The P2 Registry
Where the Semantic Web and Web 2.0 meet Digital Preservation

http://eprints.ecs.soton.ac.uk/17556/

David Tarrant, Steve Hitchcock & Les Carr

davetaz / sh94r / lac @ecs.soton.ac.uk

School of Electronics & Computer Science
Disclaimer

This paper/talk is *not* actually about a new registry for preservation data.

The P2-Registry is simply a demonstration of what can be done with machine readable data which is published openly on the web.
Outcomes

- Linked Data
  - What? Why? How?

- Semantic Web
  - Machine Understanding for Linked Data

- P2-Registry
  - What we can do with this data

Where the Semantic Web and Web 2.0 meet format risk management: P2 registry
http://eprints.ecs.soton.ac.uk/17556/
Linked Data

- Why?
  - Reduce redundancy
  - Facilitate re-use
  - Maximize discovery
  - Community of publishers
  - Enables trust to be related back to source

Tom Heath - How to Publish Linked Data on the Web
http://sssw09.org/invited-talks.php
Wikipedia & dbpedia

- http://dbpedia.org/resource/San_Francisco → Thing
- http://dbpedia.org/data/San_Francisco → RDF data
- http://dbpedia.org/page/San_Francisco → HTML page

Linked Data – Connect Distributed Data across the Web
http://linkeddata.org
Linked Data - The Technology

- **URIs & URLs**
  - These become one and the same (sort of)
  - i.e. when you go to a URI it should resolve to a useful URL related to that URI

- **HTTP / HTML**
  - HTTP headers and status codes
  - HTML link alternate tags
4 Rules of Linked Data

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using standards.
- Include links to other URIs so that they can discover more things.
URIs & URLs – More Examples

- http://dbpedia.org/resource/San_Francisco
- http://dbpedia.org/data/San_Francisco
- http://dbpedia.org/page/San_Francisco

- http://eprints.ecs.soton.ac.uk/17556/
- http://eprints.ecs.soton.ac.uk/cgi/export/17556/XML/
- http://eprints.ecs.soton.ac.uk/cgi/export/17556/DC/

- http://www.nationalarchives.gov.uk/pronom/fmt/18
- http://www.nationalarchives.gov.uk/pronom/fmt/18.xml
The Semantic Web

- Data comes as Facts (according to that domain)
- Facts are represented by Triples

Technology Stack

- RDF
- OWL/RDFS

RDF/XML Syntax Specification (Revised)
http://www.w3.org/TR/rdf-syntax-grammar/

OWL Web Ontology Language Overview
http://www.w3.org/TR/owl-features/
RDF & OWL/RDFS

- RDF enforces the requirement to use namespaces for everything!
- RDF limits the data model to that of simply containing triples.
- OWL/RDFS provide a means to represent your RDMS model and validation tools which sit on top in RDF
Example RDF Graph

http://www.ecs.soton.ac.uk/people/dt2

foaf:name

David Tarrant

foaf:knows

http://www.ecs.soton.ac.uk/people/drn05r

RDF Primer
http://www.w3.org/TR/rdf-primer/
OWL/RDFS Example

- **Machine Readable!**

```xml
<owl:Class rdf:ID="WhiteWine">
  <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:about="#Wine" />
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasColor" />
      <owl:hasValue rdf:resource="#White" />
    </owl:Restriction>
  </owl:intersectionOf>
</owl:Class>
```

OWL Web Ontology Language Guide
http://www.w3.org/TR/owl-guide/
Core RDFS and OWL

- **rdf:type**
  - The subject is an instance of a class (URIs)

- **rdfs:label & rdfs:comment**
  - Human readable fields (Text)

- **rdfs:subClassOf**
  - The subject is a subclass of another class (URIs)

- **rdfs:domain & rdfs:range**
  - The domain and range of values for this subject. (URIs)

- **owl:sameAs**
  - The subject URI can be considered to represent the same as object URI.
The P2-Registry

- Is a registry which caches data available on the web (dbpedia and pronom)
- Provides a set of RESTful services and SPARQL interface to enable cross domain queries
Data Translation from PRONOM

http://nationalarchives.gov.uk/pronom/Format/617

Portable Document Format

http://nationalarchives.gov.uk/pronom/Identifier/application/pdf

1.4

application/pdf

MIME

http://p2-registry.ecs.soton.ac.uk/pronom/Format/617(.html)
Making Links (Ontology)

http://p2-registry.ecs.soton.ac.uk/pronom/SoftwareLink
SPARQL

- SPARQL is the query language standard for data represented in RDF.

```
select distinct ?x ?y where {
    ?y rdf:type http://nationalarchives.gov.uk/pronom/SoftwareLink
}
```

- Before alignment with dbpedia this returned 19 results. After it returned 70.

