

WEBINAR FOR DATA PROVIDERS

data.
europa
academy 

From theory to action: Automatic data publishing

- 20 January 2023
- 10.00 – 11.30 CET

Introduction



Simon Steuer
*Publications Office of the EU
Head of Sector*



Simon Dutkowski
*Fraunhofer FOKUS,
Digital Public Services*

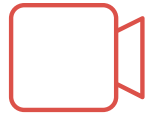


Torben Jastrow
*Fraunhofer FOKUS,
Digital Public Services*



Bart Hanssens
*Data.gov.be,
Belgium Federal
Public Service Policy
and Support (BOSA)*

Rules of the game



The webinar will be recorded



Please mute yourselves during the webinar



Please reserve 3 min after the webinar to help us improve by filling in our feedback form



For questions, please use the Teams chat. We will respond to your questions during the Q&A.

Agenda

10.00 – 10.05	Opening (OP)
10.05 – 10.15	How to describe your metadata with DCAT-AP (TJ)
10.15 – 10.20	What does automatic publishing mean for data.europa.eu? (TJ)
10.20 – 10.35	Two ways of automated data publishing provided by data.europa.eu (SD)
10.35 – 10.55	Dos and don'ts for the automated data publishing (SD)
10:55 – 11:10	Data publishing on data.europa.eu (BH – Guest Speaker)
11.10 – 11.25	Q&A
11.25 – 11.30	Closing and feedback

DCAT-AP

How to describe your metadata using DCAT-AP

DCAT-AP

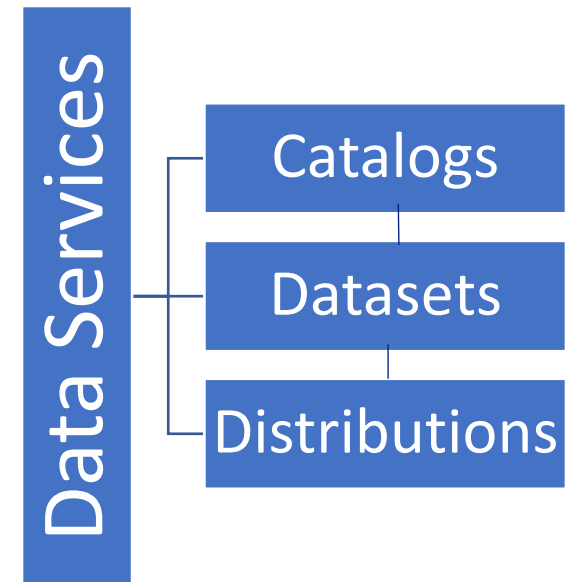
- Data Catalog Vocabulary - Application Profile
- Specification for describing public sector datasets in Europe.
- Used as the data model in data.europa.eu
- 117 catalogues in data.europa.eu are harvested in DCAT-AP
- The profile is based on Linked Data principles and the Resource Description Framework (RDF) & DCAT.

DCAT-AP

- It is designed to increase interoperability and allows the user to search for Open Data across multiple portals.
- The standard is constantly refined and currently published in version 2.1.1.
- Learn more about DCAT-AP at <https://data.europa.eu/en/academy/dcat-and-dcat-ap-interoperability-data-model>

