European Commission

10 MARCH 2022

Guidelines on identifiers in the context of DCAT-AP

DIGIT.D2 - Interoperability.

interoperable europe

Welcome – Let’s introduce ourselves

The SEMIC team:

- Seth Van Hooland
- Pavlina Fragkou
- Bert Van Nuffelen
- Makx Dekkers

Motivation of the webinar

The alignment on the expectations on the usage of identifiers has been

- identified as an implementation issue (already for a long time)
- set by DCAT-AP focus groups an important priority
    - During last webinar (WG 21 Oct 2021) it was raised as a topic for future work
    - issues on github are still unresolved (#187, #141)
- dataspaces are emerging

Agenda

01

State of play

Identifiers

Use cases

Existing guidelines

02

Proposals

dct:identifier

adms:identifier

03

Next steps

State of play – Generic aspects of Identifiers

Identifiers design principles

Principle 1

- Assigned by the ‘responsible’ of the entity
- The assignee knows the lifecycle of the entity

Attention point: the expectation is that the responsible assigns one identifier to the same entity. So in two different systems the same identifier!

Identifiers design principles

Principle 2

- The identifier is **persistent**

Attention point: Identifiers that refer to old, historic, not anymore maintained entities should be deprecated instead of deleted.

Identifiers design principles
Principle 3

- The identifier is **dereferenceable.** This means that one can retrieve just on the basis of the knowledge of the identifier the core information about the entity to which this identifier refers.

- Alternative wording: provide the context in which the identifier is an identifier.

Attention point: use a universal known protocol to achieve this.
By preference HTTP/HTTPS.

Use cases

Use cases – to facilitate processing

To make processing

- Deterministic
- Reliable
- Efficient
- Idempotent
- Avoid mistakes
- …

Use cases – to facilitate networking

- ownership/responsible
  - Identifier can aid to clarify ownership or identify the responsible.
- cross reference
  - When referring to another entity (the dataset X is derived from dataset Y)
- stability in evolution
  - The information associated with the dataset is evolving over time. E.g. a new license is imposed, distributions are changed, etc.

Use cases – to facilitate portals

- Data portals want to provide a nice coherent representation of all the datasets in their catalogue.

- The UI framework imposes technical requirements on the identifiers used. All entities need an identifier.

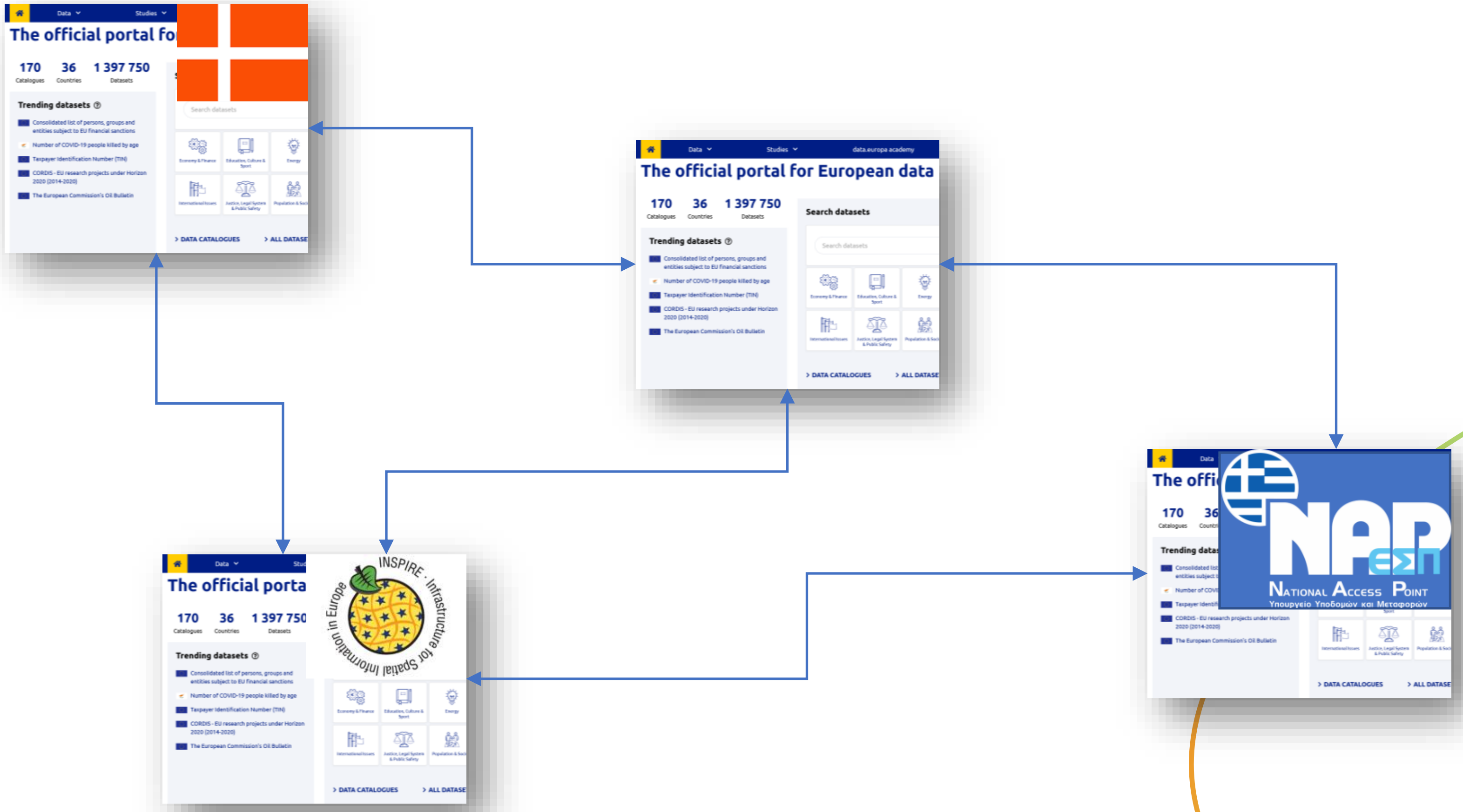
Use cases – harvesting

Harvesting is the process of aggregating source catalogues into a singe larger catalogue

