2016-06-03

StatDCAT-AP Meeting 5

SEMIC Phase 7

Meeting Minutes

Date: 17/06/2016

Webinar: StatDCAT-AP virtual meeting			
Venue	Adobe Connect	Meeting date	2016-06-03
Author	Stefanos Kotoglou	Meeting time	10:00 – 12:00
Reviewed by	Makx Dekkers	Issue date	2016-06-10
Status	Accepted	Version	0.04

1. ATTENDEES

Name	Abbreviation	Organisation
Agnieszka Zajak	AZ	Publications Office of the EU
Alan Vask	AV	Marketing and Dissemination Department, Estonia
Anastasia Dimou	AD	iMinds Data Science Lab, Belgium
Athanasios Karalopoulos	AK	European Commission
Aurelien Bonnet	AB	IWEPS, Belgium
Bert Van Nuffelen	BN	TenForce, Belgium
Chris Nelson	CN	Metadata Technology Ltd., UK
Denis Groflis	DG	European Commission
Gregor Boyd	GB	Scottish Government
Jan Drovak	JD	euroCRIS
Makx Dekkers	MD	Member SEMIC team, editor
Marco Pellegrino (co-chair)	MP	Eurostat - European Commission
Martial Menard	MM	European Commission
Nikolaos Loutas	NL	PwC EU Services
Norbert Hohn (co-chair)	NH	Publications Office of the EU
Rob Davidson	RD	Office for National Statistics (UK)
Soren Roug	SR	European environment agency

Stefanos Kotoglou	SK	PwC EU Services
Uros Milosevic	UM	TenForce, Belgium
Valentina Janev	VJ	Institute Mihajlo Pupin, Serbia
Vassilios Peristeras	VP	European Commission
Willem Van Gemert	WG	Publications Office of the EU

AGENDA

ID	Description
1.	Opening, agenda, tour de table
2.	Outcomes Rome meeting, 13 May 2016
3.	Objectives of this meeting
4.	Overview of DCAT-AP and StatDCAT-AP
5.	Extensions: Dimensions and attributes, Quality aspects, Visualisation, Other extensions
6.	Mapping theme vocabulary (Eurostat à MDR)
7.	SDMX Transformation mechanism
8.	Any other business - Conclusions
9.	Next steps

2. OPENING, AGENDA, TOUR DE TABLE

MP presented the agenda for the meeting, and invited all the participants to present themselves. A full list of all the participants of the working groups' meetings is available on Joinup.

3. OUTCOMES ROME MEETING, 13 MAY 2016

MD presented the key points and decisions taken at the face to face meeting in Rome. All these key points were discussed during that day's meeting (03/06/2016). The key points were the following:

- There are several decisions for extensions, and further discussion on proposals for RDF terms should take place.
- The SDMX transformation mechanism will be moved to the Annex of the specification. However, the working group showed slight preference for the use of a Metadata Structure Definition (MSD).

- Publications Office will propose mappings from Eurostat themes to Metadata Registry (MDR) data themes.

4. OBJECTIVES OF THIS MEETING

MP presented the objectives of the meeting:

- Agree on proposed extensions
- Review mapping of themes Eurostat to Metadata Registry (MDR)
- Reach common understanding of proposed SDMX transformation mechanism
- Identify further issues for the future
- Prepare for public review period

5. OVERVIEW OF DCAT-AP AND STATDCAT-AP

MD presented the DCAT-AP data model. The model has four main entities: Catalogue, Catalogue Record, Dataset and Distribution.

MD mentioned that the StatDCAT-AP includes additional optional properties for Dataset and Distribution. More specifically, the classes of the extended DCAT-AP model should include the following attributes:

Class	Additional optional properties
Dataset	dqv:hasQualityAnnotation qb:attribute or stat:attribute qb:dimension or stat:dimension schema:population stat:numSeries stat:statUnit
Distribution	dct:type

