



VTL and StatDCAT: two new standards interacting with the SDMX information model (overview)

Marco Pellegrino
Eurostat, Unit B.5
Data and Metadata Standards and Services

Two “new” standards?

VTL = Validation and Transformation Language
(building on the SDMX I.M. for transformation)

StatDCAT-AP = Application Profile of the Statistical variant of DCAT
(W3C Recommendation for the exchange of descriptions of datasets between open data portals)

The main VTL goals

- Define and preserve V&T rules
- Exchange and share V&T rules
- Apply V&T rules in automated processes

Taking care of
making VTL applicable to several standards
(e.g. SDMX, DDI, GSIM and possibly others)



A very challenging target!

Governance and Standards Alignment

VTL will be maintained by the SDMX TWG

- Task Force composed of members of the SDMX TWG and SWG (Statistical Working Group) and other experts involved in DDI, GSIM and SDMX design and evolution

Has already produced some feedback to GSIM for next version

- VTL can be mapped against SDMX
- VTL can be directly utilized by DDI in those places where computations are included
- VTL could be used in CSPA services where processing is performed
- As GSIM processing Rules

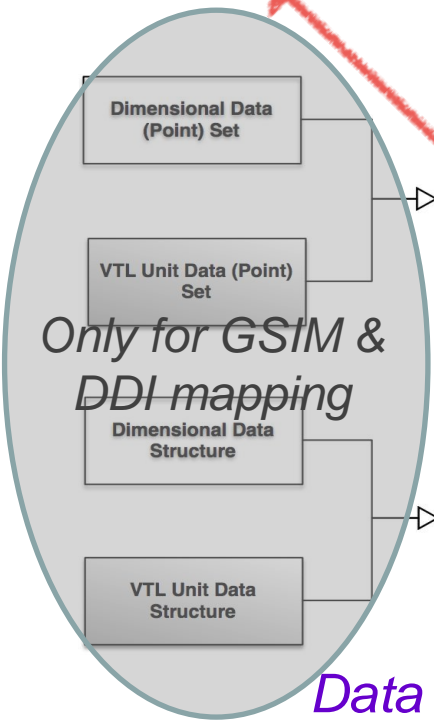
SDMX – VTL mapping

Dataflow Definition

Observation

Measure Dimension

Dimension other than Measure Dimension

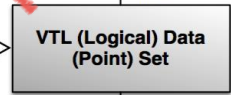


Only for GSIM & DDI mapping

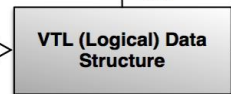
Data Structure Definition



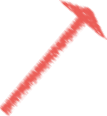
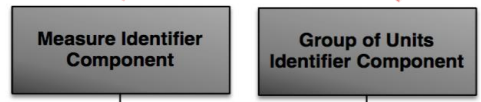
0..N
has ▲
1..1



0..N
structured by ▼
1..1

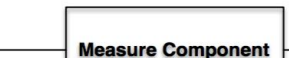


has



1..N

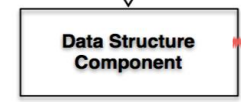
0..N



Dimension Component

Primary Measure

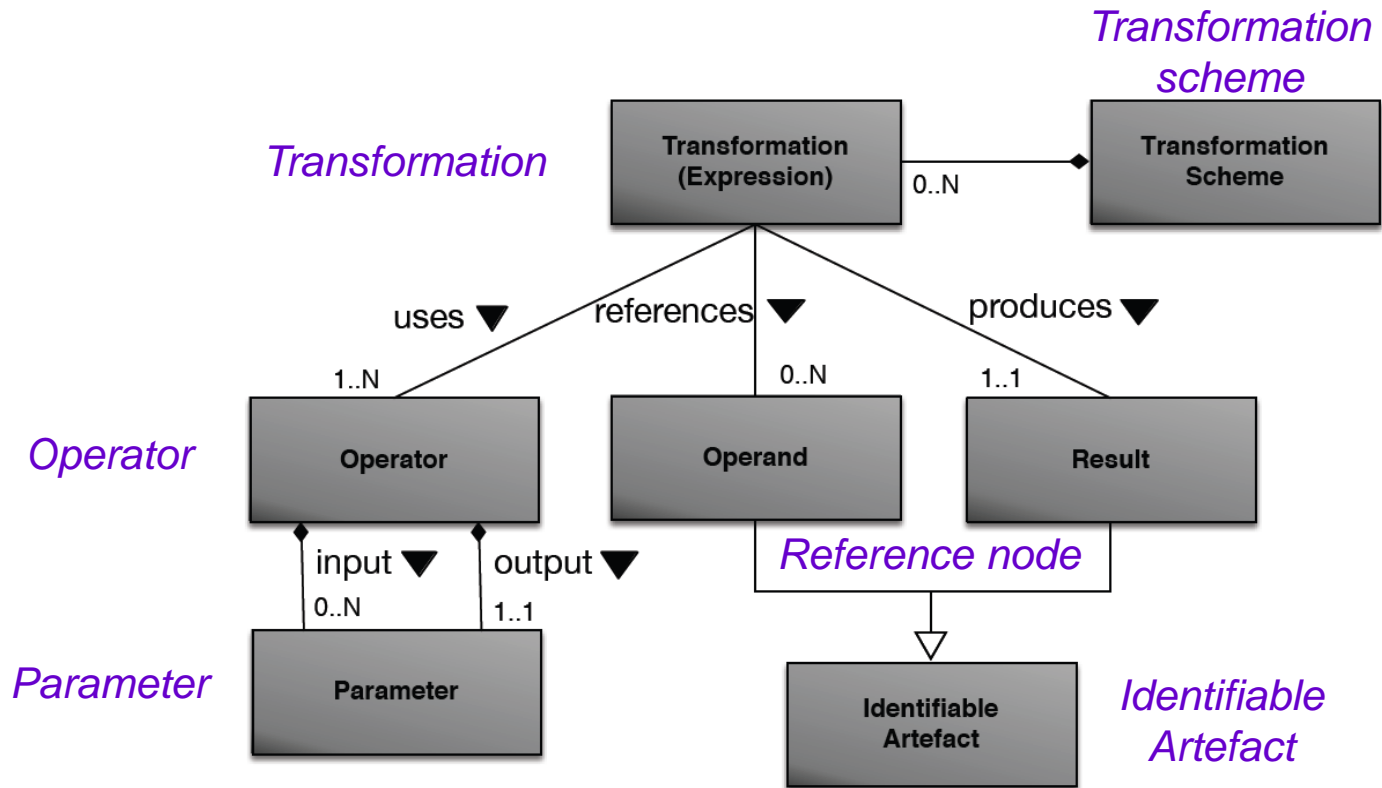
Data Attribute



Component

- White: same artefact as in GSIM 1.1
- Light grey: similar to GSIM 1.1
- Dark grey: additional detail (in respect to GSIM 1.1)