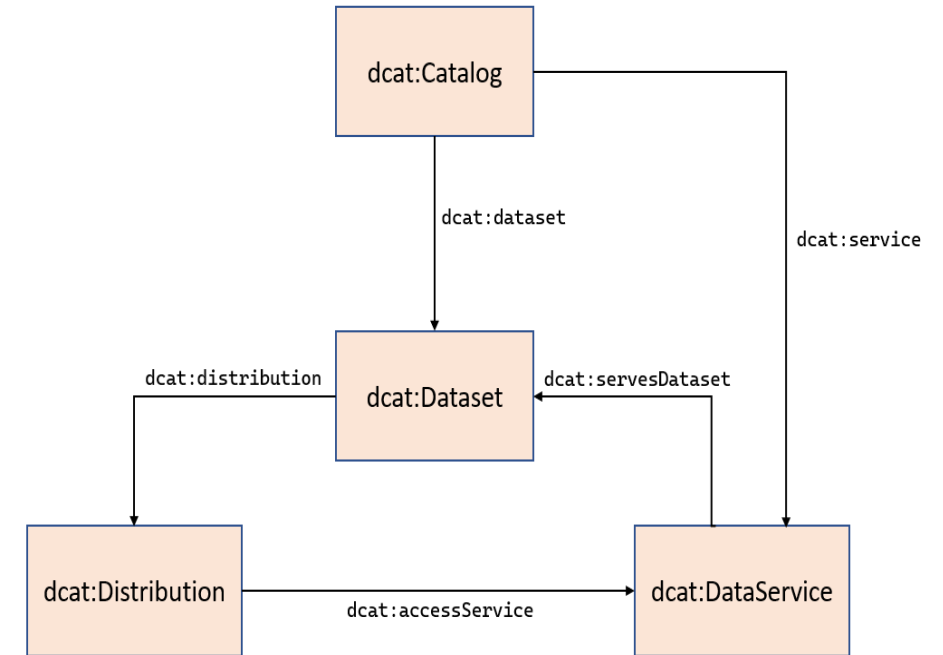
DCAT-AP

- In essence, DCAT-AP consists of four core classes.
- Catalogs combine a set of Datasets and each Dataset can have multiple Distributions.
- A Distribution represents the actual data and provides access via direct download.
- A Data Service represents a service that provides access to the actual data for a dataset or a whole catalogue.
- Further DCAT-AP specific classes are for instance Catalog Record or Category.
- Each class consists of a plethora of properties:
 - Catalogue – 18 properties
 - Dataset – 35 properties
 - Distribution – 23 properties
 - Data Service – 7 properties



Data Service

- Distribution: static link to files
- Data Service: points to endpoint with dynamic access
 - E.g. current traffic data
 - On-demand data export
- Data Service can have a Distribution or a Catalogue as parent



DCAT-AP

- Data Types and Languages

Title	dct:title	rdfs:Literal	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name.	1..n
-------	-----------	--------------	---	------

- Sub Data Structures

creator	dct:creator	foaf:Agent	This property refers to the entity responsible for producing the dataset	0..n
---------	-------------	------------	--	------

- Vocabularies

dct:format	Distribution	EU Vocabularies File Type Named Authority List ²⁵	http://publications.europa.eu/resource/authority/file-type	
------------	--------------	--	---	--

Vocabularies

- DCAT-AP includes mandatory controlled vocabularies.
- Most of them are published by the Publications Office of the EU.
- EU Vocabularies:
<https://op.europa.eu/en/web/eu-vocabularies>
<https://op.europa.eu/en/web/eu-vocabularies/authority-tables>

The screenshot shows the EU Vocabularies website interface. The top navigation bar includes the Publications Office of the European Union logo, a search bar, and language options. The main content area is titled "EU Vocabularies" and features a breadcrumb trail: "EU Vocabularies > Controlled vocabularies > Authority tables". Below this, there is a section for "Authority tables" with a filter box and a list of tables. Each table entry includes a "View" button and a "Name" column. The tables listed are: "Access right", "Accessibility", "Accreditation", and "Accreditation decision".

