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 openly 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.

Introduction

The World Wide Web is recognised as the fastest growing publication medium of all time, now containing well over 1 trillion URLs (Alpert and Hajaj 2008) and still growing exponentially according to figures taken from reports by Google, Yahoo and Netcraft. As a result we face problems in both finding and being able to use all the available data. In this paper we focus on maximising the value of data published on the web, specifically in the area of digital preservation and file format registries. The core outcome of this work is to demonstrate how some emerging web publishing techniques can lead to the ability to construct a set of powerful and flexible services focused on digital preservation.

Publishing of data is one of the core features of the Web 2.0 (OReilly 2007) and Semantic Web (Berners-Lee, Hendler, and Lassila 2001) initiatives, and both have shown that users are willing to share information and collaborate on the Web on a large scale. This can be seen with the success of Wikipedia, and the take up of blogging and social networking services designed to build links between people on the web. These are just a few examples that are now publishing information in machine-readable formats such as RSS or ATOM that can be customised and displayed in ways to suit the consumer of the information.

To process data automatically for structured consumption, machine readability is only the first step. The next step is machine understanding, where not only is the data split into concepts, but these concepts are understood and aligned with other concepts. This is at the core of the Semantic Web. By using techniques from the Semantic Web, this paper demonstrates the simplicity of aligning data available on the web from services such as PRONOM, a file format registry produced by the National Archives in the UK (Brown 2005), and DBpedia, a linked data version of Wikipedia, such that seemingly complicated searches across these services can now be performed with a single request.

Being able to query data from a disparate set of services requires some form of caching of the data available at those services. The P2 Registry essentially provides this cache by storing data in a model-free, unstructured database on top of which many services are built to manipulate the data. The registry automatically harvests information from various defined data 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 such as institutional repositories.

Information in the registry is made available through a set of user and programming interfaces (APIs) that are designed to present information on resolving format risk analysis. By providing both high-level summary interfaces, where the searches are hidden from the user, as well as the search interface itself, ensures that the end user has the greatest level of flexibility when it comes to using the data known by the registry. This paper presents both interfaces and give examples of how the high-level interfaces are constructed from a few simple queries through the API.

The paper is structured to outline the entire process, from good publication techniques on the web, to the construction of the high-level services for the P2 Registry. The first sections look at the background to linked data and the Semantic Web. The linked data section focusses on the importance of following four simple rules for publishing on the web. Then we look at how techniques from the Semantic Web can be used to provide understanding of linked data and the use of ontologies. We then briefly look at existing technologies which are designed to aid digital preservation, including registries of data pertaining to file formats and related tools. In this section we also look at a few of the services built on top of these registries with the aim to show later how the P2 registry can compliment these. The main body of this paper outlines the P2 Registry and its interfaces for importing, processing and presenting data. We look at how the P2 Registry is able to directly import and cache linked data in the form of RDF, how this is aligned using a series of simple ontologies and how it is queried using simple searches. Finally, the wider uses of the P2 Registry, its implications for digital preservation and possible further development are considered.

This work began in the JISC Preserv 2 project (http://preserv.eprints.org) as a response to the perceived limitations of the available tools for file format analysis, and continues in the JISC KeepIt project (http://preservation.eprints.org/keepit/). Both projects are concerned with managing and preserving the contents of institutional and digital repositories, with the former focussed on the development of preservation tools and services, while the latter is working with repository managers to apply these tools to exemplar preservation repositories.

The most important aspect of this paper is to emphasise the power of a community and the sheer volume of data it can publish cooperatively. The P2 Registry brings this data together so that it can be used to answer questions on digital preservation.

Linked Data

Organisation of data on the web has proved to be a real problem over the years. As people start to realise the importance of linkable data, however, we are starting to see better use of one the web's simplest technologies, the Uniform Resource Locator (URL). A URL represents the location of "something" on the web. Aligning the principles of the URL with that of giving everything on the web a URI (Uniform Resource Identifier) empowers users to be able to link directly to very specific parts of the web, those which now provide data about "things".