SDMX - VTL mapping (transformations)



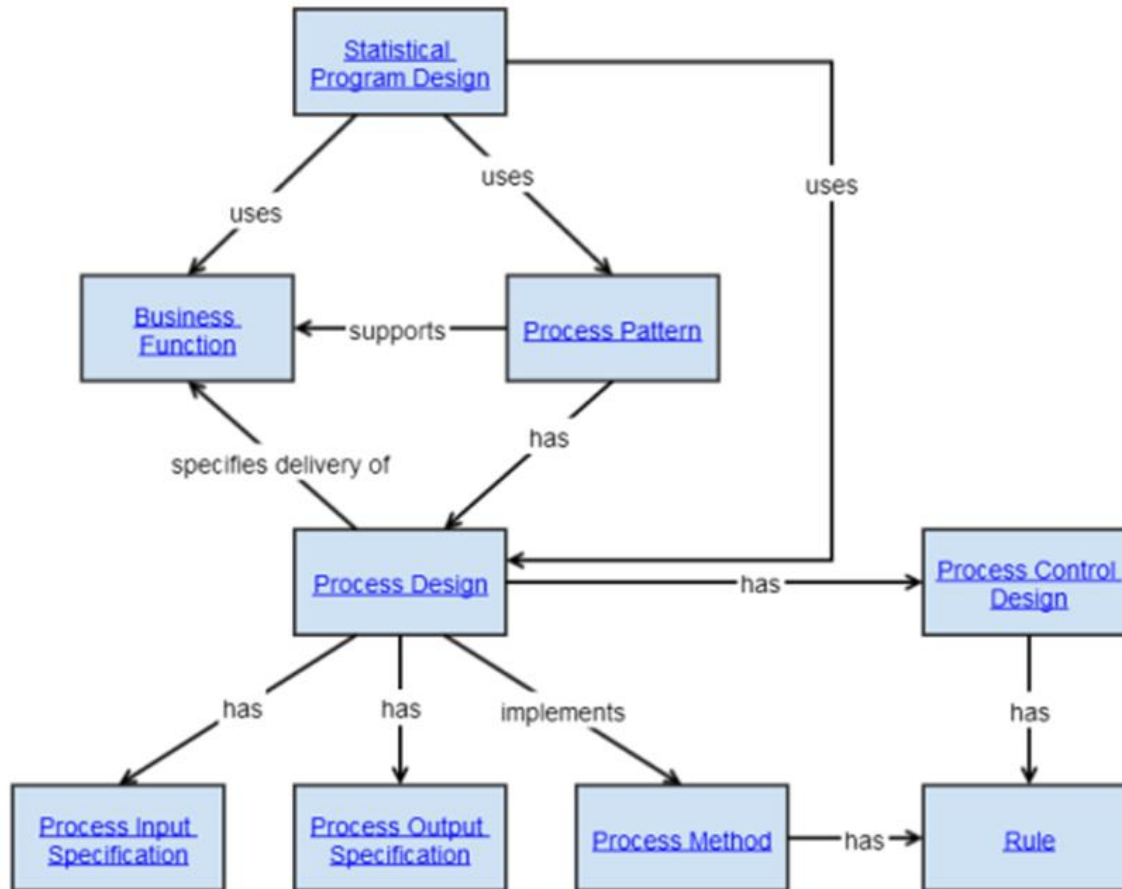
Transformation model

- It exists in SDMX, but not in GSIM and DDI
- It allows defining calculations through mathematical expressions
- It does not allow cycles (same structure than a spreadsheet)

Process model

- It exists in SDMX, GSIM, DDI and other standards (e.g. BPM)
- It allows defining calculations through a process
- It allow cycles (like a procedural programming language)

Process Method and Rules



- **VTL 1.0: published in March 2015**
(http://sdmx.org/?page_id=5096)
 - VTL part 1 (General description)
 - VTL part 2 (Library of Operators)
 - eBNF (Extended Backus-Naur Form) Technical notation
- **VTL 1.1: in progress**
 - More operators
 - Reusability of rules, language redesign
- **SDMX implementation: in progress**
 - Mapping of SDMX and VTL artefacts
 - Messages for exchanging VTL rules
 - Registry for storing VTL rules
 - Web services for retrieving VTL rules

VTL 1.1 public review

October 2016

VTL 1.1 (General part and Reference manual) will be published on the SDMX web site at <https://sdmx.org>

October to December

Public review

February 2017

Publication of the final version of VTL 1.1

Decision gates on the adoption of VTL as the standard validation language in the different constituencies

Comments and suggestions for improvement: twg@sdmx.org

StatDCAT-AP

A Common Layer for the Exchange of Statistical Metadata in Open Data Portals

Marco Pellegrino
Eurostat, Unit B.5
Data and Metadata Standards and Services

The challenge: data silos

- The data landscape consists of many data silos:
 - Statistical data, Geospatial data, Legal data, Research data, Archival data
 - Etc. etc.
- Many of these silos build portals harvesting information
 - <http://ec.europa.eu/eurostat/data/database>
 - <http://inspire-geoportal.ec.europa.eu>
 - <http://eur-lex.europa.eu>
 - <http://www.ecb.europa.eu/stats/html/index.en.html>
 - <http://stats.oecd.org>
 - <https://www.openaire.eu>
 - <https://www.archivesportaleurope.net>
 - <http://www.europeana.eu>
- Plus: These portals serve their goal for a specific audience
- Minus: No easy way to discover data across domains

The proposed solution

- Bringing together data from the multitude of domains in one '**general data portal**' to expose domain-specific data
- Using a cross-domain description standard that is able to capture a **core set** of characteristics of domain-specific data:
 - DCAT Application Profile for data portals in Europe**
- **Extension** of cross-domain standard for additional features of domain-specific data: **GeoDCAT-AP, StatDCAT-AP**
- **NB:** Local systems and domain-specific portals continue to use domain-specific standards: approach based on **export of metadata** according to cross-domain standard
- Creating a high-level index of domain-specific resources for the purpose of **discovery**