The second screenshot shows a detailed view of a "Concept scheme" for the "File type" dataset. It includes the following information:

- Version: 20221214-0
- URI: <http://publications.europa.eu/resource/authority/file-type>
- Type of dataset: Name authority list

Below this information is a table with columns: Code, Label, Valid since, Valid until, Predecessor, Successor, and Definition. The table contains the following data:

Code	Label	Valid since	Valid until	Predecessor	Successor	Definition
ZZ	7z	2000-01-01				7z is a compressed archive file format that supports several different data compression algorithms.
AAB	AAB	2021-08-01				An Android App Bundle (AAB) is a package file format used by the Android operating system to deliver applications to Google Play.
AAC	AAC	1997-01-01				Advanced Audio Coding (AAC) is an audio coding standard for lossy digital audio. AAC generally achieves higher sound quality than MP3 at the same bit rate.
AKN4EU	AKN4EU file	2019-06-20				AKN4EU file is an Akoma Ntoso XML file conforming to the AKN4EU (Akoma Ntoso XML) schema. AKN4EU files can be part of an AKN4EU ZIP file. A special AKN4EU file is the AKN4EU ZIP file.
AKN4EU_ZIP	AKN4EU ZIP	2019-06-20				AKN4EU ZIP is a ZIP archive containing an AKN4EU manifestation of a document. It must contain at least one AKN4EU file starting with a dot (.), and may contain other associated files and subfolders.
APK	APK	2008-09-23				Android Package (APK) is the package file format used by the Android operating system to distribute and install apps, mobile games and middleware.
APPX	AppX	2012-10-26				AppX is a file format used to distribute and install apps on Windows 8.x and Windows 10 IoT Core.

Example

```
a dct:Location .
<https://piveau.io/set/distribution/1>
a
  dcat:accessService
  dcat:availability
  dct:title
  dct:description
  dcat:accessURL
  dcat:downloadURL
  dcat:downloadURL
  dcat:mediaType
  dct:issued
  dct:modified
  dct:format
  dcat:compressFormat
  dcat:packageFormat
  dct:license
  dct:language
  dct:language
  dcat:spatialResolutionInMeters
  dcat:temporalResolution
  odrl:hasPolicy
  dcat:Distribution ;
  <https://piveau.eu/set/service/test-service> ;
  [ a
    skos:Concept ;
    skos:prefLabel "stable" ; ] ;
  "Example Distribution"@en ;
  "This is a example Distribution"@en ;
  <http://accessurl.com> ;
  <http://download.de/file1.csv> ;
  <http://download.de/file2.pdf> ;
  <http://www.iana.org/assignments/media-types/text/csv> ;
  "2015-08-27T22:00:00Z"^^xsd:dateTime ;
  "2018-04-03T11:48:21.950626Z"^^xsd:dateTime ;
  <http://publications.europa.eu/resource/authority/file-type/CSV> ;
  <http://www.iana.org/assignments/media-types/application/gzip> ;
  <http://www.iana.org/assignments/media-types/application/gzip> ;
  <http://europeandataportal.eu/ontologies/od-licenses#CC-BY-SA3.ONL> ;
  <http://publications.europa.eu/resource/authority/language/ENG> ;
  <http://publications.europa.eu/resource/authority/language/DEU> ;
  "10"^^xsd:decimal ;
  "P20M"^^xsd:duration ;
  [ a
    odrl:Policy ;
```

```
@prefix dcatapde: <http://dcat-ap.de/def/dcatde/> .
@prefix senias: <http://senias.de/ns/sis#> .

<https://piveau.io/catalogue/test-catalog>
a
  dcat:Catalog ;
  dct:type
  dct:title
  dct:description
  dct:publisher
  foaf:homepage
  dct:language
  dct:license
  dct:issued
  dct:modified
  dcat:themeTaxonomy
  dct:spatial
  dct:hasPart
  dct:isPartOf
  dcat:catalog
  creator
  rights
  dataset
  <https://piveau.io/set/service/test-service>
  dcat:DataService ;
  endpointURL
  endpointDescription
  servesDataset
  accessRights
  description
  license
  title
  <https://piveau.io/set/data/test-dataset>
  dcat:Dataset ;
  type
  title
  title
  language
  dct:language
  dct:description
  dct:description
  dcat:theme
  dcat:theme
  dct:subject
  dcat:distribution
  stat:attribute
  <https://piveau.eu/def/creator> ;
  [ a
    dct:RightsStatement ;
    rdfs:label "public" ] ;
  <https://piveau.eu/set/data/test-dataset> .

  <https://piveau.eu/set/service/test-service>
  dcat:DataService ;
  endpointURL <http://example.com/service/api/endpoint> ;
  endpointDescription <http://example.com/service/api/manual> ;
  servesDataset <https://piveau.eu/set/data/test-dataset> ;
  accessRights <http://publications.europa.eu/resource/authori
  description "This is an example Data Service"@en ;
  license <http://europeandataportal.eu/ontologies/od-lic
  title "Example Data Service"@en .

  <https://piveau.io/set/data/test-dataset>
  dcat:Dataset ;
  type <https://piveau.eu/def/type/dataset> ;
  title "Example Dataset"@en ;
  title "Beispieldatensatz"@de ;
  language <http://publications.europa.eu/resource/a
  <http://publications.europa.eu/resource/a
  dct:language
  dct:description
  dct:description
  dcat:theme
  dcat:theme
  dct:subject
  dcat:distribution
  stat:attribute
  <https://piveau.io/set/distribution/1> ;
  <http://some.attribute.l.example.com> , <
```