Consider a simple real world example, say, booking a holiday over the internet. You and a travel partner are online, chatting to each other over email and browsing a travel website. You find a nice hotel and then copy and paste the browser link in an email to show your partner. The problem is that the website uses session- and path-based browsing, which means that your partner opens the link to be greeted by a page containing a session error. This could be avoided if the hotel page had a URI that could be referenced independently regardless of the path taken to find it.

Essentially this is the goal of Web 2.0, to put discoverable data online. This means that data should remain online in a static location, be well annotated and also link to other resources. Establishing static URIs for resources is the first of four rules for publishing on the linked data web (Berners-Lee 2006):

1. Use URIs as names for things
2. Use HTTP (web) URIs - thus they are also URLs
3. Provide useful information in useful formats, e.g. RDF
4. Include links to other URIs

Rules 2-4 emphasise that if someone goes to your URI on the web then it would not only exist but also tell you something about itself and link to other related items. The way this differs from many current web publishing techniques, however, is the publication of data, either alongside or instead of human-readable web pages.

In a series of tutorials Heath explains how to publish linked data on the web ((Heath 2009) and (Bizer et al. 2008)). In these he also looks at the many ways to serialise and provide data about "things". Once again the key is the use of URIs and URLs for identification and location. Figure 1 outlines the basic principles behind data publishing on the web where we start with a URI representing a "thing". From this URI many serialisations can be accessed which expose the same data in different forms. For instance a plain HTML page would be the default location displayed by web browsers. Alternative (HTTP code 303) versions (or serialisations) of the same data such as XML, RSS or RDF versions might also be offered.

If we look at linked data from the view of digital preservation, specifically file formats, each format would be represented by a URI. From this URI you would be able view information about the format, get alternative versions of the information, such as in XML, and, most importantly, be able to follow links to other similar file types or types with similar properties. To a certain extent PRONOM and DBpedia provide this functionality. For this reason these two services were used as a starting point for the P2 Registry.

Semantic Web

While Web 2.0 has focused on getting readable data from the web in a linked data fashion, the Semantic Web realises that even at this point there is still a problem with data deluge. The Semantic Web introduces the requirement for data publishers to add context and encourages the creation of formal descriptions for concepts, terms, and relationships

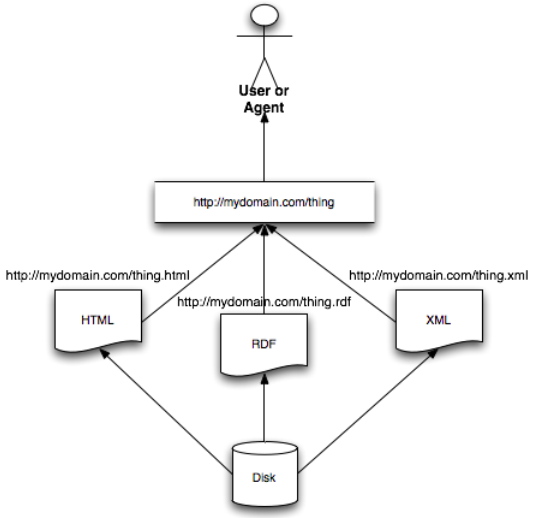

Figure 1: URIs and URLs, alternative serialisations

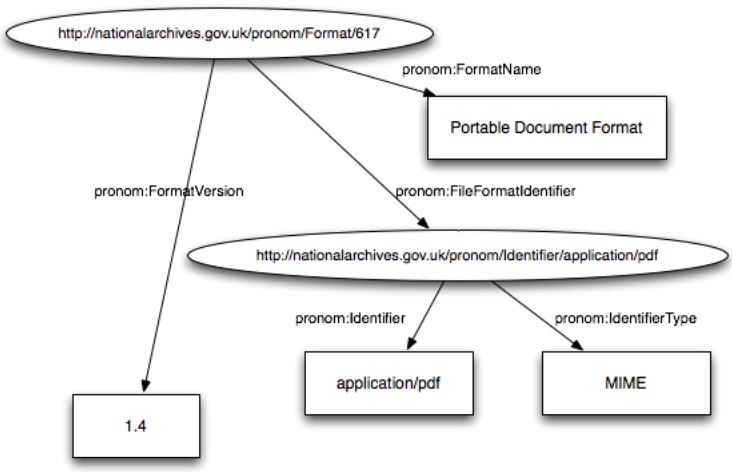

Figure 2: RDF graph relating to file format data

within each knowledge domain. This was systems can be envisaged which having understanding of real world concepts, as outline in Berners-Lee's original Semantic Web paper (Berners-Lee, Hendler, and Lassila 2001).Much like the structure used in a relational database model, the Semantic Web encourages publication of a glossary or terminology with the data such that the "model" can be understood and thus kept constant.