6. EXTENSIONS: DIMENSIONS AND ATTRIBUTES, QUALITY ASPECTS, VISUALISATION, OTHER EXTENSIONS

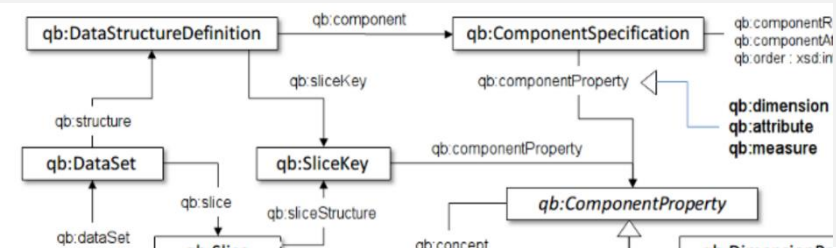
6.1. Extensions: dimensions and attributes

MD mentioned that there is a requirement for exposing information about:

- Dimensions, e.g. observations related to sex, age, etc.
- Attributes, e.g. observations expressed in certain units.

MD presented the two alternative options for the StatDCAT-AP:

Option 1 - Re-use properties **qb:dimension** and **qb:attribute** from Data Cube Vocabulary:



- Expected values: URI of qb:DimensionProperty and qb:AttributeProperty.
- However, these properties are not directly attached to qb:Dataset.
- This might lead to confusion for datasets that are published as qb:Datasets.

Option 2 - Define new properties **stat:dimension** and **stat:attribute** in StatDCAT-AP namespace:

- Allows semantics to be precisely defined to meet StatDCAT-AP requirements.
- Expected values can still be: URI of qb:DimensionProperty and qb:AttributeProperty.
- For Data Cube datasets, values can be derived (copied) from qb:dimension and qb:attribute.

MD mentioned that the structure of qb:dimension might be confusing. If we use it for datasets, then dimension is used in different way than it was meant to be used. There might be a confusion if people start converting datasets to Data Cube.

MD inquired whether the participants agree on following Option 2, and whether we should rename the 2 properties for avoiding any confusion.

DG mentioned that it is better to reuse existing properties. Creating new properties could be drawback when linking them to the Data Cube Vocabulary.

MD mentioned that both options use the same range, which means that the expected values are the same as using the qb properties.

DG inquired whether we should describe and clarify what the differences are comparing to the qb properties.

MD agreed with the proposal made by DG.

BN suggested contacting the editors of the Data Cube Vocabulary for initiating discussions for the inclusion of an optional property in the data Cube Vocabulary. The property should create a link between qb:dataset and qb:component.

CN suggested contacting David Reynolds from Data Cube vocabulary.

The participants decided to use the stat properties (stat:dimension and stat:attribute).

6.2. Quality aspects

MD mentioned that due to time and resource constraints, we cannot fully address this issue now.

MD presented the 2 proposed phases:

- Short-term: provide mechanism to link to existing quality information in StatDCAT-AP, version 1
- Longer-term: consider integrated quality framework as basis for extensions to StatDCAT-AP, version 2

MD provided more details regarding each phase:

Short-term: annotation

- Link to existing document/webpage with quality information, or provide plain text

- Use property: dqv:hasQualityAnnotation from W3C Data Quality Vocabulary (in development)
- Expected value: URI (e.g. webpage) or plain text conformant to specification of oa:Annotation

MD mentioned that DQV is being developed by W3C. Two annotation examples were provided that will also be provided in the public review:

```
:Dataset-001 a dcat:Dataset ;
  dqv:hasQualityAnnotation :Annot-001 .

:Annot-001 a dqv:QualityAnnotation ;
  oa:hasBody <URL> ;
  oa:hasTarget :Dataset-001 ;
  oa:motivation oa:commenting .
```

```
:Dataset-001 a dcat:Dataset ;
  dqv:hasQualityAnnotation :Annot-001 .

:Annot-001 a dqv:QualityAnnotation ;
  oa:hasBody [ oa:text "Some text" ] ;
  oa:hasTarget :Dataset-001 ;
  oa:motivation oa:commenting .
```

Longer-term: Quality aspects of SIMS

- Eurostat's Single Integrated Metadata Structure includes specific quality aspects:
  - e.g. Accessibility and clarity; Quality management; Relevance; Accuracy and reliability; Timeliness and punctuality; Coherence and comparability
- This set of aspects can form the basis for future extensions to StatDCAT-AP, or even to DCAT-AP.

The WG agreed with this approach.