Some expectations on harvesting:

- Harvested datasets should be easily retrievable in the aggregation.
- Harvesters should not be required to impose cross-source requirements like sources are disjoint
- Harvesters should not contribute to the creation of duplicates
- Harvesters should not claim “ownership” of the sources. There should be ways that users of the aggregated catalogue can find back the original source.
- Harvesting should be ‘cheap’: both for dataset owners as for the harvesters

Use cases – harvesting


The official portal for European data
170 Catalogues
36 Countries
1 397 750 Datasets
Trending datasets
Search datasets
Consolidated list of persons, groups and entities subject to EU financial sanctions
Number of COVID-19 people killed by age
Taxpayer Identification Number (TIN)
CORDIS - EU research projects under Horizon 2020 (2014-2020)
The European Commission's Oil Bulletin
DATA CATALOGUES
ALL DATASETS
INSPIRE · Infrastructure for Spatial Information in Europe
NAP ΕΣΠ
National Access Point
Υπουργείο Υποδομών και Μεταφορών

Identifiers examples

- https://data.europa.eu/data/datasets/1735eaaf-afe6-4d90-af67-488c4c37b91f?locale=en
- http://data.europa.eu/88u/dataset/1735eaaf-afe6-4d90-af67-488c4c37b91f
- https://inspire-geoportal.ec.europa.eu/download_details.html?view=downloadDetails&resourceId=%2FINSPIRE-f0c91711-ece0-11e8-a08e-52540023a883_20210903-160102%2Fservices%2F1%2FPullResults%2F221-240%2Fdatasets%2F5&expandedSection=metadata
- https://data.gov.be/en/node/179577
- https://opendata.vlaanderen.be/dataset/adressen
- https://metadata.vlaanderen.be/srv/resources/resources/5c52b299-8f05-4d35-9839-a42934f1e619

Share all the same identification string 5c52b299-8f05-4d35-9839-a42934f1e619

State of play – Existing guidelines

adms:Identifier

Existing guidelines

DCAT:

- Usage guide on dereferenceable identifiers:
  - https://w3c.github.io/dxwg/dcat/#dereferenceable-identifiers

DCAT-AP:

- Guidelines on avoiding duplicates:
  - https://joinup.ec.europa.eu/release/dcat-ap-how-manage-duplicates
- Guidelines on usage of identifiers:
  - https://joinup.ec.europa.eu/release/dcat-ap-how-use-identifiers-datasets-and-distributions

Generic:

- 10 Rules for Persistent URIs:
  - https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/document/10-rules-persistent-uris

Identifier properties in DCAT(-AP)

The following properties are available in:

- dct:identifier : Literal
  - Purpose: the main identifier, a simple notation of the identifier
- adms:identifier : adms:Identifier
  - Purpose: the notation with metadata about the identifier

Identifier properties in DCAT(-AP)

The following properties are available in:

- dct:identifier : Literal
  - Purpose: the main identifier, a simple notation of the identifier
  - Usually implementers like to put restrictions on this, fitting the usage context of the portal.
- adms:identifier : adms:Identifier
  - Purpose: the notation with metadata about the identifier
  - Usually this is more open, not attractive because there are multiple identifiers and then question is how to deal with them. Often considered as to be ignored information.

Identifier properties in DCAT(-AP)

The following properties are available in:

- dct:identifier : Literal
  - Purpose: the main identifier, a simple notation of the identifier
  - Usually implementers like to put restrictions on this, fitting the usage context of the portal.
- adms:identifier : adms:Identifier
  - Purpose: the notation with metadata about the identifier
  - Usually this is more open, not attractive because there are multiple identifiers and then question is how to deal with them. Often considered as to be ignored information.
- When sharing data as RDF, then also the RDF:about is part of the identifier discussion
  - The alignment with the RDF is a separate topic.

Proposals

Approach

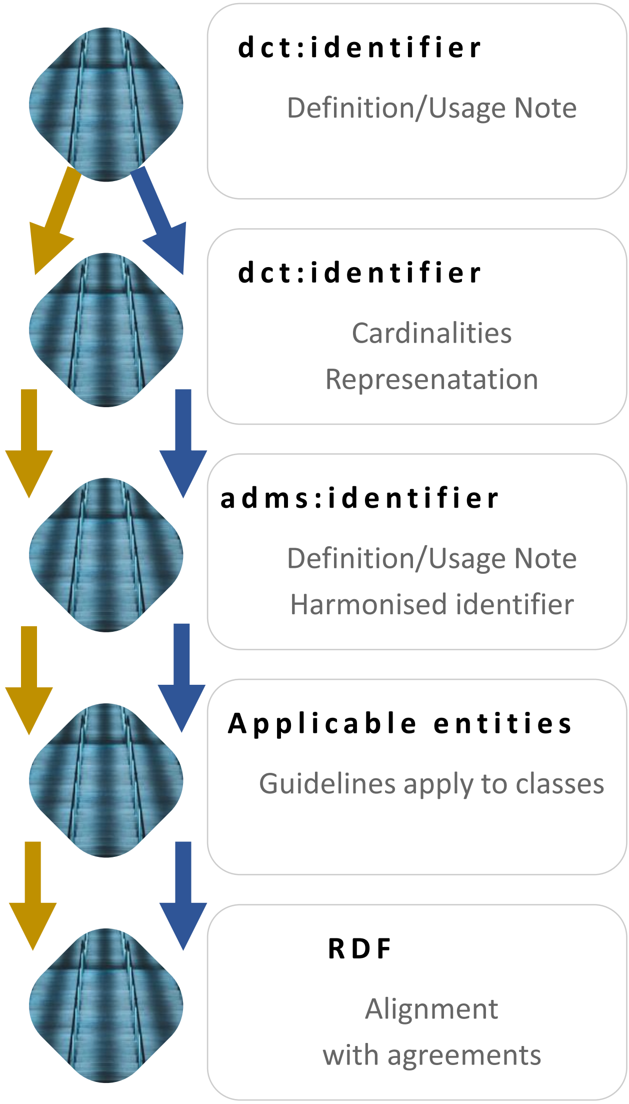
- Introduction of a discussion topic
  - As concrete as possible.
- Discussion with the WG to understand the opinions
- The proposal of late topics may be influenced by earlier discussions

