Application of LOD to Enrich the Collection of Digitized Medieval Manuscripts at the University of Valencia

José Manuel Barrueco
barrueco@uv.es
Cristina García Testal
testal@uv.es
University of Valencia (Spain)
1. The UV manuscripts collection
2. What are we trying to do
3. Implementation of LOD at the UV
   • Data sources used
   • Application development
   • How it looks like
4. Results
5. Questions for LOD consumers:
   • Data sources available
   • Quality of the data
   • Licenses used
6. Conclusions
2: The UV manuscripts collection:

- The UV ancient books collection:
  - Manuscripts: +1,100 volumes going back to the XIII century
  - Incunabula: 334 volumes
  - Printed books (XVI – XIX centuries): +40,000 volumes

- The UV has been involved in digitization projects since 2000.

- Partner in the Europeana Regia project (2010-2012):
  - EU founded project to create a virtual library with the most important European royal collections of documents from the Middle Ages to the Renaissance.
  - Bibliothèque nationale de France, Bayerische Staatsbibliothek, Herzog August Bibliothek and the Koninklijke Bibliotheek van België.
  - [http://www.europeanaregia.eu](http://www.europeanaregia.eu)
  - The UV contributes with 92 codex (Royal Library of the Aragonese Kings of Naples).
  - They have been used as test bed for this work
1: What are we trying to do:

- Explore the opportunities of LOD to enrich the collection of digitized medieval manuscripts by providing additional information about authors:
  - Name (with variations)
  - Occupation (Historian, Poet...)
  - Biography
  - Picture
  - Main works

- Integrate LOD into a productive library application:
  - book viewer for digitized materials.

- Analyze the problems faced by institutions whiling to consume LOD:
  - Availability of data sources
  - Licenses used
  - Technical issues ...
3: Implementation at the UV:

What?

- We want to provide for each author (at least):
  - Name (with variations)
  - Occupation (Historian, Poet...)
  - Biography
  - Picture
  - Main works

- Integration of the data in the book viewer
- Not storing locally any RDF data
- Working with XML -> HTML conversions on the fly
- Storing the resulting data as HTML to present to the user
- Crawling the web of data
- Starting point: VIAF
  - Including VIAF URIs in the authority records of the institutional repository

How?

Application of LOD to enrich the collection of digitized manuscripts ...
3: Implementation at the UV:

- Data sources used:
  - VIAF
  - DBpedia
  - DNB
  - IdRef
  - BNF
  - YAGO
  - KB
  - Freebase
  - OpenCyc
  - Gutendata
  - Es.dbpedia
  - UV

Application of LOD to enrich the collection of digitized manuscripts ...
3: Implementation at the UV:

- **Digitized books included in institutional repository:**
  - DSpace with locally developed book viewer:
    - METS metadata + JP2000 image files
    - XSLT (METS -> HTML)
    - IIP image server
  - Application developed (perl + xslt):
    - Input: VIAF URI for each author or contributor
    - Dereference URI:
      - Take name variations in foaf:name
      - Dereference owl:sameAs link to dbpedia if exist
        1. Take dbpedia-owl:abstract [en|es|ca]
        2. Take foaf:depiction or dbpedia-owl:thumbnail
        3. Take dbpedia-owl:occupation (follow URI until spanish label)
        4. Take dbpprop:notableWorks (follow URI to works description)
        5. Dereference owl:sameAs link to es.dbpedia if exists
    - Repeat 1-4 completing missing data
  - Output: HTML static page with the description of the author

Application of LOD to enrich the collection of digitized manuscripts…
Application of LOD to enrich the collection of digitized manuscripts...
3: Implementation at the UV:

Application of LOD to enrich the collection of digitized manuscripts
3: Implementation at the UV:

Application of LOD to enrich the collection of digitized manuscripts...
4: Results:

- 92 manuscripts used as test bed:
- 97 authors and coauthors with VIAF URIs:
- 73 main authors and 24 scribes, illuminators, miniaturists...

<table>
<thead>
<tr>
<th></th>
<th>All authors</th>
<th>Main Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name forms (from VIAF)</td>
<td>97 (100%)</td>
<td>73 (100%)</td>
</tr>
<tr>
<td>Biography (from dbpedia)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>37 (38.14%):</td>
<td>37 (50.68%):</td>
</tr>
<tr>
<td></td>
<td>33: in English and Spanish</td>
<td>33: in English and Spanish</td>
</tr>
<tr>
<td></td>
<td>4: only in English</td>
<td>4: only in English</td>
</tr>
<tr>
<td>Picture (from dbpedia)</td>
<td>31 (31.95%)</td>
<td>31 (42.46%)</td>
</tr>
<tr>
<td>Occupation (from dbpedia &amp; es.dbpedia)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 (8.24%):</td>
<td>8 (10.95%):</td>
</tr>
<tr>
<td></td>
<td>7 spanish label</td>
<td>7 spanish label</td>
</tr>
<tr>
<td>Main works (from dbpedia &amp; es.dbpedia)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 (5.15%):</td>
<td>5 (6.84%):</td>
</tr>
<tr>
<td></td>
<td>2 URIs to works description</td>
<td>2 URIs to works description</td>
</tr>
<tr>
<td>Main works (from gutendata)</td>
<td>10 (10.30%)</td>
<td></td>
</tr>
</tbody>
</table>