Automatic Publishing

What it means for data.europa.eu

Requirements for automatic publishing

- Metadata has to be
 1. On a space that can be accessed
 2. In a Format that can be understood
- Only Metadata can be automatically published, not the actual data

Catalogue Access

- Portal API
 - e.g. <https://api.dane.gov.pl/catalog>
- File on the Web
 - e.g. https://raw.githubusercontent.com/Fedict/dcat/master/all/datagovbe_edp.xml.gz
 - (compression possible)
- SPARQL Endpoint
 - e.g. <https://data.gov.cz/sparql>
- CSW/INSPIRE
 - <http://gdk.gdi-de.org/gdi-de/srv/eng/csw>

```
https://api.dane.gov.pl/catalog

:prefix adms: <http://www.w3.org/ns/adms#> .
:prefix dcat: <http://www.w3.org/ns/dcat#> .
:prefix dct: <http://purl.org/dc/terms/> .
:prefix foaf: <http://xmlns.com/foaf/0.1/> .
:prefix hydra: <http://www.w3.org/ns/hydra/core#> .
:prefix owl: <http://www.w3.org/2002/07/owl#> .
:prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
:prefix skos: <http://www.w3.org/2004/02/skos/core#> .
:prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
:prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

.:bn1 dct:identifier <http://sws.geonames.org/798544/> ;
      rdf:type dct:Location .

.:bn0 rdf:type vcard:Kind ;
      vcard:fn "KPRM" ;
      vcard:hasEmail <mailto:kontakt@dane.gov.pl> .

.:bn3 rdf:type hydra:PagedCollection ;
      hydra:itemsPerPage 20 ;
      hydra:lastPage "https://api.dane.gov.pl/catalog?sort=id&page=102" ;
      hydra:nextPage "https://api.dane.gov.pl/catalog?sort=id&page=2" ;
      hydra:totalItems 2033 .

<http://publications.europa.eu/resource/authority/country/POL> rdf:type dct:Location ;
      skos:inScheme <http://publications.europa.eu/resource/authority/country> .

<http://publications.europa.eu/resource/authority/data-theme> dct:title "Data theme"@en ,
      "Kategoria danych"@pl ;
      rdf:type skos:ConceptScheme .

<http://publications.europa.eu/resource/authority/data-theme/AGRI> rdf:type skos:Concept ;
      skos:inScheme <http://publications.europa.eu/resource/authority/data-theme> ;
      skos:prefLabel "Agriculture, fisheries, forestry and food"@en ,
      "Rolnictwo, rybołówstwo, leśnictwo i żywność"@pl .

<http://publications.europa.eu/resource/authority/data-theme/ECON> rdf:type skos:Concept ;
      skos:inScheme <http://publications.europa.eu/resource/authority/data-theme> ;
      skos:prefLabel "Economy and finance"@en ,
      "Gospodarka i finanse"@pl .

<http://publications.europa.eu/resource/authority/data-theme/EDUC> rdf:type skos:Concept ;
      skos:inScheme <http://publications.europa.eu/resource/authority/data-theme> ;
      skos:prefLabel "Education, culture and sport"@en ,
      "Edukacja, kultura i sport"@pl .

<http://publications.europa.eu/resource/authority/data-theme/ENVI> rdf:type skos:Concept ;
      skos:inScheme <http://publications.europa.eu/resource/authority/data-theme> ;
      skos:prefLabel "Environment"@en ,
      "Środowisko"@pl .

<http://publications.europa.eu/resource/authority/data-theme/GOVE> rdf:type skos:Concept ;
      skos:inScheme <http://publications.europa.eu/resource/authority/data-theme> ;
      skos:prefLabel "Government and public sector"@en ,
      "Rząd i sektor publiczny"@pl .
```

