Challenges for an Open, Domain-Adapted Metadata Repository: LingHub

Martin Brümmer, Bettina Klimek, Sebastian Hellmann

AKSW Technology Lab (http://aksw.org), Institute for Applied Computer Science (InfAI), situated at the University of Leipzig

Submitted as an *In-Use Presentation* to EDF

In 2011, the European Commission published its Open Data Strategy[1] defining the following six barriers for "open public data":

1. A lack of information that certain data actually exists and is available;
2. A lack of clarity of which public authority holds the data;
3. A lack of clarity about the terms of re-use;
4. Data which is made available only in formats that are difficult or expensive to use;
5. Complicated licensing procedures or prohibitive fees;
6. Exclusive re-use agreements with one commercial actor or re-use restricted to a government-owned company.

Taking these barriers as a starting point, it becomes obvious that no universal remedy exists. Instead these high-level problems have to be broken down into smaller challenges for specific use cases and domains in order to provide adapted solutions.
In this submission, we will refine these challenges to create a blueprint for the implementation of a metadata repository called "LingHub". "LingHub" intends to collect metadata for datasets adhering to the intersection of the following four topics:

- language resources and linguistic data;
- linked data (including the precursor data sources);
- openly licensed or freely available;
- scientific data value chains (with an emphasis on provenance).

Our work originates from the discussions within the Working Group for Open Data in Linguistics (OWLG)[2]. OWLG is one of 19 working groups[3] of the Open Knowledge Foundation (OKFN), each promoting open knowledge in their specific domain. OWLG mainly exists of individual researchers organized in a grassroots movement without institutional orchestration.
The realization of "LingHub", however, is supported by the recently started LIDER EU project[4] whose goal is to support communities such as OWLG and also provide: (1) A set of guidelines and best practices for building and exploiting LOD-based resources in multimedia and multilingual content analytics and for developing NLP services on top of Linguistic Linked Data. (2) A reference architecture for Linguistic Linked Data built on top of existing and future platforms and freely available resources. (3) A long-term roadmap for the use of Linked Data for multilingual and multimedia content analytics in enterprises.

[1] http://europa.eu/rapid/press-release_MEMO-11-891_en.htm
[2] http://linguistics.okfn.org/, The Open Linguistics Working Group. Christian Chiarcos et. al., LREC 2012
[3] http://okfn.org/wg/
[4] http://lider-project.eu/

In the next section, we will define basic problems of metadata repository, which will be followed by a section, where we name concrete challenges and requirements for our use cases.
After this, we will present our research on available metadata repositories and data models and describe their shortcomings to meet the given challenges. Finally, we give an overview of already existing tools and components, which enable us to showcase the feasibility of our blueprint for an adaptable metadata repository.

Problem Statement

Although a large variety of metadata repositories and data catalogues exist for extensive and heterogeneous types of data, there is a basic framework every metadata repository is bound to: the term *data*. Taking it back to its etymological origins reveals that data is the plural form of *datum* meaning 'given'. The verb "give" has a valency of three taking a subject and two objects: someone gives something to somebody. In consequence data is bound to a trinity consisting of (1) the producer or source of the data, (2) the entity that constitutes the data itself and (3) the receiver or user of the data[5]. This trinity supplies important details for implementation, when considering the three different perspectives of metadata collection:

1. the data producer, publisher or provider is in a unique position to provide contextual information about the circumstances under which the data came into existence (be it derived or original) such as license, provenance, contributors or additional notes;
2. the data itself can be inspected for metadata, gaining insight about technical information such as file size, format, uptime, availability, validity and in case of RDF and Linked Data: ontology used, links to other datasets, class and property structure;
3. finally, the user of the data can supply valuable information on the required information and the usefulness of data categories. Users are able to give direct feedback about the quality of data and metadata. Users have a direct incentive to correct and curate data, so that either their applications work properly or they are able to answer research questions (e.g. how many open dataset are available).

Within a metadata repository no dataset exists independently. Beginning with the first topic or entity someone collected data about, an infinite reuse of this data is possible. Most of the data provided is derived from another data source and very often new datasets develop out of mash-ups of existing data. Providing metadata about a dataset consequently involves the consideration of both the actual dataset and the source data it has been derived from. This is a crucial point not taken into account by already existing metadata repositories, which often ignore provenance and also versioning. Metadata quality is another issue with the ultimate goal to rely on a gapless provenance chain.

