HDL - Towards a Harmonized Dataset Model

Ahmad Assaf[1 2], Aline Senart[2] and Raphaël Troncy[1]

[1] EURECOM, Sophia Antipolis, France. <firstName.lastName@eurecom.fr>
[2] SAP Labs France. <firstName.lastName@sap.com>

Abstract. The Open Data movement triggered an unprecedented amount of data published in a wide range of domains. Governments and corporations around the world are encouraged to publish, share, use and integrate Open Data. There are many areas where we can see the value of Open Data, from transparency and self-empowerment to improving efficiency, effectiveness and decision making. The growing amount of data constitutes the need for rich metadata attached to it. This metadata enables dataset discovery, comprehension and maintenance. Data portals, which are considered to be datasets' access points, present their metadata in various models. In this paper, we propose HDL, a harmonized dataset model based on the analysis of seven prominent dataset models (CKAN, DKAT, Public Open Data, Socrata, VoID, DCAT and Schema.org). We further present use cases that show the benefits of providing rich metadata to enable dataset discovery, search and spam detection.

Keywords: Dataset, Dataset Profile, Metadata, Dataset Model

1 Introduction

Open data is the data that can be easily discovered, reused and redistributed by anyone. It can include anything from statistics, geographical data, meteorological data to digitized books from libraries. Open data should have both legal and technical dimensions. It should be placed in the public domain under liberal terms of use with minimal restrictions and should be available in electronic formats that are non-proprietary and machine readable. Open Data has major benefits for citizens, businesses, society and governments. It increases transparency and enable self-empowerment by improving the visibility of previously inaccessible information and allowing citizens to be more informed about policies, public spendings and track activities in the law making processes. Moreover, and despite of the legal issues surrounding Linked Data licenses [7], it is still considered a gold mine for organizations who are trying to leverage external data sources in order to produce more informed business decisions [2].

Datasets should contain the metadata needed to effectively understand and use them. It is one of the Linked Data publishing best practices mentioned in [1]. The ability to automatically check this metadata helps in:

- **Delaying data entropy**: *Information entropy* is the degradation or loss that limits the information content in raw or metadata. Information entropy, data complexity and dynamicity can shorten the life span of data.

Even when the raw data is properly maintained, it is often rendered useless when the attached metadata is missing, incomplete or unavailable. Comprehensive high quality metadata can counteract these factors and increase dataset longevity [6].
- **Enhancing data discovery, exploration and reuse**: Users who are unfamiliar with a dataset require detailed metadata to interpret and analyze accurately raw data. Moreover, several prominent data portals rely on the metadata to enable search and filtering.
- **Enhancing spam detection**: Detecting spam in public data portals is increasingly difficult even with security measures like captchas and anti-spam devices. Good dataset metadata quality reflects highly on the quality of its raw data.

The value of Open Data is recognized when it is used. To ensure that, publishers need to enable people to find datasets easily. Data portals are specifically designed for this purpose. They make it easy for individuals and organizations to store, publish and discover datasets. The data portals can be be public like DataHub[3] and Europe's Public Data[4] or private like Quandl[5] or Engima[6]. The data available in private portals is of higher quality as it is manually curated but in lesser quantity compared to what is available in public portals. Similarly, in some public data portals, administrators manually review datasets information, validate, correct and attach suitable metadata information.

Data models vary across portals. Surveying the models landscape, we did not find any that offers enough granularity to completely describe complex datasets facilitating search, discovery and recommendation. For example, the DataHub[7] uses an extension of the Data Catalog Vocabulary (DCAT) [5]. This data model prohibits a semantically rich representation of complex datasets like DBpedia[8] where it has multiple endpoits and thousands of dump files with various contents in several languages [3]. Moreover, to properly integrate Open Data into business, a dataset should include the following information: i)Access information: The dataset is rendered useless if it does not contain accessible data dumps or queryable endpoints. ii)License information: Businesses are always concerned with the legal implications of using external content. As a result, datasets should include both machine and human readable license information that indicates permissions, copyrights and attributions. iii)Provenance information: Depending on the dataset license, the data might not be legally usable if there are no information describing its authoritative and versioning information. Current models underspecified these main aspects limiting the usability of many datasets.

