The P2 Registry

Where the Semantic Web and Web 2.0 meet Digital Preservation

David Tarrant, Steve Hitchcock & Les Carr

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

School of Electronics & Computer Science

UNIVERSITY OF Southampton

UNIVERSITY OF Southampton

Disclaimer

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

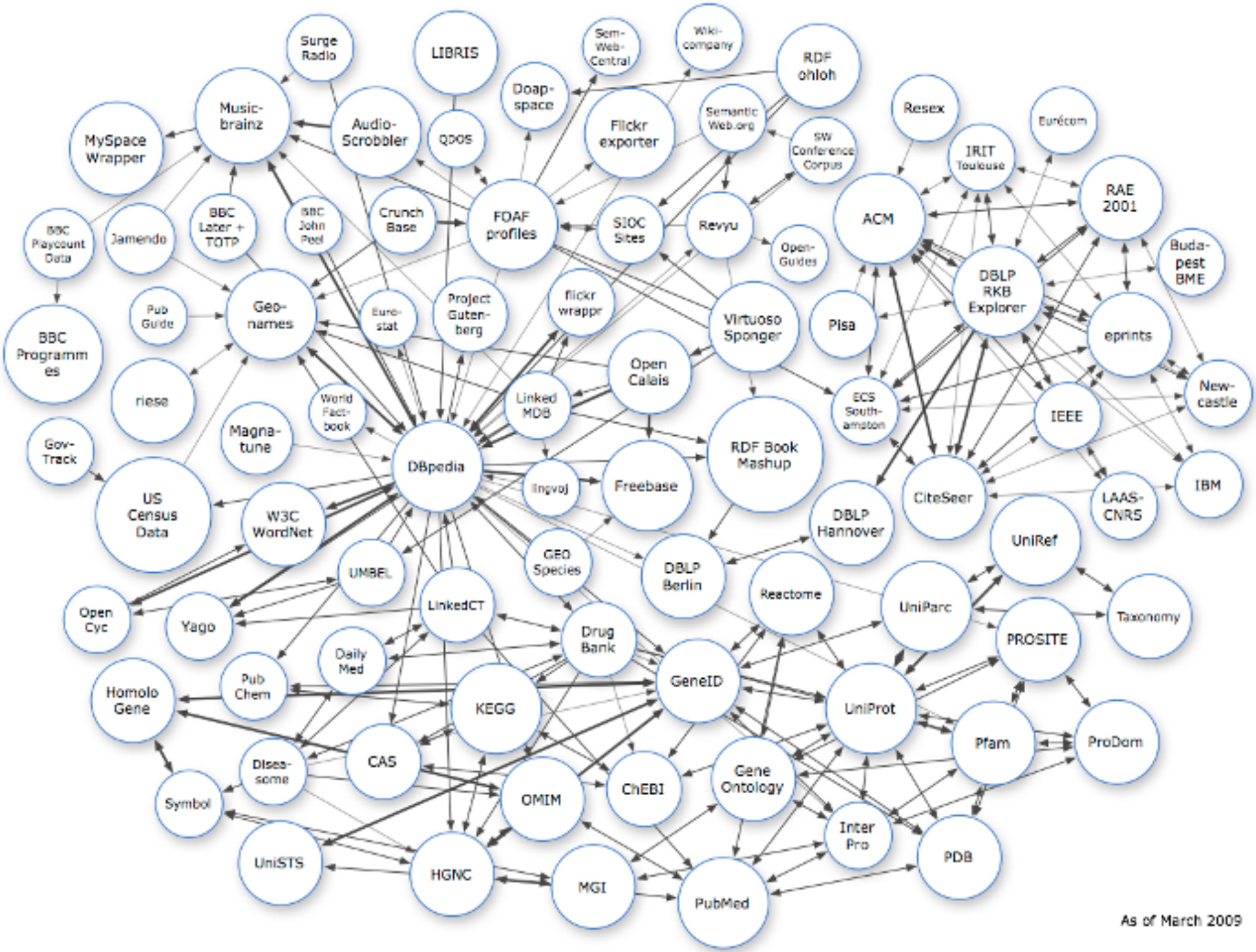
The P2-Registry is simply a demonstration of what can be done with machine readable **data** which is published openly on the web.