To keep things simple, the Semantic Web encourages information to be published as simple 'triples', where two items are related via a third element that describes the relation (Manola, Miller, and McBride 2004). In this space the triple is constructed from three URIs representing the subject, predicate and object, respectively. At the core of the Semantic Web effort is the Resource Description Framework (RDF), a markup language that extends XML. Through these extensions it makes namespaces mandatory for both resources/objects as well as for the predicates that link them. Figure 2 shows a simple set of triples describing some characteristics of a file format showing the namespacing with a ":" separator. In this representation ovals represent URIs, squares represent plain text nodes and predicates are represented by the arrows which join the nodes.

While it is obvious that the subject and object are interchangeable, e.g. "Tom isTheBrotherOf Fiona" and "Fiona isTheSisterOf Tom", it is the importance of adding context around the predicate that enables a machine to interpret the statement. Thus you could say "isTheBrotherOf hasDomain Boy" and "isTheBrotherOf hasRange Person" (based on the assumption we have also said Tom and Fiona are Boy and Girl respectively), thus allowing interpretation and inferring further information automatically.

While the Semantic Web does not define a limited set of predicates and relations to use, there are a set of well established ontologies (glossaries of terms) that allow mappings between data as well as define the concepts themselves. For the purposes of this work we focus on some of the key terms provided by the RDF, RDFS (RDF-Schema) and OWL (Web Ontology Language) namespaces:

- rdf:type - The subject is an instance of a class (URI linker)
- rdfs:label & rdfs:comment - Human readable fields (Text)
- rdfs:subClassOf - The subject is a subclass of another class (URI linker)
- rdfs:domain & rdfs:range - The domain and range of values for this subject. (URI linker)
- owl:sameAs - The subject URI can be considered to represent the same as object URI.

The advantage of using such glossaries is there are many applications developed to use semantically annotated data in RDF which can understand these concepts. These caching stores are essentially databases that are not constrained by any data model but still provide the query interface. The model builds itself as data is added or imported into the database, and because the underlying store understands terms such as owl:sameAs at the lowest level then queries are able to return implicit results. Such applications, commonly known as triple stores include Jena[1], Sesame [2] and 3Store[3]. In this work we use 3store due to previous knowledge of the platform.

To query the caching store we are using SPARQL (Simple Protocol and RDF Query Language), which is a World Wide Web Consortium (W3C) recommendation (Seabotne and Prud'hommeaux). SPARQL is much like SQL in syntax with extensions to support data in the triple-based format.

[1] Jena - jana.sourceforge.net
[2] Sesame - http://www.openrdf.org
[3] 3store - http://www.aktors.org/technologies/3store/

SPARQL is bundled with the 3store product, which also provides a web-based API for accessing the services and performing queries remotely.

Digital Preservation

Digital preservation is becoming of greater concern as we see many resources born in digital form only. A major part of digital preservation lies in the keeping the file bitstreams in tact either on disk, tape or even in the cloud (Tarrant, Brody, and Carr 2009). The other key aspect of digital preservation realises that even if the original bitstream or file is accessible in 20 years time, there is a risk that there will be no software able to read or accurately render that file type.

The first stage in active file preservation is file format identification, and this includes specific revisions and characteristics of the file. Several introductions to file formats and their selection for different situations have been but together by (Abrams 2007), this also goes some way to introducing the importance of significant properties. Significant properties of a file are a local list of the important characteristics of a file, for example in a word document the track changes may be an extremely important piece of metadata which is lost when a PDF is made. (Wilson 2007) gives an a excellent introduction and background to this area, which in the end is simply another set of metadata relating to file formats.

