StatDCAT-AP

Familiarisation Webinar

5 February 2016
ISA Programme Action 1.1

ISA

Agenda

- Practical arrangements
- Welcome and introduction
- Introduction of the participants - Round table
- Scope and approach
- Rationale/scenarios
- Earlier work and existing solutions
    - Work done in 2015
    - DCAT-AP/Data Portals ecosystem
    - SDMX/ESMS
- Your questions
- Next steps

Practical arrangements

ISA

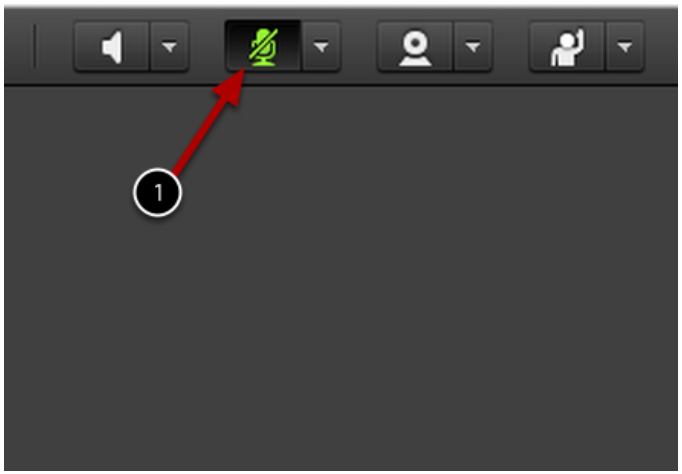
Practical arrangements

Join the meeting room:

- Go to http://ec-wacs.adobeconnect.com/stat_dcat/
- Click '**Enter as a guest**', fill in your name and click 'Enter meeting'
- Click '**ok**' to agree with the Web Conferencing Disclaimer
- Answer '**Start**' to the question: Would you like audio conferencing with this meeting?
- Enter your telephone number and have the system call you,

  OR dial your Local or International Access Number
- Dial the Conference Room Number: *1323004#

Practical arrangements



Please:

- **Mute your mic** when not speaking
  - To mute your microphone click the Microphone Icon. When muted, the icon will remain green with a slash.
- **Raise your hand** to request the floor (and help making the speaker aware of who is raising hands)
- **Use the chat box** for sharing links and making comments
- **Voting**: use the chat box (+1 / -1 / abstain)

Welcome and Roll call

The work is funded under ISA Action 1.1 of the ISA Programme on improving semantic interoperability in European e-Government systems.

Interoperability Solutions for European Public Administrations

Find out more on http://semic.eu

Round table of introductions

30 seconds each: name, affiliation, interest

Scope and approach

ISA

Scope and objectives

- Enhance interoperability between descriptions of statistical data sets and general data portals
- Develop and reach consensus on an Application Profile of the Data Catalog Vocabulary (DCAT) for description of statistical data sets with an initial focus on discovery
- Elaborate initial guidelines on the extraction of relevant metadata from the existing implementation at Eurostat and others
- Propose extensions to DCAT-AP with descriptive elements particularly useful for discovery of statistical data sets

Longer term roadmap

First stage, until July 2016

1. Connecting descriptions of statistical datasets with general open data portals through a common basic exchange format StatDCAT-AP
2. Developing guidelines for the extraction of metadata from specific implementations of statistical standards towards StatDCAT-AP

Later stages

3. Harmonising implementations of statistical standards, possibly as an extension of the basic StatDCAT profile (for the metadata level) and through the use of W3C RDF Data Cube Vocabulary (for the data level)
4. Creating a set of tools to facilitate automatic extraction and validation of metadata from data described by statistical standards into StatDCAT-AP
5. Conducting practical pilots in various stages of the above activities to test and verify approaches and solutions.

Approach

- Involvement of stakeholders from both the statistical domain and the open data domain
- Working group led by Eurostat (Marco Pellegrino) and Publications Office (Norbert Hohn) with ISA and DG CONNECT
- Editor: Makx Dekkers, SEMIC team
- Building mutual understanding in open discussion
- Striving for consensus among stakeholders towards commonly agreed specification

Rationale and Scenarios

ISA

Rationale/scenarios

The statistics community and the data (portals) community are apart, lacking a common language for describing data

Scenarios:

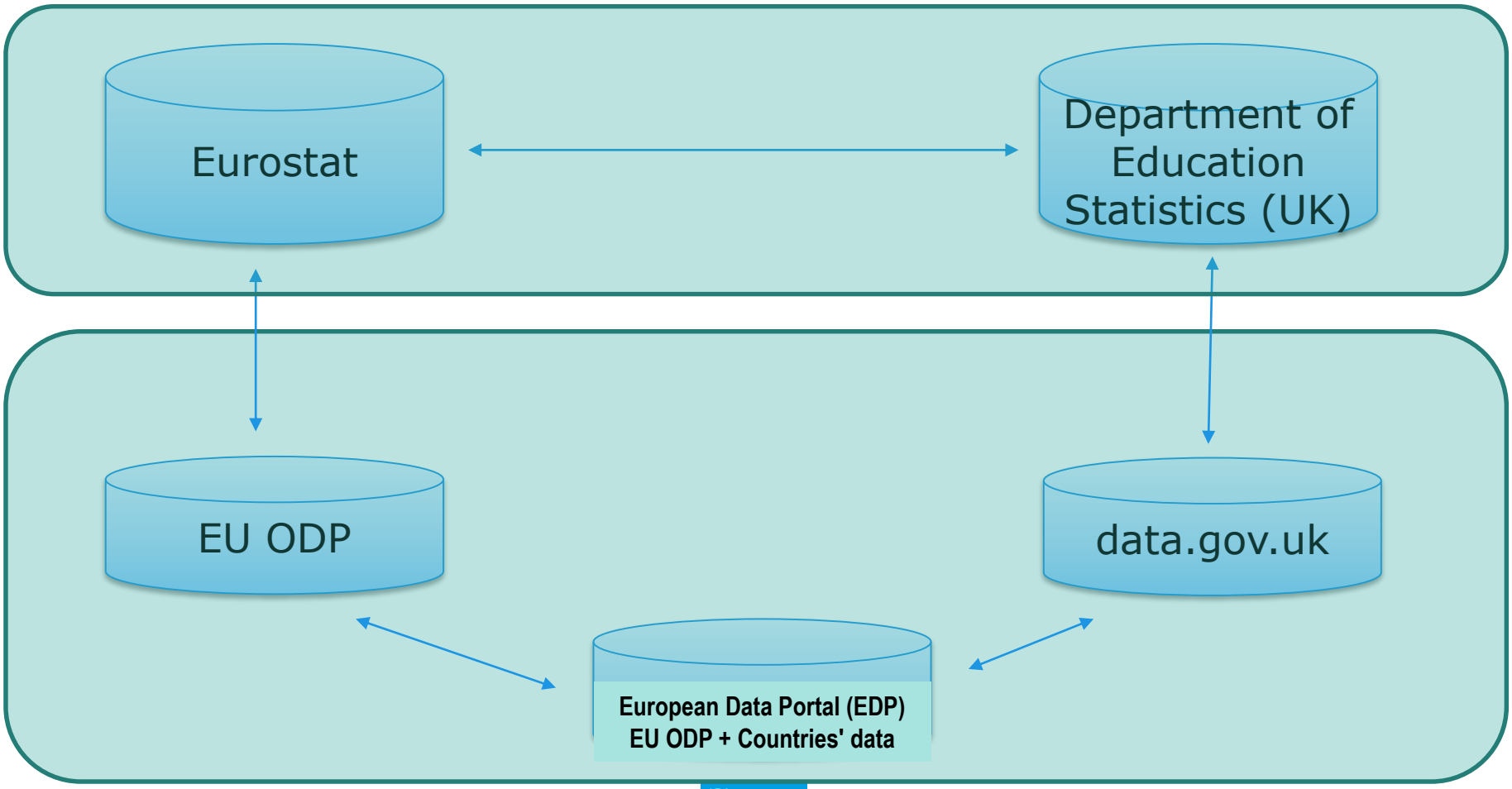
Simple exchange for retrievability and basic discovery

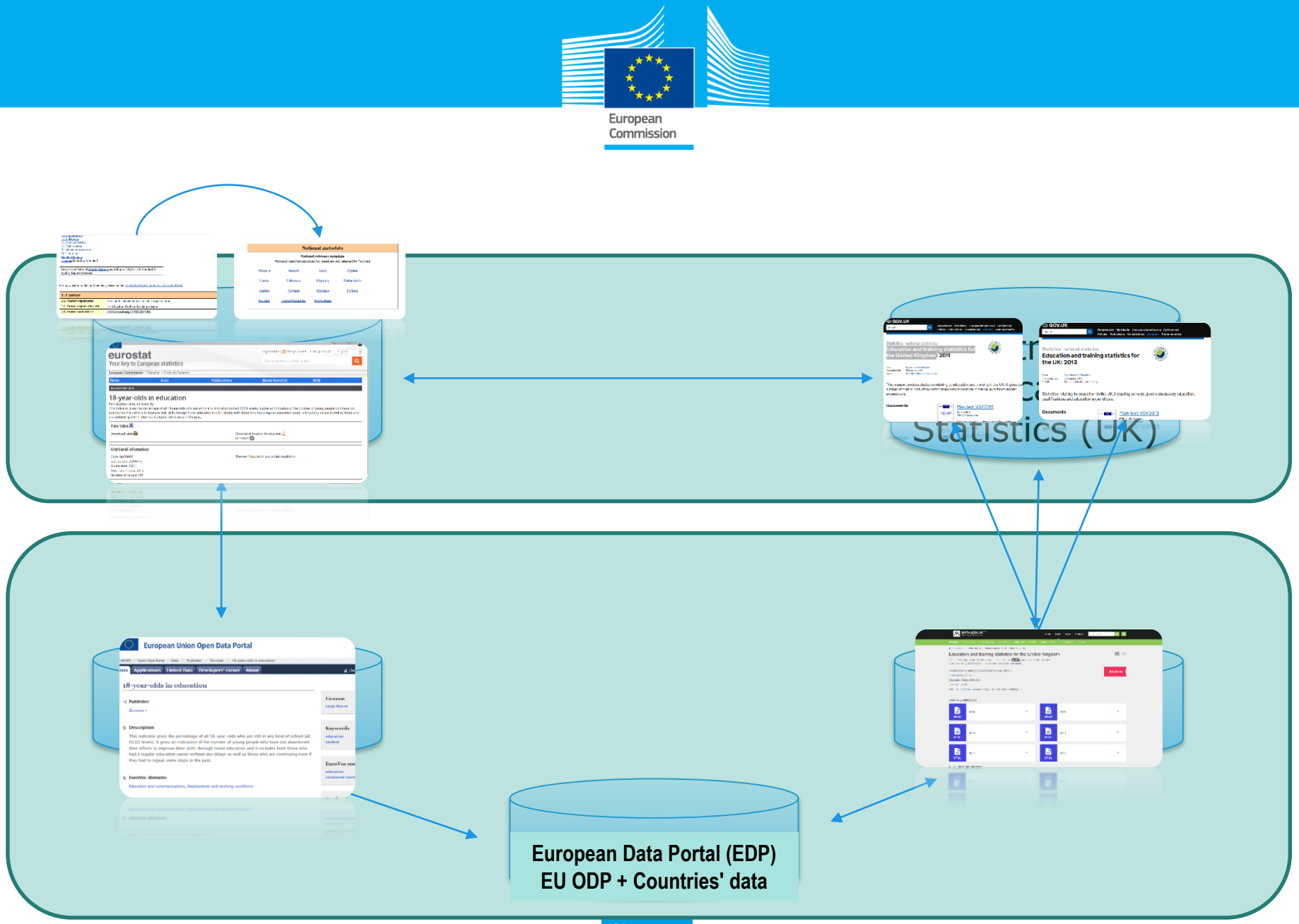
There is a need for mutual exposition of metadata descriptions in both statistical and data repositories in order to facilitate:

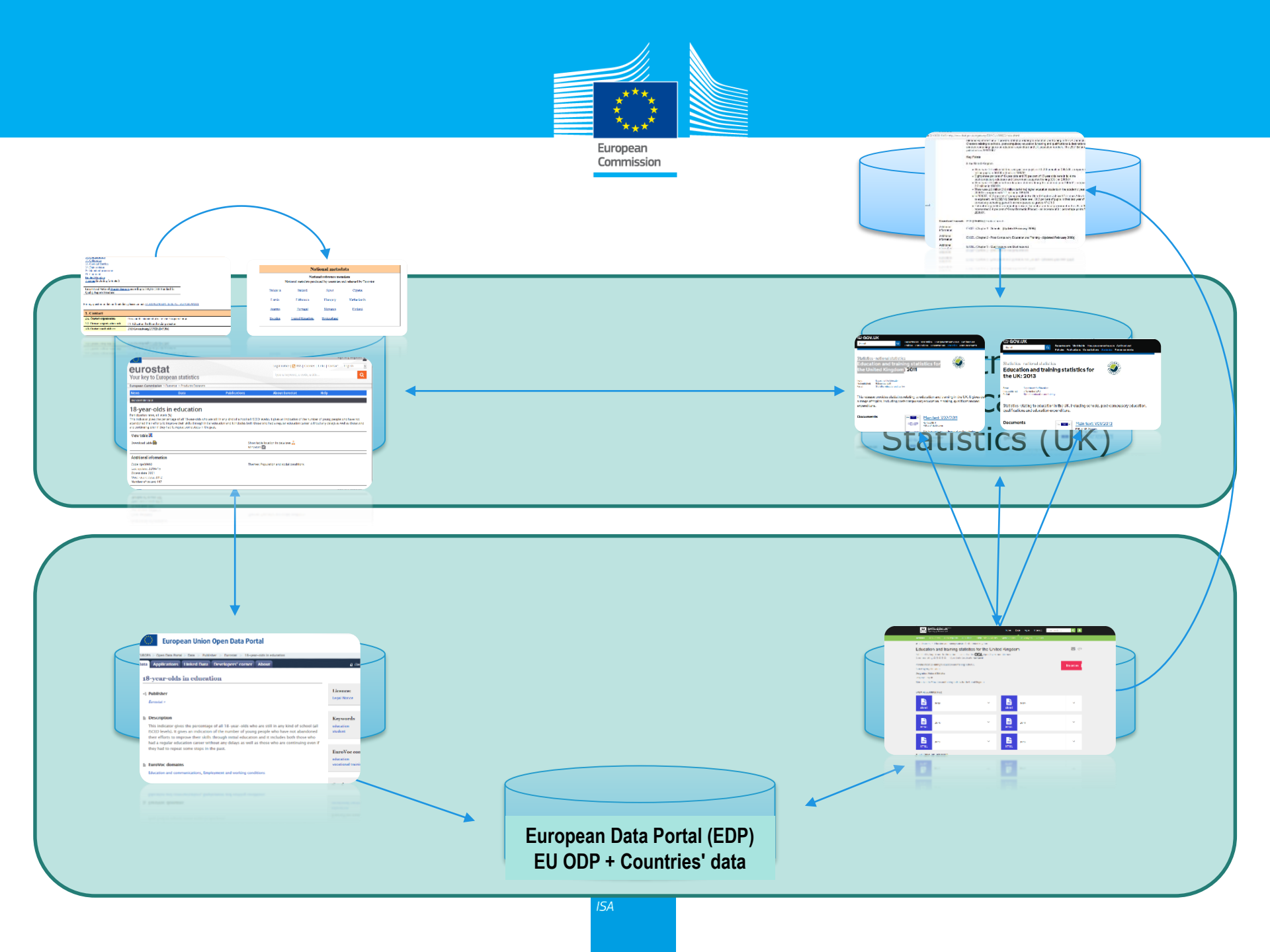
- Visibility (Retrievability), e.g. expose metadata of ECB statistics in EU Open Data Portal, and metadata of a data portal dataset in a statistical repository
- Basic discovery
- Exchange: Push-button export of StatDCAT-AP metadata from existing statistical (meta)data into data portals and vice versa, instead of ad-hoc mappings

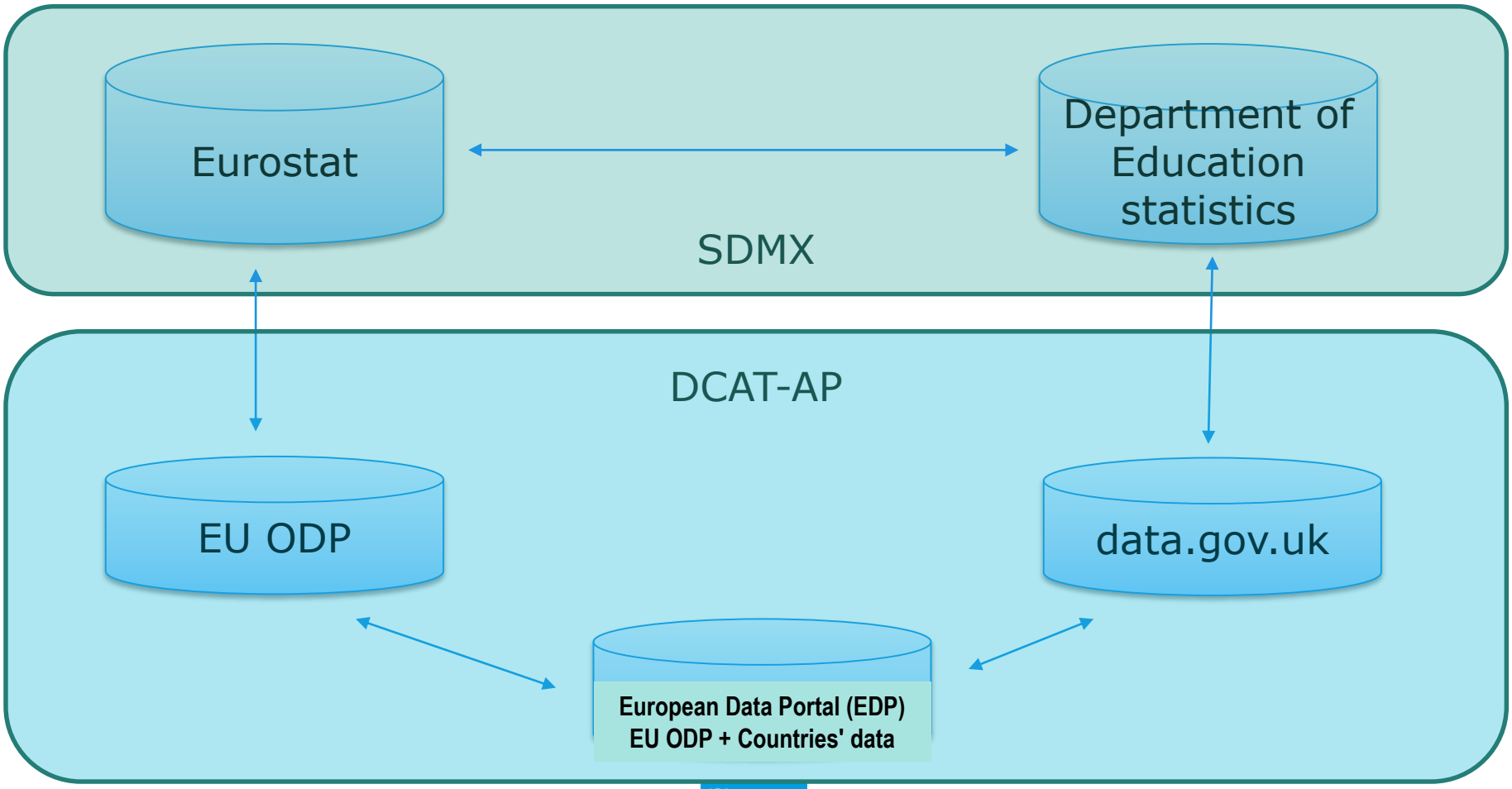
Scenarios:
Enhanced discovery thanks to an extension of DCAT AP by key statistical properties

- Enrich DCAT AP with key specific properties of statistical datasets (extension)
- Enhance discovery and filtering: Query DCAT AP based portals via key concepts of statistical metadata e.g. quality, dimensions


Eurostat
Department of Education Statistics (UK)
EU ODP
data.gov.uk
European Data Portal (EDP)
EU ODP + Countries' data


European Commission
National metadata
eurostat
Your key to European statistics
18-year-olds in education
Education and training statistics for the United Kingdom: 2011
Education and training statistics for the UK: 2013
Statistics (UK)
European Union Open Data Portal
18-year-olds in education
European Data Portal (EDP)
EU ODP + Countries' data
ISA


European Commission
eurostat
Your key to European statistics
18-year-olds in education
European Union Open Data Portal
18-year-olds in education
Statistics (UK)
European Data Portal (EDP)
EU ODP + Countries' data
ISA


Eurostat
Department of Education statistics
SDMX
DCAT-AP
EU ODP
data.gov.uk
European Data Portal (EDP)
EU ODP + Countries' data