What is DCAT-AP

- Application Profile of the **DCAT W3C Recommendation** for the exchange of descriptions of datasets between (open) data portals
- **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 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
- Funded under ISA Action 1.1 of the ISA Programme on improving semantic interoperability in European e-Government systems

StatDCAT: scope of work

- *StatDCAT-AP: extension of DCAT-AP enabling cross-portal search for statistical data sets beyond the possibilities offered by the generic DCAT-AP.*
- *Extend DCAT-AP by adding:*
 - **Metadata elements from statistical standards (e.g. SDMX)**
 - **Recommendations for use of specific controlled vocabularies**
- *Focus on use cases:*
 - **Improving discovery of statistical data sets in open data portals**
 - **Facilitating integration of statistical data sets with open data from other domains**

The public review

- Final draft of specification is available on Joinup:

<https://joinup.ec.europa.eu/node/152858>

StatDCAT-AP - Draft 4

(⌕ ★★★★★) 5/5 | 1 votes |

Description

Fourth editor's draft of the StatDCAT-AP specification has been made available for public review and discussion.

This draft is available for **public review** until 23/10/2016. The following options exist:

- include your comments directly on this page; or
- create an issue using the **Issue tracker**; or
- contact us via the **public mailing list**: stat_dcat_application_profile@joinup.ec.europa.eu.

Themes

[eGovernment](#)

Distributions

[StatDCAT-AP - Draft 4](#)

↓ PDF

Detailed presentations



Using SDMX and VTL for performing structural and content validation

Marco Pellegrino
Eurostat, Unit B.5
Data and Metadata Standards and Services

September 2016

Background

Data validation, a critical issue for the E.S.S.

Eurostat and Member States: double work or "no work"?

Inefficiencies:

- Lack of coordination
- Lack of documentation
- Lack of formalisation of validation procedures and rules
- Low harmonisation of software solutions.

Need of a comprehensive solution: portfolio of actions

- SDMX evolution: originally focused on data collection and dissemination
- From 2011 on: Supporting other stages of the statistical production process



Validation & Transformation activities



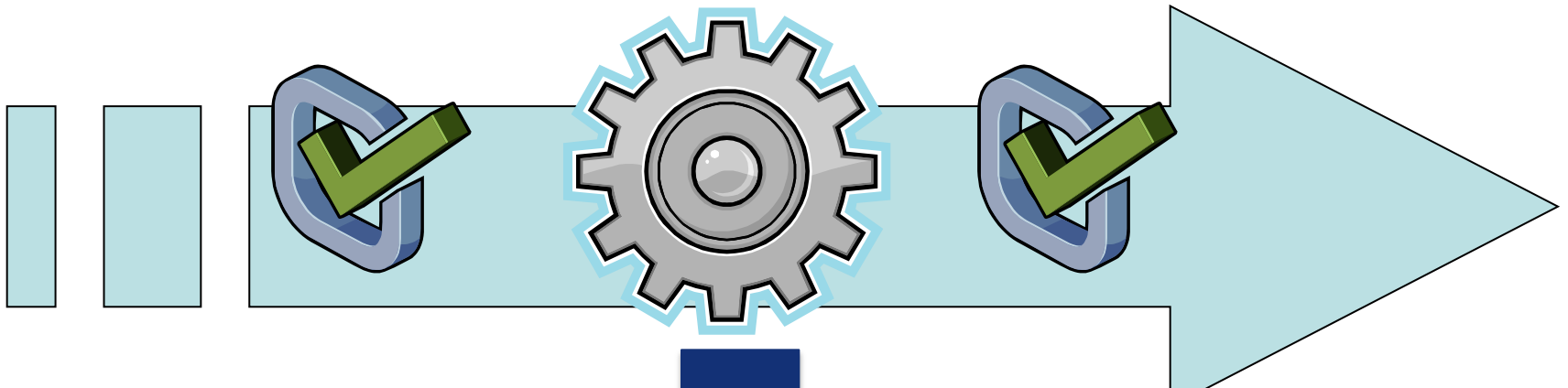
Data Validation Process

- *Before/During Transmission*
(*"First Level"*)
 - **Covered by SDMX today**

 - **Format Check (*SDMX-ML*)**
 - **Code Check (*SDMX DSD*)**
- *After Transmission*
(*"Second Level"*)
 - **Not **yet** covered by SDMX**
→ **SDMX-VTL**

 - **Detailed value check**
 - **Content check**

 - ...



The main VTL goals

- Define and preserve V&T rules
- Exchange and share V&T rules
- Apply V&T rules in automated processes

Taking care of
making VTL applicable to several standards
(e.g. SDMX, DDI, GSIM and possibly others)



A very challenging target!

Diagram of the Operators



Governance and Standards Alignment

VTL will be maintained by the SDMX TWG

- Task Force composed of members of the SDMX TWG and SWG (Statistical Working Group) and other experts involved in DDI, GSIM and SDMX design and evolution

Has already produced some feedback to GSIM for next version

- VTL can be mapped against SDMX
- VTL can be directly utilized by DDI in those places where computations are included
- VTL could be used in CSPA services where processing is performed
- As GSIM processing Rules

A language manipulates the artefacts of an IM
(IM = information model)

SDMX, DDI, GSIM have different IMs
a language for one of them wouldn't fit the others

→ ***a dedicated IM for VTL***

**designed to be very abstract and mappable to the IMs
of SDMX, DDI, GSIM (and possible others)**

Using VTL in SDMX, DDI, GSIM ...

