StatDCAT-AP

Face-to-face and virtual meeting 3

13 May 2016

ISA Programme Action 1.1

Opening, agenda, tour de table

ISA

Agenda

1. Opening, agenda, tour de table
2. Objectives of the meeting
3. StatDCAT-AP overall characteristics
4. Proposed extensions
5. SDMX-based transformation mechanism
6. Next steps

Tour de table

Objectives of the meeting

Intended outcome

- Discuss, agree potential extensions
  - Number of observations, number of data series, link to visualisation, dimensions as property, dimension as keywords, quality aspects, statistical unit, statistical population
- Discuss, agree SDMX-based transformation
  - SDMX Structural Metadata, SDMX Metadata Set
- Decide on content of specification

StatDCAT-AP overall characteristics

DCAT-AP for statistics

- Fully conformant extension of DCAT-AP
- General data portals will understand the ‘core’ of StatDCAT-AP (which is DCAT-AP)
- General data portals get opportunity to enhance services by processing the additional properties in StatDCAT-AP; after all, statistical datasets are an important collection for portals

Proposed extensions

Number of observations

- Total number of values contained in the Dataset
- Gives ‘logical size’ of the content of the dataset, as opposed to the ‘physical size’ of the data file in dcat:byteSize
- Possible RDF vocabulary term: dct:extent with normalised text, e.g.

```
:Dataset-001 dct:extent “20 observations”
```

Number of data series

- Total number of series contained in the Dataset
  - E.g. Dataset with data broken down by region (3 regions) sex (2 sexes) and age group (6 age groups) with observation values for 4 time periods has 144 observations in 36 series
- Additional 'logical size' parameter
- Possible RDF vocabulary term: dct:extent with normalised text, e.g.

```
:Dataset-001 dct:extent "36 series"
```

Link to visualisation

- Provides a link to a page where the data can be seen in a graphical or tabular representation
- Expected value: URL that opens the visualisation for the Dataset
- Does this information exist for many datasets?
- Do we know of existing RDF vocabulary terms?

