Proposal for a Session at the Berlin Workshop: First Steps towards interoperability of key open data asset descriptors

Version 0.9

Problem/issue we are going to address: Providing interoperable metadata on key asset descriptors (Suggested are: Provenance, licenses and tags).

Expected outcome of the workshop: Suggestion

1. What should be included in which form in interoperable metadata? Why are these metadata key?
2. Possible steps to achieve a standard?

Intended audience: Open Data Publishers, Maintainers of open data portals, Scientists, Re-Users of Open Data

Agenda/Moderation methods:

1. Introduction of Participants
2. Collection of topics to be discussed (Method 3-6-5)
3. Voting on Agenda
4. Discussion
5. Conclusion/Recommendations

Facilities required: projector, flip-board, markers.

Improving the Interoperability of Open Data Portals

Dietmar Gattwinkel*, Konrad Abicht+, René Pietzsch^, Michael Martin+
** Staatsbetrieb Sächsische Informatik Dienste, Riesaer Straße 7 Haus D, 01129 Dresden*
+ Universität Leipzig, Institut für Informatik, AKSW/BIS, Augustusplatz 10, 04109 Leipzig
^ eccenca GmbH, Hainstraße 8, 04109 Leipzig

In the context of the European Digital Agenda[1] governmental authorities as well as virtual communities (e.g. datahub.io) have published a large amount of open datasets. This is a fundamentally positive development, however one can observe that many different that both for the metadata and the data itself different vocabularies are in use. Furthermore, the established vocabularies are often augmented with portal specific metadata standards and published in different (local) languages. If Open Data are to be integrated and aggregated across portals, this entails a lot of effort. In this paper we present examples how the problems of interoperability (section 1) and multilingualism (section 2) could be addressed for key open data asset descriptors. We focus on the analysis of today's problems and ways to solve them.

1. Improving interoperability by using core vocabularies

There exists already a variety of core vocabularies, which can be used for the modelling of many different facts. Such core vocabularies establish a common base of terms and concepts.
By increasingly using established Vocabularies one can increase the trust in one's data, lessen the workload for its creation and assists its reuse.

1.1 - Example 1: Who is accountable for the data?

One of the most important things about open data is the question is accountable for which datasets or resources. This is obviously important from a legal point of view, but will also influence how much trust people put into it. It is also important in the attempt to de-duplicate metadata catalogues after harvesting different repositories with entries referencing identical resources with different metadata.

Accountability has many nuances:

1. Who created the data and generated the dataset?
2. Who contributed in the data creation and who contributed to the generation of the dataset?
3. Who published the data?

[1] http://ec.europa.eu/digital-agenda/en/open-data-0

Obviously these can nuances cannot be dealt with by modeling accountability with a simple character string denoting a person or organisation. A more sophisticated approach references a (RDF) resource which allows to model additional statements about the person. Figure 1 illustrates both approaches.
Figure 1 illustrates both approaches:

Abbildung 1 - Two Approaches for modelling accountability

We recommend not to model accountability with a simple character string referencing distinct entities. In this approach after every name change (e.g. after marriage, divorce of person, or some organizational reshuffle) necessitates changing the metadata for every dataset attributed to this name. Using a property to reference the individual resource of this person no change is needed at the datasets metadata. Only the respective property (lastName) of the resource has to be changed. Furthermore, additional information about the person can be added by adding properties to the resource as needed and as your data model advances[2].
In order to codify the three mentioned types of accountability we propose to use the vocabulary of the Dublin Core Metadata Initiative[3]. Specifically `dctems:creator`[4] to denote the creator, `dcterms:contributor` to reference any contributing person and finally `dcterms:publisher` to reference the publisher of the data. The referenced value will be a resource of type `foaf:Person`[5]. The type `foaf:Person` is defined in the FOAF vocabulary[6], which is designed for describing persons, their activities and their relations to other people and objects.
Alternatively, it is possible to use the PROV standard[7], which provides properties and classes to express the provenance of a piece of data. The specification defines a model for statements about the different stages of data creation, connected activities and dates. Knowledge about the provenance of data helps to evaluate it and to use it more effectively. The modelling draws on both views in order to facilitate the use by a wide audience.