by mapping their artefacts to the VTL artefacts

VTL Data Model

Organizes Data Points into Data Sets

Describes Data Structures using Structure Components

- **Measures**
- **Attributes**
- **Identifiers**

very similar to GSIM

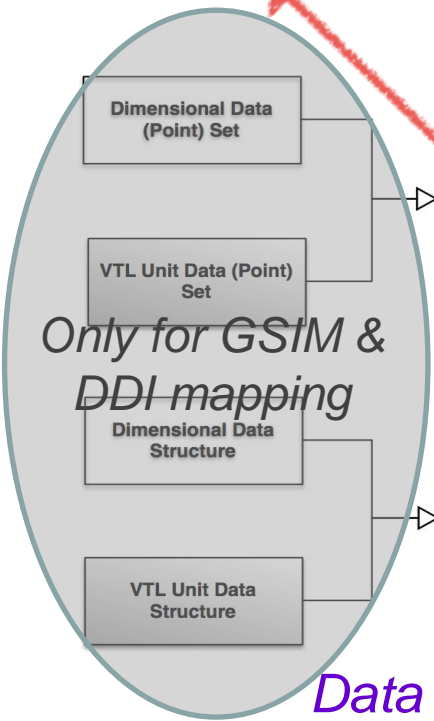
SDMX – VTL mapping

Dataflow Definition

Observation

Measure Dimension

Dimension other than Measure Dimension

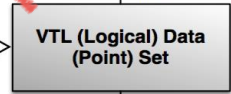


Only for GSIM & DDI mapping

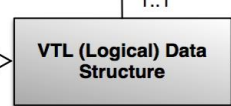
Data Structure Definition



0..N
has ▲
1..1



0..N
structured by ▼
1..1



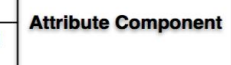
has



1..N



0..N

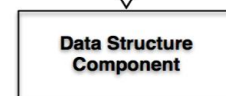


0..N

Dimension Component

Primary Measure

Data Attribute



Component

- White: same artefact as in GSIM 1.1
- Light grey: similar to GSIM 1.1
- Dark grey: additional detail (in respect to GSIM 1.1)

Transformation Model

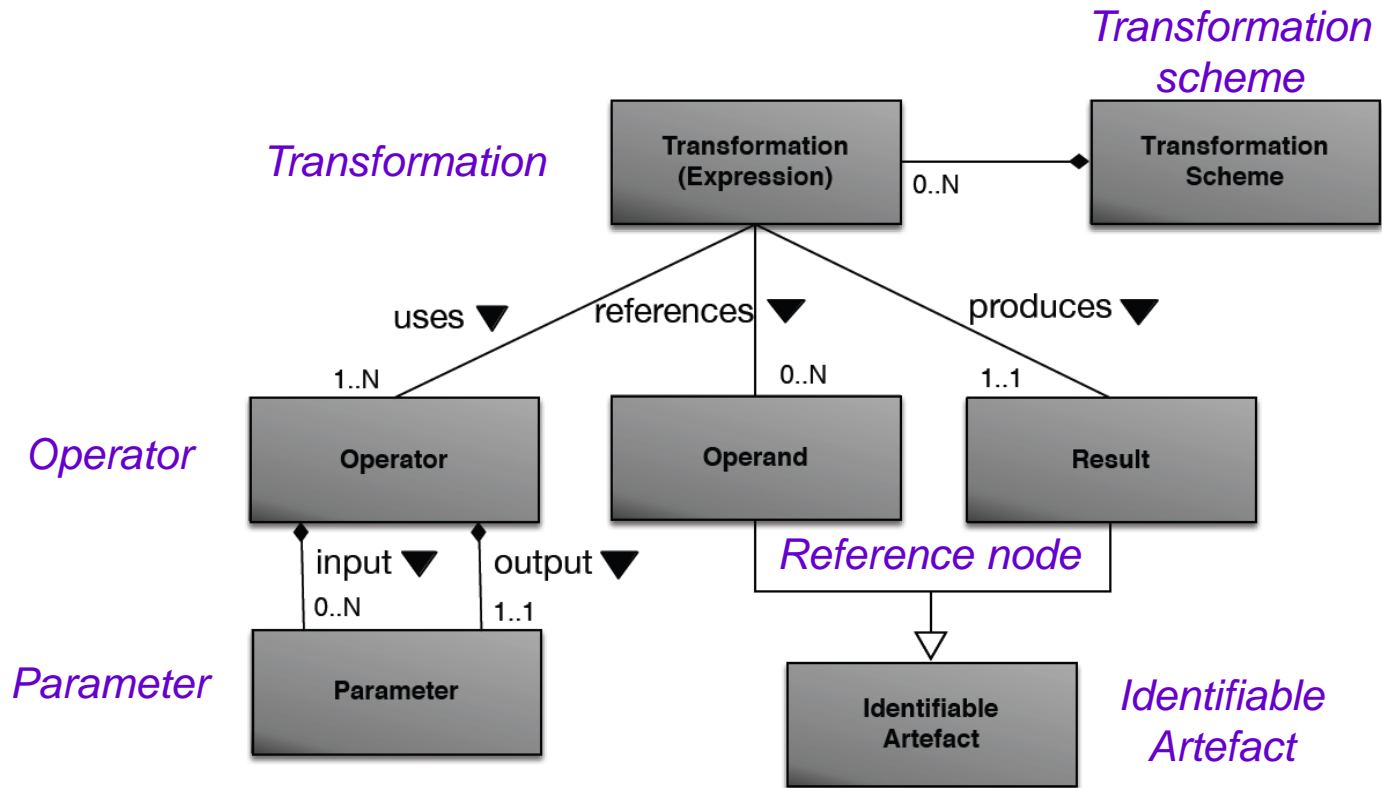
Takes a set of Transformation Expressions and organizes them into a Transformation Scheme

Each Expression has an Operand, and Operator, and a Result

- **Operands can have Parameters**
- **Operators and Results are identified by the Expression when it is executed**
- **VTL specifies the Operators and the types of Parameters**

VTL uses the SDMX Transformation model

SDMX - VTL mapping (transformations)