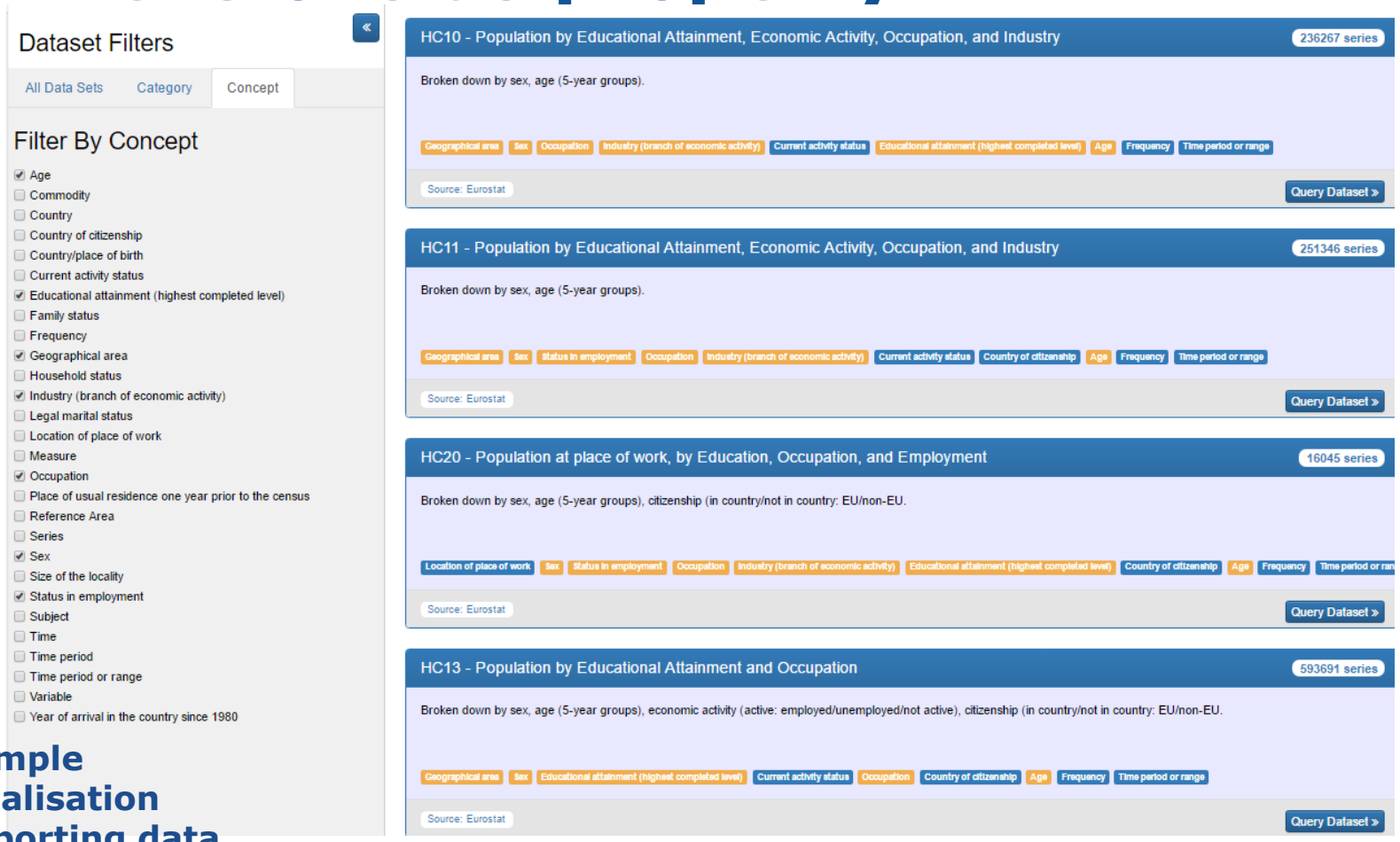
Dimensions as property

- Exposes dimensions in Dataset in structured way
  - E.g. time periods, regions, sex, income etc.
- Expected value: URI of qb:DimensionProperty
  - E.g. sdmx-dimension:timePeriod, sdmx-dimension:age
- Possible RDF vocabulary term: qb:dimension

```
:Dataset-001 qb:dimension sdmx-dimension:sex
```

European Commission

Dimensions as property



Example visualisation supporting data discovery

Dimensions as keywords

- Exposes dimensions in Dataset in text
- Provides simple mechanism to use existing DCAT property
- Expected value taken from label of the corresponding Dimension Property

```
:Dataset-001 dcat:keyword "Sex"@en
```

Quality aspects

- Possible use of W3C Data Quality Vocabulary (DQV, under development)
- Further quality characteristics from Euro-SDMX Metadata Structure (ESMS)?
- Use case: discovery or presentation?
- Text annotation or structured information?
- Existing RDF vocabulary terms?

Statistical unit

- ESMS concept STAT_UNIT:
  - Defined as "*entity for which information is sought and for which statistics are ultimately compiled*"
  - Usage note: "*list the basic units of statistical observation for which data are provided. These observation units (e.g. the enterprise, the local unit, private households,...) can be different from the reporting units used in the underlying statistical surveys*"
- Use case: discovery or presentation?
- Text annotation or structured information?
- Existing RDF vocabulary terms?

Statistical population

- ESMS concept STAT-POP:
  - Defined as: "*total membership or population or "universe" of a defined class of people, objects or events*"
  - Usage note: "*describe the target statistical population (one or more) which the data set refers to, i.e. the population about which information is to be sought*"
- Use case: discovery or presentation?
- Text annotation or structured information?
- Existing RDF vocabulary terms?