UNIVERSITY OF Southampton

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/


Surge Radio
LIBRIS
Sem-Web-Central
Wiki-company
RDF ohloh
Doap-space
Music-brainz
Audio-Scrobbler
QDOS
Flickr exporter
Semantic Web.org
SW Conference Corpus
Resex
IRIT Toulouse
Eurécom
MySpace Wrapper
BBC Playcount Data
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
RAE 2001
Buda-pest BME
DBLP RKB Explorer
Pisa
eprints
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
BBC Programmes
riese
Open Calais
ECS South-ampton
IEEE
New-castle
World Fact-book
Linked MDB
Magna-tune
Gov-Track
DBpedia
RDF Book Mashup
Ingvoj
Freebase
CiteSeer
IBM
LAAS-CNRS
US Census Data
W3C WordNet
DBLP Hannover
UniRef
UMBEL
GEO Species
DBLP Berlin
Reactome
UniParc
Taxonomy
Open Cyc
Yago
LinkedCT
Drug Bank
PROSITE
Daily Med
Pub Chem
Homolo Gene
GeneID
UniProt
KEGG
Pfam
ProDom
Disea-some
CAS
Gene Ontology
Symbol
OMIM
ChEBI
Inter Pro
UniSTS
HGNC
MGI
PubMed
PDB
As of March 2009

UNIVERSITY OF Southampton

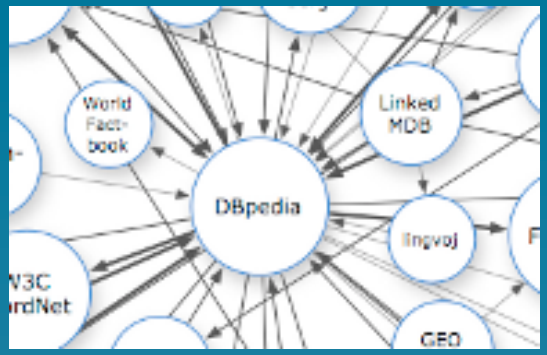
Linked Data

- Why?
  - Reduce redundancy
  - Facilitate re-use
  - Maximize discovery

  - Community of publishers
  - Enables trust to be related back to source

UNIVERSITY OF Southampton

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

UNIVERSITY OF Southampton

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

UNIVERSITY OF Southampton

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.

Linked Data, Design Issues (Tim Berners-Lee)
http://www.w3.org/DesignIssues/LinkedData

UNIVERSITY OF Southampton

URIs & URLs – More Examples

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

- http://eprints.ecs.soton.ac.uk/17556/ ← Thing
- http://eprints.ecs.soton.ac.uk/cgi/export/17556/XML/ ← XML data
- http://eprints.ecs.soton.ac.uk/cgi/export/17556/DC/ ← Dublin core data

---

- http://www.nationalarchives.gov.uk/pronom/fmt/18 ← Thing
- http://www.nationalarchives.gov.uk/pronom/fmt/18.xml ← XML data
- http://www.nationalarchives.gov.uk/pronom/fmt/18.html ← HTML page