Format

Without transformation (Highly preferred)

- DCAT-AP

With transformation/mapping

- CKAN
- uData
- Socrata
- JSON:API
- INSPIRE

Push or Pull

Two ways of automated data publishing provided by data.europa.eu

Harvesting (Pull)

- We highly recommend that you provide DCAT-AP!
- Providing your complete catalogue
 - Via API
 - Via download file
- Uniquely identify each of your dataset
- Harvesting is regularly
 - Daily, Weekly, Monthly, ... Finally, you decide how often and when.
- In case you're not providing directly DCAT-AP, support us for the transformation/mapping!
 - Properties mapping
 - Value mapping

Identification when Source is DCAT-AP

If not specified, looking for:

1. dct:identifier
2. URIRef

Configuration

- dct:identifier
- URIRef
- Prefer URIRef over dct:identifier
- Partially URIRef (E.g., '/...')

From your Identifier to our URIRef

- Datasets
<http://data.europa.eu/88u/dataset>
- Catalogues
<http://data.europa.eu/88u/catalogue>
- Catalog Records
<http://data.europa.eu/88u/record>
- Distributions
<http://data.europa.eu/88u/distribution>

your-identifier

<http://data.europa.eu/88u/dataset/your-identifier>

Identifier Duplication

- Problem: Same identifier in another catalogue. Actual examples:
 - “17” (four times)
 - open-data-greece
 - dati-gov-it
 - open-data-bulgaria
 - govdata
 - “123” (three times)
 - “public-toilets” (three times)

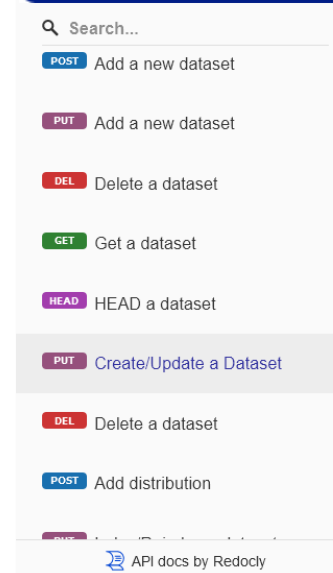
- => E.g., final URIRef could be <http://data.europa.eu/88u/17~~1>

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>  
PREFIX dct: <http://purl.org/dc/terms/>
```