SDMX-based transformation mechanism

ISA

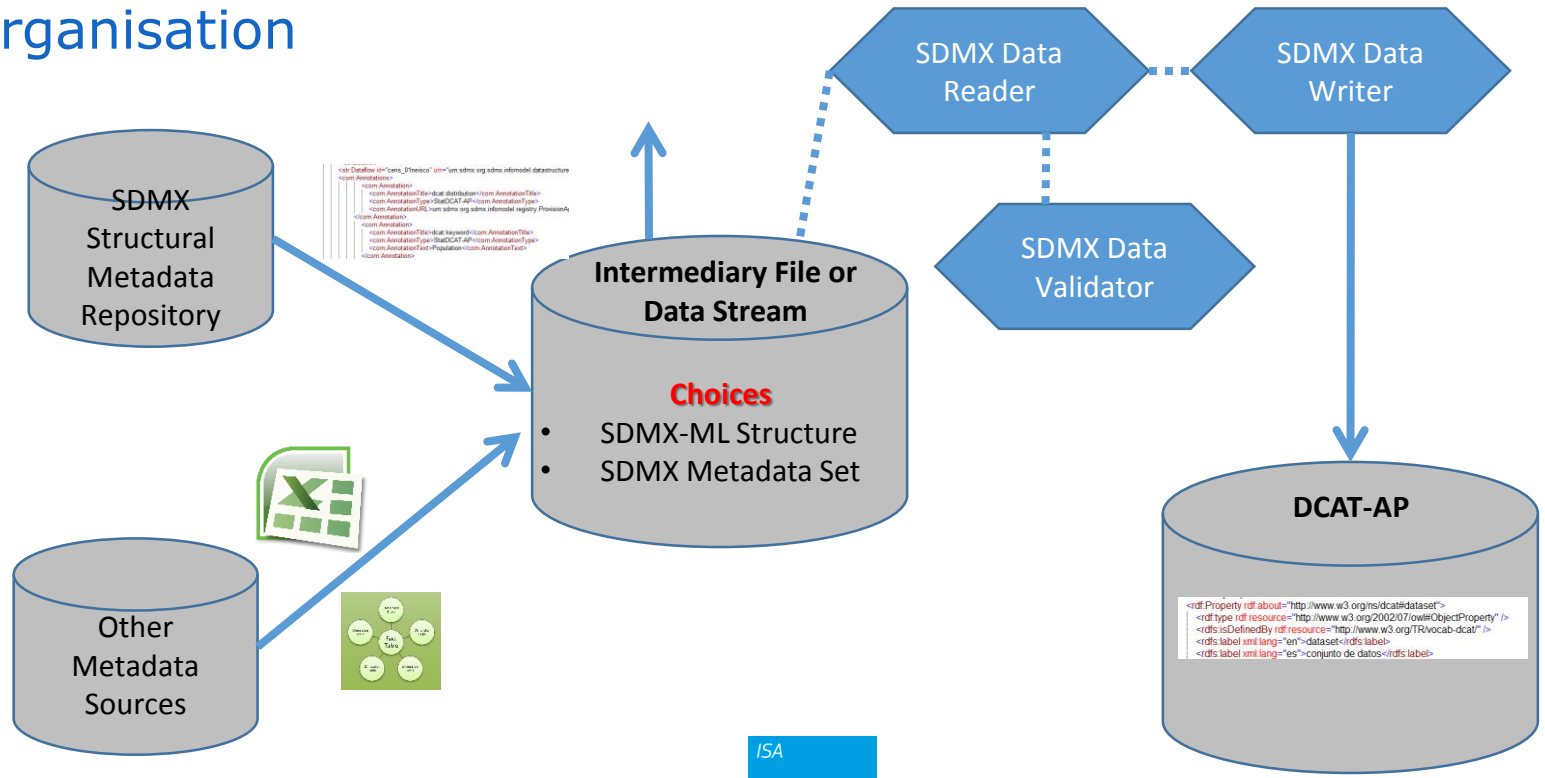
Preamble

- Organisations are free to choose how to create DCAT-AP from their systems
  - For SDMX users the specification defines a mapping between SDMX-ML structural metadata and DCAT-AP
  - For those not wishing to use SDMX, the organisation must make its own map between the metadata in its system and DCAT-AP
- If an organisation prefers to use a tool to create DCAT-AP then the development of two tools are under consideration
  - SDMX structural metadata to DCAT-AP
  - SDMX metadata set to DCAT-AP