UNIVERSITY OF Southampton

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/

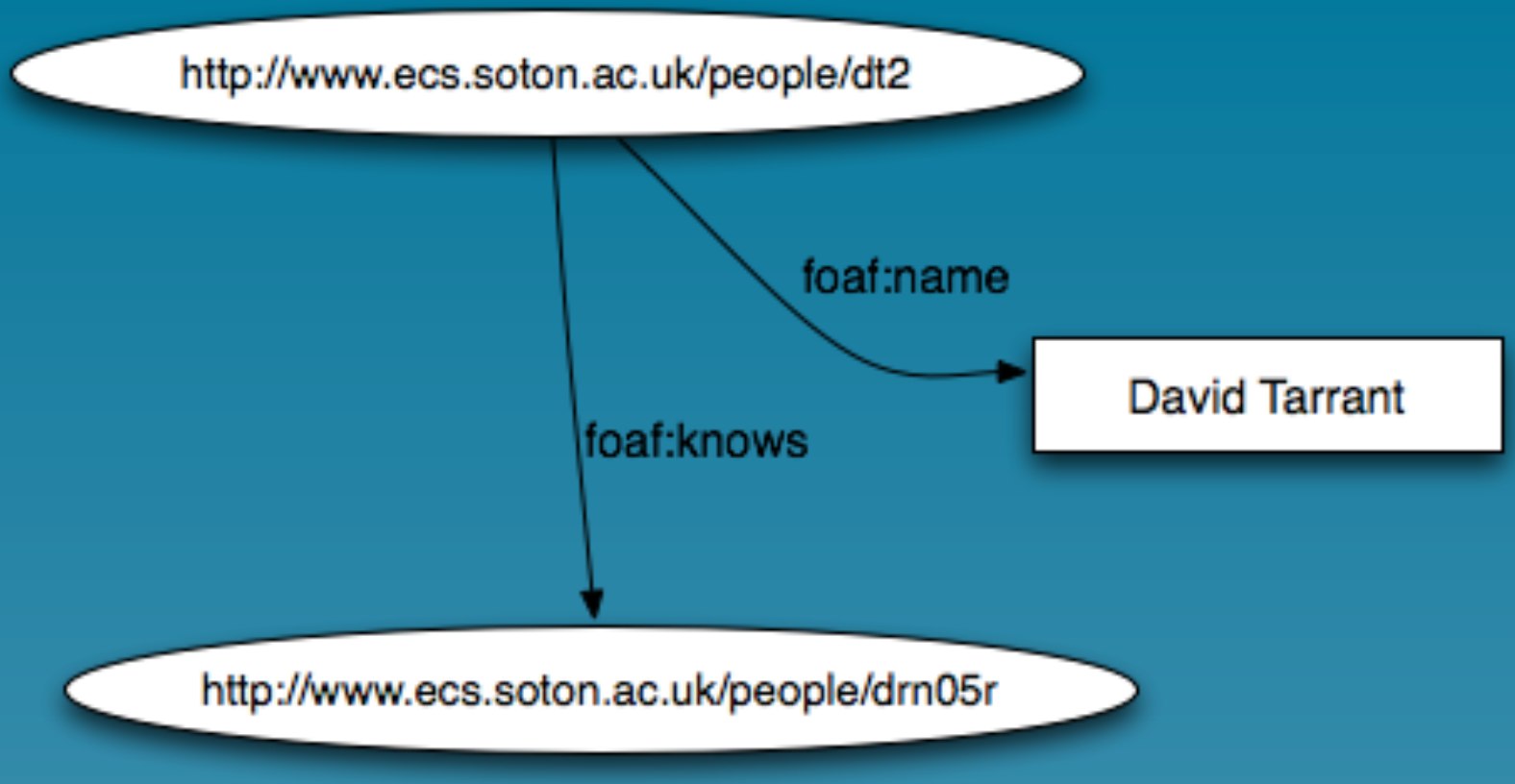
UNIVERSITY OF Southampton

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

UNIVERSITY OF Southampton

Example RDF Graph



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