Transformation model

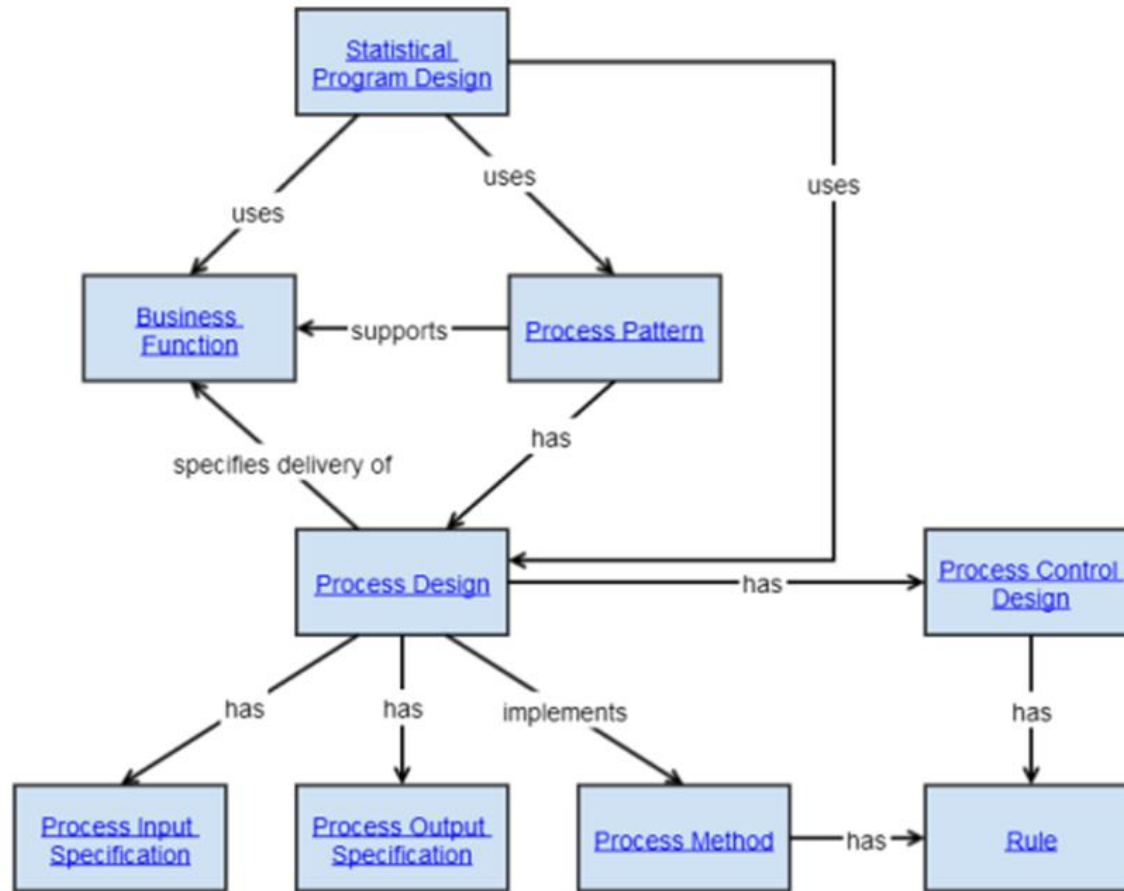
- It exists in SDMX, but not in GSIM and DDI
- It allows defining calculations through mathematical expressions
- It does not allow cycles (same structure than a spreadsheet)

Process model

- It exists in SDMX, GSIM, DDI and other standards (e.g. BPM)
- It allows defining calculations through a process
- It allow cycles (like a procedural programming language)



European
Commission



- **VTL 1.0: published in March 2015**
(http://sdmx.org/?page_id=5096)
 - VTL part 1 (General description)
 - VTL part 2 (Library of Operators)
 - eBNF (Extended Backus-Naur Form) Technical notation
- **VTL 1.1: in progress**
 - Language extensions
 - Reusability of rules, structural validation, ...
- **SDMX implementation: in progress**
 - Mapping of SDMX and VTL artefacts
 - Messages for exchanging VTL rules
 - Registry for storing VTL rules
 - Web services for retrieving VTL rules

VTL 1.0 Assessment - Results

Completeness: the language is complete (all rules proposed have been translated in VTL)

Correctness: Needs to eliminate some inconsistencies (union, keep operators)

Usability: needs to simplify some operators and introduce more statistical operators

Towards VTL 1.1

- Includes new operators, defining a set of "core" operators and a library of high-level operators
- Allows to create user functions
- Enhances the reusability of the VTL code
- SDMX specifications (e.g. for exchanging VTL rules in SDMX messages, for storing rules and for requesting validation rules from web services) in progress
- Implementation tests with some pilot domains, Integration within the ESS Validation Architecture

VTL 1.1 public review

<i>October 2016</i>	<i>VTL 1.1 (General part and Reference manual) will be published on the SDMX web site at https://sdmx.org</i>
<i>October to December</i>	<i>Public review</i>
<i>February 2017</i>	<i>Publication of the final version of VTL 1.1</i>
<i>2017</i>	<i>Decision gates on the adoption of VTL as the standard validation language in the different constituencies</i>

Contribute to VTL 1.1 !!!

**Comments and suggestions for improvement can be sent
to the SDMX Technical Working Group**

twg@sdmx.org

marco.pellegrino@ec.europa.eu



StatDCAT-AP

A Common Layer for the Exchange of Statistical Metadata in Open Data Portals

Marco Pellegrino
Eurostat, Unit B.5
Data and Metadata Standards and Services

The challenge: data silos

- The data landscape consists of many data silos:
 - Statistical data, Geospatial data, Legal data, Research data, Archival data
 - Etc. etc.
- Many of these silos build portals harvesting information
 - <http://ec.europa.eu/eurostat/data/database>
 - <http://inspire-geoportal.ec.europa.eu>
 - <http://eur-lex.europa.eu>
 - <http://www.ecb.europa.eu/stats/html/index.en.html>
 - <http://stats.oecd.org>
 - <https://www.openaire.eu>
 - <https://www.archivesportaleurope.net>
 - <http://www.europeana.eu>
- Plus: These portals serve their goal for a specific audience
- Minus: No easy way to discover data across domains

The proposed solution


- Bringing together data from the multitude of domains in one '**general data portal**' to expose domain-specific data
- Using a cross-domain description standard that is able to capture a **core set** of characteristics of domain-specific data:
 - DCAT Application Profile for data portals in Europe**
- **Extension** of cross-domain standard for additional features of domain-specific data: **GeoDCAT-AP, StatDCAT-AP**
- **NB:** Local systems and domain-specific portals continue to use domain-specific standards: approach based on **export of metadata** according to cross-domain standard
- Creating a high-level index of domain-specific resources for the purpose of **discovery**



European Commission

The European example


- European Data Portal
- Developed for European Commission DG CONNECT
- Harvesting metadata from national data portals