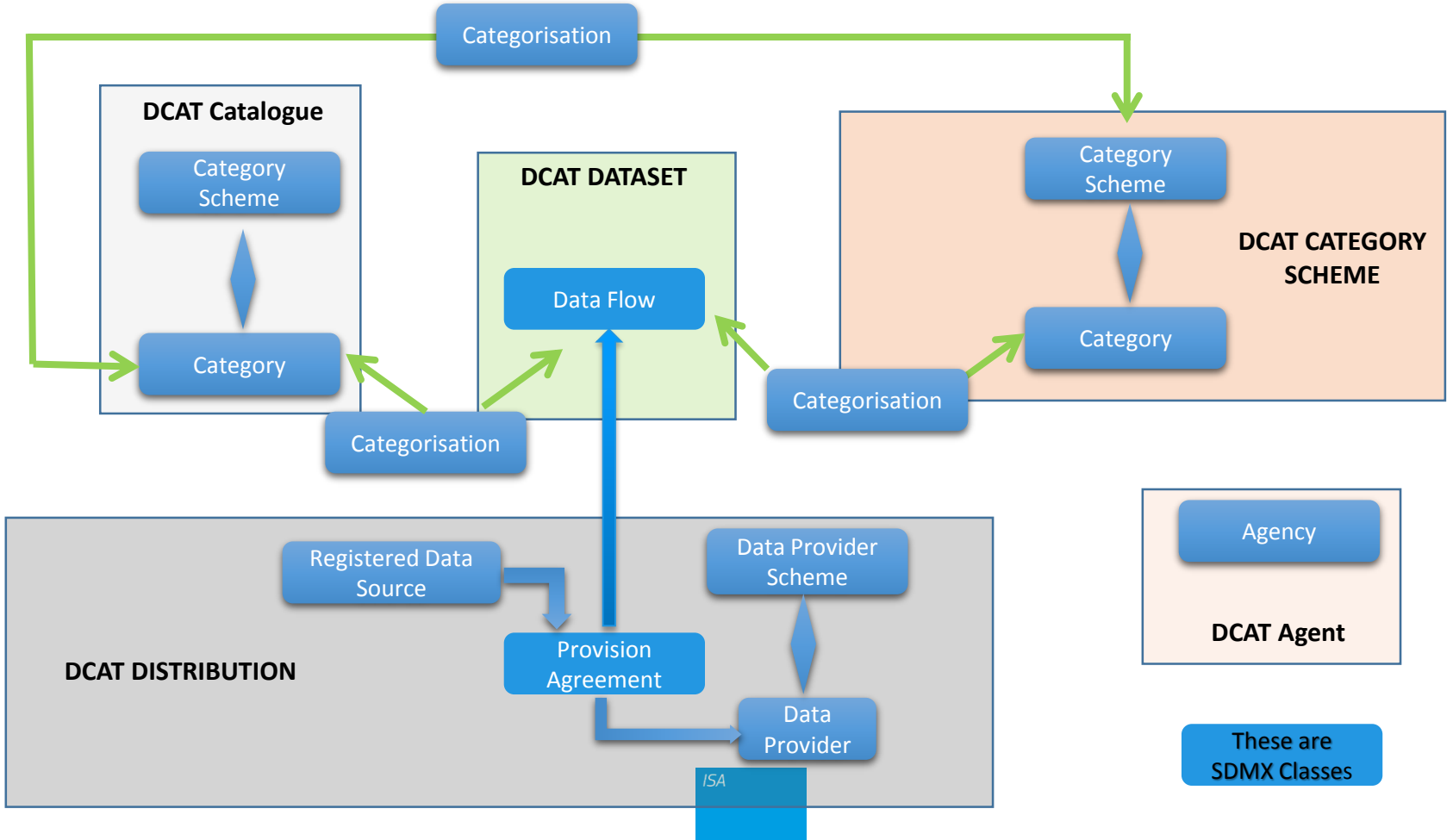
DCAT-AP Transformation Mechanism

These components can be developed in Java and .NET and integrated into SDMX systems or used in SDMX conversion tools

