Open Planets Foundation

A New Registry for Digital Preservation

Conceptual Overview

January 2011

Commissioned by:	Maurice van den Dobbelsteen, Nationaal Archief
Author:	Bill Roberts, Swirrl IT Ltd
Reviewer:	Bram van der Werf, Open Planets Foundation
Date:	28 January 2011
Version:	1.1
Distribution:	OPF Board, OPF Members and OPF website

Contents

1. Objective

The objective of the Open Planets Foundation (OPF) is to facilitate the creation of a representation information registry solution that will support the current and future digital preservation needs of memory institutions in a flexible and cost-effective way.

In Autumn 2010 the OPF published an Outline Document to facilitate discussions on this subject around the iPRES conference (available through the OPF website). Based on feedback there and thereafter this document takes the process of making that solution a reality a big step further.

At the end of this document is a planning of what the OPF will do in the short-term. Your institution is invited to join in and support this process by participating in the discussions, providing development and/or conceptual support and by working to bring your internal systems in alignment with this vision.

It encompasses a new concept for representation information registries in digital preservation called the 'registry ecosystem'. It is based on interlinking various sources of information to create interconnected 'registry collections' using Linked Data, rather than creating and maintaining a single registry (yourself).

2. Background

Few will question the need for representation information registries in digital preservation. One main driver is to have the right information for preservation planning and actions. The second driver is a need for automation as collections are growing exponentially.

As will be described in more detail later in this document, the starting point is a file format registry. Quickly one arrives at other sorts of registries, such as migration pathway and tool information, policy information, technology watch information, registries holding tools wrapped into services, a registry with an ontology of digital object properties, a registry holding viewpaths and environments for emulation etc.

In the PLANETS project the original idea of separate Characterisation and Tool Registries was stepped up to the creation of a Planets Core Registry, designed to hold all sorts of information mentioned above. Having reached that goal, we realised at the end of the project that there would never be a single institution that could maintain all that information. A follow-up course of action would have to be based on a decentralised solution, taking into account all the valuable lessons we learned during the registry work in PLANETS.

It was clear that the technological foundation used and extended in PLANETS – the PRONOM Registry developed to a stable and robust version 6.2 – was never created with a decentralised solution in mind. Thus, although functioning very well for use in internal systems at TNA, Nationaal Archief and some other institutions, a new approach had to be found.

Thinking out that new approach, whilst taking into account the recommendations from PLANETS and new research that became available last year, led to a ground-breaking new concept for representation information registries in digital preservation. It is that concept of a registry eco-system we want to share with you in this document, as an invitation to become involved.

This process is now driven by several factors.

- First of all TNA will produce a totally new PRONOM that uses the modern technology of Linked Data (a W3C standard), which opens up a host of new possibilities.
- Generally there is a tendency to move to more 'lean and mean' interconnecting systems rather than 'big systems', a tendency driven by budget constraints and by lessons from the world of cloud computing.
- Thirdly the Nationaal Archief is actively pursuing automation through registry services for its digital repository which will see full production in 2011. Nationaal Archief offered to invest in OPF research to help drive the creation of new registry solutions.
- Finally, the founding OPF members acknowledged the unique position the OPF holds to make a new concept for representation information registries in digital preservation a reality.

These factors combined mean that there is now momentum for practical results.

3. Outline

The OPF wants to drive a new concept for representation information registries in digital preservation called a 'registry ecosystem'. It is based on interlinking various sources of information to create interconnected 'registry collections' using Linked Data, rather than creating and maintaining a single registry (yourself).

In a registry ecosystem information can be acquired from various sources with different levels of trustworthiness: crowd sourcing, information from institutions you trust and your own information. As an example, one might want to source information on Word from Microsoft and on PDF from Adobe. It could be interesting to enhance one's policy information with information from a related and trusted institution. Finally, there may be information on very new or very old items that can be sourced from the crowd.

The key would be to have a 'dashboard client' software that lets you bring together this information in a 'collection'. This client will be close in appearance to a peer2peer client, which offers control over what to download from whom and what to upload to whom, but is a database in its own right. It is a registry which holds information from various sources including information from one's own existing registry. The information is interconnected and exchanged by means of Linked Data, a W3C standard that uses HTTP URIs as identifiers for entities of interest, and represents information using RDF. The dashboard client will manage the process of retrieving updates from your chosen sources of information, and publishing new or changed information in your own registry.

