# LinkedPipes DCAT-AP Viewer: A Native DCAT-AP Data Catalog[⋆]

Jakub Klímek[0000−0001−7234−3051] and Petr Škoda[0000−0002−2732−9370]

Charles University, Faculty of Mathematics and Physics
Malostranské nám. 25, 118 00 Praha 1, Czech Republic
klimek@opendata.cz

**Abstract.** In this demonstration we present LinkedPipes DCAT-AP Viewer (LP-DAV), a data catalog built to support DCAT-AP, the European standard for representation of metadata in data portals, and an application profile of the DCAT W3C Recommendation. We present its architecture and data loading process and on the example of the Czech National Open Data portal we show its main advantages compared to other data catalog solutions such as CKAN. These include the support for Named Authority Lists in EU Vocabularies (EU NALs), controlled vocabularies mandatory in DCAT-AP, and the support for bulk loading of DCAT-AP RDF dumps using LinkedPipes ETL.

**Keywords:** catalog · DCAT · DCAT-AP · linked data

## 1 Introduction

Currently, two worlds exist in the area of data catalogs on the web. In the first one there are a few well established data catalog implementations such as CKAN or DKAN, each with their data model and a JSON-based API for accessing and writing the metadata. In the second one, there is the Linked Data and RDF based DCAT W3C Recommendation [1] and its application profiles, such as the European DCAT-AP[1], which are de facto standards for representation of metadata in data portals. The problem is that CKAN has been around for a while now, and is better developed, whereas DCAT is still quite new, with insufficient tooling support, nevertheless it is the standard. At first glance, the data models seem similar, CKAN having its packages and resources and DCAT having datasets and distributions. However, a hands on experience shows us some serious compatibility issues. These are caused mainly by the fact that CKAN is not built with Linked Data in mind while DCAT depends on the Linked Data principles quite heavily, especially in its application profiles such as DCAT-AP. To bridge this gap, various extensions for CKAN are being developed such as `ckanext-dcat`[2], providing harvesting of DCAT metadata into CKAN

[⋆] This work was supported in part by the Czech Science Foundation (GAČR), grant number 16-09713S and in part by the project SVV 260451.

[1] https://joinup.ec.europa.eu/release/dcat-ap-v11

[2] https://github.com/ckan/ckanext-dcat

and export of CKAN metadata to its DCAT representation. Nevertheless, the extensions still only provide a syntactical mapping of the models, which does not solve the underlying issues which demand a different design approach to the whole software stack, especially regarding the usage of the multilingual Named Authority Lists in EU Vocabularies (EU NALs)[3].

Our primary goal when implementing the Czech National Open Data Catalog was to be compliant with standards, i.e. DCAT-AP. We had approximately 120 000 datasets harvested from local catalogs of institutions publishing open data in the Czech Republic, represented using DCAT-AP in an RDF dump file. When trying to load this data into CKAN or DKAN, we faced the following challenges:

1. **Poor bulk load performance.** The local institutions kept updating the datasets daily in an automated fashion, resulting in the need to reload most of the datasets daily, which we were unable to do using the CKAN API and the CKAN bulk load utilities. This is because CKAN is focused mainly on manual data entry.
2. **Insufficient support for license information.** The DCAT Recommendation attaches licensing information to distributions of datasets using its URL. CKAN has the license support hardwired for datasets (packages) and provides its own list of licenses identified only by their proprietary codes such as `cc-by`.
3. **Poor support of datasets with many distributions.** According to the DCAT-AP Implementation Guideline on Dataset series[4], when users are expected to be interested in the dataset series as a whole, the individual files in the series should be represented as distributions of one dataset. In our case, we had a dataset with approx. 7 000 distributions. Since in CKAN, they are represented as resources and there is no paging implemented for them, CKAN kept crashing on these datasets. A hotfix for this had to be implemented e.g. in the CKAN based European Data Portal[5], which simply limited the maximum number of resources, which is not an optimal solution.
4. **Insufficient support for controlled vocabularies.** The DCAT-AP standard mandates the usage of, often multilingual, EU NALs. Since in the RDF dump of DCAT there are links to those vocabularies, the catalog needs to be aware of them to display human readable labels for the vocabulary items, in the appropriate language.