change

no change

DISCUSSION

Flow of discussion



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities

Represenatation

adms:identifier

Definition/Usage Note

Harmonised identifier

Applicable entities

Guidelines apply to classes

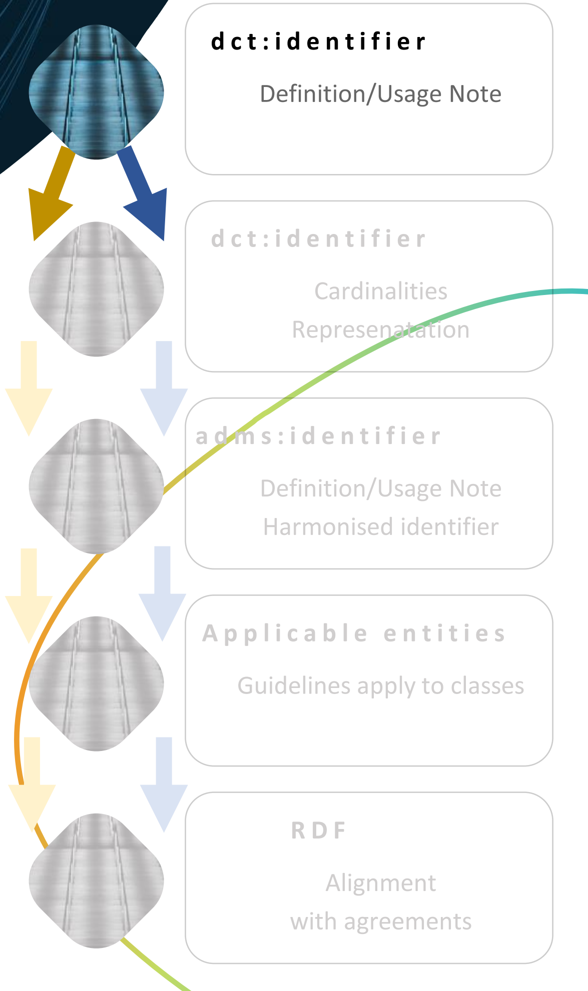
RDF

Alignment

with agreements

Based on outcome on topic 1,
the yellow or blue track will be followed

Usage Proposal 1



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities

Represenatation

adms:identifier

Definition/Usage Note

Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment

with agreements

change

Proposal 1

Usage note dct:identifier: (#187)

This property contains the main identifier for the Dataset, e.g. the URI or other unique identifier in the context of the Catalogue.

Ambiguity in the sentences:

a) the value assigned by the catalogue, or

b) the value assigned by the owner/publisher of the dataset

DISCUSSION

Assessment

The value assigned by the catalogue

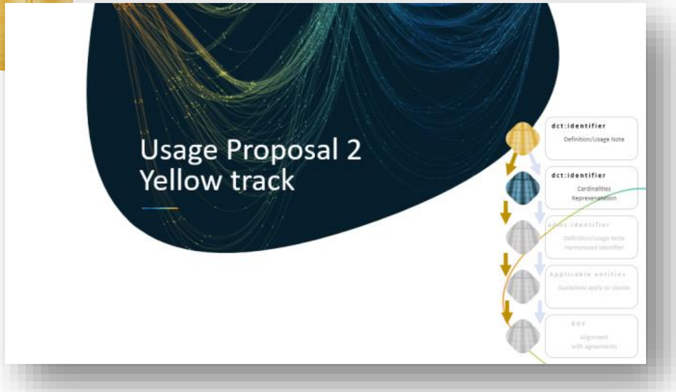
- The value is coherent in the catalogue
- Creates big impact
  - Most existing dataset descriptions will be affected
  - Harvesting must replace the value with a value assigned by the harvesting process, as the context changes. Otherwise it is not anymore the identifier in the context of the catalogue.
  - Requires to tackle a complex alignment process to ensure that interlinking of datasets is maintained. (Cross portal references)
- Most immediate benefits are for the catalogue, not for the network:
  - cross-linking challenge (to the catalogue identifier or to one of the adms:identifier's)

The value assigned by the owner/publisher of the dataset

- Minimal impact, maximum clarity
- The publishers/owners are incentivised to consider good identifier management
- Oriented towards network benefits, while catalogue benefits are secondary.

Poll – which interpretation

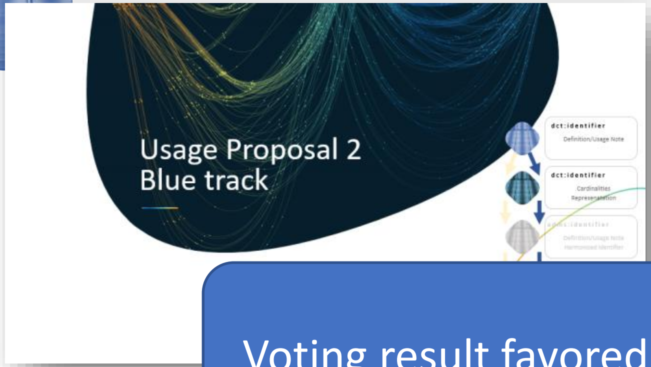
Clear yellow



Usage Proposal 2
Yellow track

- **dct:identifier** — Definition/Usage Note
- **dct:identifier** — Cardinalities, Representation
- **adms:identifier** — Definition/Usage Note, Harmonised identifier
- **Applicable entities** — Guidelines apply to classes
- **RDF** — Alignment with agreements

Clear blue



Usage Proposal 2
Blue track

- **dct:identifier** — Definition/Usage Note
- **dct:identifier** — Cardinalities, Representation
- **adms:identifier** — Definition/Usage Note, Harmonised identifier

