Reflections on: DCAT-AP Representation of Czech National Open Data Catalog and its Impact*

Jakub Klímek[0000−0001−7234−3051]

Department of Software Engineering, Faculty of Mathematics and Physics,
Charles University, Malostranské náměstí 25, 118 00 Praha 1, Czech Republic
klimek@ksi.mff.cuni.cz
https://jakub.klímek.com

Abstract. Open data is now a heavily discussed topic around the world and in the European Union. In the Czech Republic, open data is a term anchored in legislation, which includes the requirement of registration of all open data in the Czech National Open Data Portal (NODC). In the journal paper [5] we describe the NODC, its architecture, dataset registration processes including the harvesting of Local Open Data Catalogs (LODCs), proprietary XML API and its obsolete dataset viewer. Next we describe the process of transformation of the NODC metadata to the DCAT-AP v1.1 RDF representation from the data model point of view and from the technical environment point of view. We describe the dataset quality measurements computed using the new data representation and its further impact on the Linked Open Data (LOD) environment including the harvesting of the metadata by the European Data Portal (EDP). Finally, we evaluate the data transformation and publishing environment from the usability, portability, availability and performance perspectives.

Keywords: open data · catalog · DCAT-AP · Linked Data

1 Introduction

Open data is currently a hot topic among institutions of public administration and data users around the world. From the political point of view, publishing open data is important for public administration institutions to show that they are transparent and open to citizens. From the legal point of view it is important that the data is published using an open license, permitting users to use the data freely. From the technical point of view, it is important that the data is published as machine readable data in an open format, accessible on the web with minimal effort. From the point of view of potential open data users it is important that the data can be searched for and found. Finally, from the economic

* This work was supported by the Czech Science Foundation (GAČR), grant number 19-01641S.

point of view, open data is expected to support creation of new services and new business models[1]. At the intersection of all of these points of view lay open data portals of the data publishing institutions, which typically include open data catalogs where datasets can be found along with the metadata describing them. The metadata descriptions contain the necessary information about the licenses of datasets, formats of their distributions, and the textual descriptions, all of which can be used for dataset search. This results in a number of open data catalogs, typically one per each institution willing to publish open data. Therefore, a problem with discoverability of the data catalogs and the individual datasets described in them arises, and along with it a need for aggregate views over multiple data catalogs. To address this need, a standard for representation of dataset metadata, the Data Catalog Vocabulary (DCAT) [3], was developed by the W3C to enable dataset metadata exchange among data catalogs, and specifically to support hierarchies of catalogs. In the European Union, an application profile of DCAT, the DCAT-AP v1.1[2], has been developed, further specifying for instance controlled vocabularies to be used to describe datasets and their distributions. At the same time, a top level European data catalog, the European Data Portal (EDP)[3], has been developed with the intent of aggregating data catalogs from the whole EU.

In the Czech Republic, the Ministry of the Interior (MoI) is in charge of the open data agenda. The NODC is run by the MoI, and the public administration institutions can either register their datasets directly in the NODC, or, preferably, they can run their own Local Open Data Catalog (LODC) which, after registration, gets automatically harvested to the NODC. Unfortunately, until now, the MoI did not consider DCAT-AP v1.1 as a metadata publishing format. The NODC internal data model was inspired by the DCAT-AP v1.0, however, it is XML based and accessible only via a proprietary XML API. In addition, the metadata registered in the NODC was only viewable to users via a rather unfriendly user interface.

As the NODC became known and used, new requirements on its functionality arose, mainly the need to monitor dataset metadata quality, and the need to be harvested by the EDP. To address these requirements, we first transform the current NODC metadata to the DCAT-AP v1.1 RDF representation and publish it. Then we use this new representation to compute the dataset metadata quality measurements and to facilitate the harvesting of the NODC metadata by the EDP. Similar requirements are identified also elsewhere, e.g. in Serbia [4], as requirements to implement the revised European Directive on the Public Sector Information (2013/37/EU) emphasizing the role of the Linked Data approach for improved interoperability and re–use.