From the analysis and background work we find that metadata about file types, their characteristics and properties is very important in digital preservation. In turn this led to many projects, summarised by (Knight 2007), being established to collect, store and use this information. One of the most widely known registries is PRONOM (Brown 2005) which is focussing on resources which are collected from UK governmental departments.

With data available through registries such as PRONOM, there are already a great many services utilising this data. PRONOM-ROAR is an extension to the Registry of Open Access Repositories (ROAR) which remotely scans and analysis file types in a repository in order to provide that repository with a preserv profile (Brody et al. 2008). Moving forward the next problem is risk analysis and migration and while risk analysis a young research area, migration services built on top of the metadata in registries are becoming more common (Ramalho et al. 2008).

The problem now is that there are too many gaps in the current registries where information pertaining to file formats is either not present or incomplete. The aim of the P2 registry is to show how by bringing in data from the wider community allows us to fill some of these gaps and in turn encourages publication of more linked data. In turn services using this data then become much richer in their capabilities.

The P2 Registry

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. Thus the P2 Registry is using many of the technologies described in the previous sections. The key additions come in the form of the data harvester, which uses a set of import plugins to adjust data on import if needed, and the high-level interfaces that make the whole system easier to use and more powerful for the digital preservation community.

The registry automatically harvests information from various defined information sources that are published in an open and machine-readable form. Currently this service is specifically directed towards the file formats of the materials collected in digital repositories. Helpfully, DBpedia publishes data in RDF so no changes are needed to import this data into the P2 system. On the other hand, PRONOM data is only currently available in XML, and has to be parsed through an import plug-in.

The purpose of the import plugins is to normalise (remove any imposed structure) and translate source data into a set of triples represented in RDF. In the case of data from PRONOM, the importer also constructs the glossary of terms used by PRONOM to represent relations between objects. With the aim of keeping the data loosely coupled, there were no restrictions applied on the glossary as any conflicts, such as two terms meaning the same thing, could be aligned later using the owl:sameAs property. Doing this also demonstrates the total flexibility of the model-free caching store.

Having imported the data into the registry from both sources, the next stage is to link the data. This step is only required because no links were found between the data on DBpedia and PRONOM. By using the set of predicates outlined earlier, one of the first relations to be established was the link between a specific file format version in PRONOM and its generic parent data imported from DBpedia. 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.

Figure 3 shows the simple use of subclasses of various PDF formats, which was added to all PDF variations where applicable. Having done this, the number of tools found that could read a PDF format (1.4) jumped from 19 in the PRONOM registry to 70 in total in the P2 system.

Since the P2 Registry does not have a 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. For example, an ontology had to be added to the P2 Registry

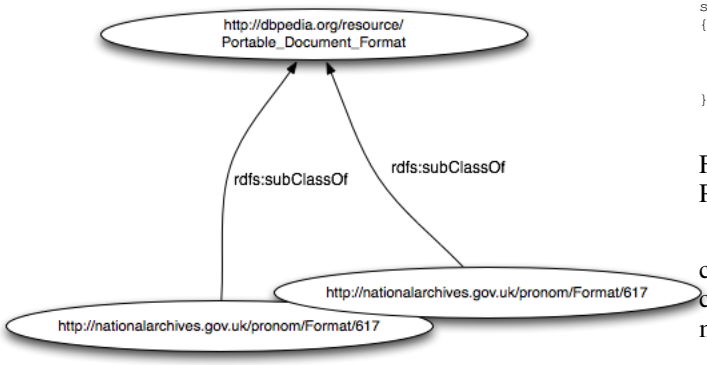

Figure 3: Simple linking example of data using RDFS

to obtain the result above where the number of tools found capable of reading PDF from a single query was greater than in the original registries.

Information imported from the original registries on software tools was specific about what the tools could do, e.g. open, save, create, render, print, etc. Performing a single query to find all tools requires the addition of further information to group these operations into one category or class. Figure 4 shows the part of the ontology constructed and added to the system to group the "SoftwareLink" class. Now it is possible to ask the registry for all software tools which have a "SoftwareLink" to the format in question. Due to all the subclasses created, PDF 1.4 will transparently include all information relevant to all PDF versions, and tools of all types will be returned by the query.