Specific challenges and requirements

Complete Metadata vs Entry Barrier: Acquiring the desired metadata is challenging, because it can increase the effort of publishing considerably. The more metadata the publisher is required to supply, the less he might be willing to do the effort. Although the effort of metadata entry can be reduced by automatic extraction, some information may only be known to the producer of the data. The trade-off between completeness of metadata and high publishing effort and entry barriers has to be carefully considered.
Incentives for Metadata Curation: For data providers it is already a lot of effort to publish data. Even more so, if they have to add information to many different metadata repositories and

[5] Lehmann, Christian (2004): "Data in linguistics." The Linguistic Review 21(3/4):275-310, p.278.

propagate updates manually as well. The risk of such statically entered metadata is that it becomes obsolete. As much metadata as possible should be automatically extracted from the datasets instead, to guarantee up-to-date information without increasing the workload for the publisher. This approach also allows to include extensive technical information like SPARQL endpoint URL and uptime, data format and validity and ontology usage. In addition to automatic extraction, distributing the load of maintenance among users to complete and consolidate data in a wiki approach is critical. Metadata repositories that supply vast amounts of any kind of data not regarding its structure and content must fear to end up as a huge data cemetery.

Metadata Coverage vs Metadata Quality: Crowdsourced data often has better coverage, but is not the same high quality as expert-curated sources. Provenance has therefore to be extended to metadata entry, allowing judgement of data and metadata quality by users and approved domain experts.

License and Attribution: Licensing is a constant issue in Linked Open Data. A compromise is needed between the needs of data publishers to have their data used correctly and their work cited accordingly and the wishes of users to access and reuse the data as freely as possible. Granular licensing and attribution information has to be explicitly included in the metadata to assure publishers to choose open instead of closed licenses. At the same time, data that is Open Access but closed license should also be included to increase repository scope and allow publishers to retain more rights if needed.

Metadata Visualisation: Metadata presentation is widely reduced to textual data in the form of "field name - field value" tuples. However, complex metadata like links between datasets and dataset categories can more effectively be represented in diagrams and images, giving the user better tools to find relevant data and understand its relations.

Granularity: Current repositories only account for dataset level metadata. However, datasets may include resources that are derived from different sources, which constitutes the need for intra-dataset provenance. RDF can be used to point into datasets and add granular provenance information if it is not provided in the original data. Too granular metadata, however, is cumbersome to maintain and may stop to be useful.

Lack of best practices and clear guidelines: Finally, we are lacking best practices and clear guidelines for modelling and publishing provenance and other metadata. Extent and desired granularity are as unclear, as the way of explicating it and conveying it to the user. There is no clear remedy for these challenges, rather, design decisions have to be grounded on the needs of producers and users alike.

Existing Repositories and Data Models

Datahub (http://datahub.io) is a platform developed by OKFN that enables users to upload, group and search data. It is build on the CKAN data management system and provides metadata about Linked Data datasets and free to edit and open for registration. By intention the metadata provided for most of the datasets is flat and simple. On the one hand, this facilitates the ease of upload and addition of datasets for providers but on the other hand it reduces the usefulness of the metadata and the accessibility of the data itself. Datasets in RDF format for example do not reference the data source they were derived from in a proper way. As a consequence datahub is incapable of adequately describing datasets with multiple source datasets and files such as the well-known DBpedia.

Another data repository is presented by **META-SHARE** (http://www.meta-share.eu/) which is part of the Multilingual Europe Technology Alliance (META) and aims at providing quality language resources. Focussing on the processing of the metadata a *strictly provider-driven* account is

taken. META-SHARE assumes a high quality of the entered data, because only scientific institutions are allowed to edit. Once the data is added to the repository no further validation of the data availability is done. For the data providers, however, it is often infeasible to update the metadata in regular intervals and there is no issue reporting by user requests.
This contributes to a rather statical way of data storage and leads to an unbalanced data repository favouring data preservation but - given the realm and possibilities of the Semantic Web and Linked Data - contributing little to an effective reuse of the data.

Current metadata models for dataset description like **VoID**[6] and the provenance ontology (**Prov-O**) [7] are fit to express informations about whole datasets. In practice, there is a need for more granular approaches, because there are many use cases that cannot be sufficiently modeled by these vocabularies. For example, the DBpedia is not one monolithic dataset, but an aggregation of 119 different language DBpedias[8]. Expressing this relation with the provenance ontology would include an "aggregation" Activity that merges the different datasets into one DBpedia. This obscures the semantic relation between the dataset and its sub datasets. It is currently also impossible to describe the single language DBpedia dump files formats sufficiently. Their MIME type, for example, can only express the "outer" data format, not the format of the contained data. Finally, existing models also lack expressivity in regards to technical descriptions, like the way content negotiation is handled.