```
SELECT (COUNT(?id) as ?count) ?id WHERE  
{  
  ?record a dcat:CatalogRecord ;  
  dct:identifier ?id  
}  
GROUP BY ?id  
HAVING(COUNT(?id) > 1)  
ORDER BY DESC(?count)
```

API (Push)

- Repository API (<https://data.europa.eu/api/hub/repo/>)
- Authentication required (Bearer Auth - JWT)
- Organizational processes still open (offline registering for authentication credentials)



Create/Update a Dataset

Create a Dataset with given id. When it already exists and has a different hash, it will be updated to the new Dataset. If it has the same hash, nothing will happen.

AUTHORIZATIONS: > *ApiKeyAuth* or *BearerAuth*

PATH PARAMETERS

id
required string

QUERY PARAMETERS

catalogue
required string
The id of the catalogue to add this dataset

data
boolean
Default: false
If set to true, callbacks for data upload are returned

REQUEST BODY SCHEMA: application/rdf+xml

Graph of the dataset
string

Responses

201 Dataset created.

API Functionality

- Identification of datasets with client-side id (origin)
- Create and/or manage catalogues
- Create and manage datasets of your catalogues
- Upload actual data
- Draft, publish, unpublish

Dos and Don'ts

Avoiding pitfalls during automated publishing

Dataset Identification



Don't

- export/provide your datasets without identifiers
- use different ways of identifying your datasets



Do

- uniquely identify each single dataset
- use always the same type of identifier for datasets, e.g. via dct:identifier

Unique Identification

On your side

- CKAN name
- OAI-PMH Identifier
- DCAT-AP
 - dct:identifier
 - URIRef
 - Partially URIRef
- INSPIRE
 - gmd:fileIdentifier

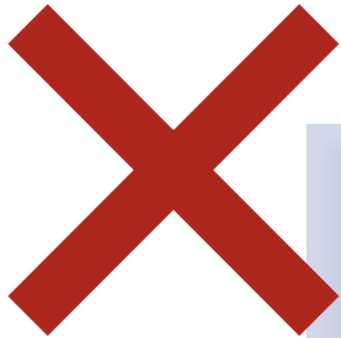
On our side

- Base URI plus your “normalized” id
- Normalization means:
 - Lowercase
 - Replacing special characters with ‘-’
 - Reducing multiple ‘-’ to one

Your-crazy-<*Special_~identifier-#1

<http://data.europa.eu/88u/dataset/your-crazy-special-identifier-1>

Syntactical Correctness



Don't

- publish syntactical incorrect presentations of your catalogue
- use “templates” to produce your metadata representation



Do

- use frameworks or tools to generate your metadata
- or apply some validation/checks of your metadata before publishing

Metadata RDF Representations

- RDF is a model to describe data in triples building a graph
- A graph can have a name, so each triple in a graph can also be a quadruple

Triples

- RDF/XML
- JSON-LD
- N-Triples
- N3
- Turtle

Quadruples

- N-Quads
- JSON-LD
- TRIX (XML for Quadruple)
- TRIG (Turtle for Quadruple)

Check Syntactical Correctness

- Presentation Level:
 - XML
 - JSON
- RDF Level:
 - Structure within presentation
 - URIRefs
- RDF Tools
 - Frameworks
 - Apache Jena
 - RDF4J
 - RDFLib (Python)
 - CLI tools
 - Apache Jena
 - Online services
 - <https://issemantic.net/>

RDF/XML Structural Examples

Incorrect RDF/XML (valid XML)

- Only one Object per Predicate allowed

```
<dcat:Dataset>
  <dcat:distribution>
    <dct:title>Title</dct:title>
    <dcat:Distribution>

    </dcat:Distribution>
  </dcat:distribution>
  <dcat:keyword>keyword</dcat:keyword>
</dcat:Dataset>
```

Correct RDF/XML

```
<dcat:Dataset>
  <dct:distribution>
    <dcat:Distribution>
      <dct:title>Title</dct:title>
    </dcat:Distribution>
  </dct:distribution>
  <dcat:keyword>keyword</dcat:keyword>
</dcat:Dataset>
```

RDF/XML Structural Examples

Incorrect RDF/XML (valid XML)

- Only Predicates are allowed in a Subject

```
<dcat:Dataset>  
  <dcat:Distribution>  
    < dct:title>Title</dct:title>  
  </dcat:Distribution>  
  <dcat:keyword>keyword</dcat:keyword>  
</dcat:Dataset>
```

Correct RDF/XML