UNIVERSITY OF Southampton

OWL/RDFS Example

- Machine Readable!

```
<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/

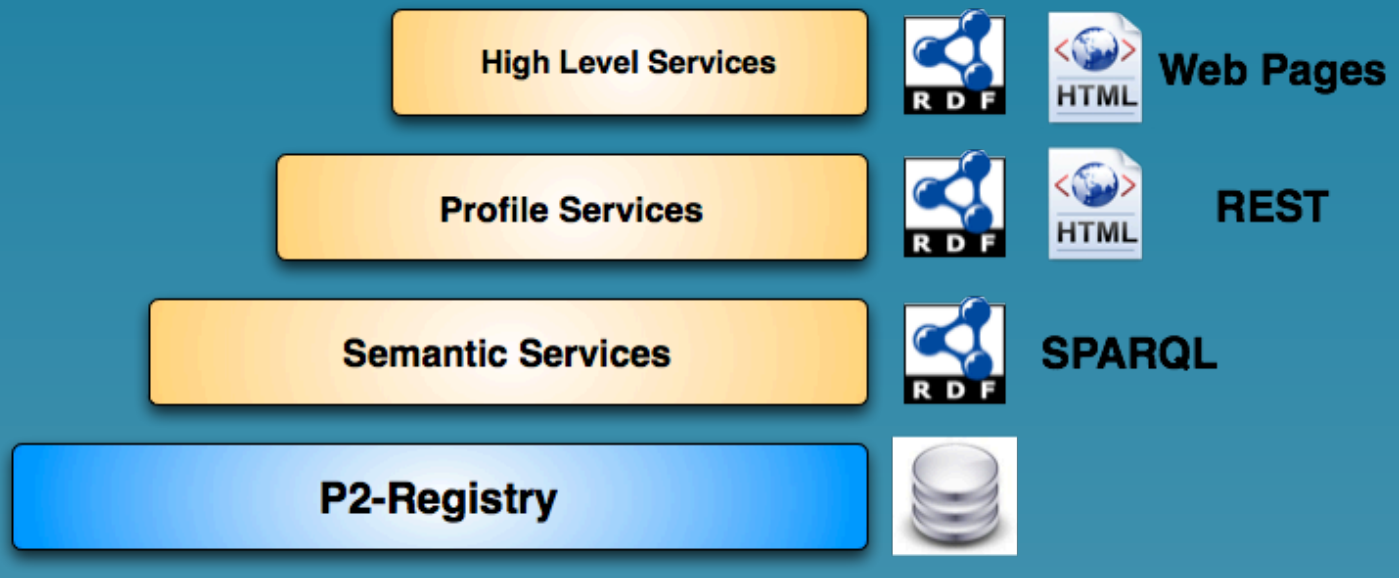
Core RDFS and OWL

- rdf:type
  - The subject is aninstance 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 rep- resent the same as object URI.

