Harvesting Open Government Data with DCAT-AP

Fabian Kirstein, 21. March 2018
The European Data Portal offers more than 800 Thousand DCAT-AP Datasets.

https://www.europeandataportal.eu/data
The European Data Portal

“Make all data published by public authorities of the European Union available in one place.”

- Project is active since 2015
- Lead by Capgemini Consulting
- 3 Areas of activity (Services)
  - **Service 1:** Conception, implementation and operation of the data infrastructure
  - **Service 2:** Support of the data providers
  - **Service 3:** Support of the reuse of the data

© ESA/NASA
Requirements (extract)

- Harvesting of national Open Data- and Geo-portals
- Support of the DCAT-AP specifications
- Translation of important metadata into 24 languages
- CKAN data platform as the foundation for the metadata catalogue
- Making metadata available via CKAN REST-API and SPARQL endpoint

Challenges:
- Effectively transforming every source data into the target format
- Combining the conservative data structures of CKAN with the Linked Data world
The Open Data Challenge

- Open Data in Europe is a highly distributed „system“
- Data formats and interfaces are fragmented
- CKAN, DKAN, CSW (INSPIRE), RSS, Socrata, OpenDataSoft etc.
- In addition multiple “dialects” of each standard may exist
- E.g. the CKAN data structure is not strict about data types and semantics

A clear target data format and structure is required!
CKAN/DCAT-AP Integration

- Mapping of DCAT-AP concepts to CKAN data structure
- Extension of CKAN data structure
- Representation of most common DCAT-AP properties
- Example:
  - dct:contactPoint has the range vcard:Kind
  - vcard:Kind defines a variety of valid properties
- CKAN is creating valid DCAT-AP RDF and stores it in the triplestore

```xml
<dct:contactPoint>
  <vcard:Organization>
    <vcard:organization-name>Bundesanstalt für Geowissenschaften und Rohstoffe</vcard:organization-name>
    <vcard:hasEmail rdf:resource="mailto:geologie.daten@bgr.de"/>
  </vcard:Organization>
</dct:contactPoint>

"contact_point": [
  {
    "type": "http://www.w3.org/2006/vcard/ns#Organization",
    "resource": "http://www.organization.resource",
    "email": "mailto:geologie.daten@bgr.de",
    "name": "Bundesanstalt für Geowissenschaften und Rohstoffe"
  }
]```
Harvesting

- Standalone harvesting software
- Each data source is requested regularly
- Harvesting of more than 75 national Open Data- and Geo-portals
  - 39 Geo portals (INSPIRE)
  - 31 CKAN portals
  - 6 DCAT-AP portals
- More than 63 Million triples! 😊
Transformation

- Transformation and harmonisation to the EDP target format is required
- For each data source a custom transformation script was created
- Two scripting languages are available: JSON and XLST
- More conformity to the standards means less error-proneness

Creating and maintaining sufficient transformation rules is an extensive and cumbersome task!
Integration

- The SPARQL endpoint and SPARQL Manager allows direct access to DCAT-AP
- Every dataset can be accessed as DCAT-AP by simply adding .n3 or .rdf to the URL
- URIs are resolved to human-readable strings in the frontend, whenever possible

In general: DCAT-AP is not recognized by the user!
Benefits of DCAT-AP for the EDP

- DCAT-AP gives a clear structure and guideline for harmonising Open Data!
- It is very consolidated and reuses existing specifications as much as possible
- Defined namespaces and vocabularies increase interoperability
  - E.g. Metadata Registry of the EU Publications Office
- In comparison with other Linked Data vocabularies it is (mostly) simple and easy to understand
- RDF supports multi-lingual data natively
Experiences

- The overall adoption of DCAT-AP still progresses slowly
- Many data providers and developers are still overwhelmed with Linked Data and RDF
- The specifications is in some points not precise enough and leaves to much room for interpretation
- More pre-defined vocabularies are required
- Clear implementation guidelines are missing

DCAT-AP is great, but needs more communication!
Outlook

- Linked Data and RDF will always be hard!
- A lot of progress is going on
  - CKAN is supporting DCAT-AP officially
  - National initiatives are evolving (DCAT-AP.de)
  - Additional profiles are developed (StatDCAT-AP)
- Developers and providers will need support
- DCAT-AP will simplify the harvesting process of Open Government Data!
- Vision: Open Data without harvesting – the federated query

The European Data Portal will move to a native DCAT-AP metadata registry in the next 12 month!
Contact

Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin, Germany
info@fokus.fraunhofer.de
www.fokus.fraunhofer.de

Fabian Kirstein, M.Sc.
Wissenschaftlicher Mitarbeiter
fabian.kirstein@fokus.fraunhofer.de
Tel. +49 (0)30 3463-7718