Institutions can continue using their own registries, but this approach will allow them to expand their registries by exchanging trusted information. For this to work it is necessary:

- that institutions adapt their registries to become interoperable with this Linked Data registry ecosystem. It will be necessary to compile guidelines on what is necessary for an existing or planned registry to be able to download data from elsewhere and make information available in return.
- to build a prototype of the dashboard client to demonstrate its operation, working towards a production version. Involvement will be required from various institutions, both on the level of providing ideas and feedback and on the level of providing resources.
- to assemble a basic working collection of registry data, as a working example and to assist smaller institutions and organisations.

4. The need for a Representation Information Registry (RIR)

The Planets report PC/3-D7, written by Adrian Brown (then of the TNA), explains the role of Representation Information Registries in digital preservation. When we use the term file format registry, in most cases this is an informal way of referring to what is in fact a representation information registry. The file format is one component of the representation information. As indicated in Figure 4 of PC/3-D7, other aspects of the representation information can include software, hardware, encodings, data dictionaries and more.

The file format often plays a special role in the representation information, as it is a way of grouping together a set of files that share the same (or similar) representation information networks and can therefore share a common preservation policy. However, note that the file format is not the only determining factor in the preservation policies of most institutions: other characteristics of a digital object, or aspects of its context or provenance, may also have an influence on how it is managed.

Furthermore, Brown notes "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."

A representation information registry helps us answer key questions in digital preservation:

- given a digital object, do I have the ability to render it and understand its contents?
- If not, what software and supporting information do I need to make that possible?

The use of a representation information registry is closely related to the issue of file format identification: to know which information in the registry is relevant, it is necessary to identify what kind of file it is and the file format is usually the starting point for that identification process.

The file format registry is typically the place where signature and regular expression data for file format identification tools are stored.

5. Separating facts about formats from institutional policy

In Planets PC/3-D7, Brown lists a number of use cases for RIRs:

Reference (a passive repository of representation information)
Characterisation (recording results of characterisation processes)
Preservation Planning (analysing representation information to plan or trigger preservation actions)
Preservation Action (the registry may describe or invoke a preservation action tool)
Ingest (RIR may be used in the initial characterisation of incoming digital objects)
Dissemination (eg specifying software needed to access a disseminated digital object or invoking a preservation action tool to convert the object into a form convenient for dissemination).

All of these use cases are essential for an institution with a responsibility to preserve digital objects. The PRONOM system, and subsequently the Planets Core Registry (PCR) supported all of these use cases to varying degrees.

However these different use cases have very different profiles in terms of:

- The universality of the information (whether it is applicable to all institutions and all situations)
- The frequency with which the information is accessed
- The frequency with which the information changes
- The way that the information is created
- The way that the information is consumed in the ingest or preservation functions of a digital repository

For these reasons we should consider whether different components of the information held in the Planets Core Registry would be easier to manage if they were held in different systems.

Information about file formats and the software applications that can render them is relatively slowly varying information that only needs to be accessed relatively rarely. It can require significant amounts of research to compile the information and it has wide applicability as it is essentially factual in nature. It is therefore important to enable and encourage researchers and experts across the world to contribute and share this type of information.

6. Managing institutional policy information

PRONOM and its offspring the Planets Core Registry, hold information representing institutional policy choices, alongside factual information about file formats. An example might be a choice of which migration tool to use to convert TIFF files to JPEG. (The list of which migration tools are capable of converting TIFF to JPEG could be seen as factual information).

The characteristics of digital objects which must be preserved to maintain their authenticity may depend on the type, context or purpose to which the objects were put, and so cannot have a single 'correct' answer. Each institution will need to manage this type of information and the preservation policy registry is an appropriate place for it.

As argued in the previous section, there are reasons to separate policy information from factual information. If the policy information no longer belongs in our file format registry, then we must ask how best it should be managed.

It may be useful for institutions to exchange and compare their preservation policies, but it is essential that different institutions can make different choices. The type of information required may vary according to different approaches to workflows or different software systems and so is less generally applicable and the sharing process is therefore a lower priority.

This is an important topic that will be covered in more depth in a separate paper, to be produced in early 2011 by the National Archives of the Netherlands. This section presents a brief overview of the main requirements of a 'preservation policy registry'.

In most institutions that operate a digital object repository, the policy registry will be used by a workflow system to drive automated ingest or preservation processes, so the policy registry must have an API that can be used by other software.

