Metadata Quality: Learning from Open Data Portalwatch

Axel Polleres

twitter: @AxelPolleres

web: polleres.net
SPARQL

Open Data

Linked Data

Latest release 04-30-2018 - 1184 Datasets
Linked Data

But: **Open Data** is more than Linked Open Data...

Latest release 04-30-2018- 1184 Datasets
Open Data is a Global Trend!

- EU & Austria, but also the (previous) US and UK administration are/were pushing Open Data!
A lot of Open Data is not Linked Data

- Cities, International Organizations, National and European Portals, Int'l. Conferences:

  - Un data
  - The World Bank Open Data
  - DBpedia
  - Wienat
  - NYC Open Data
  - London.gov.uk
  - European Union Open Data Portal
  - Data.gov.uk
  - Data.gov
  - Open Data Portal Österreich
  - European Data Forum

We are aware of currently 260 active such portals worldwide.
Different portals...

[Images showing examples of data portals and datasets]

- Data.gov
- data.gv.at

These portals offer access to various datasets, including housing affordability and education spending data.
What’s the problem(s)?

- Metadata is **heterogeneous** and (partially) messy
  - Software-specific metadata (CKAN vs Socrata vs …)
  - Portal-specific metadata
  - Missing metadata (file formats, API descriptions, …)

- Metadata not available as Linked Data
  - Only partially in DCAT vocabulary
  - **No mappings** for additional metadata fields

- Poor **discoverability** of datasets
  - No content information in metadata (e.g., CSV headers)
  - Datasets’ metadata not optimized for search engines

- (Meta-)data becomes **stale/offline/outdated**
Open Data Portal Software

CKAN … http://ckan.org/

- almost "de facto" standard for Open Data Portals
- facilitates search, metadata (publisher, format, publication date, license, etc.) for datasets

- http://data.gov/
- http://data.gv.at/

- machine-processable? …
  … partially
Our solution:

- Open Data Portalwatch
  - Monitoring Metadata quality
  - Mapping to standard vocabularies
  - Enriching Metadata to improve search
1) Monitoring and QA over evolving data portals

<table>
<thead>
<tr>
<th>Date</th>
<th>Portals</th>
<th>CKAN</th>
<th>Socrata</th>
<th>ODSof</th>
<th>DCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/2015 [1]</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Demo:

http://data.wu.ac.at/portalwatch/portal/data_gov/1818
2) Mapping to Standard vocabularies & Linked Data

- Mapping & Heuristic Enrichment
  - DCAT
  - PROV
  - CSVW
  - Schema.org

- Enable uniform access:
  - SPARQL endpoint
  - Linked Data & Memento Protocol

Finally:

3) Why is Search in Open Data a problem?
Why is Search in Open Data a problem?

https://www.youtube.com/watch?v=kCAymmbYlvc

Structured Data in Web Search by Alon Halevy

Open Data Search is hard...

a) No natural language „cues“ like in Web tables...

b) Existing knowledge graphs don’t cover the domain of "Open Data"

c) Open Data is not properly geo-referenced
Some starting points:

- First baby steps on building an Open Data Knowledge Graph:

- Ongoing work to enable **spatio-temporal search** in Open Data e.g. in our project communidata.at (just submitted to JWS)
Demo:

http://data.wu.ac.at/odgraphsearch/

Idea:

Link OD to geospatial and temporal Knowledge bases:
- Geonames
- OpenStreetMap
- Perio.do
- Wikidata
- ...
Take-home messages:

- Scalable monitoring of metadata quality and evolution is possible and useful

- Meta-Data Quality can be improved by IE and looking into the data

- Linked Data and standard protocols (SPARQL, Memento)
  - helps to provide an integrated view
  - Enable combination with other sources to improve search
Other Ongoing Projects (data.wu.ac.at)

WU Open Data Portal
WU lectures, rooms and organizations
data.wu.ac.at is an Open Data portal where you can find data about lectures, rooms and organizations at WU.
121 datasets

Open Data Portal Watch
Monitoring & exposing portals' metadata
Open Data Portal Watch assesses the evolution of the (meta) data quality of about 260 Open Data portals over since September 2014.
259 portals

DBpedia Wayback Machine
Extract past DBpedia versions
The DBpedia Wayback Machine aims at providing the wayback functionality for DBpedia based on the revisions of their Wikipedia article.

Jupyter Notebook Server
Programming & Documentation
Notebook documents are documents which contain both computer code (e.g. python) and human-readable rich text elements.
Only available within local WU Vienna network

CSV Engine
Search & enrich CSVs
The CSV Engine is a collection of tools and services for processing and enriching CSV files.

Open Data AT Assistant
Search chatbot for Austrian datasets
The assistant will help you to explore the content of the austrian open data portals: data.gv.at and opendataportal.at.