Earlier work and existing solutions

Work done in 2015

ISA

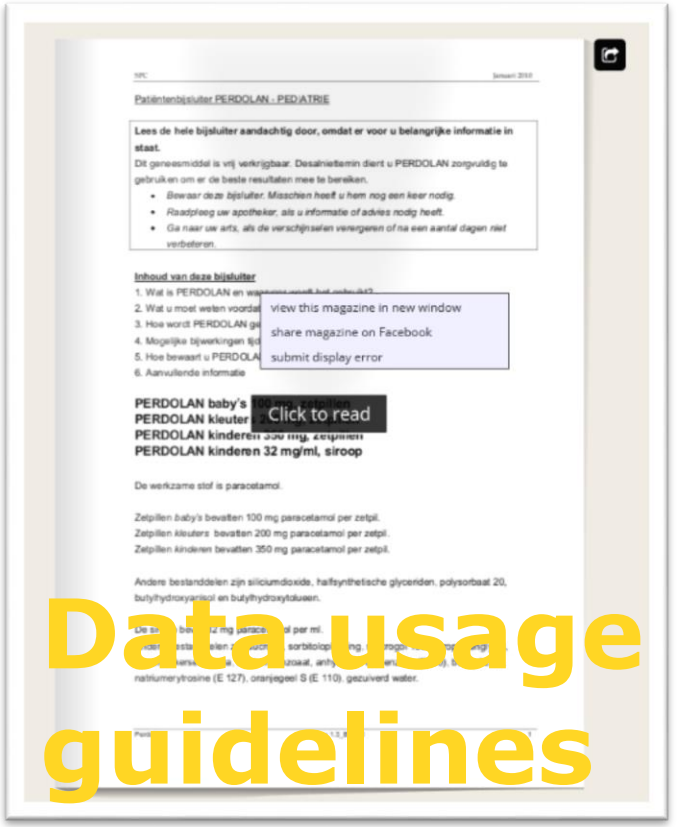
Preparation activities in 2015

Main driving question: *what information is being disseminated in the current data exchanges within the statistical ecosystem that can be related to DCAT-AP?*

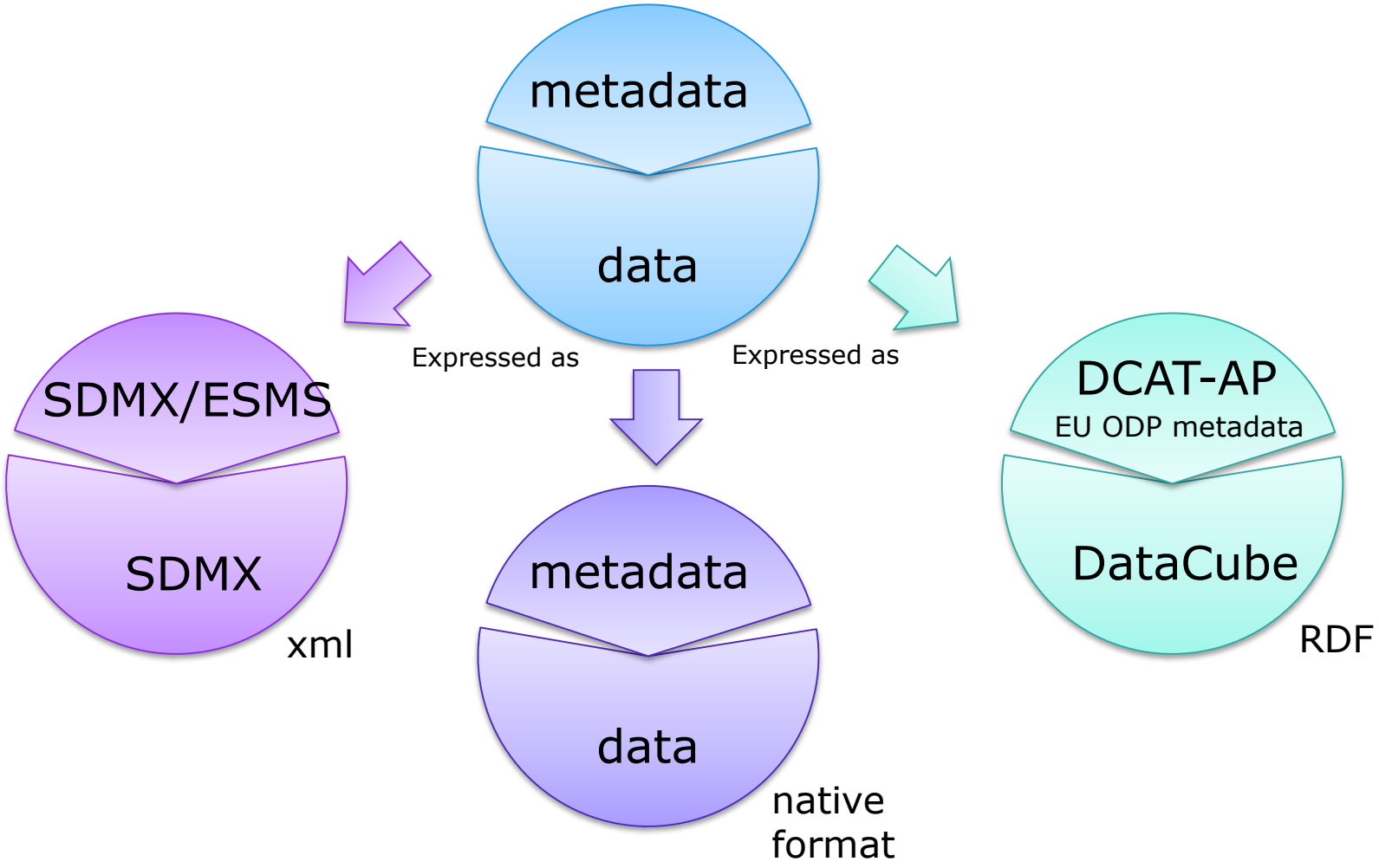
- analysis of data exchange formats such as SDMX, DDI, DataCube, CSV's
- analysis of statistical data portals: Eurostat, OECD, Belstat, ...

We realized the need for a terminological framework to bridge communities

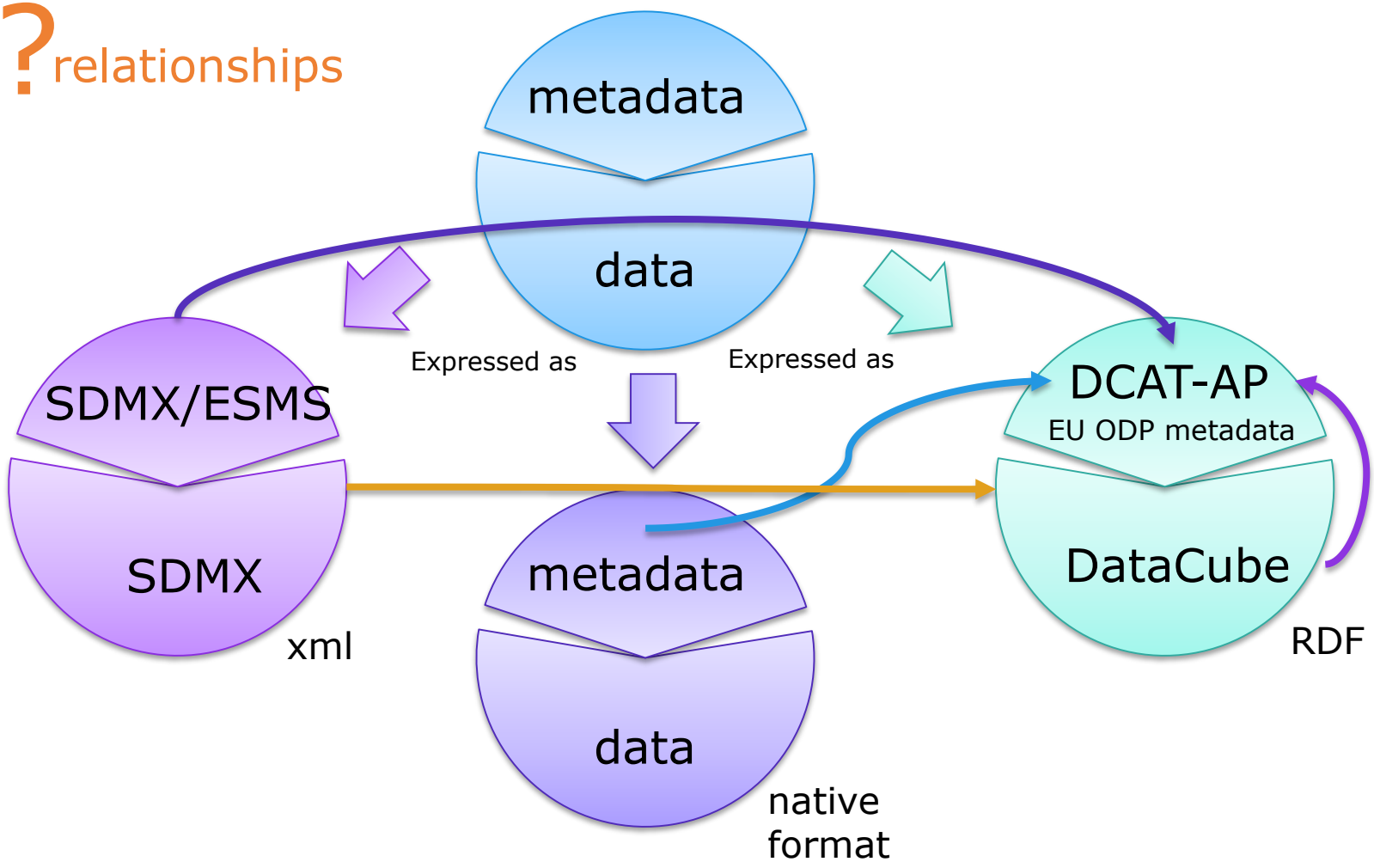
A metaphor



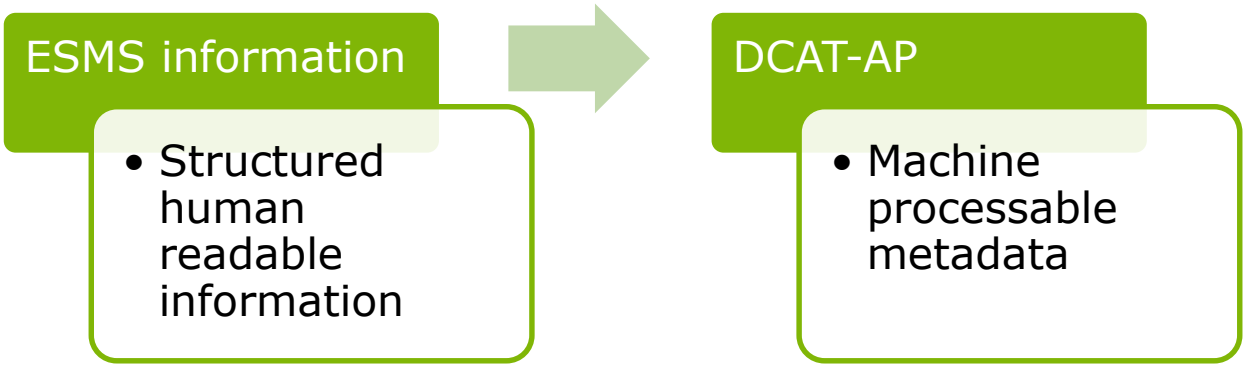
ISA


metadata
data
Expressed as
Expressed as
SDMX/ESMS
SDMX
xml
metadata
data
native
format
DCAT-AP
EU ODP metadata
DataCube
RDF