The information in the policy registry is likely to be required every time a file is processed, therefore the speed with which the information can be accessed will be important. The workflow automation system must have rapid access to the contents of the policy registry, whether directly via the API, or via a cache of some sort.

The choices made regarding preservation strategies and the tools to implement them may have a significant impact on the ability of an institution to fulfil its obligations to preserve digital material. This implies the need for a formal process for populating the policy registry, ensuring that each aspect of preservation policy is backed up by sufficient research and a careful assessment of options. The tools and methodologies developed in Planets, notably PLATO and the Testbed, can play an important role in this process.

Note that the preservation policy registry will depend on information from file format registries, for example on which tools exist that can convert from file format A to file format B.

7. Registries and the ingest process

A key use of registries is during the process of ingesting a set of new digital objects to a digital repository. Typical steps in an ingest process include:

1. A set of new digital objects is received for ingest
2. Each object is processed to identify the file format (and possibly to apply other characterisation tools). The file format registry is typically the authoritative location to store file identification signatures, but the identification tool will probably need to keep its own copy of all signatures in order to work efficiently.

3. Based on the file format (and possibly other measured properties), check whether a policy is in place to deal with objects of this type. This will involve consulting the institution's preservation policy registry. In simple terms a policy for a class of digital objects might be "accept as is", or "reject" or "convert to format X and then accept" or "accept and create additional access copy in format Y"
4. If a suitable policy exists then apply it
5. If no policy exists, then a process must be initiated to create one. This is likely to involve expert human input, rather than being an automatic process. It will involve consulting the format registry (or possibly multiple distributed format registries) to find out for example which applications can be used to render an object, or which preservation action tools are available to transform it to an alternative format. The institution needs to make a choice about which format registries to consult. This involves awareness of which registries are available and some kind of evaluation of trustworthiness.

Therefore consulting the format registry can be a relatively rare occurrence but the possibility to be notified when a format registry is updated would allow an institution to update any local cache of file identification signatures or other information.

As hardware and software exist and rendering or migration tools may become obsolete, an institution may need to modify its preservation policies. This is another example of when it would be useful to be updated of changes to format registries.

8. Collecting and sharing information about file formats

There is a core of common formats which will cover a large proportion of material received by typical memory institutions such as archives and libraries. However there is a 'long tail' of rarer or more specialist file formats which may nonetheless be important. The overall effort to maintain information on all these formats is considerable, too much in general for any single institution to manage.

Therefore it is essential that the work of researching and documenting file formats can be shared amongst many institutions. A mechanism must be in place for experts on specific formats to easily share their knowledge with the broader community. This information may be produced by memory institutions, by academic researchers, by software vendors, by specific expert groups or by individual independent experts or interested parties.

Any system of format registries should be designed to enable and promote sharing of information from this broad spectrum of contributors and that should be the prime consideration in determining the design of a system to exchange information about file formats.

What is the minimum we have to agree on to enable efficient sharing of information around file formats?

To help answer that question, we can consider the main use cases. The two most important use cases are:

Use Case 1: Enter and publish new information on a file format
Actor: a digital preservation researcher or file format expert **Precondition:** the researcher has information on a file format that they want to make available to others, for example that Format X can be reliably rendered by Application Y.
Check whether suitable identifiers already exist for Format X and Application Y. If not, then new identifiers need to be created according to the identifier scheme of the particular format registry that the researcher is using. Note that we assume that each format registry operator can create their own identifiers. If other registries are known to have different identifiers for this format, then a link should be made to those identifiers, noting that they refer to the same format. If other registries are known to hold information about this format, then a link should be made to inform users that further information is available at other locations. The researcher adds the new piece of format information to their format registry. Depending on the policy of the organisation operating the registry, it may be necessary for the new item of information to be reviewed or approved. Update the online information about this format. If the registry maintains a list of subscribers, then inform those subscribers that new information is available.

Use Case 2: Look up information on a file format
Actor: a digital preservation or digital repository manager in a memory institution **Precondition:** the user has a file of a known format, because it has been successfully identified by a file format identification tool.
The memory institution maintains a list of registries that it trusts. (It may be useful to have an 'order of preference' or different levels of trust for different registries). Using the file format identifier produced by the file format identification tool, look up those registries to see what information is available. Entries in each registry may include pointers to additional related information in other registries. The user can choose whether to consult those registries too. Retrieve information on (for example) which applications can render this format. This will require knowledge of the data model and exchange format used by the registry, so the user can interpret the information available. Apply this information to make a decision on what to do with the file, possibly updating a local policy on how to manage files of this type.