Application of LOD to enrich the collection of digitized manuscripts ...
How to know the data sources available?
- Need for registries of linked data sets
- The datahub (http://datahub.io)

What datasets are available for use?
- September 2011: Bizer et al. [1] identified 295 linked open datasets
- October 2013: 8,920 datasets, 891 of them are LOD (10%)
- That sounds good! but ..... 
  - bioportal ontologies: 244 datasets
  - rkb-explorer: 55 datasets
  - ~ 594 LOD data sets

5: Questions for LOD consumers:

- What is the scope of the available datasets?
5: Questions for LOD consumers:

- Are they compliant with best practices for data provisioning?

RDF links pointing at other data sources:

<table>
<thead>
<tr>
<th>Out-Links</th>
<th>Datasets (Sep 2011)</th>
<th>Datasets (Sep 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No links</td>
<td>30 (10.17 %)</td>
<td>212 (23.79 %)</td>
</tr>
<tr>
<td>up to 1,000</td>
<td>90 (30.51 %)</td>
<td>243 (27.27 %)</td>
</tr>
<tr>
<td>1,000 to 10,000</td>
<td>58 (19.66 %)</td>
<td>190 (21.32 %)</td>
</tr>
<tr>
<td>10,000 to 100,000</td>
<td>45 (15.25 %)</td>
<td>135 (15.15 %)</td>
</tr>
<tr>
<td>100,000 to 1,000,000</td>
<td>43 (14.58 %)</td>
<td>69 (7.74 %)</td>
</tr>
<tr>
<td>more than 1,000,000</td>
<td>29 (9.83 %)</td>
<td>42 (4.71 %)</td>
</tr>
<tr>
<td></td>
<td>295</td>
<td>891</td>
</tr>
</tbody>
</table>

Provide dataset-level metadata:

<table>
<thead>
<tr>
<th></th>
<th>Datasets (Sep 2011)</th>
<th>Datasets (Sep 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiD descriptions</td>
<td>95 / 295 (32.20 %)</td>
<td>222 / 891 (24.91 %)</td>
</tr>
<tr>
<td>Sitemaps</td>
<td>53 / 295 (17.97 %)</td>
<td>87 / 891 (9.76 %)</td>
</tr>
<tr>
<td>voiD description or sitemaps</td>
<td>109 / 295 (36.95 %)</td>
<td>243 / 891 (27.27 %)</td>
</tr>
<tr>
<td>voiD description and sitemaps</td>
<td>66 / 891 (7.40 %)</td>
<td></td>
</tr>
<tr>
<td>nothing</td>
<td>186 / 295 (63.05 %)</td>
<td>648 / 891 (72.72 %)</td>
</tr>
<tr>
<td></td>
<td>295</td>
<td>891</td>
</tr>
</tbody>
</table>

Application of LOD to enrich the collection of digitized manuscripts …
5: Questions for LOD consumers:

- Licenses used to distribute the data:
  - In which way can we use the data sets?
  - Are they really open?

- 197 (22.33%) datasets without license information
- 694 (77.89%) with some type of license

<table>
<thead>
<tr>
<th>License type</th>
<th>Datasets (Sep 2013)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undefined license model (open)</td>
<td>287 (41.48%)</td>
</tr>
<tr>
<td>Creative Commons</td>
<td>277 (39.91%)</td>
</tr>
<tr>
<td>Open Data Commons</td>
<td>92 (13.25%)</td>
</tr>
<tr>
<td>Undefined license model (not open)</td>
<td>49 (5.49%)</td>
</tr>
<tr>
<td>UK Open Government Licence</td>
<td>23 (3.31%)</td>
</tr>
<tr>
<td>ukcrown-withrights</td>
<td>6 (0.86%)</td>
</tr>
<tr>
<td>GNU Free Documentation License</td>
<td>6 (0.86%)</td>
</tr>
<tr>
<td>General Public License</td>
<td>2 (0.28%)</td>
</tr>
<tr>
<td>apache</td>
<td>1 (0.14%)</td>
</tr>
<tr>
<td></td>
<td>694</td>
</tr>
</tbody>
</table>
As consumers of LOD we would ask for:
- comprehensive registries of data sources
- comprehensive metadata at data set level
- licenses following any of the available models (CC, DC, ...)
- more owl:sameAs links to interconnect data islands
- ... more data sets

As librarians implementing an application of LOD:
- we have been able to easily develop an application to integrate LOD in a collection of our institutional repository
- Enrich the collection of manuscripts providing biographical information for almost half of the authors
- Our future plan:
  - to extent the coverage to other materials starting by early printed books
Thanks for your attention!

more information: http://somni.uv.es

José Manuel Barrueco
barrueco@uv.es

Cristina García Testal
testal@uv.es