? relationships



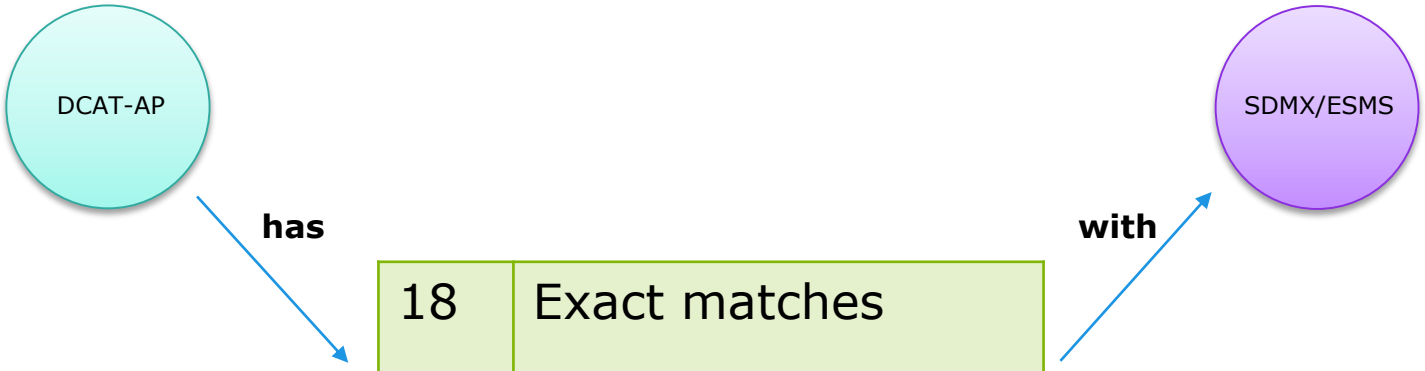
Via SDMX "reference metadata" (ESMS*)



- **High overlap** for ***trivial*** info: title, description, contactperson, publisher
- **Growing gap** (from some possible mapping to no correspondence at all) for ***contextual*** info

* ESMS: Euro-SDMX Metadata Structure, used at Eurostat and in the ESS for documentation about statistical data, its production process, methodology and quality.

Via SDMX “reference metadata” (ESMS)



18	Exact matches
9	Close matches
11	Broaden matches
6	Narrow matches

5 unmatched

40 unmatched

Conclusions (1/2)

- Only one unique property has been spotted:
  - The number of observations as measure of the size
- Some properties might get a more generic definition in the statistical context:
  - Geographical & temporal coverage are specific cases of dimensions
- Grouping & relationships between datasets common practice:
  - Statistical datasets are shared by subsets (slices)
  - Dataset get updated (data series)

Conclusions (2/2)

- DataCube, DDI RDF notation refer to dcat, dcterms as metadata structures
- SDMX data cube structure definitions offer information that can be exploited by DCAT-AP
- ESMS is not as-is transformable into DCAT-AP
- Statistical data portals overlap highly with DCAT-AP information

Existing solutions

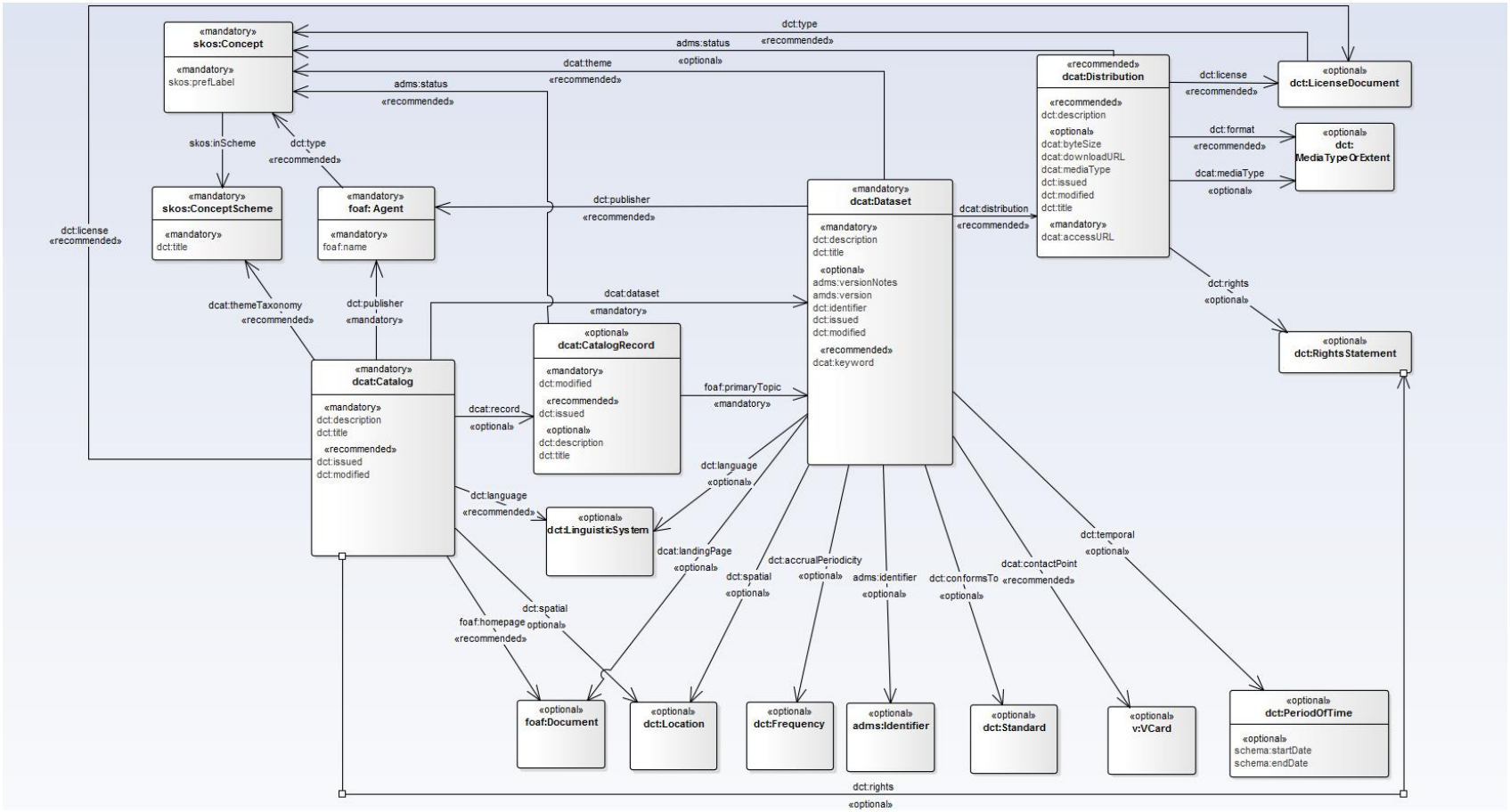
DCAT-AP
Application Profile for exchange of dataset descriptions between data portals in Europe

What is DCAT-AP

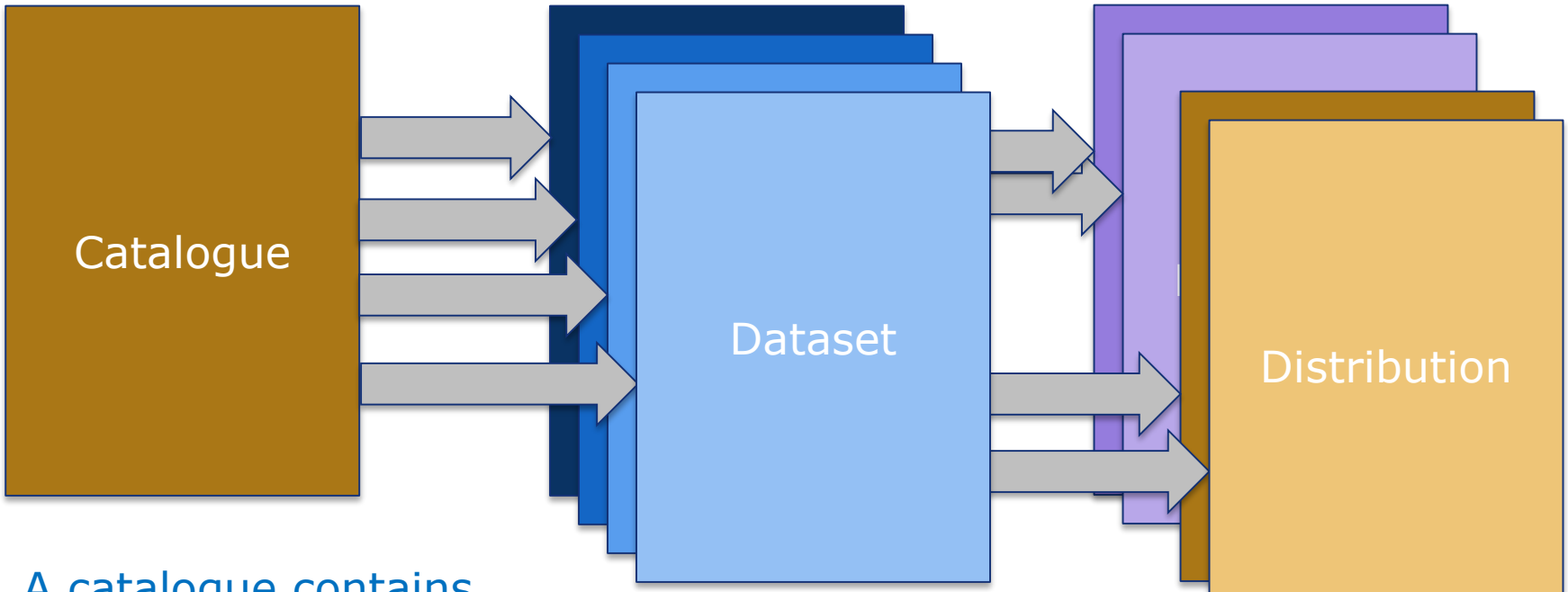
- **DCAT-AP** is an Application Profile of the W3C DCAT Recommendation to be used for the exchange of descriptions of datasets between data portals in Europe
- **DCAT** was developed by the Government Linked Data Working Group at W3C in 2012-2013 as an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web
- **DCAT-AP** was developed by the SEMIC activity under the ISA programme in 2013 and revised in 2015 for specific use in Europe, among others to support the European Data Portal

