Preservation requirements in a deposit system, *IBM/KB Long-term Preservation Study Report series, 3*
www.kb.nl/hrd/dd/dd_onderzoek/reports/3-preservation.pdf