In this paper, we propose HDL, a harmonized dataset model that addresses the shortcomings of existing dataset models by based analyzing seven prominent

[3] http://datahub.io
[4] http://publicdata.eu
[5] https://quandl.com/
[6] http://enigma.io/
[7] http://datahub.io
[8] http://dbpedia.org

dataset models (CKAN, DKAT, Public Open Data, Socrata, VoID, DCAT and Schema.org). We further present use cases that show the benefits of providing rich metadata to enable dataset discovery, search and spam detection.

The remainder of the paper is structured as follows. In Section 2, we present our classification for the different metadata information. In Section 3, we present the existing dataset models used by various data portals. In Section 4, we describe our proposed model and suggest a set of best practices to ensure proper metadata presentation and we finally conclude and outline some future work in Section 5.

2 Metadata Classification

A standard dataset metadata model should contain information about four sections:

- **Resources**: Distributable parts containing the actual raw data. They can come in various formats (JSON, XML, RDF, etc.) and can be downloaded or accessed directly (REST API, SPARQL endpoint).
- **Tags**: Provide descriptive knowledge on the dataset content and structure. They are used mainly to facilitate search and reuse.
- **Groups**: A dataset can belong to one or more group that share common semantics. A group can be seen as a cluster or a curation of datasets based on shared categories or themes.
- **Organizations**: A dataset can belong to one or more organization controlled by a set of users. Organizations are different from groups as they are not constructed by shared semantics or properties, but solely on their association to a specific administration party.

Upon examining the various data models, we grouped the metadata information into four main types. Each section discussed above should contain one or more of these types. For example, resources have general, access, ownership and provenance information while tags have general and provenance information only. The four types are:

- **General information**: General information about the dataset. e.g., title, description, ID, etc. This general information is manually filled by the dataset owner. In addition to that, tags and group information is required for classification and enhancing dataset discoverability. This information can be entered manually or inferred modules plugged into the topical profiler.
- **Access information**: Information about accessing and using the dataset. This includes the dataset URL, license information i.e., license title and URL and information about the dataset's resources. Each resource has as well a set of attached metadata e.g., resource name, URL, format, size, etc.
- **Ownership information**: Information about the ownership of the dataset. e.g., organization details, maintainer details, author, etc. The existence of this information is important to identify the authority on which the generated report and the newly corrected profile will be sent to.

- **Provenance information**: Temporal and historical information on the dataset and its resources. For example, creation and update dates, version information, version, etc. Most of this information can be automatically filled and tracked.

3 Dataset Models

There are many data portals hosting a large number of private and public datasets. Each portal present the data based on the model used by the underlying software. In this section, we present the result of our landscape survey of prominent data portals and their models.

3.1 DCAT

Data Catalog Vocabulary (DCAT), a W3C recommendation established designed to facilitate interoperability between data catalogs published on Web [5]. The goal behind DCAT is to increase datasets discoverability enabling applications to easily consume metadata coming from multiple sources. Moreover, they foresee that aggregated DCAT metadata can facilitate digital preservation and enable decentralized publishing and federated search.

DCAT is an RDF vocabulary defining three main classes (`dcat:Catalog`, `dcat:Dataset` and `dcat:Distribution`). We are interested in: 1)`dcat:Dataset` class which is a collection of data that can be available for download in one or more formats and 2)`dcat:Distribution` class which represents the accessible form of a dataset e.g., RSS feed, REST API, SPARQL endpoint, etc..

DCAT-AP for for data portals in Europe is a specification that re-uses terms from DCAT, ADMS, etc., and adds more specificity by identifying mandatory, recommended and optional elements to be used for a particular open data catalogue. Studies conducted by EU commission (Vickery, 2011) have shown that businesses and citizens are facing difficulties in searching and reusing data sets from public sector. Therefore, the availability of a unified method to describe data sets in a machine-readable format with a small number of commonly agreed metadata could largely improve the co-referencing and interoperability among different data catalogues. DCAT-AP is developed under this context and is expected to be applied across Open Data portals in EU countries.