European Commission

DCAT-AP Data model



DCAT model overview

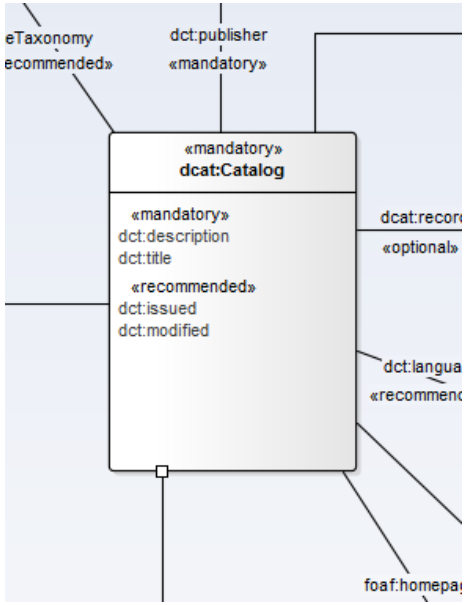


A catalogue contains one or more datasets

A dataset has one or more distributions

DCAT-AP model elements

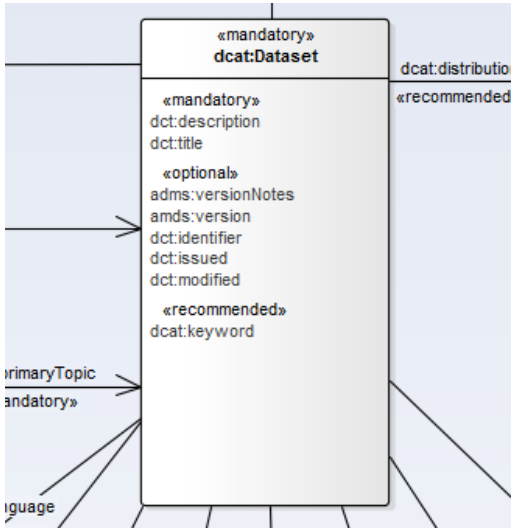
Catalogue



- Catalogue: a curated collection of metadata about datasets

DCAT-AP model elements

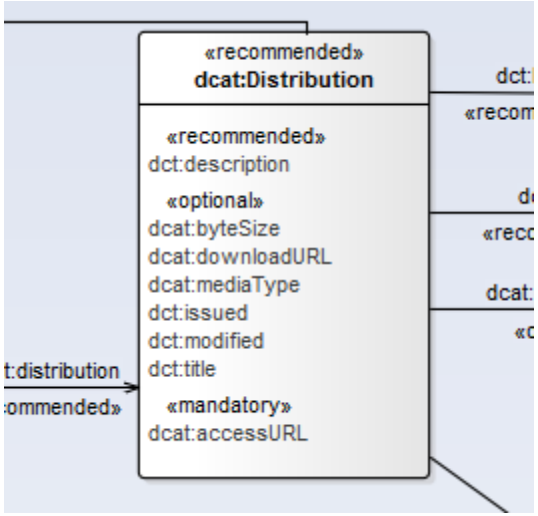
Dataset



- Dataset: A collection of data, published or curated by a single agent, and available for access or download in one or more formats

DCAT-AP model elements

Distribution



- Distribution: a specific available form of a dataset. Each dataset might be available in different forms, these forms might represent different formats of the dataset or different endpoints

Main aspects of DCAT-AP

- DCAT-AP provides a **common target** for **exchange of metadata**
- It is applicable **across domains** as it does not limit the kinds of datasets that can be described
- Its objective is to support exchange of metadata for the main purpose of **discoverability**
- As such, it only describes the characteristics of datasets that are relevant for **cross-domain discovery**
- Additional characteristics for datasets in particular domains can be specified in **extension profiles** (e.g. GeoDCAT, StatDCAT)

Existing ecosystem

- Several data portals support DCAT-AP
  - European Data Portal
  - National data portals, e.g. Netherlands, Sweden, Italy
- Several national profiles have been developed
  - DCAT-BE
  - DCAT-AP Switzerland
  - DCAT-AP-NL
- One domain-specific profile has been defined so far
  - GeoDCAT-AP for exchange between geoportals and general open data portals

GeoDCAT-AP

- **Extension of DCAT-AP** for describing geospatial datasets, dataset series, and services
- Basic use case: **to make spatial datasets, data series, and services searchable on general data portals**, thereby making geospatial information better searchable across borders and sectors.
- RDF syntax binding for the union of metadata elements defined in the main domain-specific standards (core profile of ISO 19115:2003 and the INSPIRE Directive).

GeoDCAT mapping specification

INSPIRE	Obligation	ISO 19115 Core	Obligation	DCAT-AP	GeoDCAT-AP
Resource	C	Dataset	M	Yes	Yes
		temporal)			
Lineage	M	Lineage	O	Yes	Yes
Spatial representation type	M	Spatial representation type	O		Yes
Encoding	M	Distribution format	O	Yes	Yes
Spatial resolution	C	Spatial resolution of the dataset	O		Yes (but as free text)
Responsible organisation	M	Dataset responsible party	O	Partially (only 3 of the 11 responsible party roles are supported)	Yes
Resource locator	C	On-line resource	O	Yes	Yes
Coordinate reference system; Temporal reference system	M; C	Reference system	O		Yes

GeoDCAT usage notes

INSPIRE	Obligation	ISO 19115 Core	Obligation	DCAT-AP	GeoDCAT-AP
Resource	C	Dataset	M	Yes	Yes
		temporal)			
Lineage	M	Lineage	O	Yes	Yes
Spatial	M	Spatial	O		Yes

Annex II DETAILED USAGE NOTES AND EXAMPLES

This annex contains further usage notes and examples on the mappings summarised in Section 5 of the GeoDCAT-AP specification.

II.1 Resource title - *Dataset title

The content of the element 'resource title' can be represented in RDF as a plain literal, and by using property dct:title.

This binding may also include the specification of the language by using attribute @xml:lang [XML]. The language to be specified is the one indicated by element metadata language, mapped to the language identifiers defined by IETF BCP 47 [BCP47].

Example

```
# Resource metadata in GeoDCAT-AP

[]  dct:title "Forest / Non-Forest Map 2006"@en.
```

Existing solutions

SDMX

The Information Model

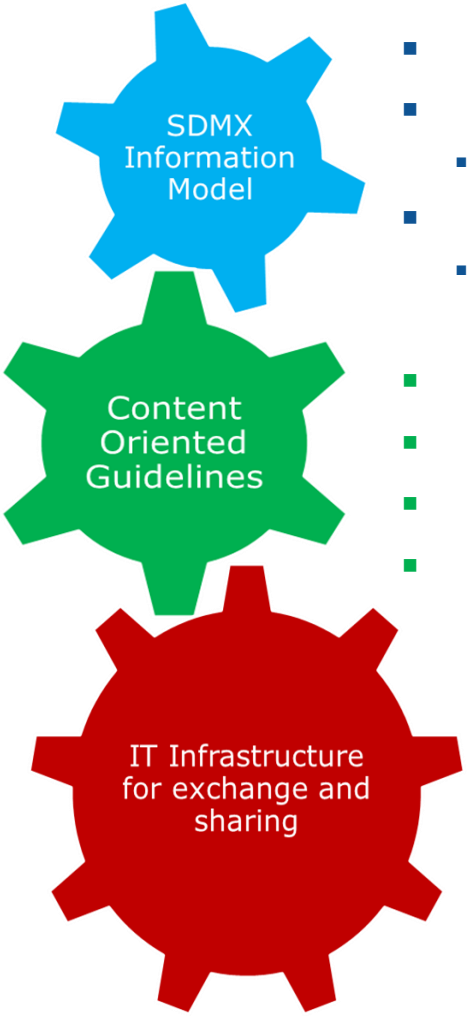
Statistical Data and Metadata eXchange

SDMX

ISO IS 17369

World Bank

MAIN COMPONENTS



- Describes statistics in a standard way
- Objects and their relationships
  - **Data and Metadata Structures and formats, Concepts, Code Lists**
- Central management and standard access
  - **SDMX Registry, SDMX Web Services**

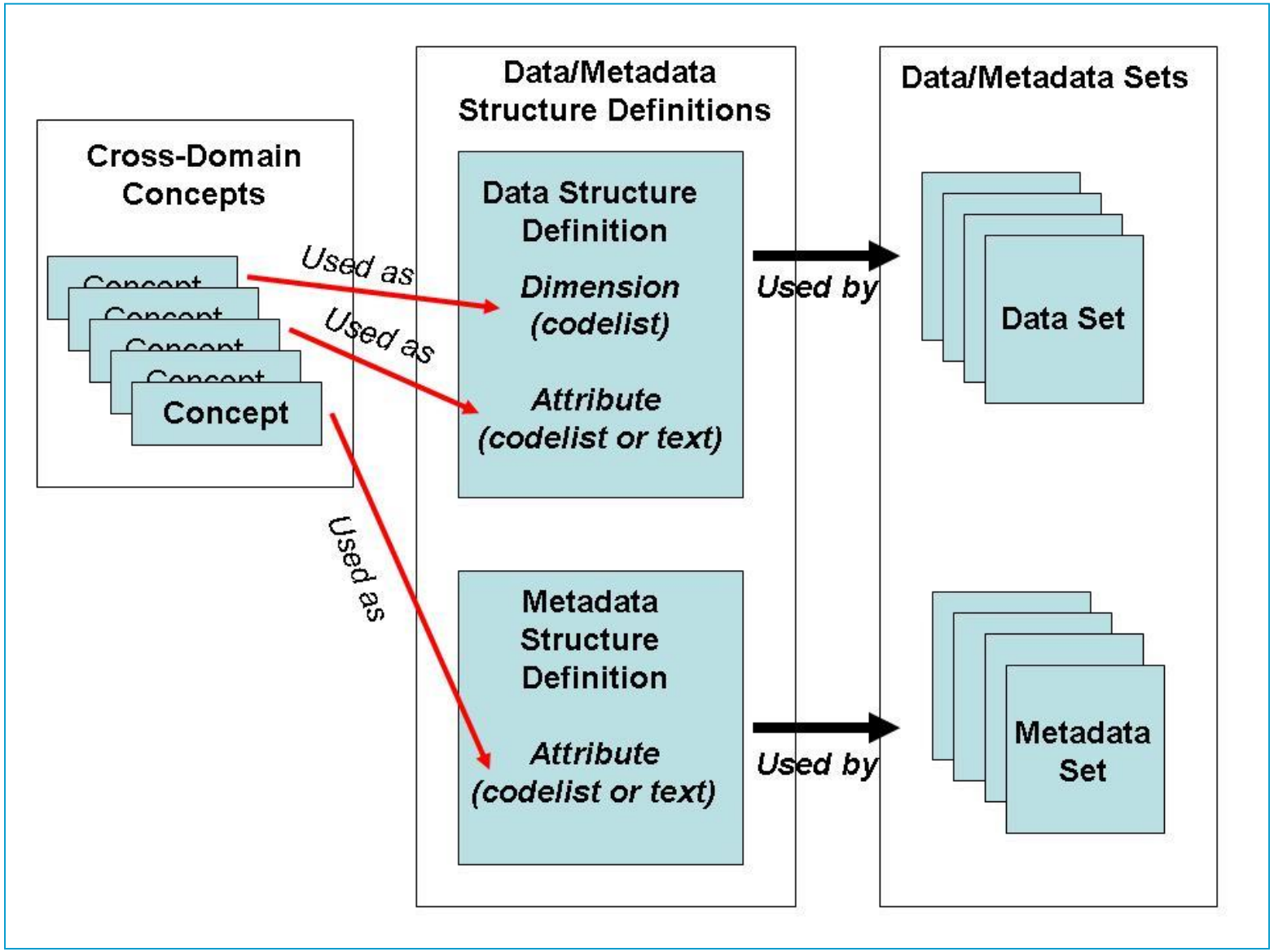
- Cross Domain Concepts
- Cross Domain Code Lists
- Statistical Domains
- SDMX Glossary (ex Metadata Common Vocabulary)