SPARQL Query Language for RDF
http://www.w3.org/TR/rdf-sparql-query/

P2-Registry SPARQL Endpoint
http://p2-registry.ecs.soton.ac.uk/SPARQL/
The P2-Registry

- The registry understands OWL & RDFS and hence it transparently follows subClass and sameAs links when queries are performed.

- Returned document also returns the relation at the profile level
Profile Services

- Profile services provide views on data
- You can create a view by simply specifying a set of fields to include/exclude.

http://p2-registry.ecs.soton.ac.uk/risk_analysis/default/617 ← Thing
http://p2-registry.ecs.soton.ac.uk/risk_analysis/default/617.rdf ← RDF data
http://p2-registry.ecs.soton.ac.uk/risk_analysis/default/617.html ← HTML page
High Level Services

- Actively process the data to final output
- This includes applying local policy
- These are examples of what could be done with the data and are not part of the core functionality of the registry
High Level Service Example

Risk Analysis - Portable Document Format (v1.3) (Default Profile)

Portable Document Format (v1.3)

- **Format Age**: Your format is 10 years old and there are 3 newer formats, the latest of which is PDF (1.6) (Released: 01 Jan 2004).
- **Software Tools (Open)**: 1 tool can Open your format.
- **Software Tools (Save)**: 1 tool can Save your format.
- **Format Documentation**: Documentation exists for this format.
- **Documentation Quality**: Documentation is complete and of a high standard.
- **Rights**: Format is proprietary.

**Portable Document Format**

- **Format Age**: Your format is 16 years old but is the latest known version of this format.
- **Ubiquity**: Format is most widely adopted of type.
- **Stability**: Format is not backwards compatible, but versions change infrequently.
- **Identification Type**: Format can be positively identified (specific).
- **Format Type**: It is not possible to obtain the original document in the original context using this format.
- **Complexity**: Medium complexity format.
- **Software Tools (Open)**: 14 tools can Open your format.
- **Software Tools (Save)**: 39 tools can Save your format.

---

**Risk Score: 3.73**

Total = 41 / 11 properties

**How is this calculated?**

The data you see here has all come from the Preserv2 registry and more specifically the risk analysis service. Available here in RDF the risk analysis services selects specific information from the registry according to a profile (in this case the default one) and outputs in in RDF. This page displays a summary of this data which has also been processed to find a score relating to this data.

Each piece of select data is either about the format itself or its related supertype format, e.g. PDF 1.6 is a type of PDF. From this point data is handled in 4 ways with all final risk levels being either low (green), medium (orange) or high (red). To calculate the final risk score low risks are worth 1 point, medium - 5 points and high - 10 points. The total is then divided by the number of properties which counted towards this score to give the final risk score. Items with issues through them are not counted due to better or more accurate overriding information being available in a different category.

The risk boundaries are:

- <3.51 = Low Risk
- >3.50 and <7.00 = Medium Risk
- >=7.00 = High Risk
Real World Application

Preserv 2

Formats/Risks

This EPrints install is referencing a trial version of the risk analysis service. None of the risk scores are likely to be accurate and thus should not be used as the basis for a program of action.

High Risk Objects

- OLE2 Compound Document Format

Medium Risk Objects

- Microsoft Powerpoint Presentation (Version 97-2002)

Low Risk Objects

- Portable Document Format (Version 1.4)
- Portable Document Format (Version 1.3)
- ZIP Format
More people using linked data

- RKBExplorer
  http://www.rkbexplorer.com

- BBC Music & BBC Programmes
  http://www.bbc.co.uk/music
  http://www.bbc.co.uk/programmes

- ACM, Citeseer & Web of Science

- EPrints, Dspace, Fedora
The Future

- More people publishing linked data!
- Migration pathways and review data.

Adding the rest of Digital Preservation

Digital Preservation: Logical and bit-stream preservation using Plato, EPrints and the Cloud

http://eprints.ecs.soton.ac.uk/17962/
“The coolest thing to do with your data will be thought of by someone else”
Thank You

Preserv2 / KeepIt funded by

JISC

David Tarrant, Tim Brody & Les Carr
davetaz / sh94r / lac @ecs.soton.ac.uk
School of Electronics & Computer Science