UNIVERSITY OF Southampton

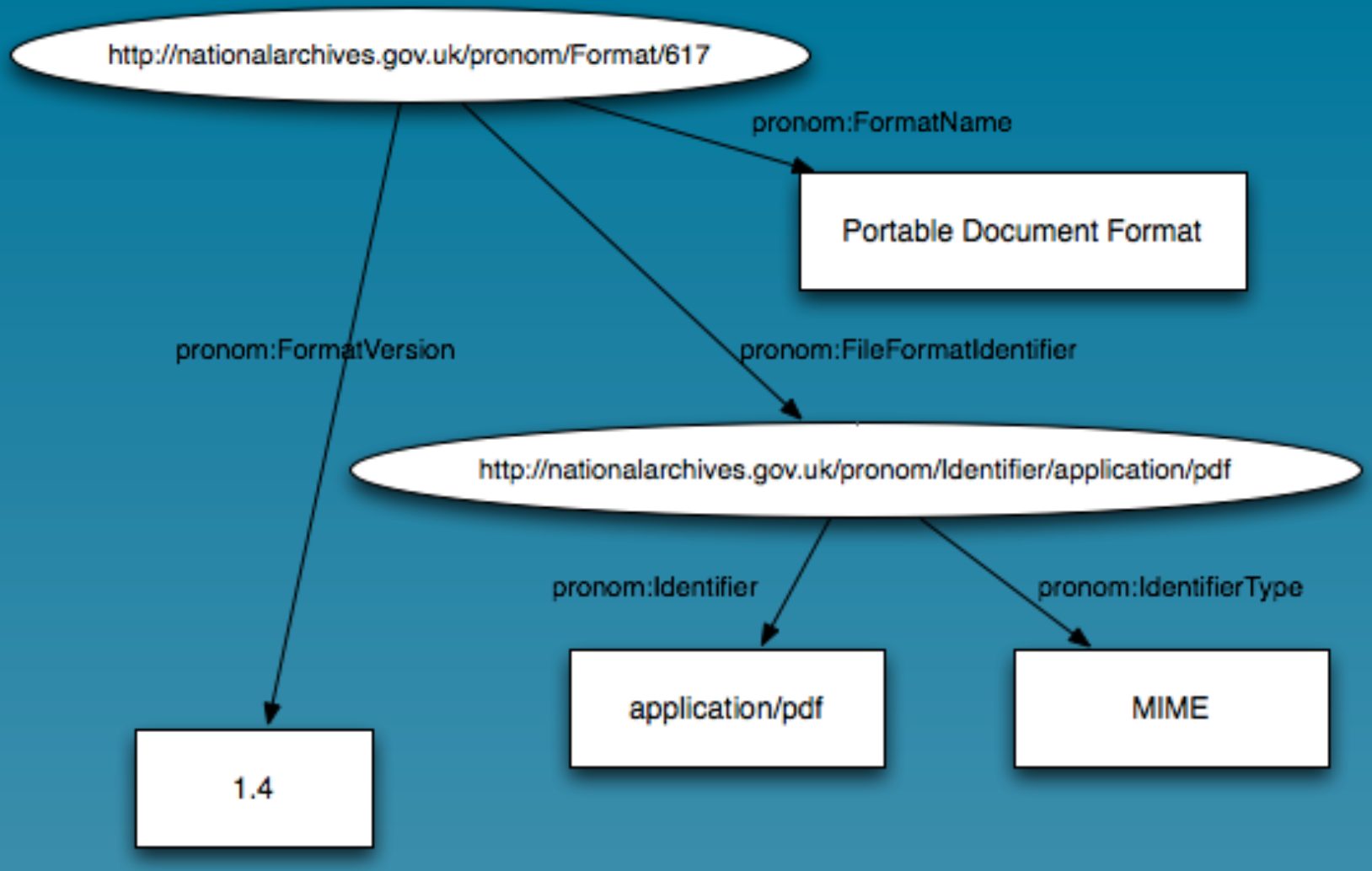
The P2-Registry

- 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



UNIVERSITY OF Southampton

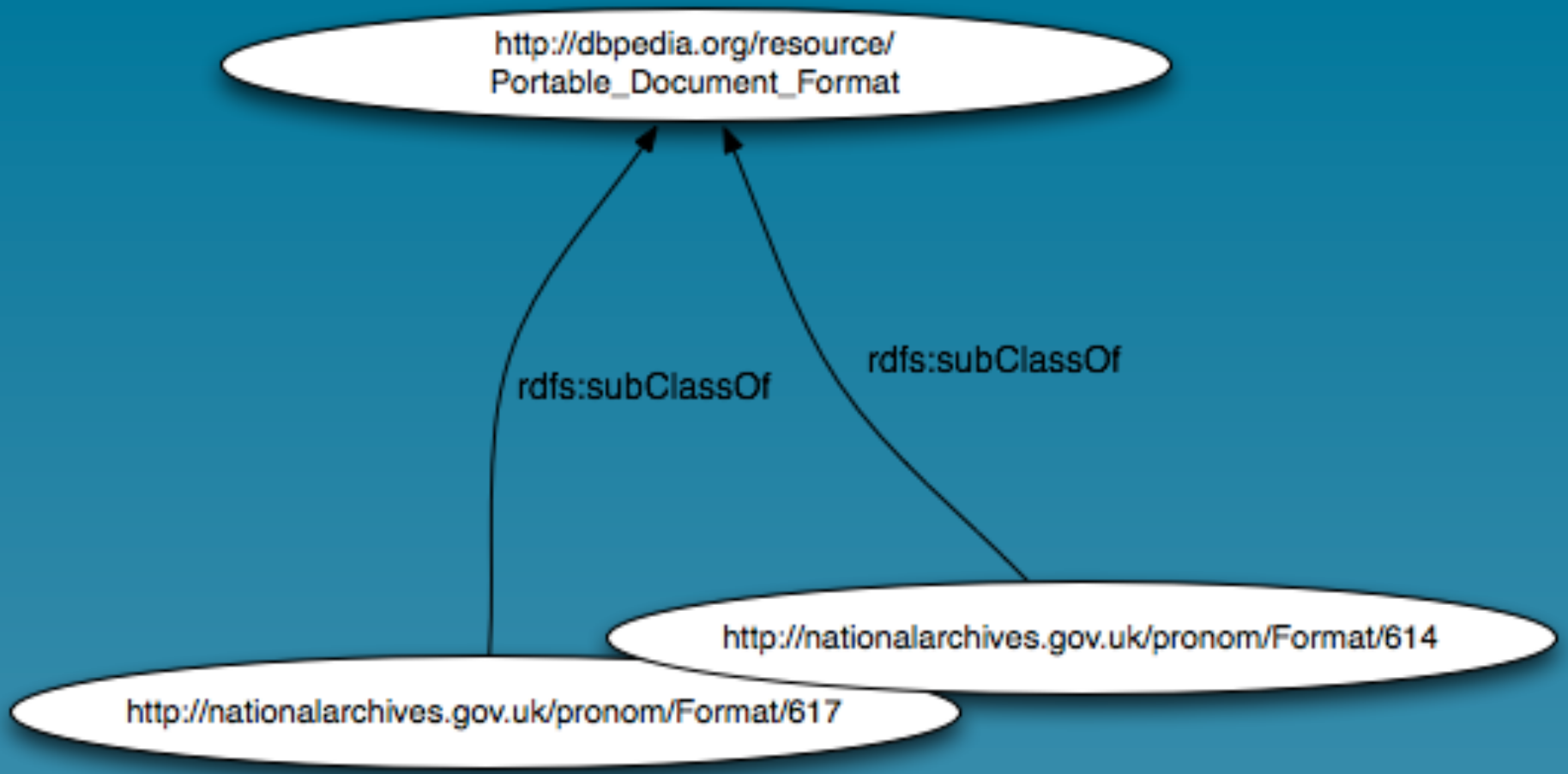
Data Translation from PRONOM



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

UNIVERSITY OF Southampton