These issues led us to believe that there is a need for a new data catalog software built with native Linked Data, DCAT-AP and controlled vocabularies support, focused on automated loading of larger numbers of datasets from data preparation pipelines. In this demo we present the LinkedPipes DCAT-AP Viewer, our solution to addressing these issues.

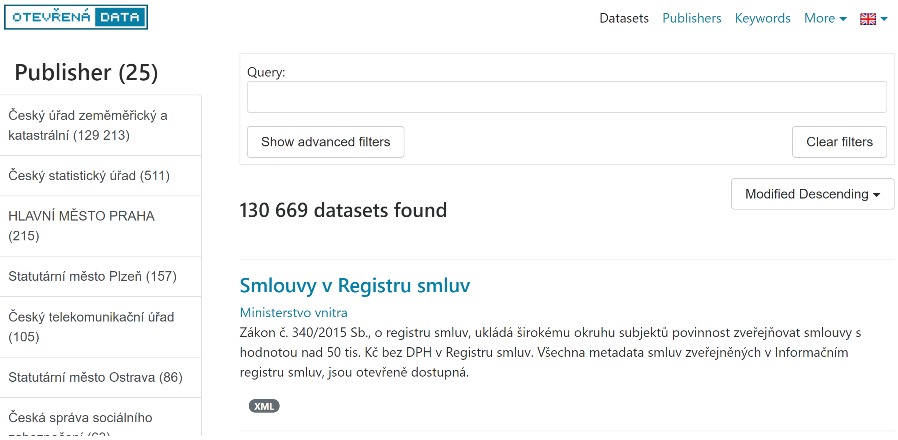
[3] https://publications.europa.eu/en/web/eu-vocabularies/authority-tables

[4] https://joinup.ec.europa.eu/release/dcat-ap-how-model-dataset-series

[5] https://www.europeandataportal.eu/

## 2 LinkedPipes DCAT-AP Viewer demonstrated features

The LinkedPipes DCAT-AP Viewer (LP-DAV) is a Node.js[6] and Bootstrap[7] based application (see Figure 1) using Apache Solr[8] for search capabilities. The dataset records themselves can be stored either in a SPARQL endpoint such as the Openlink Virtuoso RDF store[9], or for increased performance in Apache CouchDB[10] document store. It is open-source and it is developed on GitHub[11].



Publisher (25)

Český úřad zeměměřický a katastrální (129 213)

Český statistický úřad (511)

HLAVNÍ MĚSTO PRAHA (215)

Statutární město Plzeň (157)

Český telekomunikační úřad (105)

Statutární město Ostrava (86)

Česká správa sociálního

130 669 datasets found

Smlouvy v Registru smluv

Ministerstvo vnitra

Zákon č. 340/2015 Sb., o registru smluv, ukládá širokému okruhu subjektů povinnost zveřejňovat smlouvy s hodnotou nad 50 tis. Kč bez DPH v Registru smluv. Všechna metadata smluv zveřejněných v Informačním registru smluv, jsou otevřeně dostupná.

**Fig. 1.** LinkedPipes DCAT-AP Viewer deployed as Czech National Open Data Catalog

Its main features which will be presented during the demonstration session on real world deployment of LP-DAV as the Czech National Open Data Catalog[12] are:

1. LP-DAV is built for DCAT-AP v1.1 with Linked Data in mind
2. Multilingual user interface and multilingual EU NALs support
3. Paging support for datasets with many distributions
4. LinkedPipes ETL (LP-ETL) [2] pipeline for bulk loading (see section 3)
5. Keyword tag cloud search

[6] https://nodejs.org/
[7] https://getbootstrap.com/
[8] https://lucene.apache.org/solr/
[9] https://github.com/openlink/virtuoso-opensource
[10] https://couchdb.apache.org/
[11] https://github.com/linkedpipes/dcat-ap-viewer
[12] https://data.gov.cz

## 3 Data preparation pipeline in LinkedPipes ETL

The loading of data from a DCAT-AP RDF dump file can be done using the supplied LinkedPipes ETL [2] pipeline[13]. Since the EU NALs are unfortunately still not dereferenceable, they need to be loaded to Apache CouchDB using a separate code list loading pipeline[14].