It can be seen from the above that for efficient information sharing, we need to agree on a core data model and one or more exchange formats.

With this distributed publishing format, it is however *not* necessary to agree on:

- Where to publish information
- Identifiers for particular file formats or applications
- Whether a particular piece of information about a format is correct or not.

In most cases of course, it is preferable if there is agreement on such matters, but the advantage of the distributed approach is that the creation and sharing of new information is not unnecessarily delayed by the need for all parties to agree on everything.

9. Outline of a solution

The minimum functions for a file format registry are:

- Select or assign identifiers to file formats, software applications etc
- Distribute information in an agreed interchange format.

Whilst numerous methods for distributing information are possible, the obvious and most useful approach would be to publish information on the web and to distribute information in a format that is easy to process in software applications.

Useful additional functions of a format registry could include:

- Publish a list of which formats the registry holds information on
- Cross-linking to other identifiers for the same format
- Cross-linking to other registries that hold further information on the same format
- Notify subscribers of updates to the registry contents
- A review and approval mechanism to ensure quality of information in the registry
- User friendly data editing and maintenance tools

It would also be useful to develop client software for retrieving information from multiple format registries. This could include user-friendly applications and software libraries to simplify integrating information from file format registries in other software (for example digital repository workflow tools).

How distributed should be the network of file format registries? At one end of the spectrum we could have a single centralised registry, at the other everyone could publish their own facts. The best compromise is between those two extremes. Relying on a single centralised registry is simple for data consumers but puts an unacceptable requirement for coordination amongst producers of format information. File format registries could be maintained by institutions such as archives or libraries; or they could be organised by interest groups or software vendors.

Data model

Past work on PRONOM and the Planets Core Registry has provided a detailed data model which could form the core of a future agreed approach [DataModel]. The PRONOM 7 data model is relatively complex and includes information which would normally be regarded as institutional policy, in addition to factual information about file formats. Other aspects of the data model are specific to the PRONOM implementation and would need to be generalised (and could probably be simplified in the process).

Therefore it would be possible to define a simplified version of this data model as a starting point for a generic data model.

Exchange format

The exchange format for file format registries should use a well-established file format that is easy to parse and process in software applications and which is capable of representing the data model.

PRONOM already has an XML format for data exchange, which could be used or adapted for a future standard. However, it is recommended that use of RDF (Resource Description Format) should be seriously considered. The potential advantages of RDF in this context are:

- It can be combined with RDF Schema and OWL allowing the data to be linked to a precisely defined machine processable representation of the data model.
- It fits well with web-based publishing approaches (see next section)

- Its graph-based structure and open-world approach are well suited to distributed publishing by multiple organisations.

RDF has a number of standard (or de facto standard) serialisation formats, notably RDF/XML, Turtle and N-Triples.

Publishing method

The obvious approach to publishing this information is to use the web. Files of any format could be made available for download via a website. However, the Linked Data [LD] is very well suited to the requirements outlined here, as identified by the Planets Core Registry Future Vision Document [FutureVision].

Linked Data follows these core principles:

- Use URIs as names for things (in our case, identifiers for file formats, software applications etc)
- Use HTTP identifiers so that people can look up those names using the web
- When someone looks up a URI, provide useful information using the standards (in particular RDF)
- Include links to other URIs so they can discover more things (in our case this would involve linking to other information about or related to the file format).

TNA has already begun work to publish a subset of the PRONOM information as Linked Data, which could form an excellent starting point for a broader initiative.

Assigning identifiers

The essential aspect of an identifier is that it should only identify one thing. With a distributed publishing model, the most convenient approach is if each registry operator can create their own identifiers when required (as opposed to a single central issuer of identifiers). This can lead to more than one identifier pointing to a single format – resources can have more than one 'name'. This has some downsides, but is a key point in enabling distributed publishing and hence in getting more people involved in creating and sharing useful information about file formats.

There are simple mechanisms in RDF for indicating that different identifiers are equivalent and for linking between different sources of information about the same resource.

10. Managing data from multiple distributed registries

This document describes an approach which enables organisations to publish information on file formats via multiple distributed registries. To use this information requires knowing which registries exist, making a choice of which registries to use and gathering the information of interest in a convenient to use form.