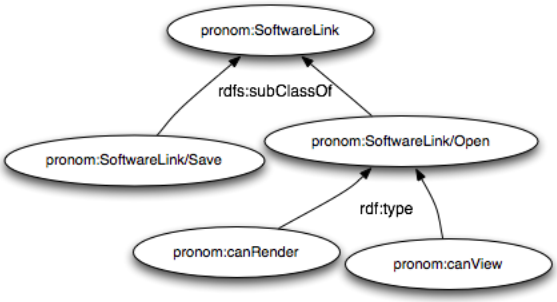

Figure 4: Software category subclassing in RDF

Figure 5 shows the SPARQL query, passed to the registry through the query interface, that returns the list of softwares able to manipulate a specific format. As can been seen, a SPARQL query consists of sets of triples where variables are represented by anything preceded with a question mark. In this case we are looking for ?x, where ?x is the URI of the software and has a SoftwareName ?name. ?x is then related to our format (617) through a SoftwareLink predicate.

SPARQL provides the base level interface to the data

```
select distinct ?name ?y where
{
    ?x ?y <http://nationalarchives.gov.uk/pronom/Format/617> .
    ?y rdf:type <http://nationalarchives.gov.uk/pronom/SoftwareLink> .
    ?x pronom:SoftwareName ?name
}
```

Figure 5: SPARQL query to find software compatible with PDF v1.4

contained in the registry. It is possible and necessary to construct a number of higher-level interfaces to allow easier manipulation and browsing of the data.

APIs and Interfaces

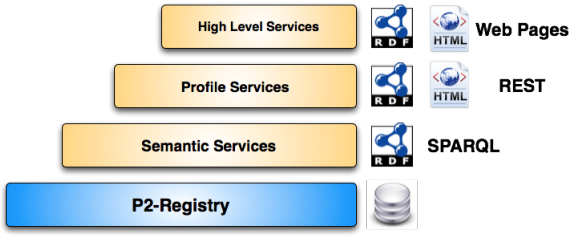

Figure 6: P2 Registry design layer cake

This section looks at the overall design and set of interfaces available at each layer in the P2 Registry. Figure 6 shows the layer cake which became the the specification for the design of the registry. At the lowest level the P2 Registry is a caching database. On top of this sits a SPARQL query service with direct access to the RDF stored within the registry. This semantic layer is designed purely for use by other services and agents which can harvest the data and results of queries for their own use.

Above this layer sits a set of services that perform some form of manipulation on the data before use. This translation will either be serialisation to provide the same data in different formats, e.g. XML and HTML, or summation services that combine data to build new profiles of concepts and objects in the registry. Taking a lead from the linked data guidelines by (Heath 2009) (Figure 1), the P2 registry exposes URIs with related URLs to obtain the same data in different formats such as HTML, XML and RDF. The profiling layer is also where RESTful services have been built to manage the registry and to import further data.

Finally, the high-level services are designed to hide the data while providing key information and interacting directly with physical users. Although some RDF can still be obtained at this level, these services are designed simply to demonstrate what can be done on top of the other services. High-level interfaces available include a data browser which uses the hubs and authorities algorithm to rank "URIs", a risk-profile analysis service which uses a set of rules added to the registry to provide a risk score for a particular format, and a migration pathway interface that provides

information on tools that can translate one format to another.

Searching

Even with the registry focussed on data specific to digital preservation, in particular file formats, there are still just under 44,000 statements in the current registry. Among these statements is data from PRONOM plus any available data from DBpedia related to the PDF formats. Searching thus becomes a key activity in order to gain familiarity with the contents of the registry, as such there also needs to be a way to order search results putting the most relevant first.

The P2 Registry provides a simple triple-based search interface, where each URI has been ranked to produce search results of greatest relevance. Using the hubs and authorities algorithm (Kleinberg 1999), each URI is viewed as a node linking to many other nodes/URIs, and search results return the highest scoring of each. Basically, by searching we are looking for a central node in the graph, thus a search for PDF will return the PDF MIMEtype node.