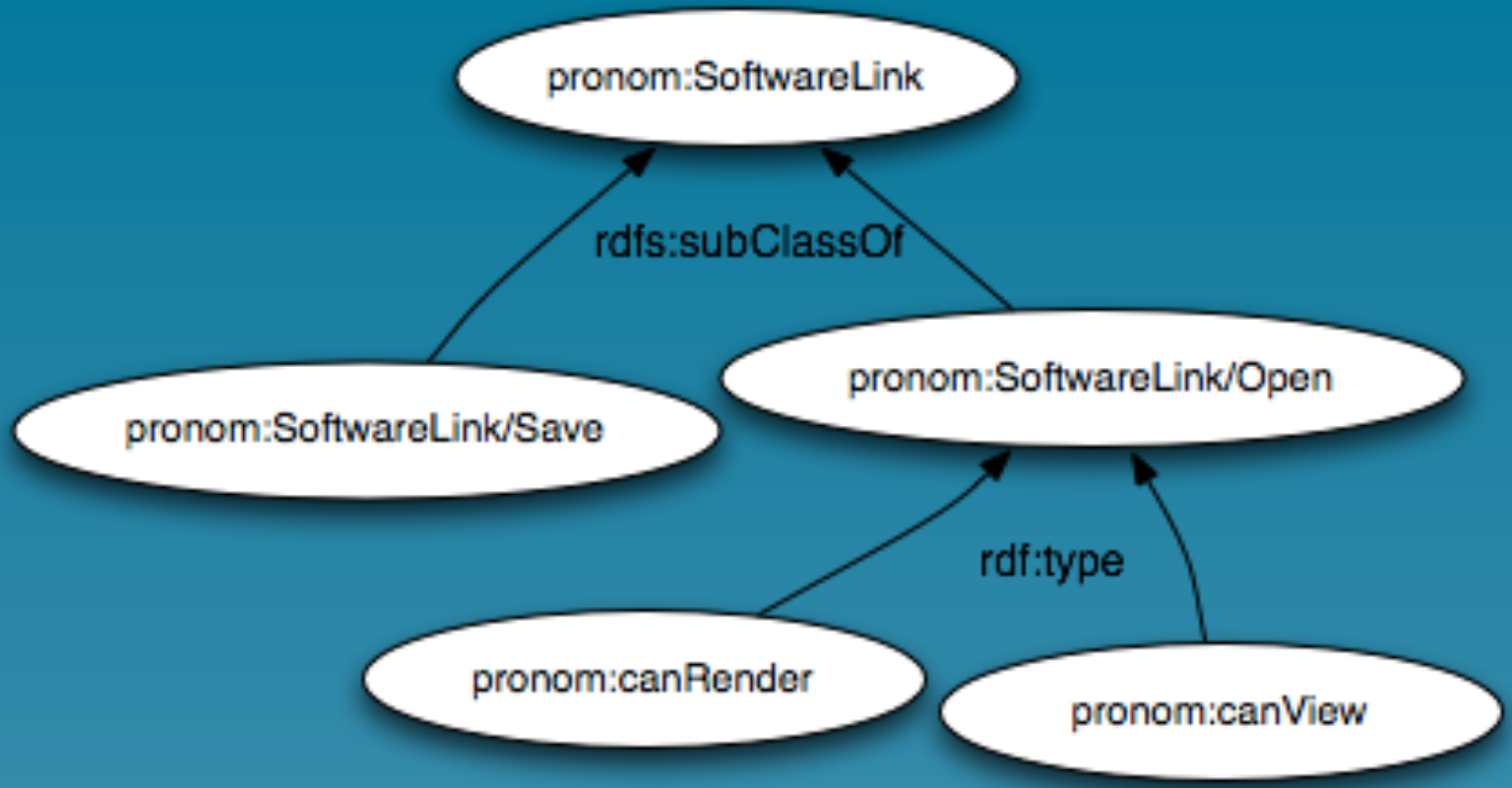
Making Links (Facts)



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

UNIVERSITY OF Southampton

Making Links (Ontology)



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

UNIVERSITY OF Southampton

SPARQL

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