Voting result favored blue track

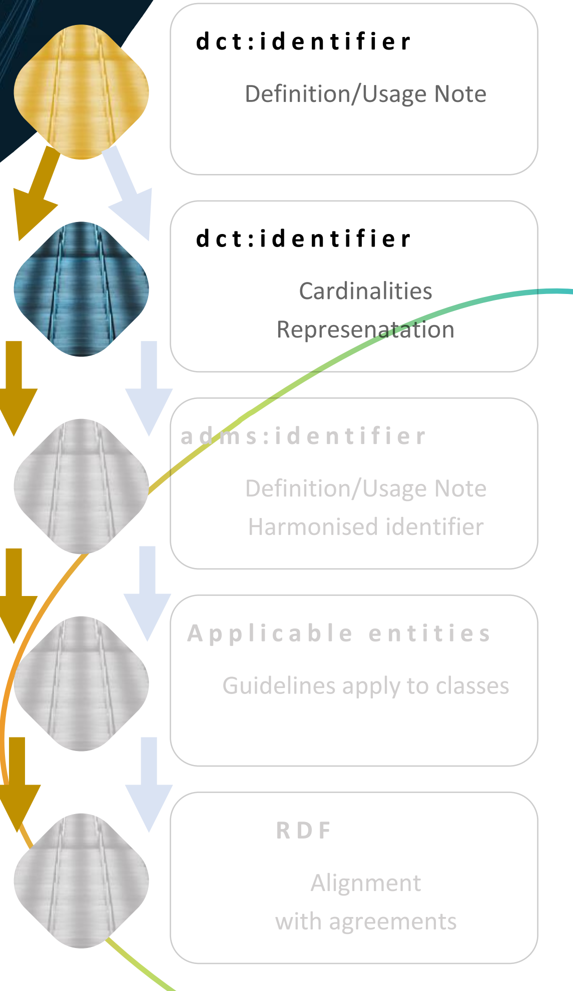
Undetermined

Flow of discussion

- **dct:identifier** — Definition/Usage Note
- **dct:identifier** — Cardinalities, Representation
- **adms:identifier** — Definition/Usage Note, Harmonised identifier
- **Applicable entities** — Guidelines apply to classes
- **RDF** — Alignment with agreements

Based on outcome on topic 1,
the yellow or blue track will be followed

Usage Proposal 2
Yellow track



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities
Represenatation

adms:identifier

Definition/Usage Note
Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment
with agreements

Yellow track

Editorial Note:

The preparation of the the yellow track has been removed from the published slides as the WG decided to follow the blue track.

European Commission

interoperable europe

innovation ∾ govtech ∾ community

Stay in touch

[(@InteroperableEU) / Twitter]

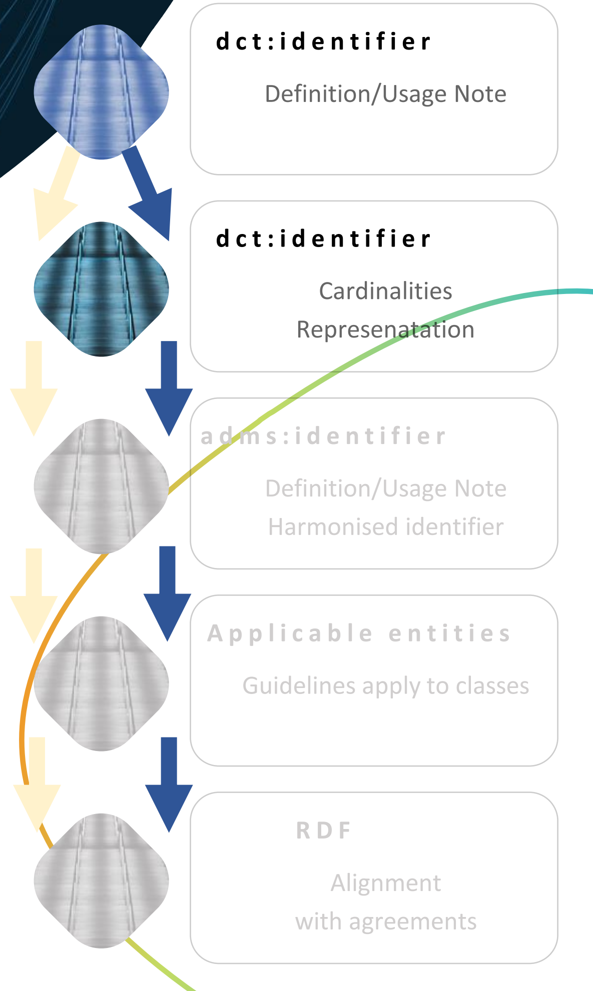
[Interoperable Europe - YouTube]

[Interoperable Europe | LinkedIn]

✉ DIGIT-INTEROPERABILITY@ec.europa.eu

https://joinup.ec.europa.eu/collection/interoperable-europe/interoperable-europe

Usage Proposal 2
Blue track



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities
Represenatation

adms:identifier

Definition/Usage Note
Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment
with agreements

no change

Proposal 2

Min cardinality dct:identifier: 0

Motivation:

- The value is the value provided by the dataset owner

Impact:

- No impact, as an enforcement (min card 1) will invalidate 50% of the data.Europa.eu datasets.

change

DISCUSSION

Proposal 2

Max cardinality dct:identifier : 1

Motivation:

- Usage note states “the main value assigned by the owner”
- No discussion on which is the “identifier” to be used, as there cannot be made distinction between different notations.

Impact:

- Harvesters should not add more dct:identifier
- Gradually this can be enforced.

no change

Proposal 2

Current usage note on dct:identifier implicitely advices to use high quality identifier, preferably a properly managed IRI.

