Open Planets Foundation

A New Registry for Digital Preservation

Outline proposal for discussion

Nationaal Archief
(National Archives of the Netherlands)

September 2010

Commissioned by:	Maurice van den Dobbelsteen, Nationaal Archief
Author:	Bill Roberts, Swirrl IT Ltd
Reviewer:	Bram van der Werf, Open Planets Foundation
Date:	7 September 2010
Version:	1.1
Distribution:	- OPF Board (Adam Farquhar, Hans Jansen, Ross King, Max Kaiser, Jacqueline Slats) - Barbara Sierman, Manfred Thaller, Clive Billenness, Bjarne Andersen, Christen Hedegaard - Andy Jackson, Sven Schlarb

Contents

1 Objective

To create a representation information registry that will support the current and future digital preservation needs of memory institutions in a flexible and cost-effective way.

2 Limitations of the PLANETS Core Registry

During PLANETS a number of limitations of the PCR were identified. These were described in the requirements for PCR Version 3 and in the “PLANETS Core Registry: Future Vision” document.

The main points can be summarised as:

- We need a distributed approach that allows collaboration between different organisations, to share the effort of maintaining a registry and to exchange knowledge and best practices.
- We need an effective way for different institutions to assign identifiers, without the risk of name clashes.
- We need to be able to separate facts about formats (or other items of representation information) from institutional policies, processing rules for a repository or other preservation planning material.
- We want the flexibility to easily add or link new kinds of information, such as results of Testbed experiments.
- We want the ability to use ‘faceted classification’, as used by GDFR.

Our opinion is that the limitations of PCR will be difficult and expensive to overcome. The PCR has been useful and has helped us learn a lot about what we need from a registry. But now is the time to apply that learning to develop a new registry system that better meets our needs.

3 High level requirements for a new registry

The key requirements of a new registry are as follows:

- store information on file formats, technical environments, pathways, software applications, preservation characterisation and action tools
- decentralised implementation and governance, that allows institutions to collaborate and share information, without being dependent on a single external source
- clean separation between file format information, public policy information and private policy and other information
- ability to select (multiple) sources of trusted information
- open source
- robust, scalable and secure

- easy to integrate with other systems, such as a digital repository, testbed or preservation planning tool.
- use Linked Data as the core technology for the solution, as it is based on World Wide Web Consortium standards and allows a clean separation of the data in the registry, the data model, and the application that supports the registry

4 How will the registry be used

The activities around a format registry can be roughly divided in three:

- Creating (or importing) and maintaining data
- Using data in the process of managing a digital collection or repository
- Sharing information with other institutions

The registry needs to ensure that these three activities can be reliably achieved.

5 First steps

The first step should be to create a simple working version to demonstrate feasibility and help with further definition of requirements. We hope to be to reach agreement soon on priorities, so we can get started quickly on producing an initial prototype quickly. We envisage that the prototype should focus on:

- creating identifiers,
- defining and publishing core data about file formats,
- illustrating how this information can be used.

Any user interface software will be kept simple in this first version.

Before starting, we anticipate a process of defining and prioritising the requirements. However, a possible list of the main components and activities in a first version could be:

- Create a pattern for assigning identifiers for file formats and other elements of representation information.
- Define an ontology, describing the main kind of information we want to hold, concentrating initially on file formats.
- Create data for an initial set of file formats, drawing on the existing contents of the PCR.
- Host this data on the web as Linked Data.
- Create a simple website for people to explore the data
- Illustrate how the information could be used with a digital repository, for example by reviewing the ingest processes of the NA Digital Depot.

6 Future developments

Of course there will be many other features that could be added and we envisage an ongoing programme of development. Review of the initial prototype should help with understanding of the most important issues and setting of priorities for the next stage. Possible features for the future include:

- An application to make it easy to update and edit data without detailed knowledge of the technology.
- The ability to query the data in more complex ways
- An update feed for notifying users of changes to the data in the registry.
- Tools to assist with merging data from external sources into your own local registry.
- Software libraries in common programming languages to assist with using the registry data in other applications.
- Access control features to restrict access to private data within a registry.
- Extend the coverage of the registry to include more types of representation information.
- Integration with the PLANETS Testbed.
- Integration with PLATO.

7 Getting involved

The Nationaal Archief (the National Archives of the Netherlands) has a direct business need for a new registry as described here and plans to invest in it. However we feel deeply that the best approach is to cooperate with other organisations in the digital preservation community, to bring about an international file format registry solution.

We would like to set up a small group of partners to contribute to producing a working prototype on a short time scale. That may then inspire others to join at a later stage.

The Open Planets Foundation (OPF) is ideally suited to facilitate this process. As one of the founding members of the OPF, the Nationaal Archief has proposed this to the OPF as a key initial activity for that organisation. Initially we would like to invite other OPF partners who also have a need for a practical file format registry, to join with us to make this idea a reality.

We invite you to attend a side-meeting for interested parties at the iPres conference in Vienna this month. This is not part of the official proceedings of the conference, but an opportunity to have many potentially interested partners in one place. The meeting will be hosted by Bram van der Werf, director of the OPF.

If you are unable to come to the meeting, please e-mail any comments or questions to Bram (bram@openplanetsfoundation.org).

8 Background documents

I. PLANETS:
  - PC/3-D7: White Paper: Representation Information Registries
  - PC/3-D25: PCR Future Vision Document
  - PC/3-D20: PCRV3: Software Requirements Document

II. Linked Data: http://www.w3.org/DesignIssues/LinkedData.html

III. GDFR faceted classification: http://www.gdfr.info/docs/GDFR-Classification-1_0_5.pdf