(Geo)DCAT-AP – Status, Usage, Implementation Guidelines, Extensions

HMA-AWG Meeting
ESRIN (Room D)

20. May 2016

Uwe Voges (con terra GmbH)
GeoDCAT-AP – European Data Portal

- **European Data Portal (EDP):** central hub providing access to metadata about the data made accessible throughout Europe
  
  > www.europeandataportal.eu

- **Consortium:**

  - 450K+ metadata sets, 34 countries covered, 70+ catalogues, metadata automatically translated into 6 languages (FR, DE, EN, ES, PL, IT)
  
  - includes *geo-data* of most important european geospatial metadata catalogues with different interfaces (INSPIRE, OGC CSW, OpenSearch)
  
  - Problem: EDP is based on CKAN and Virtuoso (RDF store)
  
    > Geo-Catalogues mostly do not provide RDF (DCAT-AP) response payload
  
    > Our job: defined an architecture and made an implementation of Geo-Harvester
GeoDCAT-AP – European Data Portal

- **EDP Harvester** interact with just one(!) well defined interface
- **Interface** (OpenData-GEO - ODG) is subset of OAI-PMH spec and returns GeoDCAT-AP
- Every **ODG-Adapter Type** can harvest one Geo-interface type
  - able to process the harvesting in blocks (partitioning) and supports differential harvesting based on the metadata’s datestamp (last modified date).

- **ODG Interface** is based on REST: accepts simple „query-“ (or better harvesting-) requests with just a few parameters (e.g. startDateDateTime, endDateDateTime, startPos/maxRecs (resumption, for partitioning).
- **ODG Adapter** map ISO metadata to GeoDCAT-AP
  - classifications, schema and semantic tests and data quality improvements
  - linking to related information (e.g. INSPIRE Annex Theme to SKOS-based GEMET thesaurus, linking to Eurovoc, )
DCAT / DCAT-AP V 1.1

- **DCAT** (http://www.w3.org/TR/vocab-dcat/): RDF vocabulary for interoperability between data catalogs published on the Web
  > describing public sector datasets in EU
  > Basic use case is to enable cross-data portal search for data sets
  > Provides semantic interoperability with other applications on the basis of reuse of established controlled vocabularies (e.g. EuroVoc, Dublin Core, SDMX, etc).
GeoDCAT-AP V 1.0

- **GeoDCAT-AP**: extension of DCAT-AP for describing geospatial datasets / series / services
  - [https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/geodcat-ap-v10](https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/geodcat-ap-v10)
  - Spec was elaborated by a WG under the EU ISA programme
  - provides RDF syntax binding for ISO 19115 core and INSPIRE metadata elements
  - does not replace the INSPIRE Metadata Regulation nor the INSPIRE Metadata technical guidelines based on ISO 19115 and ISO 19119.
  - Provides INSPIRE metadata to be used in other contexts / for other use cases
- GeoDCAT-AP defines a **Core** and an **Extended Version**.
  - Core: alignments for the subset of INSPIRE metadata elements that can be mapped to DCAT-AP
    - enables harvesting and re-use of INSPIRE metadata records through DCAT-AP-conformant applications and services.
    - For a number of INSPIRE metadata elements, GeoDCAT-AP proposes the use of URI code list registers.
Extended alignments for those INSPIRE metadata elements not included in GeoCAT-AP Core

In EDP we use Core with XML encoding

Tools: http://rdf-translator.appspot.com/

JSON-LD would also be possible

XSLT script existing that transforms ISO 19139 metadata into GeoCAT-AP

For the EDP we developed our own which is better aligned with EDP requirements

GeoCAT-AP based definitions

GeoDCAT-AP
DCAT-AP V 1.0 – Implementation Guidelines (Extensions)

- How to model dataset series: e.g. dct:hasPart, dct:isPartOf
- URIs for Organizations: investigate the existence of sources
- Modelling of publisher vs. contact point
- Mapping national themes to MDR Data Theme Vocabulary
- How to describe and refer to licences
- Which identifiers to use for datasets and distributions
- How to extend (profile) DCAT-AP
- How to detect and handle duplicates (e.g. when metadata is harvested from different portals)
- Controlled keywords

- DCAT-AP currently defined as an RDF vocabulary with additional recommendations concerning cardinalities, datatypes and value vocabularies that go beyond the base DCAT recommendation.
  > Suggestion: to publish an OWL expression of DCAT-AP.
- How to model and express provenance? Source and lineage of metadata and data, use of PROV-O
- Standard Entity-ID Naming Service
- Mapping DCAT-AP to CKAN
- How to structure and partition DCAT-AP files
- Which tools exist for creating DCAT-AP. DCAT-AP API. Validation
- Relationships between an dataset and its source
- Data versioning
DCAT-AP V 1.0 – Service Access Extension