[1] https://www.europeandataportal.eu/en/highlights/economic-benefits-open-data
[2] https://joinup.ec.europa.eu/release/dcat-ap-v11
[3] https://www.europeandataportal.eu

1.1 Contributions

In the journal paper [5] we describe the NODC and its architecture, data input, data model, its API and the mechanism of harvesting the metadata from LODCs to the NODC. We describe how we transform the current metadata to the DCAT-AP v1.1 RDF representation both from the data model point of view and from the technical environment point of view. We show the metadata quality measures that we compute using the DCAT-AP v1.1 representation and we evaluate further impact this new metadata representation has, including a new frontend for viewing the dataset metadata and the effects it has on the Linked Open Data (LOD) environment. Finally, we evaluate the data transformation environment from the usability, portability, availability and performance perspectives.

1.2 Outline

The rest of this extended abstract is structured according to the original paper [5], providing an outline of the paper by summarizing the contents of each of the sections.

In Section 2 we describe the NODC, its architecture, the processes of harvesting LODCs, and its proprietary XML API. In Section 3 we describe the data transformation to DCAT-AP v1.1 and the linking to related datasets and code lists. In Section 4 we describe the technical environment used for the data transformation and publication process. In Section 5 we describe data quality measures which we compute based on the transformed data. In Section 6 we show additional impact of the published dataset by presenting some of the known usages of the published data, which include harvesting by the European Data Portal. In Section 7 we evaluate the data transformation and publishing environment according to various criteria. Finally we survey related work and in Section 8 we conclude.

2 Czech National Open Data Catalog (NODC)

The institutions in the Czech public administration are required to register their published data in the Czech National Open Data Catalog (NODC) before they can call it open data. There are two ways of registering to the NODC. For smaller institutions such as village councils there is the possibility of registering individual datasets directly in the NODC using a form to fill in all the necessary metadata. For larger institutions such as ministries or city councils, there is the possibility of registering their own local Open Data catalog (LODC), as there is an assumption that such institutions will have their own data portals anyway. Once a LODC is registered, the metadata from it is automatically and regularly harvested by the NODC, giving the institutions more flexibility in their dataset management. The registered datasets can be viewed in a rather unfriendly web

user interface, which does not provide many features known from wide-spread data catalog implementations such as CKAN[4] and DKAN[5].

3 Data modeling and transformations

Our goal is to publish the dataset metadata from the NODC according to the DCAT-AP v1.1 specification. The proprietary NODC XML API will serve as our data source. In this section of the original paper [5] the RDF vocabularies used in the transformed data and the legacy code lists and data items are showed and mapped to the European Union Metadata Registry Named Authority Lists (EU MDR NALs), now parts of EU Vocabularies[6], and the RTIAR.

4 Data transformation and publishing environment

In this section of the paper [5] we describe the technical environment used to transform the data from the original NODC proprietary XML API to a DCAT-AP v1.1 dataset published as Linked Open Data. The environment is built on open-source tools.

The transformation is done using LinkedPipes ETL [7], which needs Java[7] and Node.js[8] to run, and Git[9] and Apache Maven[10] to build the source code from the GitHub repository[11].

LinkedPipes ETL is an open-source ETL tool for production and consumption of Linked Data, which is in use by multiple organizations in the Czech public administration. It is also used in the OpenBudgets.eu platform [10] for publication and transformation of fiscal data. It runs the data transformation process as a so called pipeline. The process is run daily, as that corresponds to the periodicity of updates of the source NODC data. The pipeline has the following principal steps:

1. Get metadata from the proprietary NODC XML API
2. Get the data box ID to publisher IRI and name mapping from the List of public administration authorities dataset
3. Transform the metadata using an XSLT [2] template
4. Map the ISO8601 frequencies to the Frequency EU MDR NAL
5. Add the File Type EU MDR NAL items based on distribution MIME Types
6. Get the previous DCAT-AP v1.1 dump and compare with the current version to generate statistics about new, changed and deleted datasets using the RDF Data Cube Vocabulary (DCV) [11]