Risk Analysis

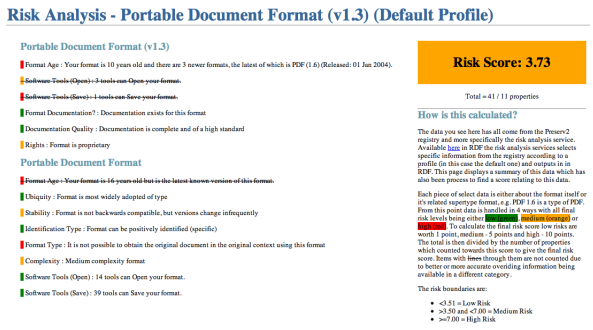

Figure 7: High Level Risk Analysis for P2-Registry

Figure 7 shows a high-level risk analysis profile for a particular format (PDF 1.3). This profile collates information from the registry's risk analysis service as RDF and then calculates a risk score according to a simple ruleset implemented on top of the registry. This ruleset consists of the data used to construct the profile, either defined by inclusion or exclusion, and then defines how to process the returned values. In figure 7 we see information presented pertaining to the format type, including documentation, stability, age, as well as the number of manipulation tools.

To generate a risk score a policy has been loaded into the registry which translates data such as age and number of tools into a low, medium or high risk category. Each category is given a score and the average of the results is the output risk score. Such policies are clearly subjective, so each policy has its own namespace in the registry. In this way policy scores can be judged on the source of the original data and its trustworthiness. This process can be managed via the REST interface.

Earlier, in the JISC Preserv 2 project, a pilot format identification and risk analysis interface was implemented in the EPrints repository software, as shown in Figure 8. The P2 Registry already supports the same set of services to provision information to this interface, however further modifications are planned to link the EPrints interface back to the registry. This would allow users to browse the contents of the registry from within the EPrints software and see how the risk scores have been constructed.

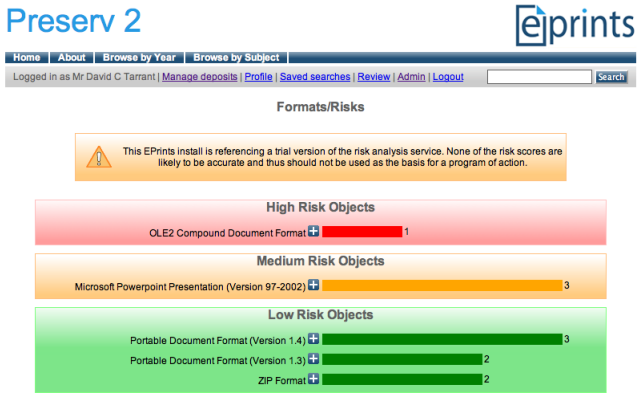

Figure 8: Risk analysis interface in EPrints

Migration Pathways

The last of the current interfaces implemented on top of the registry exists at the profile services level. It provides useful services as well as demonstrating one of the RESTful interfaces. The migration pathways service is designed for users who need to translate a file from one format to another, and the SPARQL interface is an ideal starting point for what is essentially a query. The software linking ontology outlined in Figure 4 is an essential part of this as we can simply use the "open" and "save" classes in the query, as shown in Figure 9.

Figure 9 shows a single step query which will only return software URIs that can both open and save in the input and output formats specified, respectively. SPARQL has the ability to take this query to an infinite number of steps but this may become recursive and thus is not advisable. By default the P2 Registry iterates only to two-step pathways involving an intermediate format via the RESTful service,

```
select distinct ?software where {
?software ?predicate1 ?in_format .
?predicate1 rdf:type SoftwareLink/Open .
?software ?predicate2 ?out_format .
?predicate2 rdf:type SoftwareLink/Save }
```

Figure 9: Single step migration pathways query

which can be accessed at http://p2-registry/migration-pathways?from=format1?to=format2. This URL is, in fact, a URI and can be referenced if comments are added to a particular migration pathway by users. This data can then be returned along with the actual data to advise future users on the quality of the migration and whether it should be used.