MD present the proposals for 2 new properties: Statistical unit and Statistical population:

Property: Statistical unit

- New property in StatDCAT-AP namespace: stat:statUnit
- Expected value: free text

Property: Statistical population

- Use property: schema:population, "Any characteristics of the population used in the study, e.g. 'males under 65'."
- Expected value: free text

MD inquired whether we should defer these issues to deeper discussion on quality issues.

MP mentioned that for the Statistical unit we should fully specify the quality dimension.

MP suggested using controlled vocabularies and not plain text.

AD, BN, NH and AB agree with the proposed solution provided by MP for the Statistical unit.

JD suggested using both free text and a SKOS concept as alternatives.

VJ inquired whether lining Statistical unit and theme/topic of the dataset is feasible.

MP mentioned that the goal is not to link the Statistical unit with the theme, which does not imply that there might not be connection.

GB inquired whether the Statistical unit could deal with datasets with multiple units.

AZ mentioned that the Statistical unit should allow multiple values.

AZ and MD inquired whether there is a *controlled vocabulary at Eurostat that verifies the types of Statistical unit* that can be used.

MP mentioned that there might be a controlled vocabulary from which we could develop the *aforementioned vocabulary*.

NH mentioned that the semantics for the 2 properties (Statistical unit and Statistical population) are not well defined. He suggested providing clarifications.

AZ mentioned that unit of Measure is also an important element.

CN mentioned that in a dataset, the dataset values are not always performed using the same unit of measure.

MP mentioned that we should re-evaluate whether the unit of Measure is relevant with our extension.

The participants agreed on deferring the Statistical population to deeper discussion.

6.3. Visualisation

MD mentioned that dct:Type should be repeatable, i.e. cardinality (1…n). Type would point to a term in the Distribution type Vocabulary.

SR mentioned a distribution could have more than one rdf:type. Dublin Core was defined for library cards, and dct:type describes the type of the item, e.g. book. When used in an RDF context, the rdf:type is doing the same job, and is recognised in all types of RDF implementation. rdf:type should be used instead of dct:type. See section 5.2 in http://dublincore.org/documents/dc-rdf/ "It is recommended that RDF applications implementing this specification primarily use and understand rdf:type in place of dcterms:type when expressing Dublin Core metadata in RDF, as most RDF processors come with built-in knowledge of rdf:type."

MD mentioned that after a common agreement between the participants, we should use the dct:type in our proposal. The discussion about rdf:type is not closed yet though.

6.4. Other extensions

The participants agreed on the proposed solutions for the "Number of data series" and the "Time coverage of the data series":

Issue	Solution
Number of data series	• New property in StatDCAT-AP namespace: stat:numSeries • Expected value: integer
Time coverage of the data series	• Use property: dct:temporal • Expected value: time period with schema:startDate and schema:endDate • Already in DCAT-AP; no extension necessary

7. MAPPING THEME VOCABULARY (EUROSTAT -> MDR)

WG presented the mapping:

Eurostat

Themes	Title
Theme 1	General and regional statistics
Theme 2	Economy and finance
Theme 3	Population and social conditions
Theme 4	Industry, trade and services
Theme 5	Agriculture, forestry and fisheries
Theme 6	International trade
Theme 7	Transport
Theme 8	Environment and energy
Theme 9	Science and technology

Metadata Registry (MDR)

Title		Code	Type
		general	folder
Economy and finance		economy	folder
Population and society		popul	folder
Economy and finance		icts	folder
Agriculture, fisheries, forestry,foods		agric	folder
Economy and finance		external	folder
Transport		transp	folder
Environment	Energy	envir	folder
Science and technology		science	folder

*Legend
Red clear: no equivalence
Red: partial equivalence
Orange: equivalence by combing two terms
Green: equivalence

WG mentioned that there is no alignment yet with Eurovoc. The next step is to align Eurovoc with Eurostat.

AZ mentioned that because of bad matching with some themes, the current mapping of Eurostat themes to Eurovoc, which is used on ODP, requires further elaboration.

8. SDMX TRANSFORMATION MECHANISM

MP mentioned that there are 2 approaches that are under consideration:

- SDMX structural metadata to DCAT-AP
- SDMX metadata set to DCAT-AP

CN mentioned that initially he performed a mapping of SDMX structural metadata to DCAT-AP. There was a direct map for nearly all of the mandatory properties. However, for some of the recommended properties there was no equivalent metadata in SDMX, and consequently he used the SDMX annotations for offering structural metadata for these DCAT-AP properties.

