What's up LOD Cloud
Observing The State of Linked Open Data Cloud Metadata

Ahmad Assaf, Raphaël Troncy And Aline Senart

@ahmadaassaf
LOD Cloud

- LOD Cloud is a mine of data
- The heterogeneity of sources reflect directly on the data quality
- Finding useful dataset without prior knowledge is increasingly difficult
What's up LOD Cloud

- LOD Cloud is a mine of data
- The heterogeneity of sources reflect directly on the data quality
- Finding useful dataset without prior knowledge is increasingly difficult

Demonstrate the of LOD Cloud metadata by running Roomba on the accessible LOD Cloud through datahub.io
Dataset Metadata

- Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve use or manage information resources.
- Data Portals are a curated collection of datasets metadata providing a set discovery and integration services.
- We divided the metadata information into four main types:
  - **General information**
    - e.g. title, description
  - **Ownership information**
    - e.g. author, maintainer_email
  - **Access information**
    - e.g. license_title, license_id
  - **Provenance information**
    - e.g. creation_date, version

What's up LOD Cloud - Observing The State of Linked Open Data Cloud Metadata
Metadata Information Groups

| Resource | Tag | Group | Organization |
|----------|-----|-------|--------------|
| Actual raw data that can be downloaded or accessed directly e.g. JSON, CSV, SPARQL endpoint | Descriptive knowledge about the dataset contents and structure. This can range from simple textual tags to semantically rich controlled terms | Organizational units that share common semantics. They can be seen as a cluster or curation based on shared themes/categories | Clustering or curation solely based on associations with specific administration parties |

What's up LOD Cloud - Observing The State of Linked Open Data Cloud Metadata
Roomba - An Extensible Framework to Validate and Build Dataset Profiles

Roomba addresses the challenges of automatic validation and generation of descriptive dataset profiles.
Roomba - An Extensible Framework to Validate and Build Dataset Profiles

Roomba addresses the challenges of automatic validation and generation of descriptive dataset profiles.

(i) Data Portal Identifier
   - CKAN
   - D3
   - Socrata

(ii) Metadata Extractor
    - Sampler
      - Random Sampling
      - Weighted Sampling
    - Validator
      - CSV Validator
      - RDF Validator

(iii) Instance and Resource Extractor

(iv) Profile and Report Generator
    - Topical Profiler
    - Statistical Profiler

(v) Profile Validator
    - General Profiler
    - Ownership Profiler
    - Access Profiler
    - Provenance Profiler

https://github.com/ahmadassaf/opendata-checker/
What's up LOD Cloud - Observing The State of Linked Open Data Cloud Metadata

Metadata Errors for Information Groups

- 41% of ownership information is missing or undefined
- Resources have the poorest metadata health across information groups
  - 64% general metadata
  - 100% access metadata
  - 80% provenance metadata
Top Metadata Errors

- 23.8% of these errors can be fixed automatically
- 33.33% can be fixed automatically by tools plugged into the data publishing workflow
23.8% of these errors can be fixed automatically

33.33% can be fixed automatically by tools plugged into the data publishing workflow
Top Metadata Errors

- 23.8% of these errors can be fixed automatically
- 33.33% can be fixed automatically by tools plugged into the data publishing workflow

| Metadata Field       | Error % | Section  | Error Type | Auto Fix |
|----------------------|---------|----------|------------|----------|
| group                | 100%    | Dataset  | Missing    | -        |
| vocabulary_id        | 100%    | Tag      | Undefined  | -        |
| url-type             | 96.82%  | Resource | Undefined  | -        |
| mimetype_inner       | 95.88%  | Resource | Missing    | -        |
| hash                 | 95.51%  | Resource | Undefined  | Yes      |
| size                 | 81.55%  | Resource | Undefined  | Yes      |
| cache_url            | 96.9%   | Resource | Undefined  | -        |
| webstore_url         | 91.29%  | Resource | Undefined  | -        |
| license_url          | 54.44%  | Dataset  | Missing    | Yes      |
| url                  | 30.89%  | Resource | Unreachable| -        |
| license_title        | 16.6%   | Dataset  | Undefined  | Yes      |
| cache_last_updated   | 96.91%  | Resource | Undefined  | Yes      |
| webstore_last_updated| 95.88%  | Resource | Undefined  | Yes      |
| created              | 86.8%   | Resource | Missing    | -        |
| last_modified        | 79.87%  | Resource | Undefined  | Yes      |
| version              | 60.23%  | Dataset  | Undefined  | -        |
| maintainer_email     | 55.21%  | Dataset  | Undefined  | -        |
| maintainer           | 51.35%  | Dataset  | Undefined  | -        |
| author_email         | 15.06%  | Dataset  | Undefined  | -        |
| organization_image_url| 10.81%| Dataset  | Undefined  | -        |
| author               | 2.32%   | Dataset  | Undefined  | -        |

Table 1: Top metadata fields error % by information type
Metadata Errors

Information Field

% of resources

- cache_last_updated
- webstore_last_updated
- size
- hash
- format
- mimetype
- mimetype_inner
- cache_url
- name
- webstore_url
- last_modified
- resource_type
- description
- url_type
- created
Top Metadata Errors

- 25% of the datasets access information are not clean [unavailability mainly]
- 68% of the resources access data can be fixed automatically
- 31.27% of the resources were not reachable
- 63.17% of the resources don’t have valid resource_type

  - Creating an aggregate report for resources > format:title
    - 62.16% of the datasets have defined SPARQL endpoints using the api/sparql resource format
    - 92.27% provided RDF example links
    - 56.3% provided direct links to downloadable RDF dumps
Metadata File Types Errors
Metadata Errors

- Noisiest part of the access metadata is license information
- 16.6% of the datasets don’t have license_title and license_id
- 54.44% datasets don’t have license_url
- 51.35% don’t have maintainer
- 55.21% of the datasets are missing maintainer_email | 15.06% author_email
- 80% of the provenance information is missing or undefined
- The only manual field in provenance information “version” is missing from 60.23% of the datasets
Enriched Profiles

- 1.87% of the resources have incorrect `mimetype` defined
- 4.82% of the resources have incorrect size values
- 47.49% of the datasets license information have been normalized via the manual license mapping file

† https://github.com/ahmadassaf/opendata-checker/blob/master/util/licenseMappings.json
Questions?

Ahmad Assaf

http://ahmadassaf.com
@ahmadaassaf
http://github.com/ahmadassaf