```
select distinct ?x ?y where {
?x ?y http://nationalarchives.gov.uk/pronom/Format/617 .
?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/

UNIVERSITY OF Southampton

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

UNIVERSITY OF Southampton

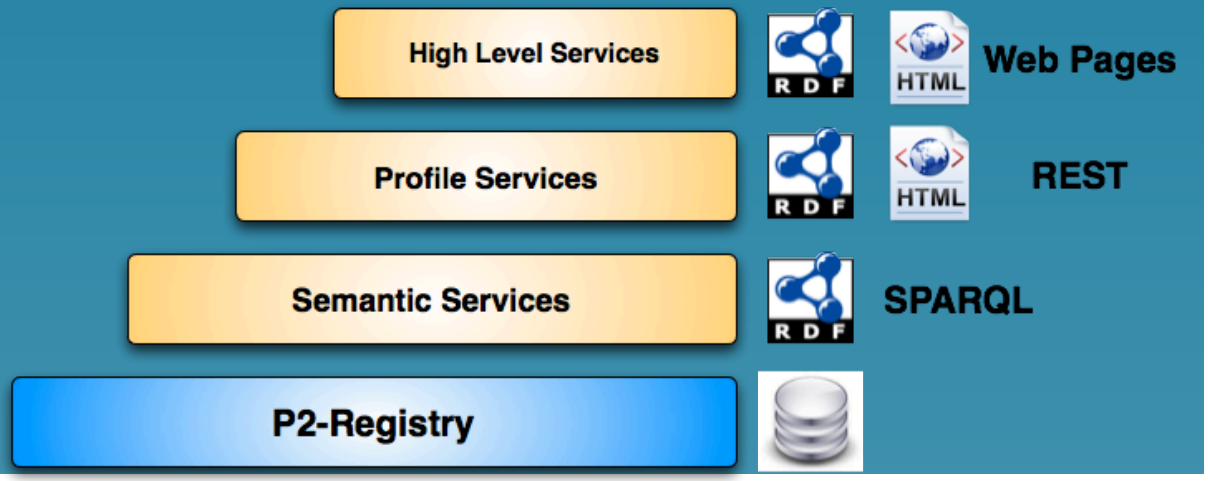
Profile Services

- 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

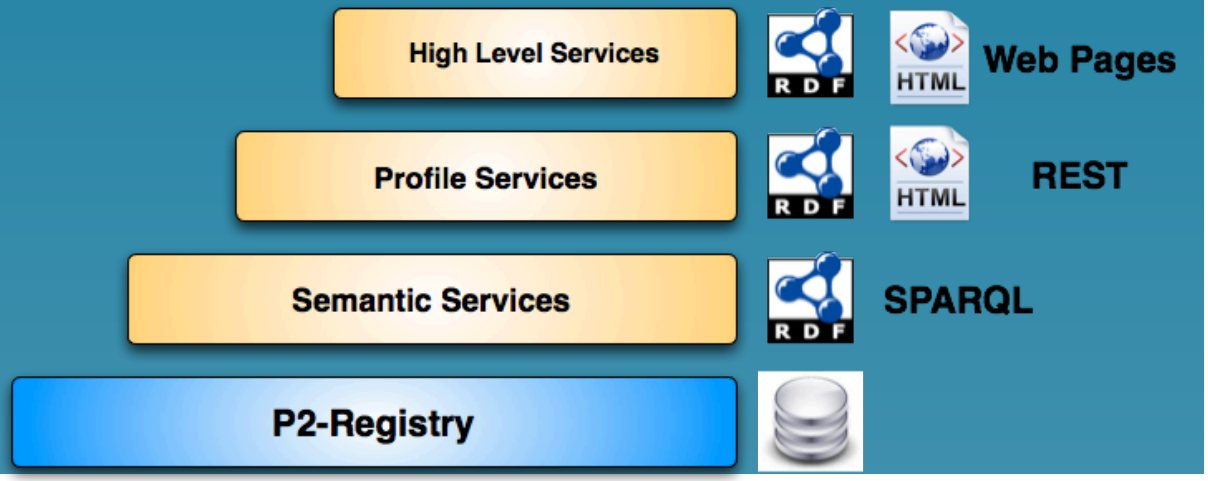
http://p2-registry.ecs.soton.ac.uk/risk_analysis/**default**/617.html ← HTML page



UNIVERSITY OF Southampton

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



UNIVERSITY OF Southampton

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) : 3 tools can Open your format.~~

~~Software Tools (Save) : 1 tools 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 process to find a score relating to this data.