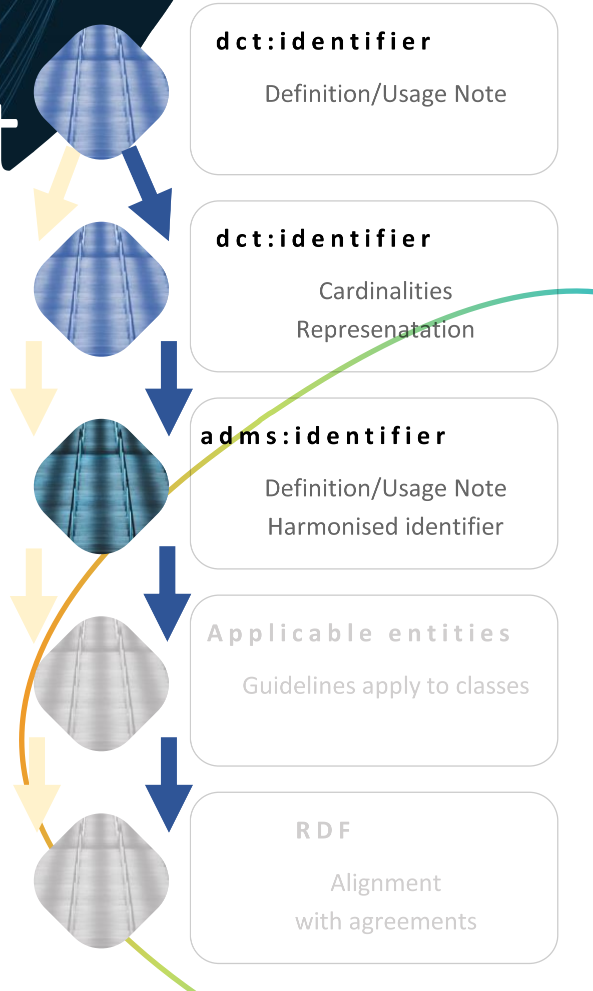
Proposal:

Maintain the advice, but maybe reformulated.

Motivation:

- A high quality identifier cannot be checked from the representation, many options possible.

Usage Proposal 3
document identifier context



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities

Represenatation

adms:identifier

Definition/Usage Note

Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment

with agreements

change

Proposal 3

```
<D1> dct:identifier “D1”.

<D1> adms:identifier [
             _:  skos:notation “D1”.
             _:  dct:creator  <Publisher>
]
```

Use adms:identifier do describe metadata about the identifier. So not only “other” identifier but information about **all** identifiers assigned.

Motivation:

- dct:identifier is a literal, without context and ownership
- adms:identifier provides means to express context, ownership of the identifier
- adms:identifier is an immer growing collection of identifiers assigned

impact:

- Seem to create duplicate information, but adms:identifier allows to distinguish identifiers based on properties, rather on value inspection.

DISCUSSION

Proposal 3 – Blue Track impact

Use adms:identifier do describe metadata about the identifier.

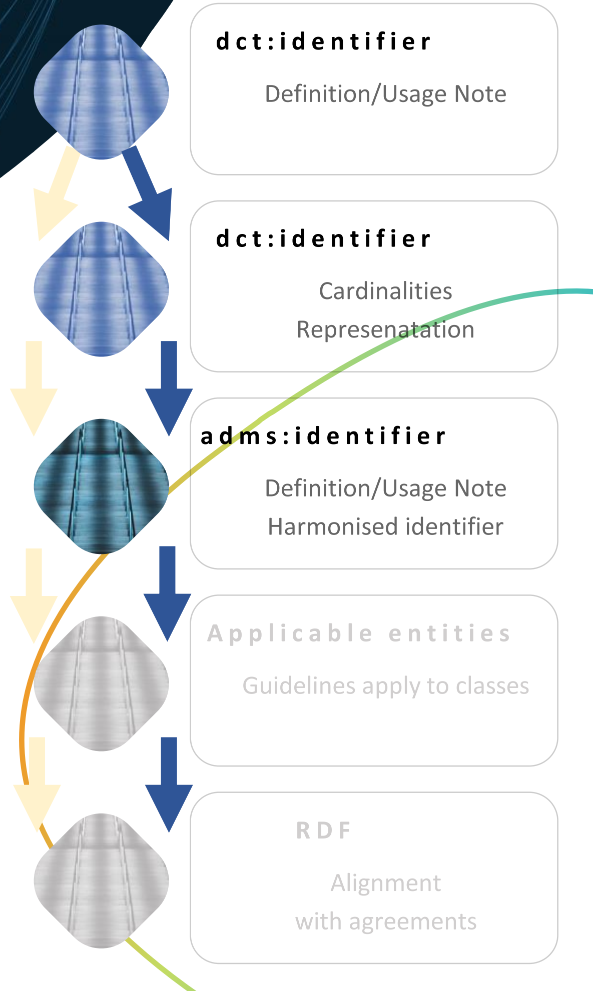
Impact:

- No impact, as the dct:identifier is not changed
- If a new identifier is assigned, then it is added to the list by the processor, so no impact.

Additional note:

- Can be used to introduce a dct:identifier when not present. In this case the “first” catalogue, close to the publisher, assigns an identifier.