Data Publisher Organisation

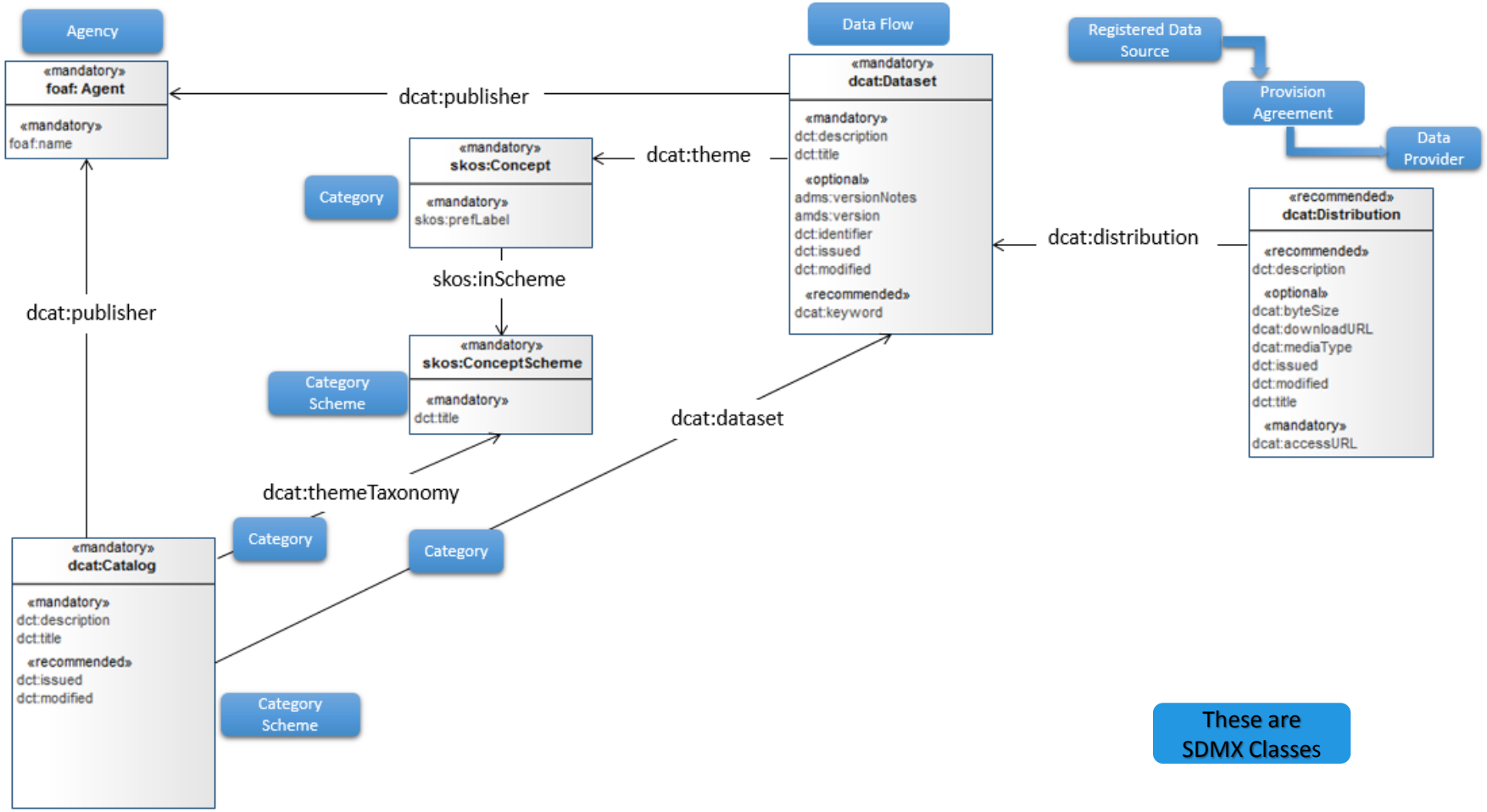


SDMX Structural Metadata mapping to DCAT-AP



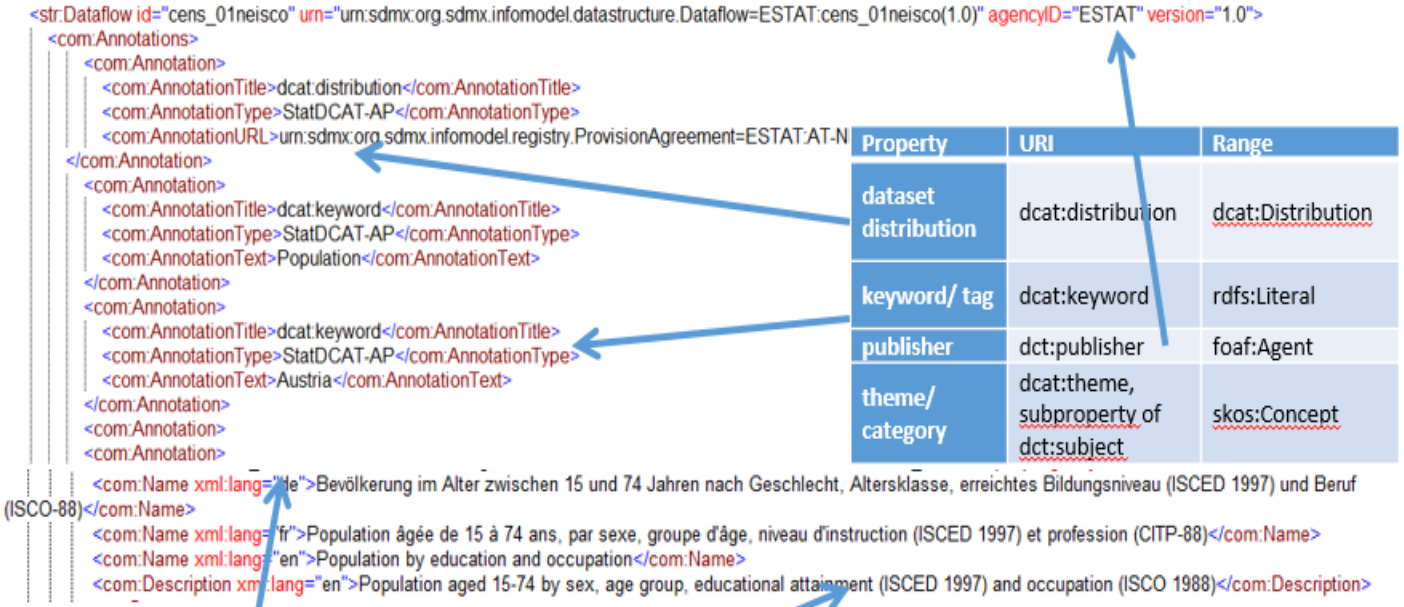
European Commission

SDMX Structural Metadata mapping to DCAT-AP



SDMX Structural Metadata (Example)

DCAT-AP Dataset



SDMX Structural Metadata - Evaluation

- Advantages
  - Familiar to organisations using SDMX
  - Can be generated easily from an SDMX Registry
- Disadvantages
- The XML can be complex and verbose
  - Annotations cannot be
    - coded (representation is restricted to text and URL)
    - hierarchical (but there is a mechanism to achieve this)
    - validated by SDMX validators (e.g. that the Title is valid)
    - given mandatory and optional status (all Annotations are optional)
  - Could create unnecessary "noise" when exchanging structural metadata with other organisations

European Commission

SDMX Metadata Set – valid content defined by Metadata Structure Definition (MSD)

Metadata Attributes Defined in MSD

[DCAT_CATALOGUE] DCAT Catalogue

- [DATASET] dcat:dataset
- [CATALOGUE_DESCRIPTION] dct:description
- [CATALOGUE_PUBLISHER] dcat:publisher
- [TITLE] dct:title
- [CATALOGUE_HOMEPAGE] foaf:homepage
- [LANGUAGE] dct:language
- [CATALOGUE_LICENSE] dct:license
- [CATALOGUE_THEME] dcat:themeTaxonomy

[DCAT_CATEGORY_SCHEME] DCAT Category Scheme

- [CATEGORY_SCHEME_TITLE] dct:title
- [DCAT_CATEGORY] DCAT Category
  - [PREFERRED_LABEL] skos:prefLabel