- Push: Provider generates and sends file to receiver
- Pull: Provider opens web service to data
- Hub: Special case of pull: receiver downloads on end-user request

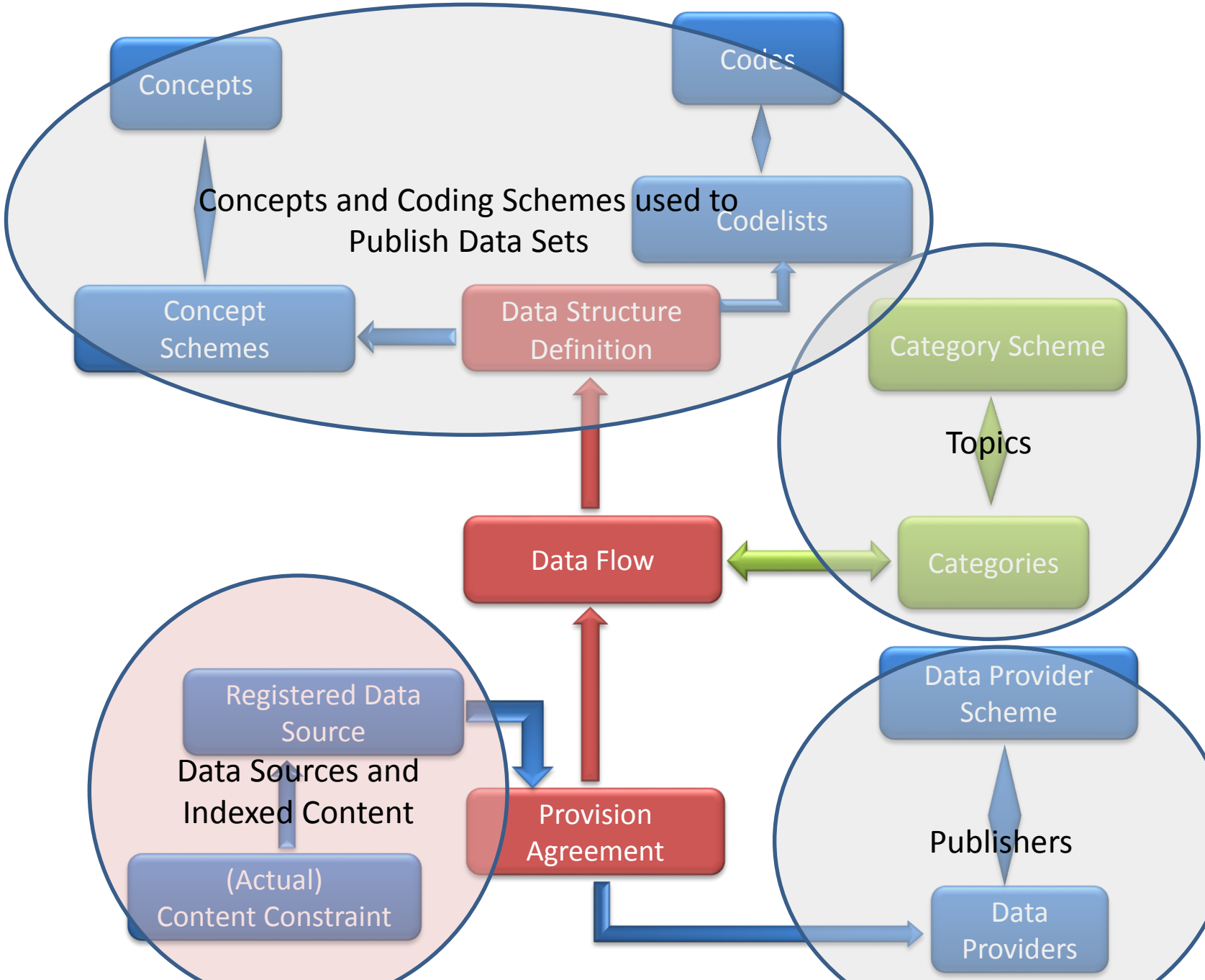
Latest version: SDMX 2.1 (2011), ISO standard in 2013

Linked Open Data is based on the SDMX Information Model
(Data Cube Vocabulary, W3C recommendation in 2012, latest version 2014)