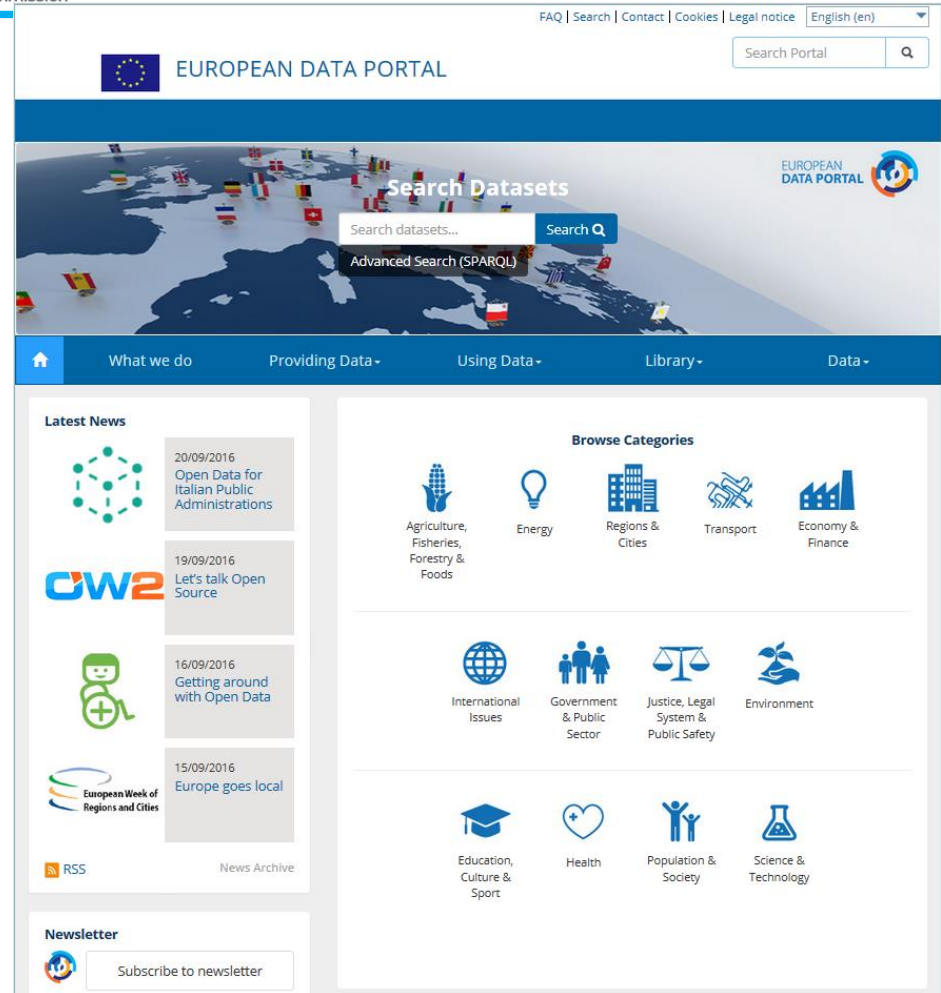
 Open Data Portal Poland
238 Datasets

 Danske Geoportal
Geo Data Portal Denmark
751 Datasets

 Data Directory - Ministry of the Interior
Geo Data Portal Greece
28 Datasets

 Data.gov.ie
Open Data Portal Ireland
3906 Datasets

 data.gov.ro
Open Data Portal Romania



The screenshot shows the European Data Portal homepage. At the top, there is a navigation bar with links for 'FAQ', 'Search', 'Contact', 'Cookies', and 'Legal notice', along with a language selector set to 'English (en)'. Below this is the 'EUROPEAN DATA PORTAL' header with a search bar. The main content area features a large map of Europe with various national flags, overlaid with a 'Search Datasets' search bar and a 'Search Q' button. Below the map is a navigation menu with categories: 'What we do', 'Providing Data', 'Using Data', 'Library', and 'Data'. The main content is divided into several sections: 'Latest News' with three articles (dated 20/09/2016, 19/09/2016, and 16/09/2016), 'Browse Categories' with icons for Agriculture, Energy, Regions & Cities, Transport, Economy & Finance, International Issues, Government & Public Sector, Justice, Legal System & Public Safety, Environment, Education, Culture & Sport, Health, Population & Society, and Science & Technology. At the bottom, there is an 'RSS' feed icon and a 'Newsletter' subscription form.

<http://www.europeandataportal.eu/>

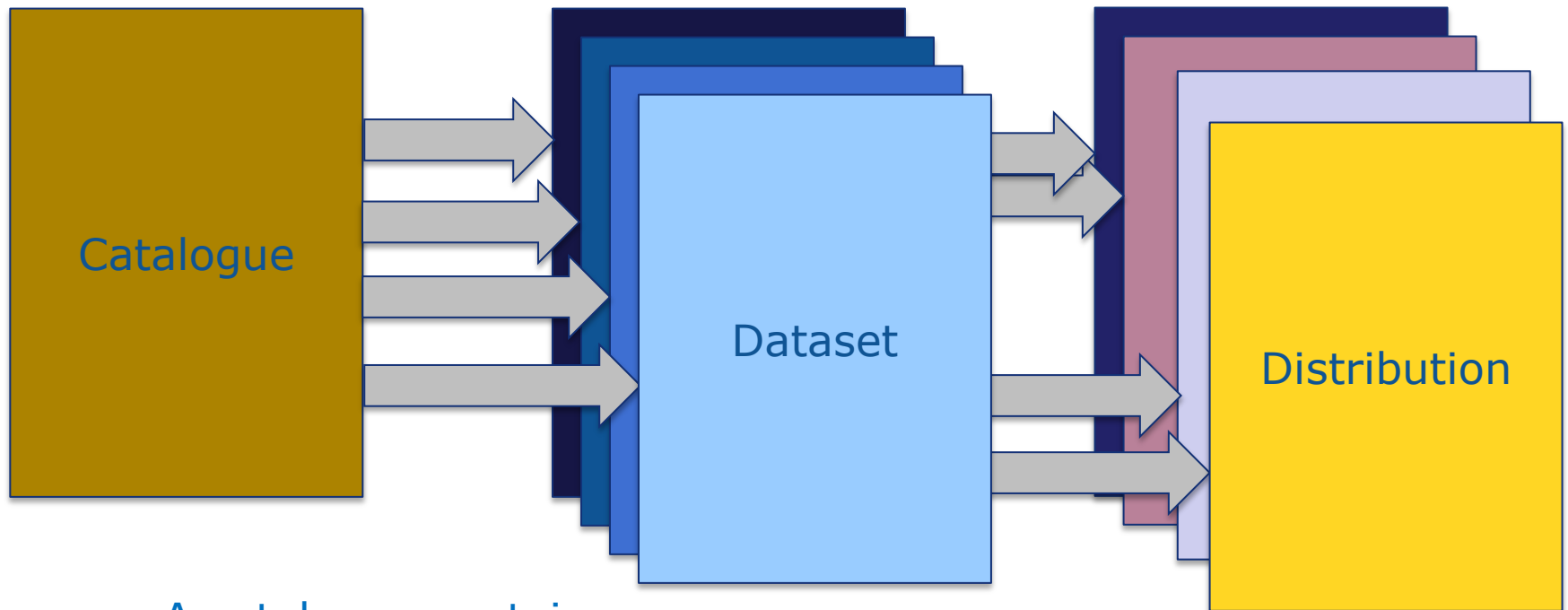
What is DCAT-AP

- Application Profile of the **DCAT W3C Recommendation** for the exchange of descriptions of datasets between (open) data portals
- **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 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
- Funded under ISA Action 1.1 of the ISA Programme on improving semantic interoperability in European e-Government systems