```
<dcat:Dataset>  
  <dct:distribution>  
    <dcat:Distribution>  
      <dct:title>Title</dct:title>  
    </dcat:Distribution>  
  </dct:distribution>  
  <dcat:keyword>keyword</dcat:keyword>  
</dcat:Dataset>
```

URIRefs Examples

- Special Characters [*space*, *<*, *>*, *^*, *|*, *`*, *{*, *}*, *“*, **, *tab*, *newline*, *return*]
 - **Invalid:** <http://example.com/path with spaces>
 - **Valid:** <http://example.com/path%20with%20spaces>
- XML Attributes (RDF/XML)
 - RDF
<http://example.com/path?param1=value1¶m2=value2>
 - RDF/XML
<http://example.com/path?param1=value1&param2=value2>

URIRefs in XML Attributes

Correct in Turtle

```
<http://example.com/path?key1=value1&key2=value2>  
  a dcat:Dataset .
```

Incorrect in RDF/XML

```
<dcat:Dataset  
  rdf:about="http://example.com/path?key1=value1&key2=value2">  
</dcat:Dataset>
```

Pagination and Sorting



Don't

- provide your catalogue as one huge chunk
- provide your datasets in an unpredictable order during pagination



Do

- offering a mechanism for pagination
- provide datasets in a sorted order during pagination

Pagination for bigger Catalogues

- HTTP/REST APIs

- Http Headers, e.g.:

Range: items=0-24

Content-Range: items 0-24/66

- Query parameters, e.g.:

../path?offset=0&limit=100

- Download DCAT-AP

- As part of the RDF
- E.g., ckanext-dcat uses Hydra Core Vocabulary

<https://www.hydra-cg.com/spec/latest/core/#client-initiated-pagination>

Sorting

- Provide sorted access during pagination
- Preferable last modified or created datasets are served last, so they do not shift the remaining datasets

=> That will minimize the risk of skipping datasets or harvesting datasets twice

Access Infrastructure



Don't

- underdesign your Infrastructure



Do

- provide enough performance when your catalogue is accessed

Harvesting Process

- Import abortion: Deletion phase missing
 - More Datasets on data.europa.eu
- Example Reasons:
 - ckanext-dcat: 500 Internal Server Error while fetching next page
 - Request timeout
 - Gateway timeouts

Infrastructure Issues

- Load Balancer
