Publishing Linked Data Requires More than Just Using a Tool

G. Atemezing, F. Gandon, G. Kepeklian, F. Scharffe, R. Troncy, B. Vatant, S. Villata
Datalift Data Camp – 09-10 October 2012

- 2 * 1-day camp
- Co-organized with data.gouv.fr
- 75 participants

Goals
- Present the Datalift platform
- Bring your dataset … we will lift it to 5 stars data, and interlink it with other datasets
### Many working groups with a mentor

<table>
<thead>
<tr>
<th>Domain</th>
<th>Illustrative scenario</th>
<th>Original dataset</th>
<th>Original format</th>
<th>Datasets for interlinking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Find translations of given names to different languages.</td>
<td>Opendata.paris.fr</td>
<td>CSV</td>
<td>DBpedia</td>
</tr>
<tr>
<td>Tourism, Culture and Events</td>
<td>Lift cultural events data in the region Picardy</td>
<td>Yellow pages, Regional Tourism Office</td>
<td>XLS</td>
<td>EventMedia</td>
</tr>
<tr>
<td>Transport</td>
<td>Lift bus stops and lines transport</td>
<td>De Lijn bus company in Flanders (Belgium)</td>
<td>CSV GTFS</td>
<td>--</td>
</tr>
<tr>
<td>Data Catalogues</td>
<td>Convert data catalogues in DCAT and use thesaurus such as EUROVOC for terms classification</td>
<td>Opendata.gouv.fr, Opendata.montpelliernumerique.fr</td>
<td>XLS CSV</td>
<td>--</td>
</tr>
<tr>
<td>Budget of collectivities</td>
<td>Convert data using DataCube, use SPARQL aggregation functions</td>
<td>Rennes, Montpellier, Toulouse</td>
<td>XLS</td>
<td>rdf.insee.fr</td>
</tr>
<tr>
<td>Geolocation</td>
<td>Interlinking and publishing different shape files with temporal data</td>
<td>OpenstreetMap France Temporal series of agricultural data (confidential data)</td>
<td>SHP CSV</td>
<td>data.ign.fr, rdf.insee.fr</td>
</tr>
<tr>
<td>Environment</td>
<td>Lift data of grapes of given parcel</td>
<td>Suez Environment INRA</td>
<td>CSV</td>
<td>DBpedia, data.ign.fr</td>
</tr>
</tbody>
</table>
**Datalift goals**

- From raw data … to 5 stars interlinked data
- One click download:
  - [https://gforge.inria.fr/scm/?group_id=2935](https://gforge.inria.fr/scm/?group_id=2935)
- Cross–platforms, triple store installed and configured
  - Sesame (default)
  - OWLIM, AllegroGraph, Virtuoso (connectors)
Datalift workflow

- Two important results:
  - rdf.insee.fr
  - data.ign.fr
Difficulties and wishes / need support

- **Vocabulary**: choose suitable vocabularies to describe the data
  - Crash course on class/property modeling

- **RDF Conversion**:  
  - Adopt a URI naming scheme to identify the data objects  
  - Simple functionality proves useful: `String2URI` module

- **Automatic detection of datasets to link to**  
  - Dbpedia is not the only dataset you want to link to
Linked Open Vocabularies (LOV)

- Selection of 260+ vocabularies (20k elements), [http://lov.okfn.org/](http://lov.okfn.org/)
  - LOV bot: [https://github.com/pyvandenbussche/lov](https://github.com/pyvandenbussche/lov)
  - LOV search: [https://github.com/pyvandenbussche/lovSearch](https://github.com/pyvandenbussche/lovSearch)
  - LOV API: [https://github.com/pyvandenbussche/lovAPI](https://github.com/pyvandenbussche/lovAPI)

- LOV module within the Datalift platform
Conclusion

- Multilingual vocabularies are important
- Hide the complexity of SPARQL with natural language QA systems such as QAKIS
- SHP files are important, need more tooling
- INSPIRE directive and W3C GLD vocabularies
- GTFS / DSPL and semantic web formats

“A little data lifted goes a long way”.