[4] https://github.com/ckan/ckan
[5] https://getdkan.org/
[6] https://publications.europa.eu/en/web/eu-vocabularies
[7] https://www.oracle.com/java/index.html
[8] https://nodejs.org
[9] https://git-scm.com/
[10] https://maven.apache.org/
[11] https://github.com/linkedpipes/etl

7. Compute metadata of the DCAT-AP v1.1 dataset itself
8. Load an index to Apache Solr
9. Load the metadata records to Apache CouchDB
10. Load the RDF TriG dump to a web server
11. Load the RDF data to the SPARQL endpoint
12. Run the pipelines computing data quality measurements (see Section 5)

5 Metadata quality measures

In this section of the paper [5] we describe quality measures monitored using the NODC loaded in a SPARQL endpoint. We distinguish two types of measures, one is based solely on what can be found in the metadata itself. The second type uses the metadata registered in NODC to try to access the linked resources, i.e. licenses, schemas, documentation and the distributions themselves, and check whether they are served correctly, e.g. with a correct Media Type. To compute both types of quality measures, pipelines in LinkedPipes ETL are used. Since the quality measures are based on the DCAT-AP v1.1 specification, they are directly reusable for other DCAT-AP v1.1 compliant datasets.

5.1 Metadata based quality measures

Since there is no validation of input data in the current NODC harvester of LODCs, there are metadata records violating DCAT-AP v1.1 constraints or constraints dictated by the Czech legislation. The first part of the quality measures in this section aims at detecting such anomalies. In addition, there are measures aiming at providing an overview of common practice. The results of these measures are published on the Czech Open Data Portal[12]. The measures are:

1. Number of distributions with unspecified license per publisher
2. Number of datasets with distributions with unspecified licenses per publisher
3. Number of datasets missing required attributes per publisher
4. List of datasets missing required attributes per publisher
5. Number of distributions with a given mime type per publisher
6. Distribution licenses per publisher
7. Number of publishers per license
8. Number of datasets with a given accrual periodicity per publisher
9. Number of datasets and distributions per publisher

These measures were selected based on the most frequently appearing errors in the metadata records. These erroneous records cause a decline in the overall metadata quality in the NODC. When the users encounter them, they tend to blame the NODC for the inconsistent looking record, therefore, it is important to us that the publisher correct their records. This is also why it is important

[12] `https://opendata.gov.cz/statistika:datova-kvalita` (in Czech only)

to consistently point out errors in the records and demand their correction. From our experience, the most effective way to achieve the correction in the (Czech) public administration is to publicly display the errors attributed to the originating publishers, along with clear instructions on how to correct the mistakes. The errors in the records also usually reveal a deeper problem with data management at the original publisher.

5.2 Web access based quality measures

The measures listed in this section check whether the resources linked from the metadata records actually exist, and whether they are served correctly. There are four types of resources linked from the metadata, i.e. distributions, licenses, dataset documentation and distribution schema. For each of the four resource types we compute a summary statistic and a list of offending resources.

Here, we list the quality measures along with the description of the individual columns in the result.

1. Statistics of unavailable dataset distributions per publisher
2. List of unavailable distributions
3. Statistics of unavailable schemas of dataset distributions per publisher
4. List of unavailable distribution schemas
5. Statistics of unavailable licenses of dataset distributions per publisher
6. List of unavailable distribution licenses
7. Statistics of unavailable documentation of datasets per publisher
8. List of unavailable dataset documentation

In addition to the availability measures described above, there is one more measure dealing with inconsistency between the distribution Media Type registered in the NODC and the Media Type returned by the web server serving the distribution.

6 Additional impact of publishing NODC as Linked Open Data using DCAT-AP v1.1

In this section of [5], we demonstrate the impact of publishing the NODC contents as Linked Open Data according to the DCAT-AP v1.1 specification besides us being able to compute the quality measures described in Section 5 by describing the effects it has on the LOD environment.