 - Different nodes serving different pages
- Updates during harvesting
 - Missing datasets
 - Twice importing datasets (not critical)
- Performance
 - GDI-DE ~450.000 Datasets last > ~8h. => around 20 datasets per second

Summary

- Make your catalogue access reliable!
- Make your catalogue access performant!
- Make sure your metadata data is correct!

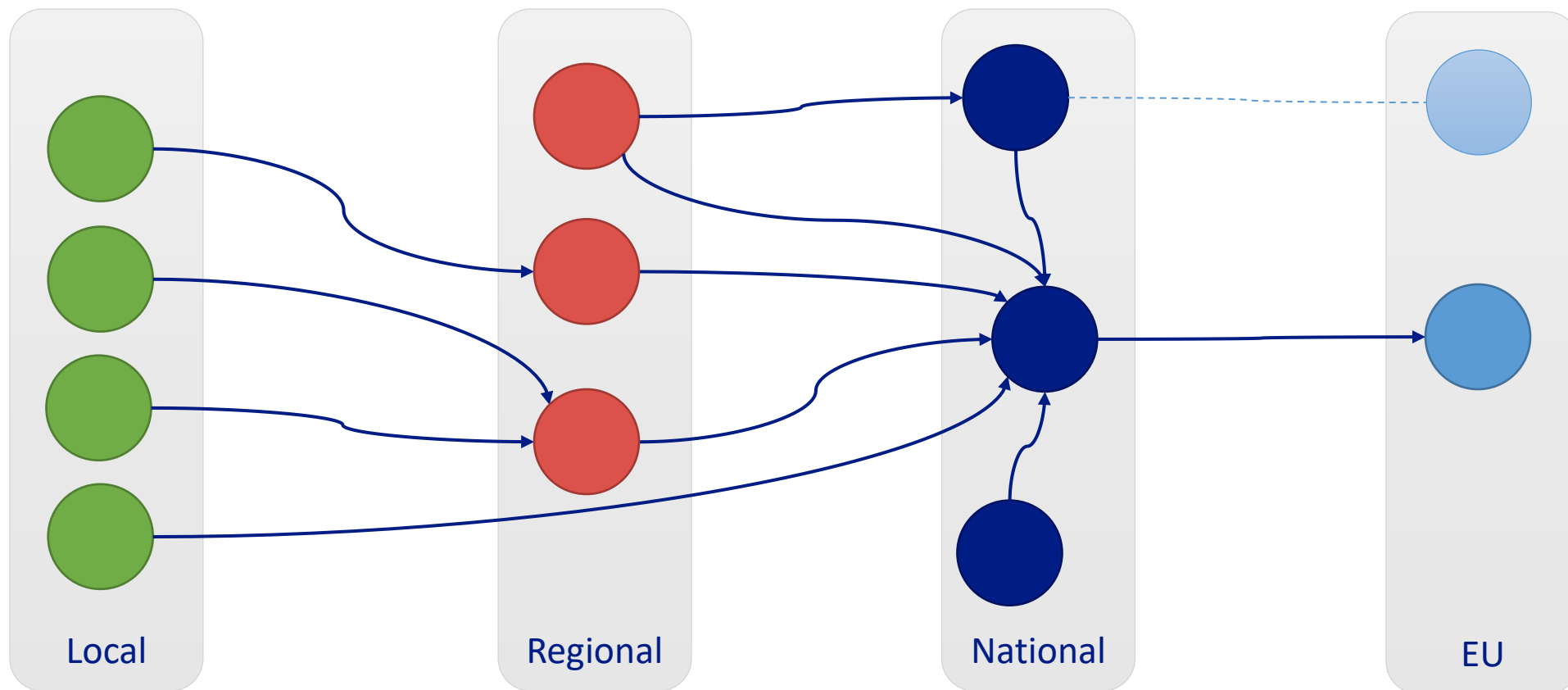
Data publishing on data.europa.eu

Data.gov.be on data.europa.eu

Data publishing on data.europa.eu

- Benefits for local / regional / national portal owners:
 - Extra visibility, administrations want to be visible cross-border
 - Extra features, e.g. machine translation

There is no such thing as “1 portal”



Indirect benefits

- Publishing on other sites creates awareness and food for thought
 - Metadata: are titles, descriptions ... clear and concise ?
 - How to automate publication and updates of (meta)data ?
 - Who to contact for more information about the data?
- (Yes, this could mean extra work, but it's worth it)

Evolution of data.gov.be

- Custom validation tools
- Almost no input in DCAT(-AP)
- Scripts and conversion tools
- Bare minimum
- Reuse of SHACL validation rules
- Roughly 50% in DCAT
- Less tooling, more SPARQL
- Focus on (Geo-)DCAT-AP 2.0

Informal community of portal owners

You're not on your own

- Support by EU data portal team and other portal owners
- Metadata specifications increasingly supported by tools / portals
- “Don't let perfect get in the way of good enough”
 - Strive for perfection, but accept that it is an iterative process

Questions ?

- opendata@bosa.fgov.be
- [#data.gov.be:matrix.org](https://matrix.org/#data.gov.be)

Questions & Answers

Stay updated!

Sign up for the newsletter: data.europa.eu/newsletter

Follow us on social media:

 [EU_opendata](https://twitter.com/EU_opendata)

 [Publications Office of the European Union](https://www.linkedin.com/company/publications-office-of-the-european-union)

 [data.europa.eu](https://www.facebook.com/data.europa.eu)

data.europa.eu The official portal
for European data



Please
provide us
your
feedback!