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)
- Higher quality metadata improves discoverability of datasets (bringing recommendations from SDMX to DCAT-AP on how to transfer metadata to a broader audience)

DCAT model overview



A catalogue contains
one or more datasets

A dataset has one or
more distributions

18-year-olds in education

Share

Publisher

Eurostat

Dct:publisher

Licence:

Legal Notice

Description

This indicator gives the percentage of all 18-year-olds who are still in any kind of school (all ISCED levels). It gives an indication of the number of young people who have not abandoned their efforts to improve their skills through initial education and it includes both those who had a regular education career without any delays as well as those who are continuing even if they had to repeat some steps in the past.

Dct:description

Keywords

education
student

Dcat:keyword

EuroVoc concepts

education
vocational training

Dcat:keyword

EuroVoc domains

Education and communications, Employment and working conditions

Dcat:theme

Resources

- DOWNLOAD Download dataset in SDMX-ML format ZIP
- DOWNLOAD Download dataset in TSV format ZIP

Dcat:distribution

Catalogue record

Added to open-data.europa.eu
2014-12-22
Updated on open-data.europa.eu
2015-05-18
Views: 486

Documentation

- VISIT PAGE ESMS metadata (Euro-SDMX Metadata structure) HTML
- DOWNLOAD ESMS metadata (Euro-SDMX Metadata structure) SDMX
- DOWNLOAD More information on Eurostat Website

Metadata URI

Suggest a dataset

Is there a dataset from the EU that you could not find in this portal?
[Please request the dataset >>](#)