---

[2] Often, in setting up open data Portals, the problem is encountered that not all necessary metadata is agreed upon, let alone gathered with the same speed. In order to lower the time-to-publish agile and flexible publishing processes are required. One practical example that helps to establish such flexible publication processes is the MetaConf-Plugin (https://github.com/AKSW/ckanext-metaconf) for CKAN (http://ckan.org/). MetaConf allows the ckan administrator to flexibly define mandatory and optional metadata fields. These fields are to be filled according to their validation settings by the datasets owner or publisher during the ckan publication process. By this means metadata definitions for your datasets could be incrementally build upon established vocabularies and possibly extended when needed.
[3] http://dublincore.org/specifications/
[4] dcterms ist das Prefix für Dublin Core Terms und steht für http://purl.org/dc/terms/
[5] foaf ist das Prefix für Friend Of A Friend und steht für http://xmlns.com/foaf/0.1/
[6] http://xmlns.com/foaf/spec/
[7] http://www.w3.org/TR/prov-primer/

1.2 - Example 2: Licence Information

Licences inform about the rights and obligations of the data user of a published data set. it is therefore necessary, to gather and present licencing information in a clear and sufficiently detailed manner. Again we advise against simply stating the name of a licence as simple character string. Instead we give preference to an elaborate and detailed description of a license in form of an individual resource (or a set of resources). The materialization of a license should take into account that rights and obligations have priority for a prospect dataset user.
A possible approach is shown in *listing 1*[8]. Based on the RDFLicense[9] vocabulary it models the creative commons Licence CC-BY-NC-ND as RDF triple with the ODRL[10] vocabulary. For example, `odrl:duty` expresses the obligation that the source must be mentioned in a work where the dataset is used and `odrl:permission` specifies the rights granted to the user of the dataset. The CC-BY-NC-ND for example grants the rights to copy and distribute the source data. But it explicitly excludes the permission for commercial use as expressed by `odrl:prohibition`.

We advise to add license information (like shown in Listing 1) as ingredient in your standard metadata information for a dataset. This ensures that your users have all relevant information specifying the granted permissions in one place without the need for further sources in order to fully understand the legal effect of the chosen license.

```
<http://purl.org/NET/rdflicense/cc-by-nc-nd3.0nl>
      rdfs:label "Creative Commons CC-BY-NC-ND Netherlands" ;
      [...]
      odrl:duty
              [ a odrl:Duty ;
                odrl:action cc:Attribution , odrl:attachPolicy
              ] ;
      odrl:permission
              [ a odrl:Permission ;
                odrl:action cc:Distribution , cc:Reproduction
              ] ;
      odrl:prohibition
              [ a odrl:Prohibition ;
                odrl:action odrl:commercialize
              ] .
```

Listing 1: DF-Triples about Creative Commons Lizenz BY-NC-ND (Netherlands)

Alternatively, there is the ODRS[11] vocabulary, which is in turn is inspired by the ORRL, among others. ODRS aims to model legal statements about the relationships of a dataset. It was

[8] Source of the snippet is http://rdflicense.appspot.com/rdflicense. CC-BY-NC-ND is the first licence of the file.
[9] http://datahub.io/es/dataset/rdflicense
[10] https://www.w3.org/community/odrl/
[11] http://schema.theodi.org/odrs/

motivated by the fact, that existing vocabularies dien differentiate between a database and its content. Furthermore it attempts to model rights statements about several datasets as well as statements about specific rights associated with an individual dataset.

2. Multilinguality Support and Taxonomy Service for Open Data Portals

Many Open Data portals yield information composed in the respective national language. In order to ease search and re-use of this data for non-native speakers we propose several solutions. The tags in the national language could be enriched by tags in other languages, using a Taxonomy Service that takes a set of given tags and returns suitable tags in target languages. The Integration could be implemented as CKAN extension, which calls the service during the creation or editing of a metadata entry. Also possible would be the subsequent integration of the multilingual tags at edit time or by live querying the Taxonomy Service when a non-native speaking user visits the dataset page.

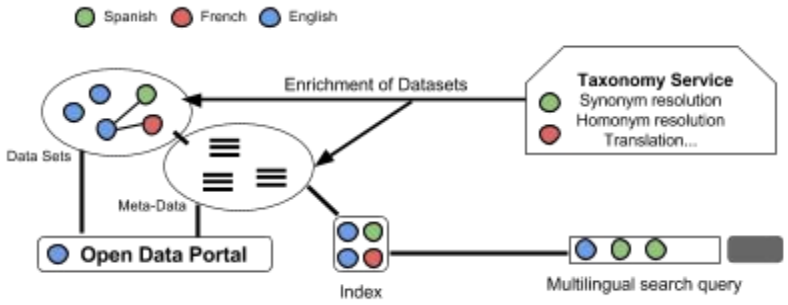

Figure 2: Illustration of taxonomy service

Besides the enrichment of the dataset metadata with additional tags we imagine a search integration. By providing a transparent multilingual search capability a user could find the best possible results for his search terms regardless of the language of the dataset. The search function is an important entry point for the exploration of a Data Portal and as such it should reveal reliable and relevant results in an acceptable response time. In order to achieve reasonable search query time efficient index and caching strategies are required. For performance reasons, search queries will only run against the local index. Queries containing foreign language tags can be expanded using using results provided by the Taxonomy Service.