Conclusions and future directions

Empowering the community of Web users to publish data and knowledge, on sites such as Wikipedia, has already demonstrated the benefits of collaborative content. Currently digital preservation has a few islands of knowledge that are beginning to publish in a linked data fashion, but there is still some way to go. By harvesting data from just two of these islands, PRONOM and DBpedia, the P2 Registry has demonstrated the benefits gained in terms of increasing the amount of knowledge available and also showing how easy it is to link this knowledge using techniques from the Semantic Web. By building some simple ontologies we have demonstrated how these islands of data can be linked by simply aligning similar concepts, or even just by saying two concepts are exactly the same.

The P2 Registry's set of high-level interfaces go some way to revealing what sort of services could be built on top of the core SPARQL API. By constructing a set of policies we have demonstrated possible ways the data could be processed and thus generating new knowledge, in this case knowledge relating specifically to file format risk analysis. From results such as the migration pathways URIs, we envisage that other services can start linking to, and commenting on, these results. Thus, a third party could state that it used a certain migration pathway, as represented by a URI, and rate the quality and experience with this to advise others who may use the service at a later date. This has parallels with the rating of items in online stores such as eBay and Amazon, and brings the P2 Registry full circle: in the first instance it consumed linked data, and by using ontologies and policies it is able to publish new linked data for others to consume.

The future of the P2 Registry is two-fold. First, as a reference platform to allow and encourage publication of preservation data as linked data on the Web. Second, better integration with third party tools such as EPrints, by enhancing and completing the high-level interfaces. Accountability and trust is a problem with open publication. In the P2 Registry this is handled by retaining the namespace and URIs from the originating information sources. In this way, if they are URLs each one can be resolved back to its creator, e.g. PRONOM or Wikipedia (via DBpedia). Although this is present in the P2 Registry, the high-level interfaces and policies are not currently detailed enough to be able to distinguish at this level, although this is not a problem via the API interfaces.

At this stage the P2 Registry has proved to be a promising and helpful platform for bringing together rich sources of linked data on file format risk information and migration pathways. This approach needs the active participation of the digital preservation community to contribute data by simply publishing it openly on the Web as linked data. By demonstrating the range of services that can be built on top of open data, it is hoped that more parties will be encouraged to make this part of their core activity and business practice, thus allowing the hard work of building preservation data registries to be distributed across the wider community.

References

Abrams, S. 2007. File formats. *Instalment of DCC, Digital Duration Manual* (1).

Alpert, J., and Hajaj, N. 2008. We knew the web was big. *http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (Retrieved September 2009)* 18:2008.

Berners-Lee, T.; Hendler, J.; and Lassila, O. 2001. The semantic web, scientific american. *Scientific American* 284(5):34–43.

Berners-Lee, T. 2006. Linked data. *w3c Design Issues*.

Bizer, C.; Heath, T.; Idehen, K.; and Berners-Lee, T. 2008. Linked data on the web (ldow2008).

Brody, T.; Carr, L.; Hey, J.; Brown, A.; and Hitchcock, S. 2008. PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *International Journal of Digital Curation* 2(2).

Brown, A. 2005. Automating preservation: New developments in the pronom service. *RLG DigiNews* 9(2):1093–5371.

Heath, T. 2009. An introduction to linked data. *Semantic Web Summer School (SSSW2009)*.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5):604–632.

Knight, G. 2007. File format typing and format registries. 18:2008.

Manola, F.; Miller, E.; and McBride, B. 2004. Rdf primer. *W3C recommendation* 10.

OReilly, T. 2007. What is web 2.0: Design patterns and business models for the next generation of software.

Ramalho, J.; Ferreira, M.; Faria, L.; Castro, R.; Barbedo, F.; and Corujo, L. 2008. Roda and crib a service-oriented digital repository. *Proceedings of iPRES 2008 Fifth International Conference on Preservation of Digital Objects*.

Seabotne, A., and Prud'hommeaux, E. Sparql query language for rdf. *W3C Recommendation, http://www.w3.org/TR/rdf-sparql-query*.

Tarrant, D.; Brody, T.; and Carr, L. 2009. From the Desktop to the Cloud: Leveraging Hybrid Storage Architectures in your Repository. *Proceedings of Open Repositories Conference 2009*.

Wilson, A. 2007. Significant properties report. *InSPECT Work Package 2.2*.