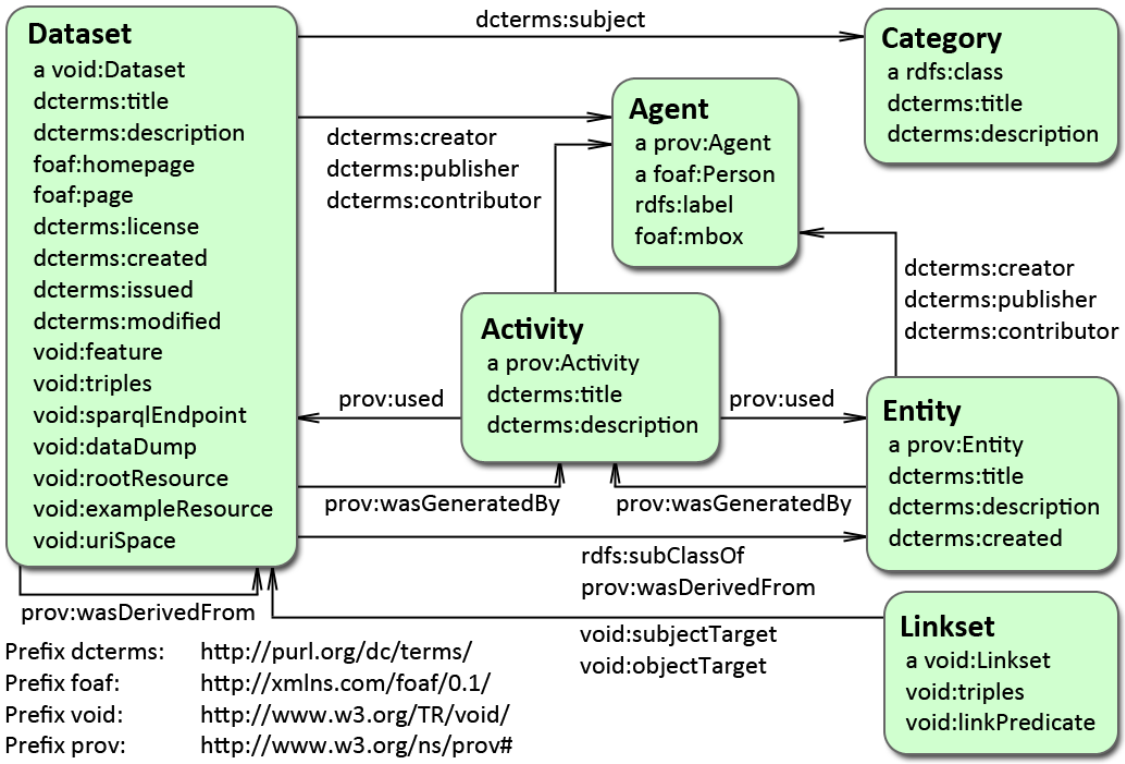

Figure 1: State of the art metadata model combining VoID and Prov-O

[6] http://www.w3.org/TR/void
[7] http://www.w3.org/ns/prov#
[8] http://downloads.dbpedia.org/3.9/

Existing Components and Outlook

A domain-adapted metadata repository has to face and solve these challenges if it is set to be a relevant and useful alternative to existing solutions. To tackle the presented issues, it can rely on a number of existing technologies and components. Ontowiki (http://ontowiki.net) can serve as a basis to represent the metadata to the user in various, domain-adaptable ways via its site extension and make it accessible to editing via its wiki functionality. The effort of metadata curation, coverage and quality judgement can be distributed between the user and the producer. LODStats (http://stats.lod2.eu/) can be used to analyze the provided data and extract technical and statistical metadata automatically such as number of triples, links to other datasets and ontologies used. Setting guidelines for metadata annotation within the datasets, dataset level metadata such as author, maintainer and provenance can also be automatically extracted, decreasing the overall manual effort and providing better maintainability. Further technical data can be retrieved by using the SPARQL Endpoint Status tool (http://sparqles.okfn.org/) to monitor SPARQL endpoint URLs and uptimes.

Contributor Names and short CVs

Martin Brümmer (AKSW, Universität Leipzig, Germany, bruemmer@informatik.uni-leipzig.de) has started as a researcher at AKSW technology lab in Dec. 2013. He is a contributor to the NLP2RDF and the DBpedia Project and was co-chair of the Multilingual Linked Data for Enterprises (MLODE) 2012 workshop. He contributed to the development of the Linguistic Linked Open Data Cloud (http://linguistics.okfn.org/resources/llod/). His research focus is on Linguistic Linked Open Data, NLP in the Semantic Web and Open Government Data.

Bettina Klimek (AKSW, Universität Leipzig, Germany, klimek@informatik.uni-leipzig.de) recently finished her Master studies in Linguistics and is now working as a researcher within the AKSW research group. Her main interests concern the usability of Linked Data with respect to both data providers and users and the adoption of Linked Data technology in Linguistics.

Sebastian Hellmann (AKSW, Universität Leipzig, Germany, hellmann@informatik.uni-leipzig.de, http://bis.informatik.uni-leipzig.de/SebastianHellmann) finished his Master thesis in 2008 at University of Leipzig and is currently a research fellow for the AKSW research group and also a member of the LOD2 EU and the LIDER project. He is contributor, co-founder and leader of several open source projects including DL-Learner, DBpedia, OWLG and NLP2RDF. Sebastian is author of over 15 peer-reviewed scientific publications and was chair at the Open Knowledge Conference in 2011, the Workshop on Linked Data in Linguistics 2012, the Linked Data Cup 2012 and the Multilingual Linked Data for Enterprises 2012 workshop. Early December 2013, he submitted his thesis “Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data” http://tinyurl.com/sh-thesis