2.1 Who is the Taxonomy Service Provider?

This is an important questions and we see three different approaches, each of which looks promising, but at the same time has some limitations.

- An existing organisation like the Open Data Institute (ODI) or the Open Knowledge Foundation could operate the service and maintain the taxonomies. This approach

would limit the scope for action to a small circle of experts, but their decisions would affect a large number of users.

- A community driven approach involving governmental agencies and professionals could increase the variety in the taxonomies at the cost of higher administration and coordination efforts.
- The third option is a largely automated solution. Algorithms could harvest existing portals for their use of taxonomies and derive the Service Taxonomy based on them. This solution scales better than the previous but needs an initial and ongoing evaluation of the used algorithms and the achieved taxonomy's data quality.

Eventually we like to point to the problem of trusting a third party service in determining the semantics and interpretation of ones own data? This questions is relevant both for administrative bodies and companies, which very often have very precise ideas about how they use terms and concepts.

3. Conclusion

The purpose of this document is to open up the discussion of possible improvements in the interoperability of key descriptors. We have presented some possible approaches. Others can be suggested. What is necessary is to limit the scope of any such endeavour to suggestions that can be implemented with sufficient ease within a limited amount time. Technical solutions must take into account the stability and long-term availability of reference data. Therefore we have tried to aim for solutions that balance established methods, existing standards and our own experiences. It is obvious as well, that Interoperability will not be achieved by core vocabularies alone. In addition to the technical side of things there is the human, which must be taken into account, too. But the vocabularies mentioned here ideally reflect the concrete decisions by many people and can therefore help lower still existing barriers for the re-use of public sector information.

Facilitator: Dietmar Gattwinkel

Projektleiter Open Government Data | Project Manager Open Government Data

STAATSBETRIEB SÄCHSISCHE INFORMATIK DIENSTE | SAXON IT SERVICES
Riesaer Straße 7 Haus D | 01129 Dresden
www.opendata.sachsen.de |dietmar.gattwinkel@sid.sachsen.de | opendata@sid.sachsen.de

Background
Dietmar Gattwinkel is Project Manager of the Open Government Data Project in the Free State of Saxony. Into this project he brings 12 years of experience in setting out Saxony's Web Strategy and a strong involvement in the overall e-government process. He is also representing Saxony in the project group implementing Germany's open data portal “govdata.de”. Prior to his work for the government he worked for a geo marketing company that pioneered the reuse of PSI in Germany. He holds a master’s degree in Communications, Law and Philosophy from Johannes-Gutenberg-Universität Mainz.

Contributor: Konrad Abicht

Wissenschaftlicher Ingenieur | Scientific Engineer

LEIPZIG UNIVERSITY, Institute of Computer Science, BIS/AKSW
Augustusplatz 10 | 04109 Leipzig
www.aksw.org/KonradAbicht | abicht@informatik.uni-leipzig.de

Background
Konrad Abicht works as scientific engineer at AKSW research group of the Leipzig University. His work topics are Linked Open Data / Semantic Web and modern Web technologies. He holds a master’s degree in Computer Science from Leipzig University. Besides working as an engineer, he is also a freelancer and entrepreneur.

Contributor: René Pietzsch

Techn. Projektleiter des Open Data Portal Leipzig | Techn. Project Manager Open Data Portal Leipzig

eccenca GmbH
Hainstraße 8 | 04109 Leipzig
www.eccenca.de | rene.pietzsch@eccenca.com

Background
René Pietzsch is Technical Project Manager of the Open Data Portal Leipzig and Product Owner at eccenca GmbH. He holds a diploma in Computer Science.

Contributor: Michael Martin, Ph.D.

Leader of research group AKSW/Emergent Semantics at Leipzig University

---

LEIPZIG UNIVERSITY, Institute of Computer Science, BIS/AKSW
Augustusplatz 10 | 04109 Leipzig
www.aksw.org/MichaelMartin | martin@informatik.uni-leipzig.de

Background

Michael Martin works as post-doc at Leipzig University and is group-leader of AKSW/Emergent Semantics research group. He has over 10 years experience in the Linked Data Web and Semantic Web.