[DCAT_DATASET] DCAT Dataset

- [DATASET_DESCRIPTION] dct:description
- [DATASET_TITLE] dct:title
- [CONTACT_POINT] dcat:contactPoint
  - [CONTACT_PHONE] Contact phone
  - [CONTACT_EMAIL] Contact email
- [DISTRIBUTION] dcat:distribution
- [KEYWORD] dcat:keyword
- [DATASET_PUBLISHER] dcat:publisher
- [DATASET_THEME] dct:theme

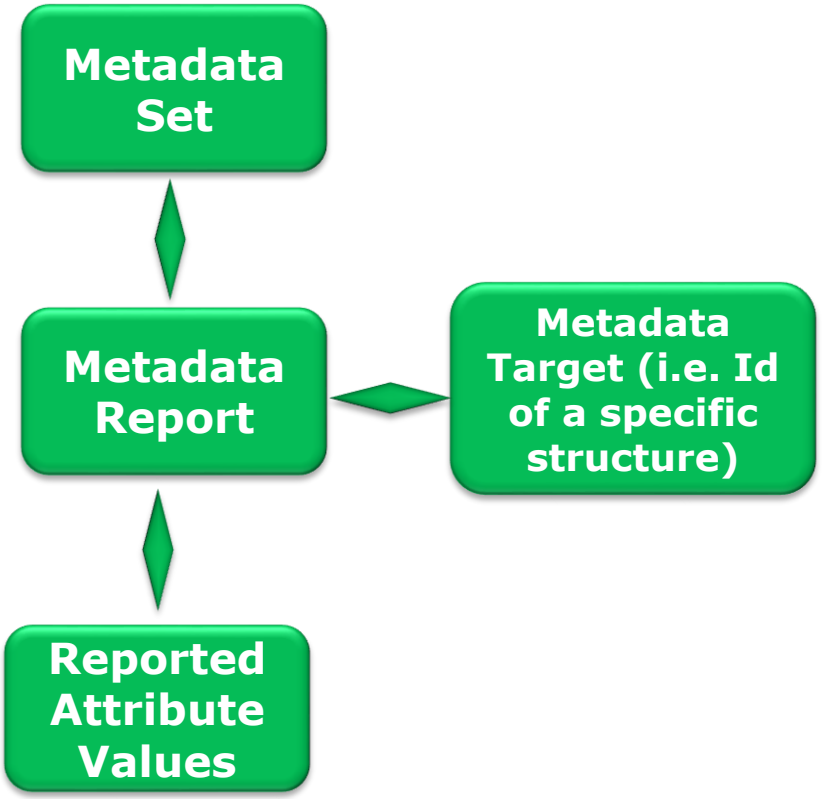
[DCAT_DISTRIBUTION] DCAT Distribution

- [ACCESS_URL] dcat:accessURL
- [DISTRIBUTION_DESCRIPTION] dct:description
- [DISTRIBUTION_FORMAT] dct:format
- [DISTRIBUTION_LICENSE] dct:license

[DCAT_AGENT] DCAT Agent

- [AGENT_NAME] foaf:name
- [AGENT_TYPE] dct:type

SDMX Metadata Set - Structure




```
<gen:ReportedAttribute id="DCAT_DATASET">
  <com:StructuredText xml:lang="en" xmlns:com="http://www.sdmx.org/resources/sdmxml/schemas/v2_
  <gen:AttributeSet>
    <gen:ReportedAttribute id="DATASET_DESCRIPTION" value="Extended description for Populati
    <gen:ReportedAttribute id="DATASET_TITLE" value="Population aged 15-74 by sex, age group,
    <gen:ReportedAttribute id="CONTACT_POINT" value="Dissemination">
      <gen:AttributeSet>
        <gen:ReportedAttribute id="CONTACT_PHONE" value="+352431034320"/>
        <gen:ReportedAttribute id="CONTACT_EMAIL" value="dissemination@ec.europa.eu"/>
      </gen:AttributeSet>
    </gen:ReportedAttribute>
    <gen:ReportedAttribute id="DISTRIBUTION">
      <com:StructuredText xml:lang="en" xmlns:com="http://www.sdmx.org/resources/sdmxml/schema
    </gen:ReportedAttribute>
    <gen:ReportedAttribute id="KEYWORD" value="Population"/>
    <gen:ReportedAttribute id="KEYWORD" value="Austria"/>
    <gen:ReportedAttribute id="KEYWORD" value="Census"/>
    <gen:ReportedAttribute id="DATASET_PUBLISHER" value="ESTAT"/>
    <gen:ReportedAttribute id="DATASET_THEME" value="urn:sdmx:org.sdmx.infomodel.categorysc
  </gen:AttributeSet>
</gen:ReportedAttribute>
```