MP mentioned that there is a slight preference for MSD but we don't make a specific proposal in the specification.

MD mentioned that the transformation mechanism is included in the Annex of the specification. The transformation mechanism is also open for discussion.

9. NEXT STEPS

ID	Description	Owner	Due date
1.	To involve external commenters and mailing lists in the public review period.	ISA	Done
2.	To share the document with the codelists (controlled vocabulary that verifies the types of Statistical unit), and initiate a discussion with the OP.	Eurostat	07/06/2016
3.	To circulate a doodle poll for the final meeting of the WG.	ISA	07/06/2016
4.	To develop couple of examples for each extension, and to describe and clarify what the differences are comparing to the qb properties.	ISA	10/06/2016
5.	To clarify the possible use of qb:dimension and qb:attribute by contacting David Reynolds from Data Cube vocabulary	ISA	10/06/2016

6.	Prepare the final draft of the specification, based on the discussions. For some discussion, there will not be conclusions (to be discussed during the public review period).	ISA	20/06/2016
7.	To re-evaluate whether the unit of Measure is relevant with our extension.	ISA	20/06/2016
8.	To discuss with the dissemination team of Eurostat and the OP any quality issues regarding the mapping of the themes between Eurovoc and Eurostat.	Eurostat	20/06/2016
9.	To announce the public review period via the mailing list.	ISA	27/06/2016

10. CHAT

The chat history has been cleared

Valentina Janev: Good morning

Athanasios KARALOPOULOS: Hello Valentina

Marco Pellegrino: Good morning everybody, welcome to the last call before the public review!

Alan Vask (Statistics Estonia): Hello!

Aurélien Bonnet (IWEPS): Good Morning

Rob Davidson: good morning all!

Martial Menard: Good morning

OP- ODP Agnieszka Zajac: Good morning. Norbet has just come

Jan Dvorak: good morning from Prague

Willem van Gemert (OP): Good morning from Luxembourg

Aurélien Bonnet (IWEPS): I will use chat

Alan Vask (Statistics Estonia): Hi

Alan Vask (Statistics Estonia): I can hear

Jan Dvorak: hello, Jan Dvorak from euroCRIS - research information

Martial Menard: Hello

Athanasios KARALOPOULOS: You can find a list of all the members of the WG participating in the calls here: https://joinup.ec.europa.eu/sites/default/files/isa_field_path/statdcat-ap_participants_v0.02_1.pdf

Søren Roug: Hello, Søren Roug from the European Environment Agency coming invia chat

Valentina Janev: Hello, working for Institute Mihajlo Pupin, Serbia. Good collaboration with the Serbian Statistical Office

Valentina Janev: no, option 2 ok

OP- ODP Agnieszka Zajac: For us Option 2 is fine

Gregor Boyd 2: happy to go with yur preference

Nikos Loutas, PwC: +1 for option 2

Alan Vask (Statistics Estonia): I lean towards option 2 also

Rob Davidson: which specification?

Rob Davidson: ah, no - I wasn't QB author

OP- ODP Agnieszka Zajac: questio odp

OP- ODP Agnieszka Zajac: Hi, the question was precisely about it: https://www.w3.org/TR/vocab-dqv/. Thank you Makx for explaining it

Bert Van Nuffelen 2: + agree for defer these to a deeper discussion

Bert Van Nuffelen 2: If unit can be coded: I propose we create an example and what is the effort to insert it.

Anastasia Dimou - iMinds: +1 not for frere test

Anastasia Dimou - iMinds: free*

Rob Davidson: +1 no free text

Gregor Boyd 2: How would stat:statUnit deal with datasets with multiple units?

Uroš Milošević 2: +2 no free text

Gregor Boyd 2: cool, thanks

Valentina Janev: how we are going to link statistical unit and theme/topc of the dataset?