Usage Proposal 4
Harvester creates harmonised identifiers



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities
Represenatation

adms:identifier

Definition/Usage Note
Harmonised identifier

Applicable entities

Guidelines apply to classes

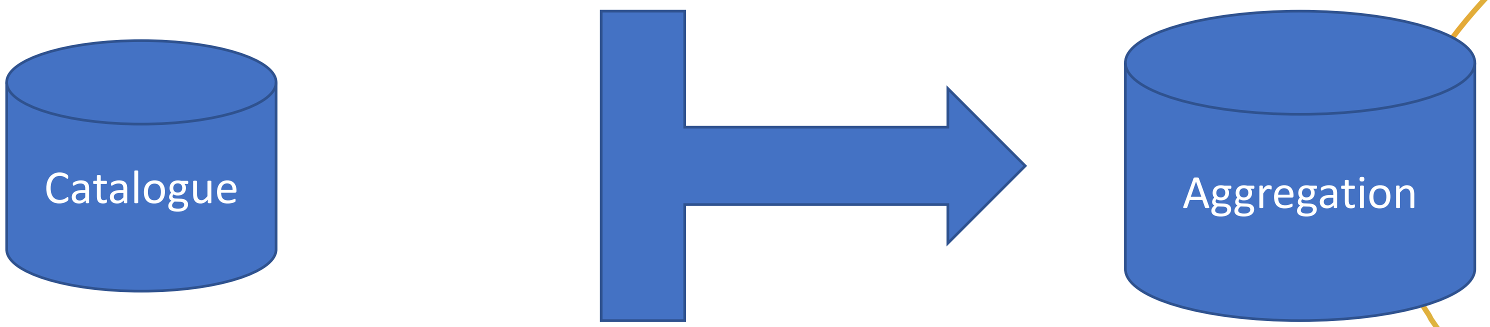
RDF

Alignment
with agreements

Proposal 4: Harvesters introduce harmonised identifiers

This proposal addresses for harvesters and associated portals the use cases:

```
<D1> dct:identifier “D1”.   <D1> dct:identifier “D1”.
```



Proposal 4: Harvester introduce harmonised identifiers

This proposal addresses for harvesters and associated portals the use cases:

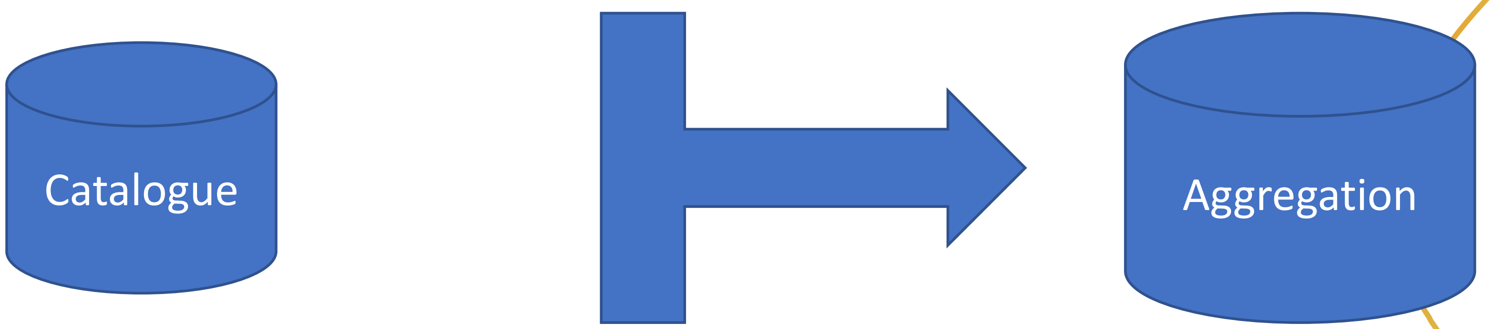
```
<D1> dct:identifier “D1”.

<D1> adms:identifier [
          _:  skos:notation “D1”.
          _:  dct:creator  <Catalogue>
]
```

```
<D1> dct:identifier “D1”.

<D1> adms:identifier [
          _:  skos:notation “HARM(D1)”.
          _:  dct:creator  <Aggregator>
]

<D1> adms:identifier [
          _:  skos:notation “D1”.
          _:  dct:creator  <Catalogue>
]
```



Harmonised Identifier

An **Harmonised Identifier** is an identifier created by the aggregator (harvester) to ensure that the aggregated data elements have an identifier all in the same representation.

An harmonised identifier is added to the element as an alternative identifier (adms:identifier) with the appropriate metadata. Minimally the creating agent is added.

An harmonised identifier must be shared to the next harvesting layer. So that cross references can be made.

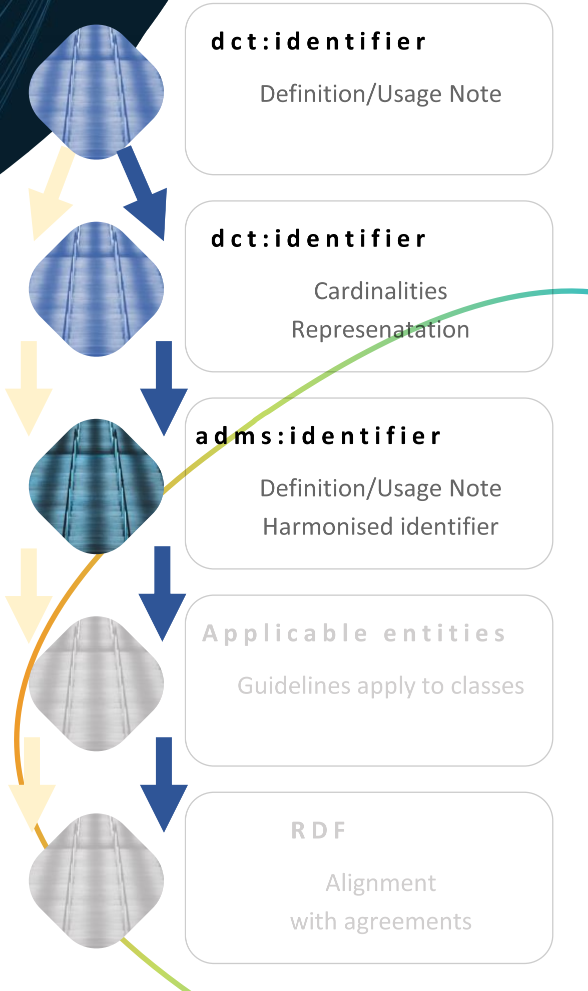
Does not change the source metadata.

DISCUSSION

Pro/contra

- Pro:
  - No impact on the source metadata
  - Simple, additive approach
  - An identifier conform the UI framework can be introduced
  - Can be used to bridge the harvesting case when no dct:identifier.
- Contra:
  - Aggregated catalogues must query through the harmonised identifier:
  - Select * where { <id> a dcat:Dataset}
  - Select * where { ?s a dcat:Dataset. ?s adms:identifier ?hid. ?hid skos:notation <id>. }

Usage Proposal 5
Extend adms:identifier



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities
Represenatation

adms:identifier

Definition/Usage Note
Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment
with agreements

DISCUSSION

Decompositition of an identifier

Extend adms:Identifier with properties to decompose the identifier in components.