SDMX Metadata Set – Example mapping

```
<gen:ReportedAttribute id="DCAT_DATASET">
  <com:StructuredText xml:lang="en" xmlns:com="http://www.sdmx.org/resources/sdmxml/schemas/v2_1/common">&lt;p
  <gen:AttributeSet>
    <gen:ReportedAttribute id="DATASET_DESCRIPTION" value="Extended description for Population aged 15-74 by s
    <gen:ReportedAttribute id="DATASET_TITLE" value="Population aged 15-74 by sex, age group, ed ational attainm
    <gen:ReportedAttribute id="CONTACT_POINT" value="Dissemination">
      <gen:AttributeSet>
        <gen:ReportedAttribute id="CONTACT_PHONE" value="+352431034320"/>
        <gen:ReportedAttribute id="CONTACT_EMAIL" value="dissemination@ec.europa.eu"/>
      </gen:AttributeSet>
    </gen:ReportedAttribute>
    <gen:ReportedAttribute id="DISTRIBUTION">
      <com:StructuredText xml:lang="en" xmlns:com="http://www.sdmx.org/resources/sdmxml/schemas/v2_1/common">&lt;p&gt;&lt;a href="urn:sdmx:org.sdmx.i
    </gen:ReportedAttribute>
    <gen:ReportedAttribute id="KEYWORD" value="Populat n"/>
    <gen:ReportedAttribute id="KEYWORD" value="Austria"/>
    <gen:ReportedAttribute id="KEYWORD" value="Census"/>
    <gen:ReportedAttribute id="DATASET_PUBLISHER" value="ESTA "/>
    <gen:ReportedAttribute id="DATASET_THEME" value="urn:sdmx:org.sdmx.infomodel.categoryscheme.Category=ESTAT:MDR_THEMES(1.0).SOCI"/>
  </gen:AttributeSet>
</gen:ReportedAttribute>
```

Property	URI
description	dct:description
title	dct:title

Property	URI
contact point	dcat:contactPoint

Property	URI
dataset distribution	dcat:distribution
keyword/ tag	dcat:keyword
publisher	dct:publisher
theme/ category	dcat:theme, subproperty of dct:subject

SDMX Metadata Set - Evaluation

- Advantages
  - Simple XML structure
  - Attributes can be:
    - assigned any type of representation (e.g. coded, text, HTML, Boolean etc.)
    - hierarchical
    - validated
    - usage status can be mandatory or optional
  - The Attribute Set can reference any object that can be identified (e.g. Dataflow, Provision Agreement, Category Scheme)
  - Is separate from the structural metadata so does not affect the structural metadata components
  - If present, a Metadata Attribute can be "presentational", just giving structure to child attributes

SDMX Metadata Set - Evaluation

- Disadvantages
  - Not always well understood by SDMX users (may result in some reluctance to use this mechanism)
  - Not widely used