Valentina Janev: yes but hey are somehow connected

Aurélien Bonnet (IWEPS): removing free text can be be restrictive

OP- ODP Agnieszka Zajac: I think the statUnit should allow for multiple values

Aurélien Bonnet (IWEPS): yes free + coded could be interesting

Anastasia Dimou - iMinds: sorry but was it clarified why statUnit and not a more generic unitMeas□ure?

Anastasia Dimou - iMinds: ah ok!

OP- ODP Agnieszka Zajac: Is there any code list in Estat that could be used for it?

Rob Davidson: Ah, I was confused by 'unit' too. We mean 'population member'?

Jan Dvorak: Re Aurelien Bonnet: I'd suggest to allow both free text and a SKOS concept as alternatives

Anastasia Dimou - iMinds: yes I still don't understand why e.g. enterprise can't be the unit of measure but I'm also not a domain expert.

OP- ODP Agnieszka Zajac: we agree that unit of measure is important

Anastasia Dimou - iMinds: but it's usually preferred when we □there is a finite number of possible options to define an entity that clarifies them. the range may not be determined but the property should be object property.

Anastasia Dimou - iMinds: @Jan Dvorak: what could be the problem if the statUnit I use is not a skos:Concept?

Søren Roug: A distribution can have more than one rdf:type

Willem van Gemert (OP): +1 By when would you need the new value added to the distribution-type table?

Søren Roug: :Visual-001 a dctype:Image

Søren Roug: or... :Visual-001 a dctype:Image, dcat:Distribution

Søren Roug: Dublin Core was defined for library cards, and dct:type descripes the type of the item - e.g. book. When used in an RDF context, the rdf:type is doing the same job, and is recognised in all types of rdf. I think rdf:type should be used instead of dct:type. See section 5.2 in http://dublincore.org/documents/dc-rdf/ "It is recommended that RDF applications implementing this specification primarily use and understand rdf:type in place of dcterms:type when expressing Dublin Core metadata in RDF, as most RDF processors come with built-in knowledge of rdf:type."

Jan Dvorak: @Anastasia Dimou: indeed it could be anything

Søren Roug: What you're really doing is to use dct:type to describe what is a the end of dcat:accessURL

Rob Davidson: soren makes a good point.

Uroš Milošević 2: it does make it a much stronger statement, though

Anastasia Dimou - iMinds: +1 for Søren Roug comment

Jan Dvorak: need to leave - apologies and good bye

Anastasia Dimou - iMinds: specs allows to have MUST and SHOULD, here we can propose one of the two, but refer to 5.2 of dcterms and leave room dct:type too

Uroš Milošević 2: a good analogy to the rdf:type vs dct:type discussion is owl:sameAs vs skos:exactMatch

OP- ODP Agnieszka Zajac: If we want to provide the information also about the type of visualisation (eg. image, interactive, inforgraphic) - how could it be done? will rdf:type allow for it?

Anastasia Dimou - iMinds: exactMatch is usedto link two concepts, indicating a high degree of confidence that the concepts canbe used interchangeably across a wide range of information retrieval applications. owl:sameAs is used when we are sure that two conecepts/entities are exactly the same real world concept

OP- ODP Agnieszka Zajac: The current mapping of ESTAT themes to Eurovoc that is used on ODP required going to lower level. The reason for it was similar to the current problems - no good match for some themes

Uroš Milošević 2: @Anastasia: Exactly. Yet people don't understand the repercussions of misuse of owl:sameAs (see http://sameas.org). I would go with dct:type instead of rdf:type for the same reason

Anastasia Dimou - iMinds: @Uroš Milošević we agree ;)

Uroš Milošević 2: apologies Marco :)

OP- ODP Agnieszka Zajac: For the public review, apart from the explanation of the semantic, could we have 1-2 examples for each extension taking as a basis some datasets from Estat?

OP- ODP Agnieszka Zajac: yes, we can hear you

Aurélien Bonnet (IWEPS): Godd-bye, thank you

Makx Dekkers: thanks for good discussion!

Willem van Gemert (OP): Thank you all!

Valentina Janev: thanks a lot, bye

Rob Davidson: thanks - good progress.bye!

OP- ODP Agnieszka Zajac: Thanks a lot, Bye

Anastasia Dimou - iMinds: thank you too, bye!

Gregor Boyd 2: thanks bye

Vassilios Peristeras, DIGIT/ISA 2: Bye and thanks