- Service based data access: I’m working on this...
- Datasets often not provided as fix downloadable items (as DCAT presumes)
- Instead: provided via standardized access services allowing dynamic access
- Reasons: e.g. huge datasets requiring data storage and complex processing
- (Geo)DCAT-AP does not have appropriate means to model this
- As service based access is a distribution channel -> dcat:Distribution appropriate class to model service access
- Simple accessURL not sufficient, to cover service based use cases. Instead:
  > URL-template including parameters (described by name, type, constraints,...)
  > URL may point to service or to a service-description(!)
  > Different protocol bindings: HTTP/GET/KVP, HTTP/POST/XML, SOAP,…
  > Format + Compression(!)
  > ServiceType: e.g. download, view,…Service Standard: e.g. OGC WMS 1.3.0

EUMETSAT High Rate SEVIRI Level 1.5 – Meteosat Second Generation (MSG) - served as OGC Web Map Service (Esri ArcGIS)

Client: European Data Portal - OGC WMS
Client (based on con terra map.apps)

Request (sample): http://solutions.conterra.de/arcgis/services/BIDS2016/BIDS2016_SEVIRI_09082015_17_00_00/ImageServer/WMSServer?
service=WMS&version=1.3.0&request=GetMap&CRS=CRS:84&bbox=-30.21,21.92,-10.05,42.406235&width=1200&height=710&layers=0&styles=default&format=image/png32
**DCAT-AP V 1.0 – Service Access Extension**

- **First approach:** link to OSDD, Capabilities, WSDL,...

```xml
<dcat:distribution>
  <dcat:Distribution>
    <dct:type rdf:resource="http://www.someServiceRegistery.eu/serviceTypes/XYZDownload/1.0"/>
    <dct:title lang="en">Downloads parts of dataset ABC via specific download service XYZ</dct:title>
    <dct:downloadService rdf:resource="http://www.iana.org/assignments/media-types/application/xml"/>
    <dct:downloadService rdf:resource="http://www.iana.org/assignments/media-types/application/json"/>
    <dct:downloadService rdf:resource="http://www.iana.org/assignments/media-types/application/gzip"/>
    <ns:service>
      <ns:HTTPService>
        <l:title>The service is not directly accessible but via the service endpoint described as URL-templates in an OpenSearch Description Document</l:title>
        <l:title>we cannot currently include these URL-templates directly here because for OSDD there is not an RDF-model available</l:title>
        <dct:downloadService rdf:resource="http://www.iana.org/assignments/media-types/application/opensearchdescription+xml"/>
        <ns:binding rdf:parseType="Resource">
          <rdfs:label>HTTP/GET</rdfs:label>
        </ns:binding>
      </ns:HTTPService>
    </ns:service>
  </dcat:Distribution>
</dcat:distribution>
```

- **Problem:** with service description in external file, the provider has to create and process this file on a web-accessible endpoint.
- **Better:** include description into DCAT-AP. But then we need to develop (or use -if somewhere already existing) such an RDF-based representation."
DCAT-AP V 1.0 – Service Access Extension

```xml
  <ShortName>XYZDownloadService</ShortName>
  <Description>The XYZDownloadService......................</Description>
  <Tags>DownloadService XYZ</Tags>
  <Contact>u Boges@con terra.de</Contact>
  <Url type="application/xml" rel="results" template="http://www.someXYZDownloadService.eu/datasetABC/download.xml?lang={language}&amp;amp;box={geo:box}&amp;amp;startDateTime={time:startDate}&amp;amp;endDate={time:endDate}?"/>
</OpenSearchDescription>
```

```
<parameters:
  Parameter name="lang" value="{language}" title="Two letters language code according to ISO 639-1"/>
  Option value="en" label="English"/>
  Option value="de" label="Deutsch"/>
</parameters:Parameter>

<parameters:
  Parameter name="box" value="{geo:box}" title="Defined by west, south, east, north' coordinates of longitude, latitude, in decimal degrees (EPSG:4326)" pattern="^[0-9.\-]+"/>
</parameters:Parameter>

<parameters:
  Parameter name="startDateTime"
  value="{time:startDate}"
  title="Beginning of the time slice. Format should follow RFC-3339"
  minValue="2016-05-11T00:00:00Z"
  maxValue="2016-05-13T00:00:00Z"
  pattern="^[0-9-]{4}-[0-9]{2}-[0-9]{2}T[0-9]{2}:[0-9]{2}:[0-9]{2}\.[0-9]\+\d\dZ$"/>
</parameters:Parameter>

<parameters:
  Parameter name="endDate"
  value="{time:endDate}"
  title="End of the time slice. Format should follow RFC-3339"
  minValue="2016-05-11T00:00:00Z"
  maxValue="2016-05-13T00:00:00Z"
  pattern="^[0-9-]{4}-[0-9]{2}-[0-9]{2}T[0-9]{2}:[0-9]{2}:[0-9]{2}\.[0-9]\+\d\dZ$"/>
</parameters:Parameter>
```

```xml
<Url type="application/xml" rel="results" template="http://www.someXYZDownloadService.eu/datasetABC/download.json?lang={language}&amp;amp;box={geo:box}&amp;amp;startDateTime={time:startDate}&amp;amp;endDate={time:endDate}?"/>
```

```
<parameters:
  Parameter name="lang" value="{language}" title="Two letters language code according to ISO 639-1"/>
  Option value="en" label="English"/>
```