Motivation:

- A UI framework requires only the uuid instead of the full URI (bridging software/data formats) (string manipulation of identifiers should not be enforced as best practice)
- Difference between version aspects versus versionless
- Avoids the creation of an additional adms:identifier which only consists of the component

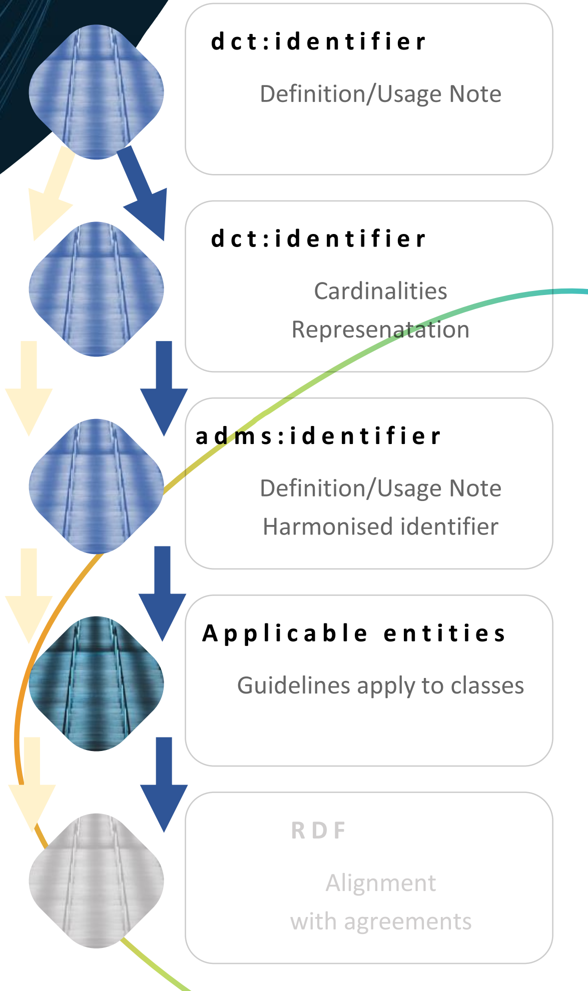
```
<D> adms:identifier [
        _:  skos:notation “context:uuid(d1)”.
        _:  dct:creator  <CV-harvester>
        _: m8g:namespace “context”.
        _: m8g:localIdentifier “uuid(d1)”
        _: m8g:versionIdentifier “<harvestingti
        _: dct:issued “<harvestingtime>”
]
```

Questions:

- Adding to adms?
- Which are the components?

Usage Proposal 6
Applicable to enti

NOT DISCUSSED in webinar



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities
Represenatation

adms:identifier

Definition/Usage Note
Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment
with agreements

change

Applicable to which entities (#141)

DISCUSSION

The previous discussed guidelines should apply to the following entities:

- **Dataset**
- **Data Service**

Possibly for

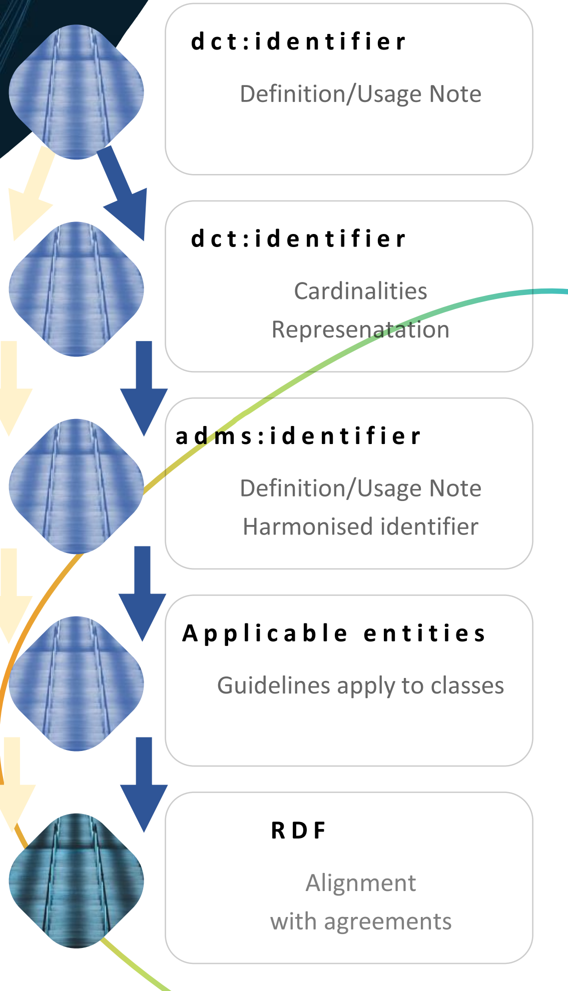
- Distribution
- Agent

- Catal
- Catalog

NOT DISCUSSED in webinar

Usage Proposal 7
RDF and dct:id

NOT DISCUSSED in webinar



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities
Represenatation

adms:identifier

Definition/Usage Note
Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment
with agreements

DISCUSSION

RDF as data sharing format

Proposal for usage note

The entity’s URI (not blank node) in the RDF format is also the value of dct:identifier, and vice versa.

Impact assessment: ***blue track assigned by publisher/owne***

- This statement enforces that publisher must u ditional hurdle to supply a value.

NOT DISCUSSED in webinar

On harvesting:

- No im

Our advic

Since it forc to use URIs as identifiers the statement is not so attractive

RDF as data sharing format

Proposal for usage note

The entity's URI (not blank node) in the RDF format is one of the values in dct:identifier or adms:identifier, and vice versa.