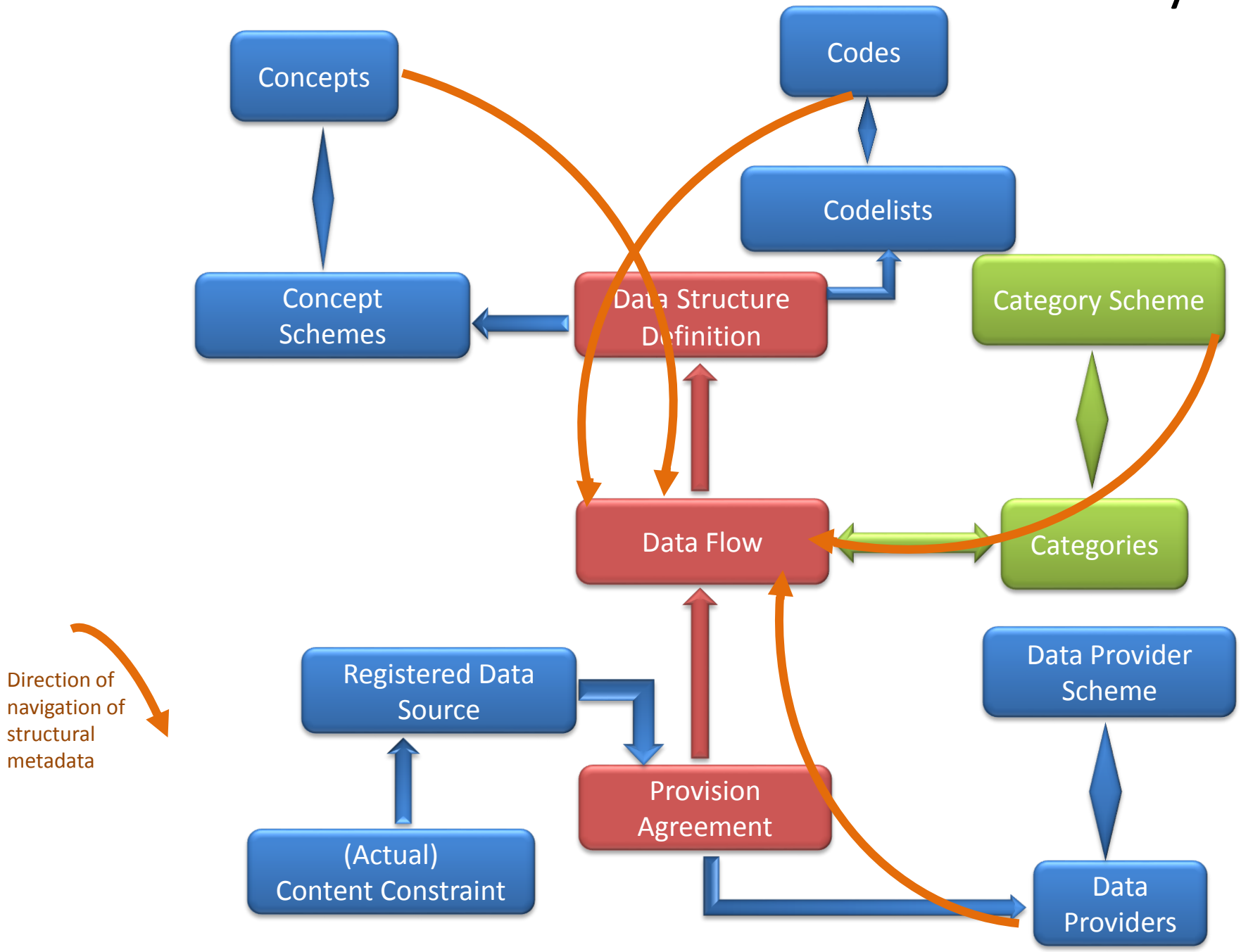
SDMX: re-using common concepts



SDMX Information Model: Schematic View



SDMX Information Model: Data Discovery



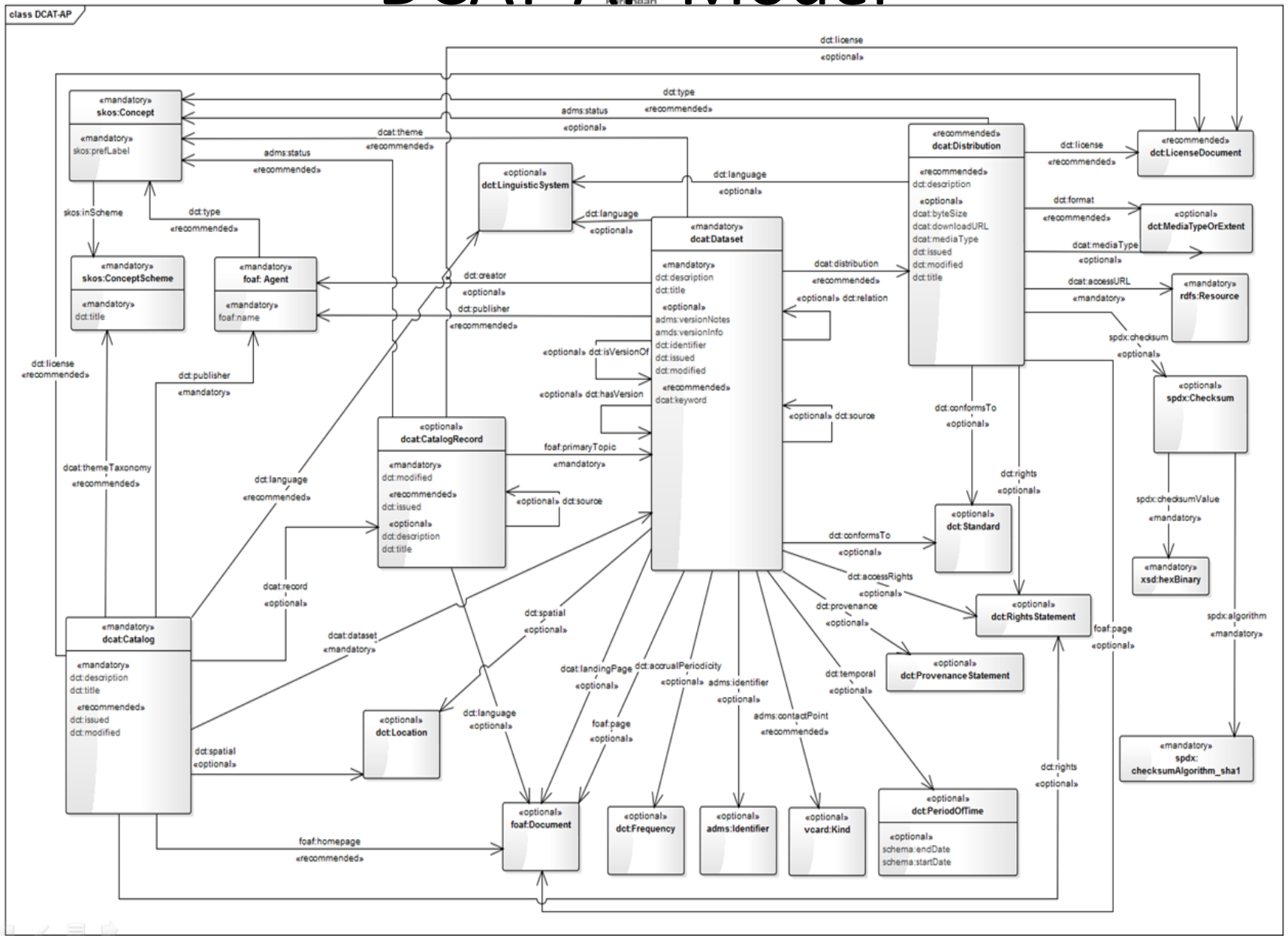
SDMX Dataflow

- Conceptually a Dataflow is a Dataset without the actual observation values
- It is the definition of data that can exist in terms of its:
  - structure
  - publishers
  - genre/topic
  - valid content
- If data exists then it is linked to
  - Data sources
  - Index of actual content
    - The data discovery system can retrieve this from the registry
- DCAT-AP Dataset contains much of the information available with the SDMX Dataflow

SDMX and DCAT-AP

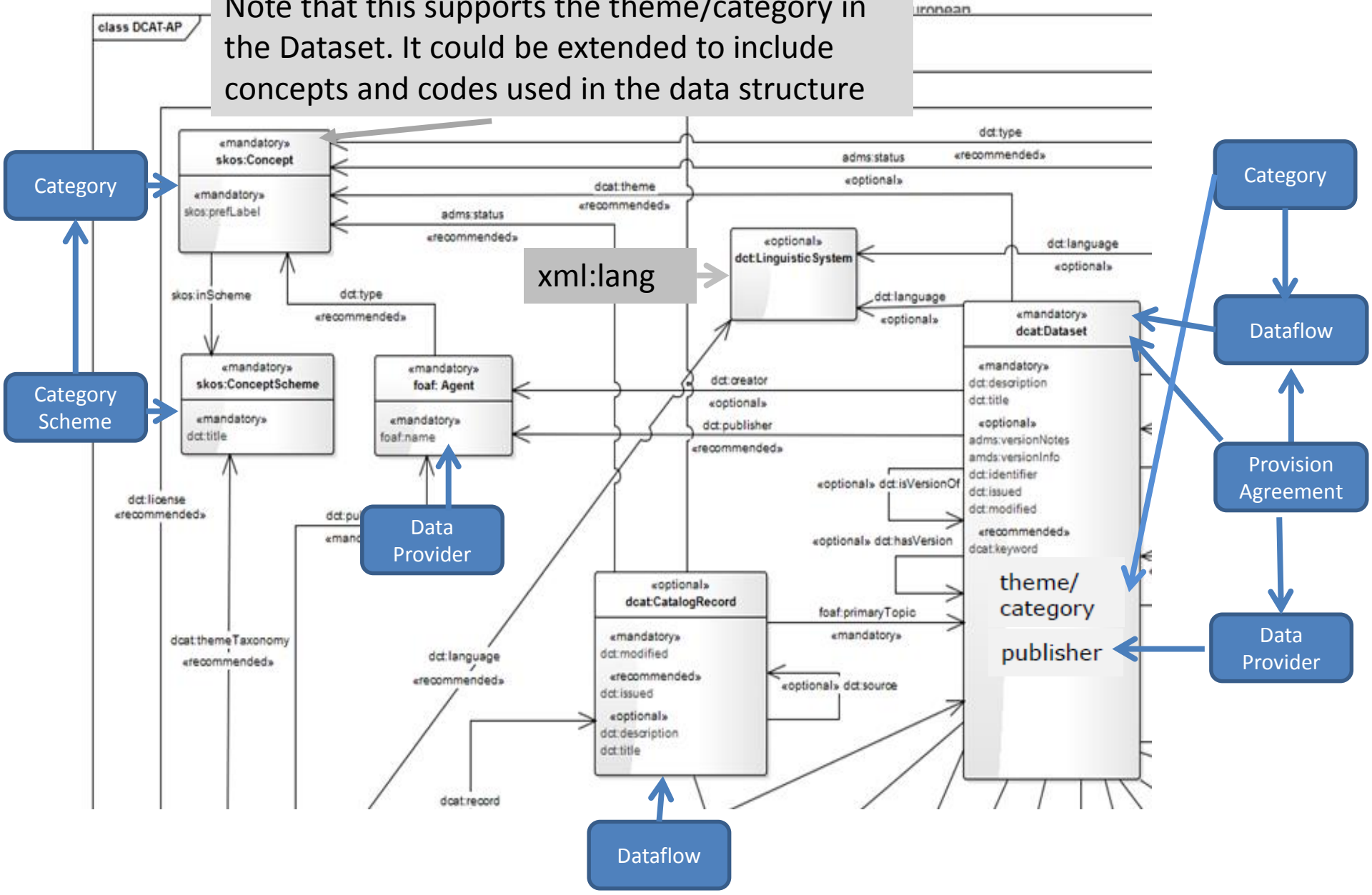
Mapping of SDMX Information Model Classes to DCAP-AP Model Classes