6.1 Promotion of Linked Open Data and usage of standardized vocabularies

Theoretical advantages of LOD described in existing literature are not convincing enough to the representatives of public administration institutions regarding publishing their data as LOD. We are using the example of the NODC and others, such as the Czech Social Security Administration [6], and the infrastructure used to process the data and publish it as DCAT-AP v1.1 to convince other institutions that publishing LOD is possible without excessive resources.

6.2 Harvesting by the European Data Portal

A clear added value of the DCAT-AP v1.1 representation of the NODC data is the ability to be harvested by the European Data Portal (EDP), a well-known pan-European open data catalog, using a native LOD way. In fact, this was beneficial not only to the Czech publishers, as their metadata got published on the European level, but also to the developers of the European Data Portal. This is because the Czech NODC was the first European data portal to publish the metadata using DCAT-AP v1.1 in RDF natively, i.e. as an RDF data dump, dereferencable IRIs and a SPARQL endpoint. At the same time, the Czech NODC is the largest catalog in EDP.

6.3 Ability to use LinkedPipes DCAT-AP Viewer as frontend

The original NODC viewer was quite unfriendly to the users. However, the proprietary nature of the published metadata made it hard to convince someone to develop an alternative frontend. Thanks to the transformation of the data to DCAT-AP v1.1 [8] we are now able to use the LinkedPipes DCAT-AP Viewer which is friendlier (based on System Usability Scale (SUS)[13]) and offers more functionality than the original, including multilingual user interface exploiting the multilingual EU MDR NALs where possible, full text search, keywords word cloud, handling of large numbers of distributions, etc.

6.4 Proof of need for a Linked Data Consumption Platform

In our research group we are focusing not only on LOD publishing, but also LOD consumption, where we identified a distinct lack of tools for actually using LOD that is published [9]. This lack of tools is proven every time we publish a new dataset, as users are saying that they do not know how to consume LOD and that they require data in formats they are used to, i.e. CSV, JSON and XML files. The case with NODC was no different. After publishing the data in RDF and in the SPARQL endpoint, some users demanded CSV exports of the data, even though they could be obtained using a simple SPARQL SELECT query. The users need a tool we call a Linked Data Consumption Platform (LDCP), which would help them in consuming LOD, and ideally, it would be easier to use than tools for consumption of CSV, JSON and XML files thanks to the benefits LOD brings, and it may even not require any specific LOD related knowledge. The users who demand those non-LOD representations of data then serve us as motivation for our LDCP related efforts.

7 Environment evaluation

In this section of [5], we evaluate the LOD publishing environment described in Section 4 used by the Ministry of the Interior of the Czech Republic (MoI) to

[13] `https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html`

prepare and publish the DCAT-AP v1.1 NODC RDF dataset. We evaluate the environment using so called *quality attributes* as introduced in [1]. We evaluate the following quality attributes:

- usability
- portability
- availability
- performance

The environment integrates various open-source tools. We do not evaluate each individual tool but the environment as a whole.

8 Conclusions

In the paper [5] we describe the current architecture of the Czech National Open Data Catalog (NODC), starting from manual data entry, Local Open Data Catalogs (LODCs) harvesting, data storage to data publication using its proprietary XML API and a rather unfriendly viewer. Next we described the vocabularies and codelists used in transformation of the NODC data to its DCAT-AP v1.1 RDF representation. We described the technical environment used for the data transformation and data quality measures that can be computed using the RDF data representation, both using LinkedPipes ETL. We describe the additional impact of publishing NODC using DCAT-AP v1.1 and evaluate the data transformation environment both from the perspective of the transformation designers and from the perspective of the data users.

During the time of writing the paper, the MoI changed their supplier of the implementation of the original NODC, making it permanently unavailable, as the new supplier was not able to take over the original implementation. This nevertheless showed another benefit of publishing open data, as our copy of the NODC used to demonstrate the transformation to DCAT-AP v1.1 and the browsing of the data in LinkedPipes DCAT-AP Viewer remained the only existing publicly available NODC instance. Currently, it is already running as the official NODC instance at `https://data.gov.cz`.

