Where the Semantic Web and Web 2.0 meet format risk management: P2 registry

David Tarrant, Steve Hitchcock and Les Carr
School of Electronics and Computer Science
University of Southampton
Southampton
UK
SO17 1BJ
dct05r,sh94r,lac @ecs.soton.ac.uk

The Web is increasingly becoming a platform for linked data. This means making connections and adding value to data on the Web. As more data becomes available and more people are able to use the data, it becomes more powerful. An example is file format registries and the evaluation of format risks. Here the requirement for information is now greater than the effort that any single institution can put into gathering and collating this information. Recognising that more is better, the creators of PRONOM, JHOVE, GDFR and others are joining to lead a new initiative, the Unified Digital Format Registry. Ahead of this effort a new RDF-based framework for structuring and facilitating file format data from multiple sources including PRONOM has demonstrated it is able to produce more links, and thus provide more answers to digital preservation questions - about format risks, applications, viewers and transformations - than the native data alone. This paper will describe this registry, P2, and its services, show how it can be used, and provide examples where it delivers more answers than the contributing resources.

So-called 'Web 2.0' tools and services have shown that users are willing to share information and collaborate on the Web on a large scale. This can be seen with people now publishing data through blogging services that are readable by both humans and machines in formats such as RSS and ATOM. Web 2.0 has also been the main driver for social networks that build links between people on the Web.

The Semantic Web has a similar set of aims to that of Web 2.0 and requires the addition of context to information to improve machine understanding of data. The Semantic Web encourages the creation of formal descriptions for concepts, terms, and relationships within each knowledge domain. The key concept we can use from the Semantic Web is the idea of simple triples of information where two items are related via a third element that describes the relation. In turn this relation can be arbitrary and could itself be described with other triples. Thus begins an unstructured but highly connected graph of information.

Both the Semantic Web and Web 2.0 recognise the importance of a simple set of Web publishing rules known as the linked data web. Linked data focuses on giving everything a resolvable URL which not only gives data about that object but also links to further information allowing the user to discover more. By utilising Web 2.0 to gather information published by the whole community, the Semantic Web technologies to add context and the four rules of the linked data web, we can build services aimed at digital preservation on top of an extensible registry of information.

The P2-Registry is essentially a Semantic Web system, backed by a model-free, unstructured RDF registry upon which ontologies and profiles can be applied to manipulate the data. The registry automatically harvests information from various defined information sources that are published in an open and machine readable fashion. Currently this service is specifically directed towards the file formats of the materials collected in digital repositories, and has data from the PRONOM registry and dbpedia (a linked data-specific export of wikipedia).

A series of import plugins are used to add data to the registry. These plugins normalise (removing any imposed structure) and translate the data into a set of triples represented in RDF. These triples are then imported to the registry. To link the data a series of search interfaces have been developed to help the user to find or provide semantic relations linking two objects.

Because the registry has no set data model it can import any amount of data from many sources and allow the data to be queried directly. The key feature of the registry is the ability to import arbitrary ontologies that can be used both to infer new facts from existing information as well as to align (in the case where two concepts are similar or the same in nature) information already in the registry.

Information in the registry is made available through a set of user and programming interfaces (APIs) that are designed to present information on format risk analysis and the resolution of these. Less effort is thus required in connecting and adding to this dataset of unstructured linked data.

On top of these base services a series of queries can be constructed to reveal key preservation information about any format in the registry. These results can then be displayed and directly exported to other services which wish to use them. Services available currently include risk analysis exports which provide only the data relating to possible risks, risk level profiles which actually process the data and provide a risk score as per a default P2 profile, and migration pathway exports. All data can be viewed directly in HTML and exported in XML and RDF.

Figure 1 shows a high-level risk level profile for a particular format, which collates information from the registries risk analysis service as RDF and then calculates a risk score according to a simple ruleset implemented on top of the registry.

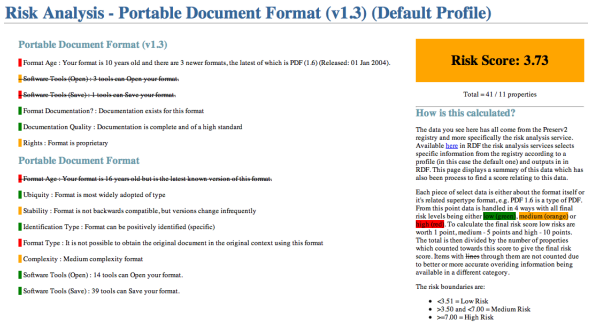

Figure 1: High Level Risk Analysis for P2-Registry

The full paper will outline the features of the registry in more depth, including how to use ontologies to manage, infer and align the data contained in the registry. It will also look at how these services can be used on the EPrints software platform to perform risk analysis of the files in an institutional repository. Repository managers are provided with an interface where the risks relating to files in their repository are represented using a simple traffic light scale. Most importantly,despite the simplicity of the interface , users do not lose the ability to delve deep into the information to discover exactly how the risk level was calculated.

In early tests the P2 registry has proved to be a promising and helpful platform for creating a rich data source of linked data on file format risk information. This approach needs the active participation of the digital preservation community to contribute data, to ensure the quality of the data and to build trust in a collaborative resource.