DCAT-AP Model

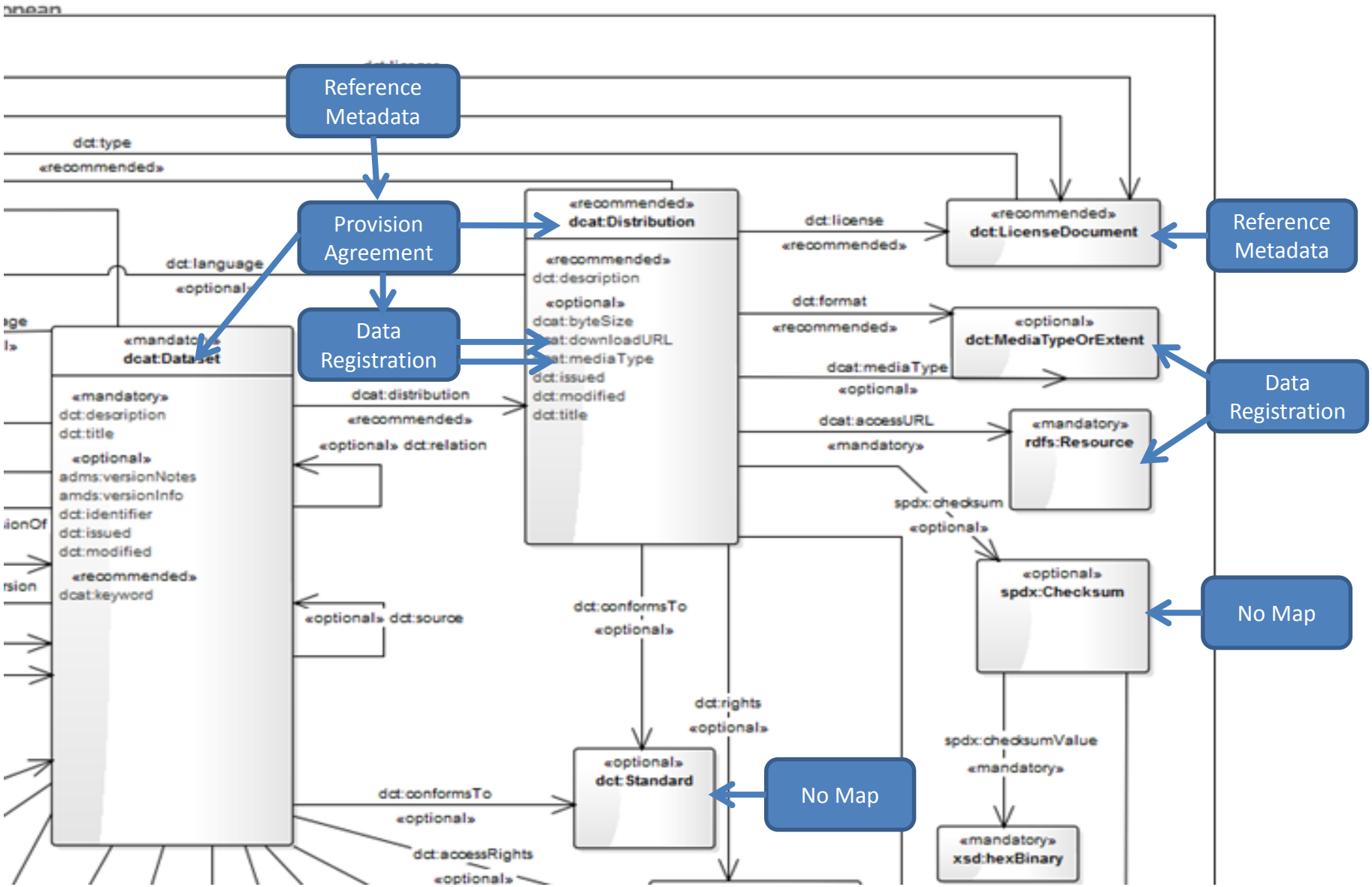


Map of SDMX to DCAT-AP

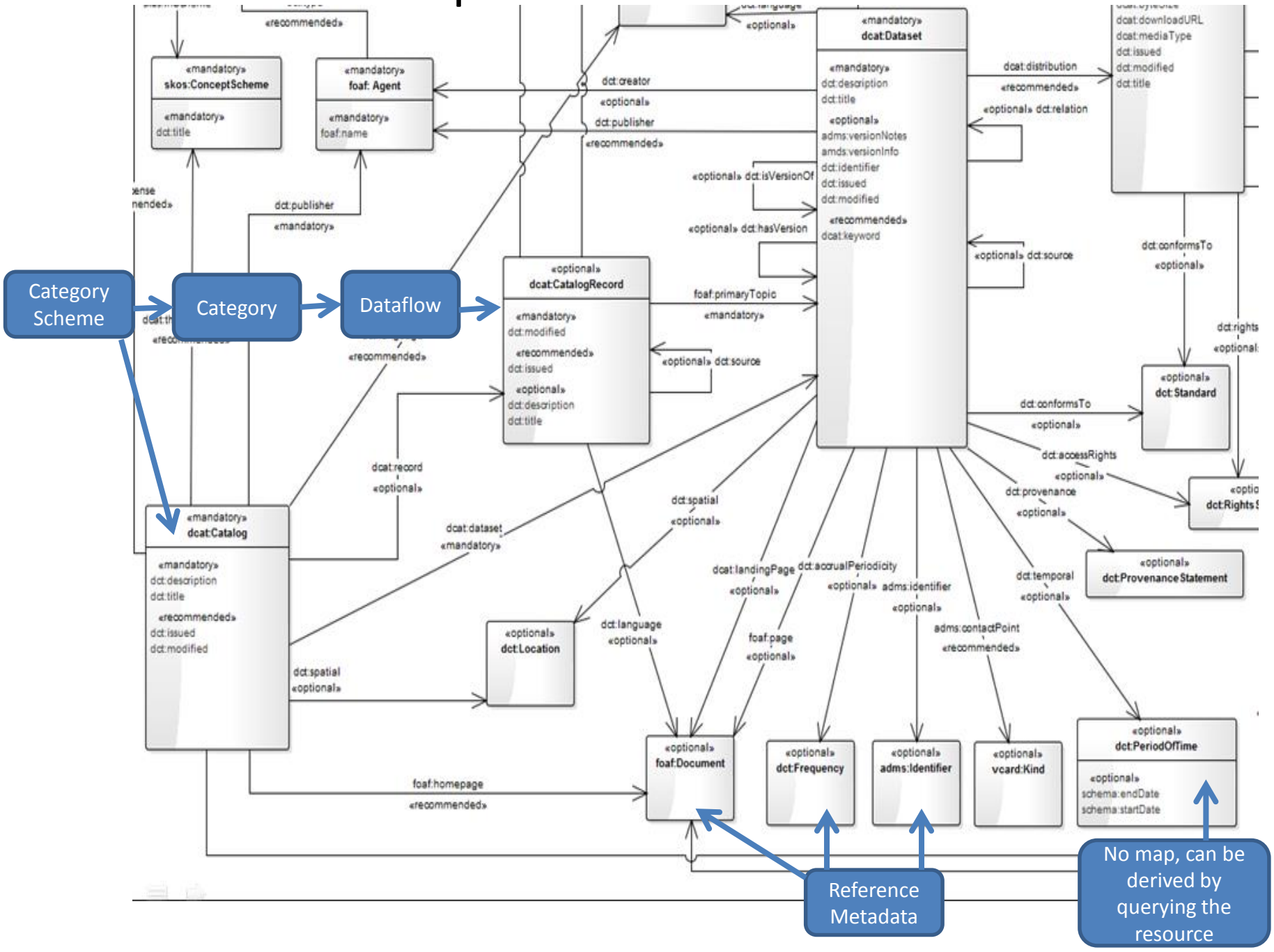
Note that this supports the theme/category in the Dataset. It could be extended to include concepts and codes used in the data structure



Map of SDMX to DCAT-AP

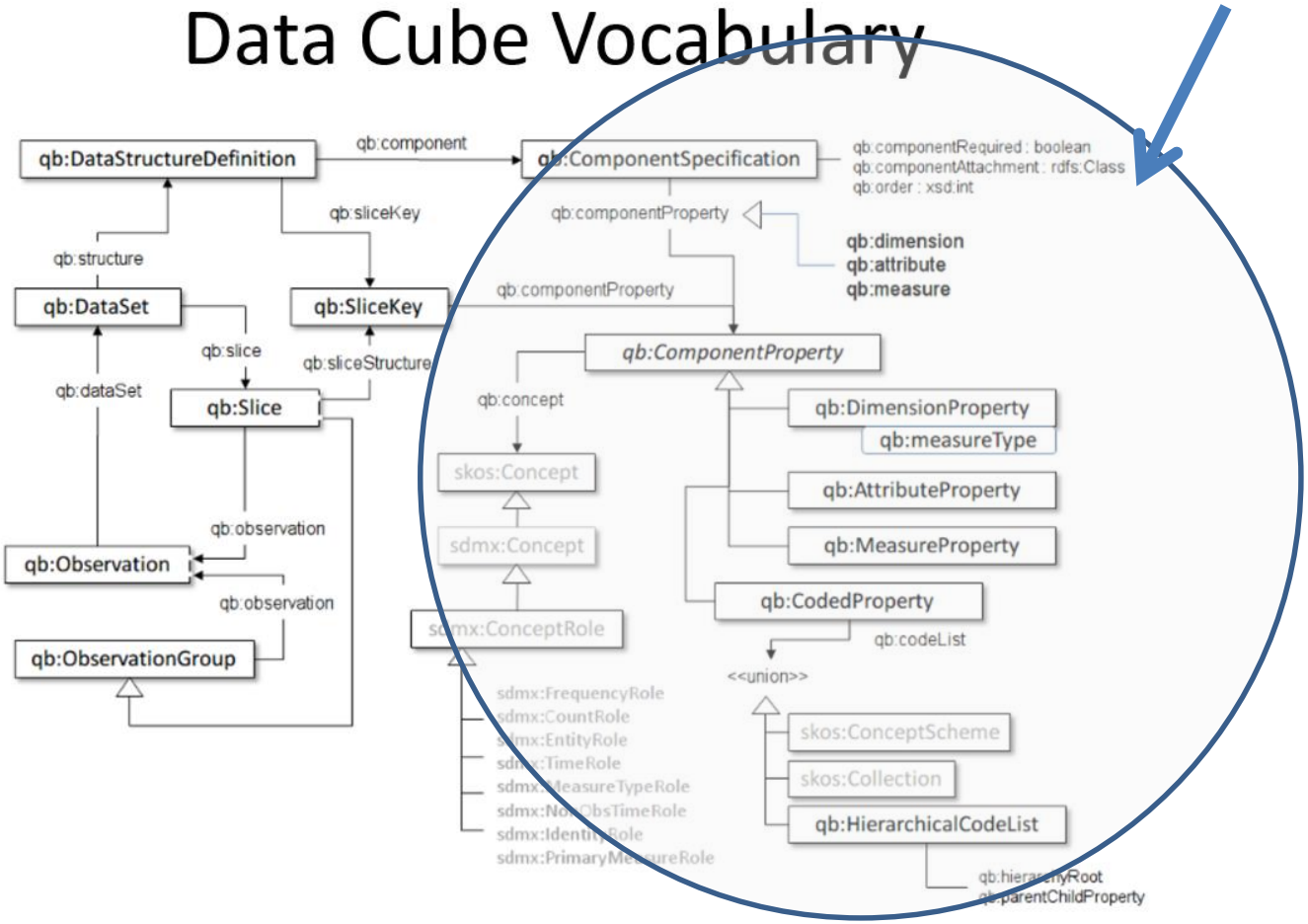


Map of SDMX to DCAT-AP



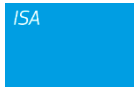
Data Structure Metadata used to Discover Data

Data discovery metadata



Your Questions


European
Commission


ISA

Next steps

Time plan

- **December 2015**: invitations to stakeholders, set up collaboration infrastructure
- **January 2016**: collect requirements and suggestions
- **5 February 2016**: Familiarisation Webinar

- **February 2016**: first draft based on initial analysis and issues raised
- **Early March 2016**: first virtual WG meeting to discuss first draft
- **March 2016**: second draft based on discussions and decision in the WG
- **Late March 2016**: second virtual WG meeting to discuss second draft
- **April 2016**: third draft, preparing for public review
- **Late April 2016**: third virtual WG meeting to discuss final draft for public review
- **May and June 2016**: public review period
- **July 2016**: fourth virtual WG meeting, discuss and resolve public comments received, publication of StatDCAT-AP

Next meetings

Proposed periods for the next meetings:

- Monthly meetings
- Doodle polls will be shared to schedule meetings:
    1. Early March
    2. Late March/early April
    3. Late April/early May
    4. Late June/early July
- Drafts of specifications to be shared ahead of meetings
- Discussions to take place through electronic means

Collaboration Infrastructure

StatDCAT-AP release page

- Joinup page where the implementation guidelines are published: https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/description

Mailing list

- Publicly archived mailing list on Joinup used for submitting change requests on the draft specification:
  stat_dcat_application_profile@joinup.ec.europa.eu

Issue tracker

- WG Members can log issues on the issue tracker or via the mailing list
- WG Members can comment on the issues that are already logged: https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/issue/all

European Commission

Save the date!

12 May 2016
Rome – Italy

Stay tuned at:
https://joinup.ec.europa.eu/node/148436

ISA

European Commission

joinup

Project Officers Vassilios.Peristeras@ec.europa.eu
Athanasios.Karalopoulos@ec.europa.eu

Visit our initiatives

ADMS ASSET DESCRIPTION METADATA SCHEMA

DCAT APPLICATION PROFILE FOR EUROPEAN DATA PORTALS

ADMS.SW

REGISTERED ORGANIZATION VOCABULARY

eGOVERNMENT CORE VOCABULARIES

CORE PUBLIC SERVICE VOCABULARY

CORE LOCATION VOCABULARY

CORE PERSON VOCABULARY

Get involved

Follow @SEMICeu on Twitter

Join the SEMIC group on LinkedIn

Join the SEMIC community on Joinup

ISA