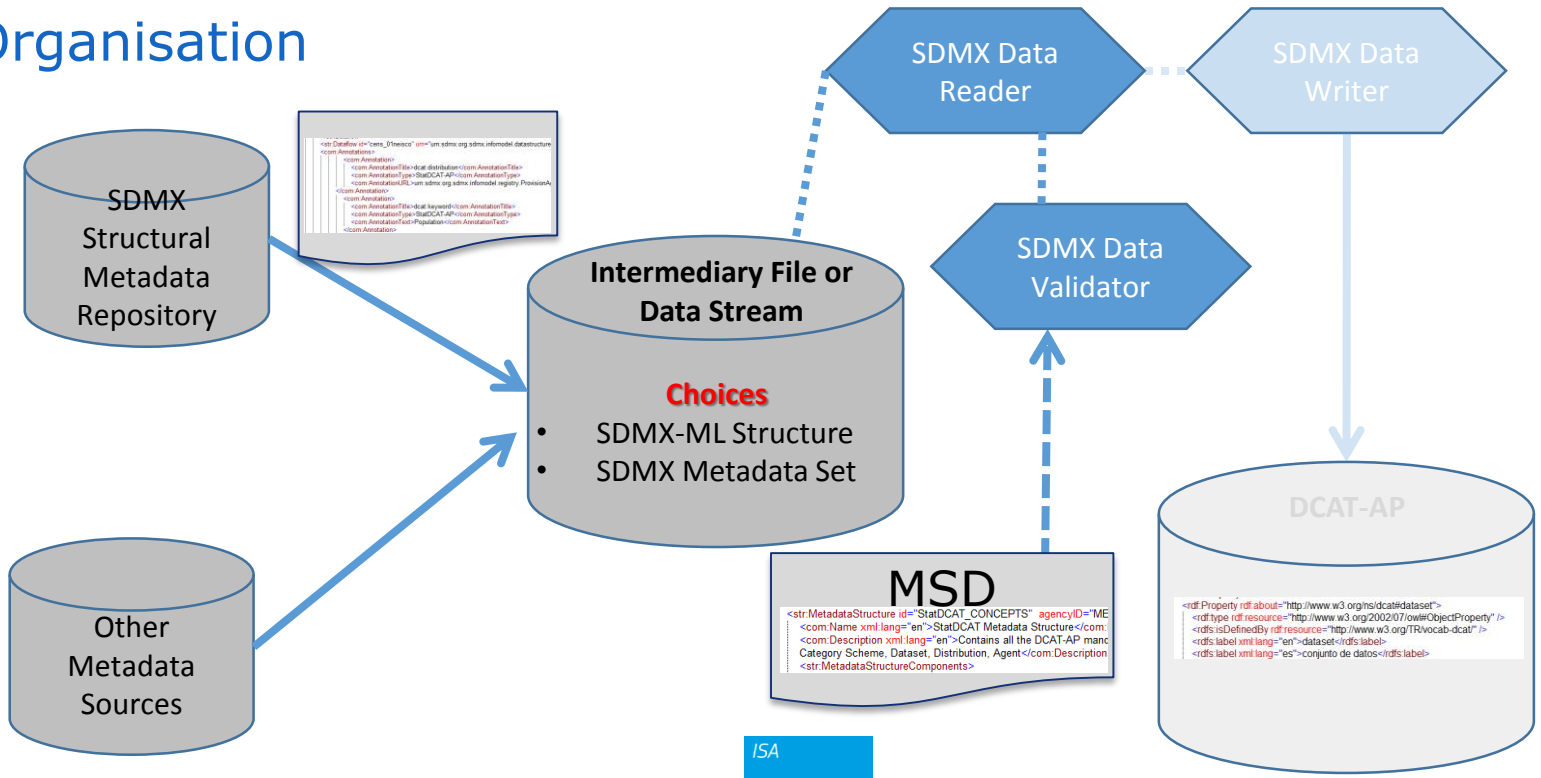
Transformation Mechanism - Summary

- Transformation mechanism is optional
- Of the two options
  - SDMX structural metadata will suit organisations using SDMX, especially those using an SDMX Registry
  - SDMX Metadata Set will suit organisations wishing to output STAT-DCAT metadata to a simple XML format
    - especially those not familiar with RDF and RDF vocabularies
- Metadata destined for DCAT-AP will need to be validated prior to generating DCAT-AP – ids, properties, references etc.
  - SDMX MSD can be used here irrespective of the intermediary format used

DCAT-AP Transformation Mechanism

These components can be developed in Java and .NET and integrated into SDMX systems or used in SDMX conversion tools

Data Publisher Organisation



Issues for discussion: Transformation Mechanism

- Is this of interest?
- Which format is of interest
  - Structural metadata
  - Metadata set?
- Is validation important?
  - Either as part of the transformation or just on its own (i.e. validate prior to creating DCAT-AP directly)
- Timescale – when would this need to be made available?

Q & A

ISA

More issues? Comments, questions?

Next steps

Developing deliverable

- Inclusion of extensions with proposal for RDF properties
- Consideration of specific controlled vocabularies and mapping to MDR Data themes
- Further work on SDMX-based transformation mechanisms

Planning

- **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
- **11 March 2016**: first virtual WG meeting to discuss first draft
- **15 April 2016**: second meeting; to discuss draft mapping and implementation options
- **6 May 2016**: second draft available for review, incorporating comments and further development
- **13 May 2016**: third meeting (face-to-face plus Adobe Connect) in Rome; to discuss mapping issues in practice

---

- **End of May 2016**: third draft, including full mapping proposal and usage of controlled vocabularies
- **3 June 2016**: fourth virtual WG meeting to agree schedule for public review
- **July and August 2016**: public review period
- **Mid-September 2016**: fifth virtual WG to discuss and resolve public comments received
- **End of September 2016**: approval of StatDCAT-AP version 1 for publication

Next meeting 3 June 2016 10:00-12:00

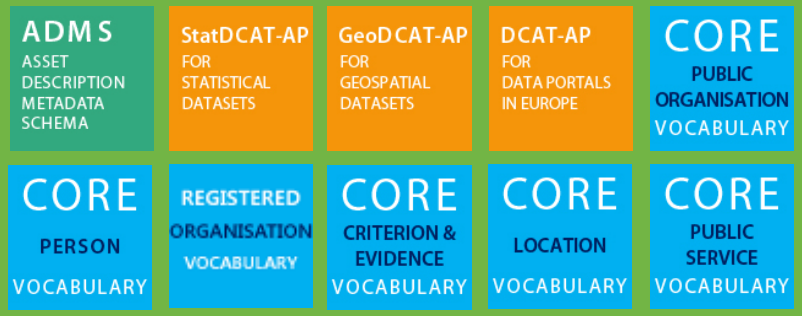
- Draft 3 will be available before the meeting
- Final proposal for extensions (RDF expression)
- Final proposal for transformation mechanism(s)
- Mapping of controlled vocabularies
- Any other issues raised by the Working Group
- Prepare for public review July – August

https://joinup.ec.europa.eu/asset/stat_dcat_application_profile/event/statdcat-ap-wg-virtual-meeting-june-3-2016

European Commission

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

Visit our initiatives



Get involved

Follow @SEMICeu on Twitter

Join the SEMIC group on LinkedIn

Join the SEMIC community on Joinup

ISA