Impact assessment: ***blue track assigned by publisher***

- This statement enforces that when a ca are minted with a URI should have this URI included in

NOT DISCUSSED in webinar

On harvesti

- Harv value should be added to adms:identifier with as agent the source

DISCUSSION

RDF as data sharing format

Proposal for usage note

The entity’s URI (not blank node) in the RDF format is one of the values in dct:identifier or adms:identifier, and vice versa.

This statement does not resolve the “priority” rule be roperties dct:identifier/adms:identifier.

And also it does not pro ers can change the URI from the entity (on input i new identifier is part of the dct:identi

But to m ised not to change the URI of the entity when resharing detection and reduces the need of identifier alignments.

NOT DISCUSSED in webinar

Our advice o proposal

This statement creates a modest impact. It fits within the usage note advice on adding harmonised identifiers.

Next steps

Next steps

Consider these changes as a **bug fix release,** as immediate impact is low.

Approach:

- Create the guidelines
- Adapt the specification
- publish a draft release on github for public review (short period)

Planning:

- Public review: during april 2022
- Publication: during may 2022

Thank you

European Commission

interoperable europe

innovation ∾ govtech ∾ community

Stay in touch

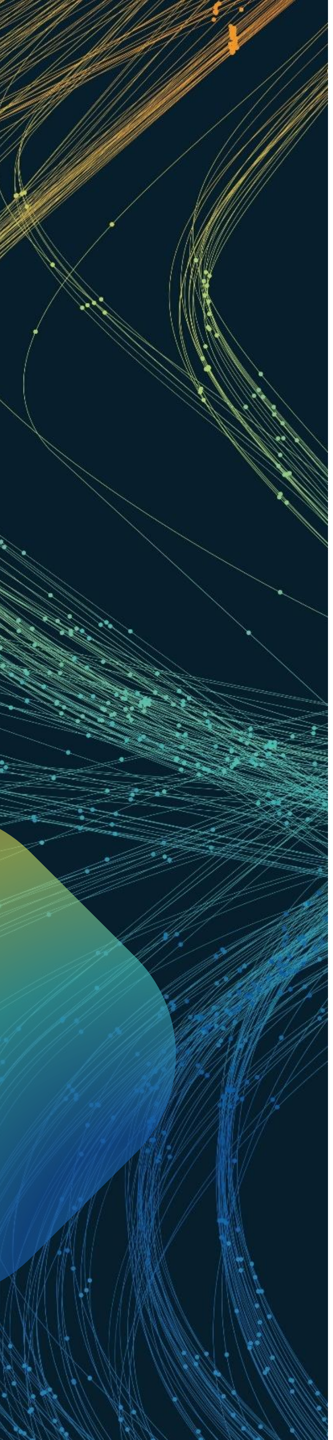
[(@InteroperableEU) / Twitter](#)

[Interoperable Europe - YouTube](#)

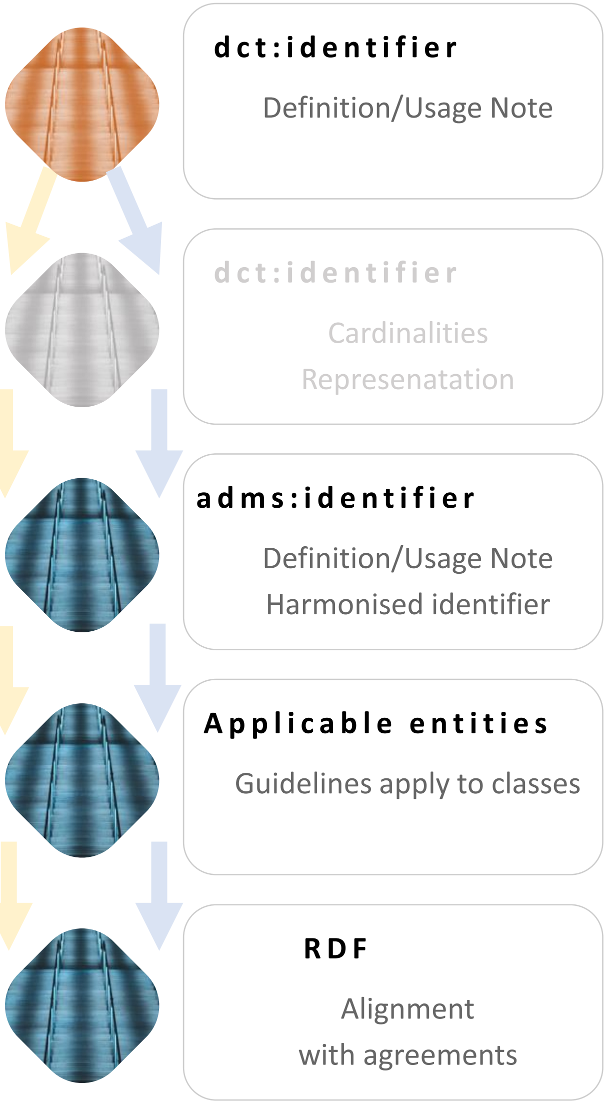
[Interoperable Europe | LinkedIn](#)

✉ DIGIT-INTEROPERABILITY@ec.europa.eu

https://joinup.ec.europa.eu/collection/interoperable-europe/interoperable-europe



Flow of discussion



dct:identifier

Definition/Usage Note

dct:identifier

Cardinalities

Represenatation

adms:identifier

Definition/Usage Note

Harmonised identifier

Applicable entities

Guidelines apply to classes

RDF

Alignment

with agreements

Based on outcome on topic 1,
the yellow or blue track will be followed

Orange track

Editorial Note:

The preparation of the the orange track has been removed from the published slides as the WG decided to follow the blue track.

European Commission

interoperable europe

innovation ∾ govtech ∾ community

Stay in touch

[(@InteroperableEU) / Twitter]

[Interoperable Europe - YouTube]

[Interoperable Europe | LinkedIn]

DIGIT-INTEROPERABILITY@ec.europa.eu

https://joinup.ec.europa.eu/collection/interoperable-europe/interoperable-europe