3.2 VoID

VoID is an "RDF Schema vocabulary for describing metadata about RDF data sets. Its primary purpose is to bridge the gap between data publishers and data consumers using an exclusive vocabulary to describe different data set attributes. The core concepts related to open data sets are: void:Dataset, void:Linkset, void:subset.

3.3 CKAN

CKAN is currently the most widely used open source data management system that helps users from different levels and domains (national and regional governments, companies and organisations) to make their data openly available. CKAN has been adopted by various levels of Open Data portals, and a few poplular CKAN instances include publicdata.eu, data.gov.uk and data.gouv.fr. CKAN provides tools to ease the workflow of data publishing, sharing, searching and management. Each data set is given its own page with a rich collection of metadata. Users can publish their data sets via an import feature or through a web interface, and then the data sets can be searched by keywords or tags with exact or fuzzy-matching queries. CKAN provides a rich set of visualisation tools, such as interactive tables, graphs and maps. Moreover, the dashboard function will help administrators to monitor the statistics and usage metrics for the data sets. Federating networks with other CKAN nodes is also supported, as well as the possibility to build a community with extensions that allow users to comment on and follow data sets. Finally, CKAN provides a rich RESTful JSON API for querying and retrieving data set information.

3.4 DKAN

DKAN29 is a Drupal-based open data platform with a full suite of cataloguing, publishing and visualisation features. Compared with CKAN, DKAN is seamlessly integrated with Drupal30 content management system, thus it can be easily deployed with Drupal and customised using different Druapl themes. The actual data sets in DKAN can be stored either within DKAN or on external sites, and it is possible to manage access control and version history with rollback. DKAN provides user analytics and data users can upload, tag, search and group data sets via a web front-end or APIs. In addition, they can also collaborate, comment, and share information via social network integration.

3.5 Socrata

Socrata provides a commercial platform to streamline data publishing, management, analysis and reusing. It integrates many useful features for both portal administrators and end users to manage, access and visualise data sets. For example, The Chicago26 and New York City27 The platform comprises a series of tools, including an open data portal which stores data in the cloud for users to access, visualise, and share. All the data sets hosted in Socrata can be accessed using RESTful API. This is accompanied by the developer site which documents how to use the Socrata API, including search and filter data sets. In Socrata, usershave the ability to to customise the data set metadata according to individual's requirements.

3.6 Schema.org

It is a collection of schemas (in RDF/Microdata format) that webmasters can use to markup HTML pages in ways recognised by major search engines. Schema.org covers many domains and there are classes and properties defined as DataCatalog and Dataset. The metadata harvester withinthe ODM project can make use of schema.org vocabulary to discover the data sets and data catalogs hosted in a certain website

3.7 Project Open Data

4 Towards A Harmonised Model

5 Conclusion and Future Work

Acknowledgments

This research has been partially funded by the European Union's 7th Framework Programme via the project Apps4EU (GA No. 325090).

References

1. C. Bizer. Evolving the web into a global data space. In *Proceedings of the 28th British National Conference on Advances in Databases*, 2011.
2. D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.
3. M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. Dataid: Towards semantically rich metadata for complex datasets. In *Proceedings of the 10th International Conference on Semantic Systems*, 2014.
4. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing linked datasets with the VoID vocabulary. Technical report, 2011.
5. J. Erickson and F. Maali. Data catalog vocabulary (DCAT). Technical report, 2014. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/.
6. Kovács-Láng. Global Terrestrial Observing System. Technical report, GTOS Central and Eastern European Terrestrial Data Management and Accessibility Workshop, 2000.
7. K. J. Prateek Jain, Pascal Hitzler and C. Venkatramani. There's no money in linked data. 2013.