Finally, we conclude the paper with a summarization of the lessons learned during our work with the MoI and publishers of open data in the Czech Republic.

References

1. Bass, L., Clements, P., Kazman, R.: Software Architecture in Practice. Addison-Wesley Professional, 3rd edn. (2012)
2. Clark, J.: XSL transformations (XSLT) version 3.0. W3C Recommendation, W3C (Jun 2017), `https://www.w3.org/TR/1999/REC-xslt-19991116`
3. Erickson, J., Maali, F.: Data Catalog Vocabulary (DCAT). W3C Recommendation, W3C (Jan 2014), `https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/`

4. Janev, V., Mijovic, V., Vranes, S.: Proposal for Implementing the EU PSI Directive in Serbia. In: Ko, A., Francesconi, E. (eds.) Electronic Government and the Information Systems Perspective - 5th International Conference, EGOVIS 2016, Porto, Portugal, September 5-8, 2016, Proceedings. Lecture Notes in Computer Science, vol. 9831, pp. 16–30. Springer (2016). https://doi.org/10.1007/978-3-319-44159-7_2, https://doi.org/10.1007/978-3-319-44159-7_2
5. Klímek, J.: DCAT-AP representation of Czech National Open Data Catalog and its impact. Journal of Web Semantics **55**, 69 – 85 (2019). https://doi.org/10.1016/j.websem.2018.11.001, http://www.sciencedirect.com/science/article/pii/S1570826818300532
6. Klímek, J., Kučera, J., Nečaský, M., Chlapek, D.: Publication and usage of official Czech pension statistics Linked Open Data. Journal of Web Semantics **48**, 1 – 21 (2018). https://doi.org/10.1016/j.websem.2017.09.002, http://www.sciencedirect.com/science/article/pii/S1570826817300343
7. Klímek, J., Škoda, P.: LinkedPipes ETL in use: practical publication and consumption of linked data. In: Indrawan-Santiago, M., Steinbauer, M., Salvadori, I.L., Khalil, I., Anderst-Kotsis, G. (eds.) Proceedings of the 19th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2017, Salzburg, Austria, December 4-6, 2017. pp. 441–445. ACM (2017). https://doi.org/10.1145/3151759.3151809, https://doi.acm.org/10.1145/3151759.3151809
8. Klímek, J., Škoda, P.: LinkedPipes DCAT-AP Viewer: A Native DCAT-AP Data Catalog. In: van Erp, M., Atre, M., López, V., Srinivas, K., Fortuna, C. (eds.) Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018. CEUR Workshop Proceedings, vol. 2180. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2180/paper-32.pdf
9. Klímek, J., Škoda, P., Nečaský, M.: Requirements on Linked Data Consumption Platform. In: Auer, S., Berners-Lee, T., Bizer, C., Heath, T. (eds.) Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with 25th International World Wide Web Conference (WWW 2016). CEUR Workshop Proceedings, vol. 1593. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1593/article-01.pdf
10. Musyaffa, F.A., Halilaj, L., Li, Y., Orlandi, F., Jabeen, H., Auer, S., Vidal, M.: OpenBudgets.eu: A Platform for Semantically Representing and Analyzing Open Fiscal Data. In: Mikkonen, T., Klamma, R., Hernández, J. (eds.) Web Engineering - 18th International Conference, ICWE 2018, Cáceres, Spain, June 5-8, 2018, Proceedings. Lecture Notes in Computer Science, vol. 10845, pp. 433–447. Springer (2018). https://doi.org/10.1007/978-3-319-91662-0_35, https://doi.org/10.1007/978-3-319-91662-0_35
11. Reynolds, D., Cyganiak, R.: The RDF Data Cube Vocabulary. W3C Recommendation, W3C (Jan 2014), https://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/