Status Quo
and
(current) limitations of Library Linked Data

Asunción Gómez-Pérez¹, Philipp Cimiano² and
Daniel Vila-Suero¹

¹Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid
http://www.oeg-upm.net
asun,dvila@fi.upm.es

²AG Semantic Computing, CITEC, University of Bielefeld
cimianog@cit-ec.uni-bielefeld.de

Acknowledgments: BNE team, DNB team, Uldis Bojars (NLL), Jodi Schneider (NUIG) and others

SWIB12 – Cologne, 28-11-2012
• The Library Linked Data Cloud
• A (library) Linked Data Life-cycle
• A collection of current limitations
• Conclusions
Library Linked Data cloud
New version of datos.bne.es by beginning of 2013 including:
- Links to digital objects
- More links to external datasets
- APIs and improved documentation
And many others (see http://thedatahub.org/group/lld)
• **Availability** of Library Linked Data is already a reality

• Many **serious efforts**: id.loc.gov, VIAF, DNB, data.bnf.fr, Bibliographic Framework Initiative etc.

• Still **several challenges** and **limitations** preventing LLD full potential
A (library) Linked Data life-cycle
A (Library) Linked Data Life-cycle

- A series of steps or phases
- Based on Villazón-Terrazas et al.*
- Others: LOD2, Datalift, etc. (see http://www.w3.org/2011/gld/wiki/GLD_Life_cycle)

A (Library) Linked Data Life-cycle

**SPECIFICATION**

**MODELLING**

**RDF GENERATION**

**LINKING ENRICHMENT**

**PUBLICATION**

**EXPLOITATION**

Definition and analysis of source data and their format, structure, etc.
A (Library) Linked Data Life-cycle

- Selection and reuse of vocabularies to represent LD.
- Creation of new local terms and mapping to existing vocabularies
A (Library) Linked Data Life-cycle

- SPECIFICATION
- MODELLING
- RDF GENERATION
- LINKING ENRICHMENT
- PUBLICATION
- EXPLOITATION

Taking source data, and vocabularies:

*Mapping (cross-walk) source data to produce RDF*
A (Library) Linked Data Life-cycle

- Discover related resources (ideally in RDF form).
- Enrich RDF data with data from other sources (e.g. substitute literals by URIs in other datasets)
A (Library) Linked Data Life-cycle

- Setup the infrastructure to expose your data to the Web (SPARQL, APIs, dumps).
- Enable discovery of your data (sitemap, voID)
- Include data provenance, license, etc.
A (Library) Linked Data Life-cycle

SPECIFICATION

MODELLING

RDF GENERATION

LINKING ENRICHMENT

PUBLICATION

EXPLOITATION

- Produce user interfaces that integrate your data (and other sources)
- Provide innovative services on top of the data
- Integrate with existing services (e.g. patron services), etc.
A collection of current limitations
• MANY source formats, encodings and schemas:
LIMITATIONS:

1. Lack of principled methods, techniques and tools to deal with heterogeneous source formats, schemas and encodings

   - metamastream, metamorph APIs from culturegraph

2. Need for analysis of the semantics of metadata schemas and the variation of their usage across libraries:

   - MARC 21: "Marc21 as Data: A start". Karen Coyle's code4lib's article*

* http://journal.code4lib.org/articles/5468
Many different models and approaches
LIMITATIONS:

1. Difficulties in using vocabularies that adapt to
   - Past and current cataloguing practices
   - Past and current library formats and schemas

2. Mapping and managing vocabularies
   1. Lack of mapping across vocabularies
   2. High manual effort and costly process
      - GND ontology,
      - Dunsire et al "Linked Data vocabulary management"*

3. Lack of multilingual vocabulary elements description
   - IFLA Namespace Multilingual efforts

RDF Generation

- Several tools, APIs and systems
- Almost each new project follows its own approach → it is still a costly process
- Participation of library experts is crucial (those who know the formats, e.g. MARC 21)
- There is not a generic mapping from source metadata schemas (due to the differences in use across libraries, different cataloguing practices)
LIMITATIONS:

1. Lack of tools and services that are easy to use by non-technical users (experts in library formats)

2. Lack of integrating tools and services that provide a full view of the mapping and RDF generation process
   - Analytics of the source data (e.g. how often is an specific MARC subfield used, how many records will an specific transformation rule affect.. )
   - Dashboard-like generation that helps to understand how the RDF data has been produced and allows to refine the mappings, find errors, etc.

http://marimba4lib.com
• Already a **high number of links at the authority level** (mostly Persons and Corporate Bodies)
  - VIAF
  - Culturegraph Authorities
  - DNB, BNF, BNE, BL

• Very positive efforts:
  - National library cataloguers adding links during the cataloguing process (e.g. DNB, BNE)
  - Cross-library collaboration for linking (British Library and DNB*)

LIMITATIONS:

1. Very limited linking at the bibliographic level

2. Lack of cross-lingual mechanisms to link resources across libraries

3. Lack of the a solid infrastructure to enable links sharing and exchange across libraries

4. No semantic enrichment of the content (textual, sound and visual) and linkage of this content to the URIs that represent the real-world entities
LIMITATIONS:

1. No extensive use of mechanisms to indicate provenance, license, last-update, etc. in a per-resource basis

2. Low usage of mechanisms to enable efficient discovery and usage like voID descriptions

3. Need for scalable and generic infrastructure to facilitate consumption of linked data:
   1. Most of the APIs are not extensively documented
   2. Not every dataset provides search over the data
   3. W3C Linked Data platform WG proposition is promising (REST access to resources and containers: paging, etc.)
LIMITATIONS:

1. Lack of integration of library linked data in
   - Library curation and cataloguing workflows
   - Existing library systems (e.g. digital library systems, patron services)

2. Need of end-user interfaces providing enriched information spaces
   - that integrate several LD sources,
   - allowing for serendipitous discovery and enhanced navigation

CultureSampo,
http://uilld2013.linkeddata.es Call for participation!
Conclusions
Conclusions

• There still exist some barriers:

  • **Organizational:** Infrastructure costs, integration of linked data into library processes, cross-library collaboration

  • **Language:** Multilingual services, cross-lingual linking and data integration

  • **Data formats and modalities:** Different digital objects and representations are rarely interlinked, cope with heterogeneity of formats and vocabularies

• But we are on the right path!
Vielen Dank!
Thank you very much!
Questions?