An Assessment of the Adoption and Quality of Linked Data in European Government Data (In-Use track)

Luis-Daniel Ibáñez [1], Ian Millard [2], Hugh Glaser [2], Elena Simperl [1]

[1] University of Southampton
[2] Seme4 Ltd
European Data Portal

- European Commission initiative
- Harvest metadata of Public Sector Information of European public data portals.
- Index and search engine
- Co-locate tools and documentation
For more details...

- Industry talk tomorrow morning
  - “The European Data Portal: Scalable Harvesting and Management of Linked Open Data”
    Fabian Kirstein, Simon Dutkowski, Benjamin Dittwald, Manfred Hauswirth
Problems

- Dataset search
- Data Integration
  - A “federation of data lakes”
- Multi-linguality
Linked Data as a solution

- DCAT-AP
  - DCAT extension for public sector
- EU Vocabularies
  - Controlled vocabularies and ontologies
- General push for Linked Data technologies
EDP wants to answer

• Are publishers using Linked Data?
• Is the Linked Data they generate of enough quality to be queried and re-used?
• If not, how to improve?
For the SemWeb community

• Uptake and acceptance of our tools
  – Public Sector was an early adopter
• Identify challenges and areas of improvement
This paper

• Quantitative study of
  – Uptake of RDF as publishing format
  – Quality of Linked Datasets in EDP
Metrics - Uptake

- Relative usage of RDF vs other formats
- Are they following recommendation when describing format in metadata?
  - How they deviate from it?
ex:anOpenDataset
   a dcat:Dataset ;
   dcat:distribution ex:aDistro .
ex:aDistro
   dct:format xxxxxxxx ;
   dcat:MediaType xxxxxxxxxx
PREFIX
ex <https://datapublisher/#> 
dcat < http://www.w3.org/ns/dcat#> 
dct <http://purl.org/dc/terms/> 
mrpo <http://publications.europa.eu/resource/authority/file-type/#> 

ex:anOpenDataset
  a dcat:Dataset ;
  dcat:distribution Ex:aDistro .

ex:aDistro
  dct:format mrpo:xxx ;

Should be from the controlled vocabulary
What we want

```
SELECT ?format COUNT(?distribution) as ?numDistros
WHERE {
}
GROUP BY ?format
```
Use of dct:Format

- 68% distributions include it (45% of datasets)
- From those including it
  - 45% OK
  - Varied errors like
    - Wrong suffix code (non-existent format)
    - A text literal
ex:aDistro
dct:format ex:distro-UUID/format .
ex:distro-UUID/format
  a dcterms:IMT ;
  rdfs:label ‘PDF’ ;
  rdf:value ‘application/PDF’ .

Valid RDF but...
Valid RDF but...

SELECT ?format COUNT(?distribution) as ?numDistros
WHERE {
} GROUP BY ?format

Won’t work...
ex:aDistro
dct:format b_node:123.
b_node:123
  rdfs:label ‘PDF’ ;

Same problem
Results - Uptake
Implications - Uptake

• Better procedures for interactive metadata required
  – EDP implemented SHACL rules and reports

• Massive number of tabular datasets
  – Help with transformation needed
Corpus

- RDF Datasets harvested by EDP
  - Could be downloaded and parsed valid.
- Main difference with previous work:
  - Not crawled, but what the EDP harvests
Quality Metrics

- Subset from previous general LOD quality studies [1,2,3]
  - Contextual
  - Representational
  - Accessibility

Metrics – Contextual

• Provision of Provenance information
  – Analysis of dct:publisher usage
Results - Contextual

• 50.3% of all datasets have dct:publisher
• 42.8% of all RDF have dct:publisher
• Relatively low
  – But much than other studies for general LOD (16%)
Metrics – Representational

- Blank nodes
- Well-known vocabularies
- Proprietary vocabularies
Results – Blank node usage

- Median
- 485 outliers

Blank nodes ratio

More blank nodes ↔ Less blank nodes
Results – Well known vocabularies

• Lots of rdf, rdfs and dcterms
  – Mostly predicates
• Few foaf, wgs84
• Almost no rss, skos, bio
  – Different from general LOD Cloud
## Results – Unknown vocabularies

### Table 5. Top-10 not well-known vocabularies by dataset percentage

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>% Datasets</th>
<th># Hosts</th>
<th># Preds</th>
<th>Deref-able?</th>
</tr>
</thead>
<tbody>
<tr>
<td>socrata.com/rdf/terms</td>
<td>52.6%</td>
<td>4</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>opendata.aragon.es/def/Aragopedia</td>
<td>13.0%</td>
<td>1</td>
<td>52</td>
<td>No</td>
</tr>
<tr>
<td>w3.org/2000/10/swap/pim/usps#</td>
<td>2.7%</td>
<td>4</td>
<td>4</td>
<td>Yes</td>
</tr>
<tr>
<td>data.press.net/ontology/stuff/</td>
<td>2.1%</td>
<td>2</td>
<td>5</td>
<td>Yes</td>
</tr>
<tr>
<td>opendata.caceres.es/def/ontomunicipio</td>
<td>1.7%</td>
<td>2</td>
<td>139</td>
<td>HTML</td>
</tr>
<tr>
<td>purl.org/ctic/infraestructuras/</td>
<td>1.1%</td>
<td>1</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>opendata.unex.es/def/ontouniversidad</td>
<td>1.0%</td>
<td>1</td>
<td>63</td>
<td>HTML</td>
</tr>
<tr>
<td>dublincore.org/documents/dcmi-box/</td>
<td>0.7%</td>
<td>1</td>
<td>4</td>
<td>No</td>
</tr>
<tr>
<td>open.vocab.org/terms</td>
<td>0.6%</td>
<td>1</td>
<td>3</td>
<td>HTML</td>
</tr>
<tr>
<td>server1.avantic.net/opendata/vocab/raw/</td>
<td>0.5%</td>
<td>1</td>
<td>206</td>
<td>No</td>
</tr>
</tbody>
</table>
Realisation