URI

<http://ec.europa.eu/eurostat/web/products-datasets/-/tps00060>

Status

Completed

Dct:identifier

Identifier

tps00060

Type of Dataset

Statistical

Dct:modified

Modified Date

2015-04-22

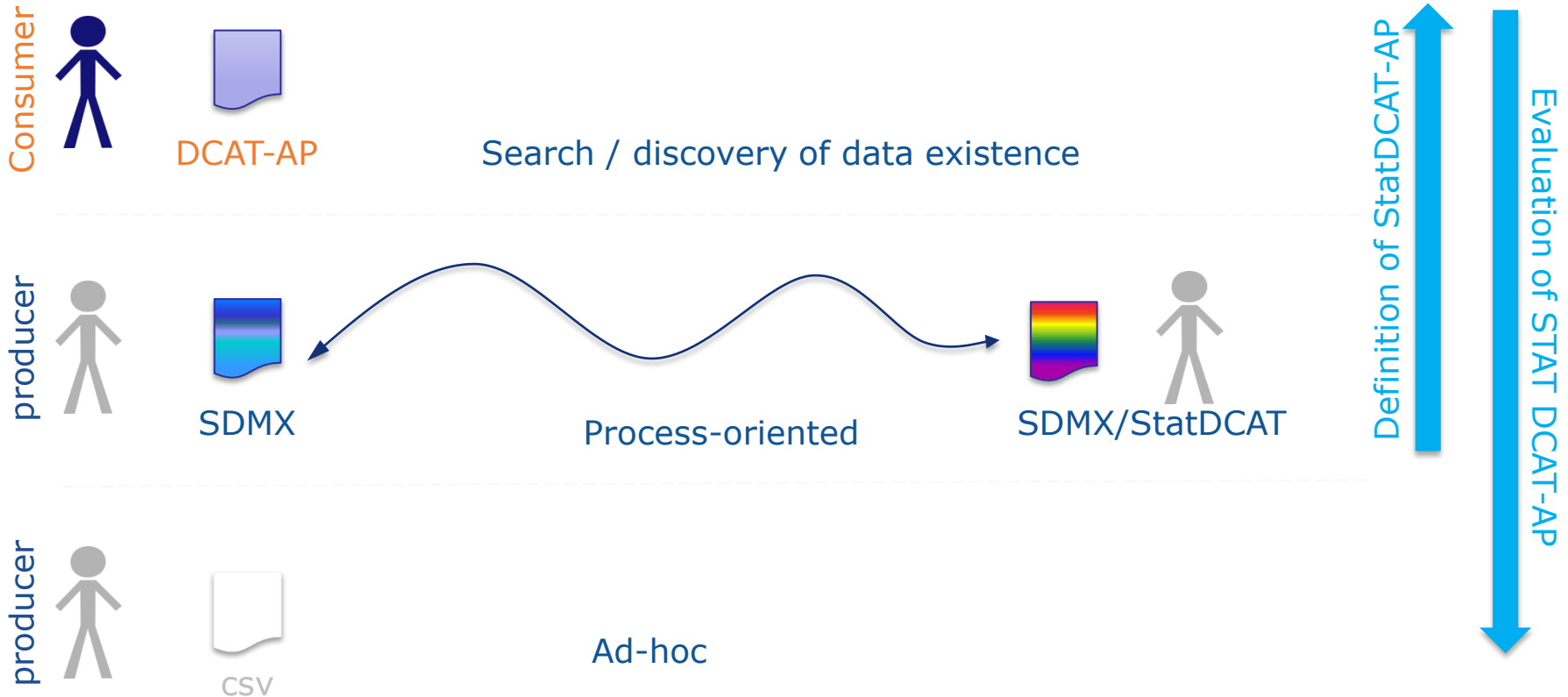
Temporal Coverage From

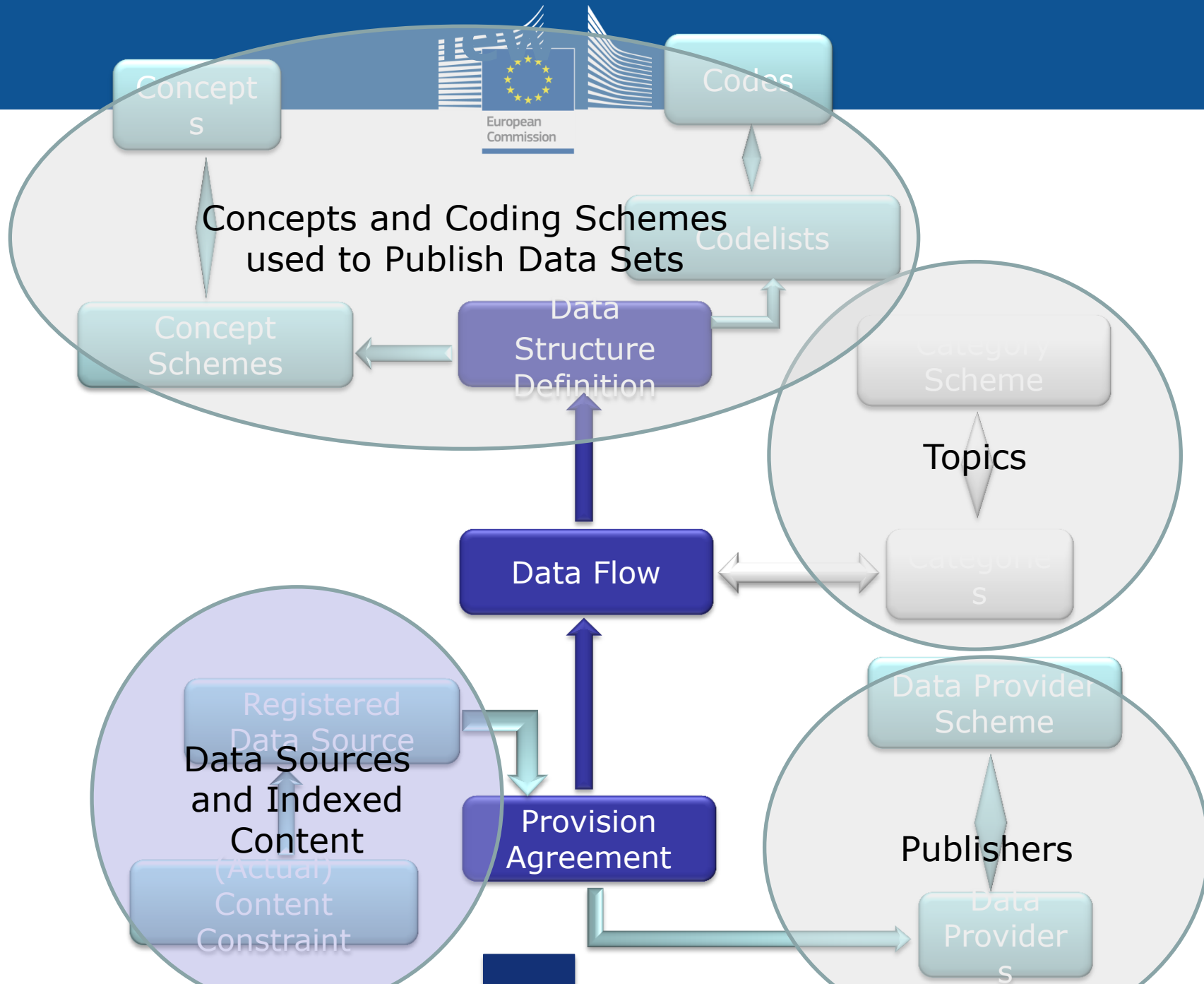
Dct:temporal

StatDCAT: scope of work

- *StatDCAT-AP: extension of DCAT-AP enabling cross-portal search for statistical data sets beyond the possibilities offered by the generic DCAT-AP.*
- *Extend DCAT-AP by adding:*
 - **Metadata elements from statistical standards (e.g. SDMX)**
 - **Recommendations for use of specific controlled vocabularies**
- *Focus on use cases:*
 - **Improving discovery of statistical data sets in open data portals**
 - **Facilitating integration of statistical data sets with open data from other domains**

Use case: StatDCAT-AP 'users'





StatDCAT-AP approach via SDMX



- The RDF Data Cube Vocabulary is based on SDMX
- The SDMX data structure definition (DSD) defines the structure of a data cube
- Data are machine-processable (see web services)
- The DSD dimensions and attributes can feature in DCAT-AP
 - Challenge: publicly published dimensions as Linked Data required



StatDCAT Work Group governance

- *Chair: Eurostat*
- *Co-chair: Publications Office*
(They represent 'owners' of the work, chair meetings and Webinars, take decisions, oversee the work of the operational team)
- *Observers:*
 - *DIGIT ISA*
 - *DG CONNECT**(Representing other stakeholders, providing advice and support)*

Stakeholders

(StatDCAT-AP and DCAT Working Groups, representatives of NSIs, int. agencies, experts in the domain of publishing statistical data, representatives of consumers such as Digital Agenda Scoreboard, EEA, representatives of the European Data Portal)

The public review

- Final draft of specification is available on Joinup:

<https://joinup.ec.europa.eu/node/152858>

StatDCAT-AP - Draft 4

(⌄ ★★★★★) 5/5 | 1 votes |

Description

Fourth editor's draft of the StatDCAT-AP specification has been made available for public review and discussion.

This draft is available for **public review** until 23/10/2016. The following options exist:

- include your comments directly on this page; or
- create an issue using the **Issue tracker**; or
- contact us via the **public mailing list**: stat_dcat_application_profile@joinup.ec.europa.eu.

Themes

[eGovernment](#)

Distributions

[StatDCAT-AP - Draft 4](#)

↓ PDF

Feedback from Public Review

POSSIBLE FUTURE FEATURES

- refinement of attributes
- data quality vocabulary (e.g. SIMS attributes)
- ...



Get involved

Joinup:

https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/description

Visit ISA initiatives

ADMS ASSET DESCRIPTION METADATA SCHEMA	StatDCAT-AP FOR STATISTICAL DATASETS	GeoDCAT-AP FOR GEOSPATIAL DATASETS	DCAT-AP FOR DATA PORTALS IN EUROPE	CORE PUBLIC ORGANISATION VOCABULARY
CORE PERSON VOCABULARY	REGISTERED ORGANISATION VOCABULARY	CORE CRITERION & EVIDENCE VOCABULARY	CORE LOCATION VOCABULARY	CORE PUBLIC SERVICE VOCABULARY

marco.pellegrino@ec.europa.eu