Each piece of select data is either about the format itself or it's 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 ~~lines~~ through them are not counted due to better or more accurate overiding 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

UNIVERSITY OF Southampton

Real World Application

Preserv 2

eprints

Home | About | Browse by Year | Browse by Subject

Logged in as Mr David C Tarrant | Manage deposits | Profile | Saved searches | Review | Admin | Logout

Search

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 1

Medium Risk Objects

Microsoft Powerpoint Presentation (Version 97-2002) 3

Low Risk Objects

Portable Document Format (Version 1.4) 3

Portable Document Format (Version 1.3) 2

ZIP Format 2

UNIVERSITY OF Southampton

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

UNIVERSITY OF Southampton

The Future

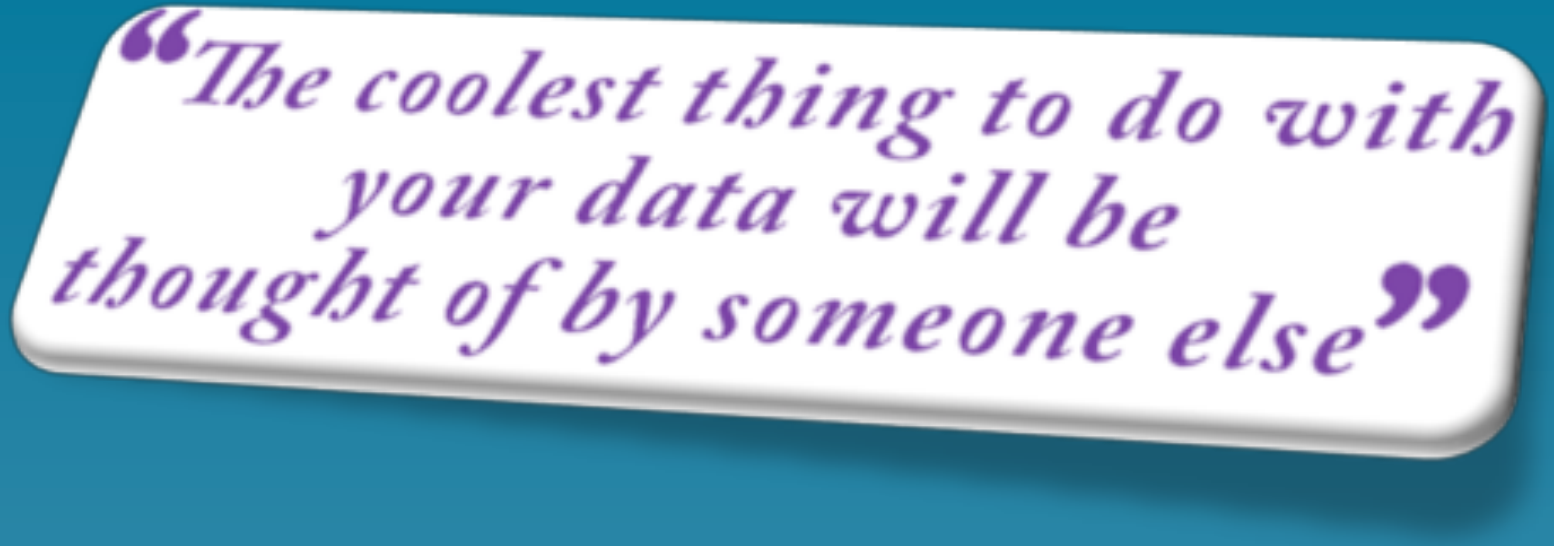
- More people publishing linked data!
- Migration pathways and review data.

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/

UNIVERSITY OF Southampton



Common Repository Interfaces Group
http://www.ukoln.ac.uk/repositories/digirep/index/CRIG

Developer Community Supporting Innovation
http://devcsi.ukoln.ac.uk/

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

University of Southampton

UNIVERSITY OF Southampton

The Complete Stack

- Lots of Uses Keeps Stuff Valuable
- Lots of Services Keeps Stuff Useful
- Lots of Description Keeps Stuff Meaningful
- Lots of Copies Keeps Stuff Safe