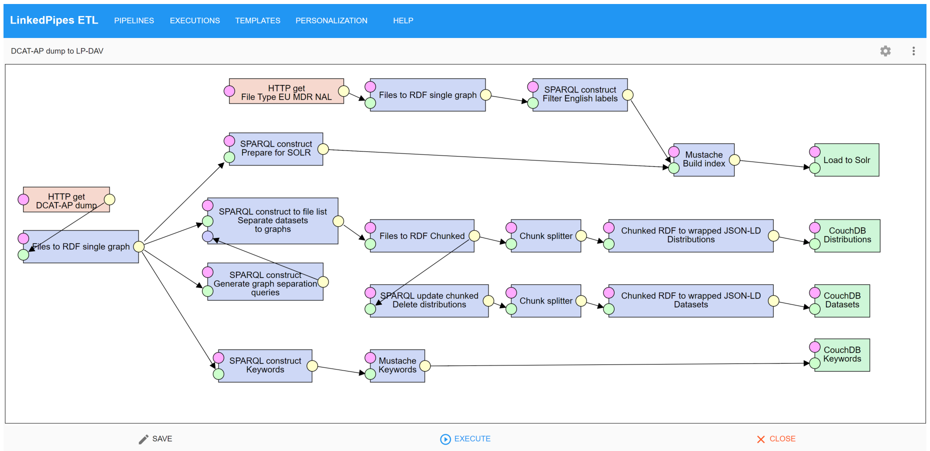

LinkedPipes ETL
PIPELINES
EXECUTIONS
TEMPLATES
PERSONALIZATION
HELP
DCAT-AP dump to LP-DAV
HTTP get File Type EU MDR NAL
Files to RDF single graph
SPARQL construct Filter English labels
SPARQL construct Prepare for SOLR
Mustache Build index
Load to Solr
HTTP get DCAT-AP dump
Files to RDF single graph
SPARQL construct to file list Separate datasets to graphs
Files to RDF Chunked
Chunk splitter
Chunked RDF to wrapped JSON-LD Distributions
CouchDB Distributions
SPARQL construct Generate graph separation queries
SPARQL update chunked Delete distributions
Chunk splitter
Chunked RDF to wrapped JSON-LD Datasets
CouchDB Datasets
SPARQL construct Keywords
Mustache Keywords
CouchDB Keywords
SAVE
EXECUTE
CLOSE

**Fig. 2.** LP-ETL pipeline preparing data for LP-DAV from DCAT-AP dump

The data preparation pipeline (see Figure 2) first loads the RDF dump and then it branches. The top branch populates the Apache Solr index to support search. The middle two branches split the data into individual DCAT dataset and distribution records, and loads them to Apache CouchDB. The bottom branch extracts the used keywords and their occurrences to support the keyword tag cloud search.

## References

1. Erickson, J., Maali, F.: Data Catalog Vocabulary (DCAT). W3C Recommendation, W3C (Jan 2014), https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
2. Klímek, J., Škoda, P., Nečaský, M.: LinkedPipes ETL: Evolved Linked Data Preparation. In: The Semantic Web - ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers. pp. 95–100 (2016), https://dx.doi.org/10.1007/978-3-319-47602-5_20

[13] https://raw.githubusercontent.com/linkedpipes/dcat-ap-viewer/develop/pipeline/dcatap2lpdav.jsonld

[14] https://raw.githubusercontent.com/linkedpipes/dcat-ap-viewer/develop/pipeline/eumdrnals2couchdb.jsonld