• Two provinces of same country, different ontologies
  – And we can’t blame them, this two really embraced Linked Data!

• No apparent alignment among themselves or among (at least that has reached EDP)
## Results – Unknown vocabularies

### Table 5. Top-10 not well-known vocabularies by dataset percentage

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>% Datasets</th>
<th># Hosts</th>
<th># Preds</th>
<th>Deref-able?</th>
</tr>
</thead>
<tbody>
<tr>
<td>socrata.com/rdf/terms</td>
<td>52.6%</td>
<td>4</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>opendata.aragon.es/def/Aragopedia</td>
<td>13.0%</td>
<td>1</td>
<td>52</td>
<td>No</td>
</tr>
<tr>
<td>w3.org/2000/10/swap/pim/usps#</td>
<td>2.7%</td>
<td>4</td>
<td>4</td>
<td>Yes</td>
</tr>
<tr>
<td>data.press.net/ontology/stuff/</td>
<td>2.1%</td>
<td>2</td>
<td>5</td>
<td>Yes</td>
</tr>
<tr>
<td>opendata.caceres.es/def/ontomunicipio</td>
<td>1.7%</td>
<td>2</td>
<td>139</td>
<td>HTML</td>
</tr>
<tr>
<td>purl.org/ctic/infraestructuras/</td>
<td>1.1%</td>
<td>1</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>opendata.unex.es/def/ontouniversidad</td>
<td>1.0%</td>
<td>1</td>
<td>63</td>
<td>HTML</td>
</tr>
<tr>
<td>dublincore.org/documents/dcmi-box/</td>
<td>0.7%</td>
<td>1</td>
<td>4</td>
<td>No</td>
</tr>
<tr>
<td>open.vocab.org/terms</td>
<td>0.6%</td>
<td>1</td>
<td>3</td>
<td>HTML</td>
</tr>
<tr>
<td>server1.avantic.net/opendata/vocab/raw/</td>
<td>0.5%</td>
<td>1</td>
<td>206</td>
<td>No</td>
</tr>
</tbody>
</table>
Realisation

• A lot of RDF in our corpus is produced from CSV through portal provider “converters”
• Conversion generates many unique URIs
  – One namespace per csv
  – Linking nightmare
Metrics – Accessibility

- Links to external datasets
  - Counted Pay Level Domains other than the publisher in the datasets
# Results - Links

<table>
<thead>
<tr>
<th>Domain</th>
<th># Datasets</th>
<th># (%) hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>w3.org</td>
<td>2467</td>
<td>55 (74%)</td>
</tr>
<tr>
<td>es.dbpedia</td>
<td>769</td>
<td>4 (5.4%)</td>
</tr>
<tr>
<td>purl.org</td>
<td>577</td>
<td>35 (47.3%)</td>
</tr>
<tr>
<td>reference.data.gov.uk</td>
<td>504</td>
<td>12 (16.2%)</td>
</tr>
<tr>
<td>data.press.net</td>
<td>123</td>
<td>2 (2.7%)</td>
</tr>
<tr>
<td>murciaturistica.es</td>
<td>122</td>
<td>1 (1.35%)</td>
</tr>
<tr>
<td>geonames.org</td>
<td>119</td>
<td>4 (5.4%)</td>
</tr>
<tr>
<td><a href="http://www.gijon.es">www.gijon.es</a></td>
<td>117</td>
<td>1 (1.35%)</td>
</tr>
<tr>
<td>schema.org</td>
<td>73</td>
<td>7 (9.5%)</td>
</tr>
<tr>
<td>dbpedia.org</td>
<td>43</td>
<td>11 (14.8%)</td>
</tr>
<tr>
<td>publications.europa.eu</td>
<td>4</td>
<td>4 (5.4%)</td>
</tr>
</tbody>
</table>
Summary

• Metadata is on good track
  – SHACL plus good comms/UI should do the trick
  – Suggestions after SHACL failures a need

• Data is not very good quality
  – Who takes responsibility of processing for linking?
Who runs the interlinking process?

Who maintains the mapping?
Summary

• Technology readiness gap?
  – Tools not yet there in full force
  – Or no one knows how to configure them

• Organisational challenge
  – Harder than a very large company
Next steps - Technical

- Keen to get our hands on the final versions of the Tabular to KG challenge contestants.
- Assess linking run by EDP
Next steps - Social

• Put data consumers in the loop
  – Validate with consumers

• Social coding == Social data-ing?
  – Patching and patch approval
  – Share re-uses and transformations?
  – Spam is a concern
Thank you!

l.d.ibanez@soton.ac.uk

www.europeandataportal.eu