This process could be greatly assisted by a 'file format dashboard' software application. A separate document will be produced in early 2011 to describe the requirements for this application, but the main points are summarised here.

- The dashboard needs a list of file format registries and the URLs where they can be accessed. This could be maintained manually, but it would be very useful to have a 'registry of registries' where available registries can be discovered.
- The preservation manager must be able to configure which of the available registries he/she would like to use. This decision will be made primarily according to the level of trust associated with each registry, but possibly also taking into account other criteria such as the technical reliability of a registry endpoint.
- The dashboard must be able to retrieve the information from the chosen registries.
- There should be tools to help the user combine information on a given file format from multiple registries, identifying and resolving any conflicts.
- The source of each item of information gathered by the dashboard should be clear.
- The dashboard should provide convenient machine-readable access to the information it holds, through an API and/or export facilities. For example the dashboard should be able to collate and export a set of file format signatures for use in a file format identification tool such as DROID or FIDO. The dashboard could also provide information on file formats to the policy registry.

11. Planning

During 2011 the OPF aims to achieve the following:

1. **Establish a core data model and exchange format**
   In discussion with others active in the area of format registries, we aim to agree a core data model for the information that a registry must hold, and a format for exchanging that information between registries. This will build on existing work where possible.

2. **Dashboard Client**
   Create a prototype of the dashboard client application, that allows your institution to create and dynamically maintain a Registry collection;

3. **Guidelines**
   Specify guidelines to make your existing or planned Registry interoperable with the proposed Registry ecosystem;

4. **Standard Collection**
   Compile a standard Registry Collection that is considered a good starting point for the needs of digital preservation. Members and the wider OPF community will discuss the ins and outs of this collection on the OPF wiki to dynamically keep this collection alive and useful.

What we ask of you:

- Firstly, please review this document and give us your feedback to help us refine our plans.

- Consider the needs of your institution in the field of representation information registries and whether the registry ecosystem we propose would be a good match to those needs.
- Let us know whether you are able to participate in making this idea a reality, and what you could contribute to this initiative.

12. Summary

A representation information registry (RIR) is an essential component of digital preservation activities. The work of gathering and organising the necessary representation information is considerable and we need to encourage and enable a broader range of contributors. We propose a distributed model of creation and publishing which we believe will support this objective. We call this model the 'registry ecosystem'.

We have put a plan in place to develop the essential components to start making the registry ecosystem a reality. This report aims to explain the key issues we need to address and how an analysis of these issues has led us to the registry ecosystem concept.

A lot of work needs to be done to go from our initial concept to an operational system. Also the approach we propose is fundamentally about collaboration between institutions with similar interests. For both of these reasons we want to engage with a broad range of institutions. The OPF will be the focus for these efforts, but we also hope to discuss with and work with other interested institutions worldwide.

Our focus is on a series of simple practical steps that can lead quickly to concrete and useful results. We invite you to get involved and help turn this idea into a reality.

Acknowledgements

There are currently several initiatives in the area of Registries for digital preservation. This document has been informed and inspired by useful conversations and previous work by many authors, institutions and initiatives, notably the Planets project and many contributions from various Planets partners, PRONOM, TNA's Linked Data version of PRONOM, GDFR, the National Library of New Zealand, Bibliothèque Nationale de France (BnF) and Southampton University.

References

[RIR]
PLANETS White Paper: Representation Information Registries, PC/3-D7,
retrieved on 28th January 2011 from:
http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf

[FutureVision]
PLANETS Core Registry Future Vision Document. PC/3-D25,
retrieved on 28th January 2011 from:
http://www.planets-project.eu/docs/reports/Planets_PC3-D24_PCRFutureVisionReport.pdf

[PCR_Req]
PLANETS Core Registry V3: Software Requirements Document, PC/3-D20,
retrieved on 28th January 2011 from:
http://www.planets-project.eu/docs/reports/Planets_PC3-D20_Software_Requirements.pdf

[DataModel]
PRONOM 7: Understanding the Data Model. Robert Trickey, Tessella
(access restricted to Planets project members)

[LD]
Linked Data,
retrieved on 28th January 2011 from:
http://www.w3.org/DesignIssues/LinkedData.html

[GDFR]
GDFR faceted classification,
retrieved on 28th January 2011 from:
http://www.gdfr.info/docs/GDFR-Classification-1_0_5.pdf