European Commission

26 APRIL 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(s)

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 earlier 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

Today: continuation of our discussion of 10 March 2022

Agenda

01

State of play (recap)

Identifiers

Use cases

Existing guidelines

02

Proposals

adms:identifier

dct:identifier

Examples scenarios

RDF guidelines

Application

03

Next steps

State of play – short recap

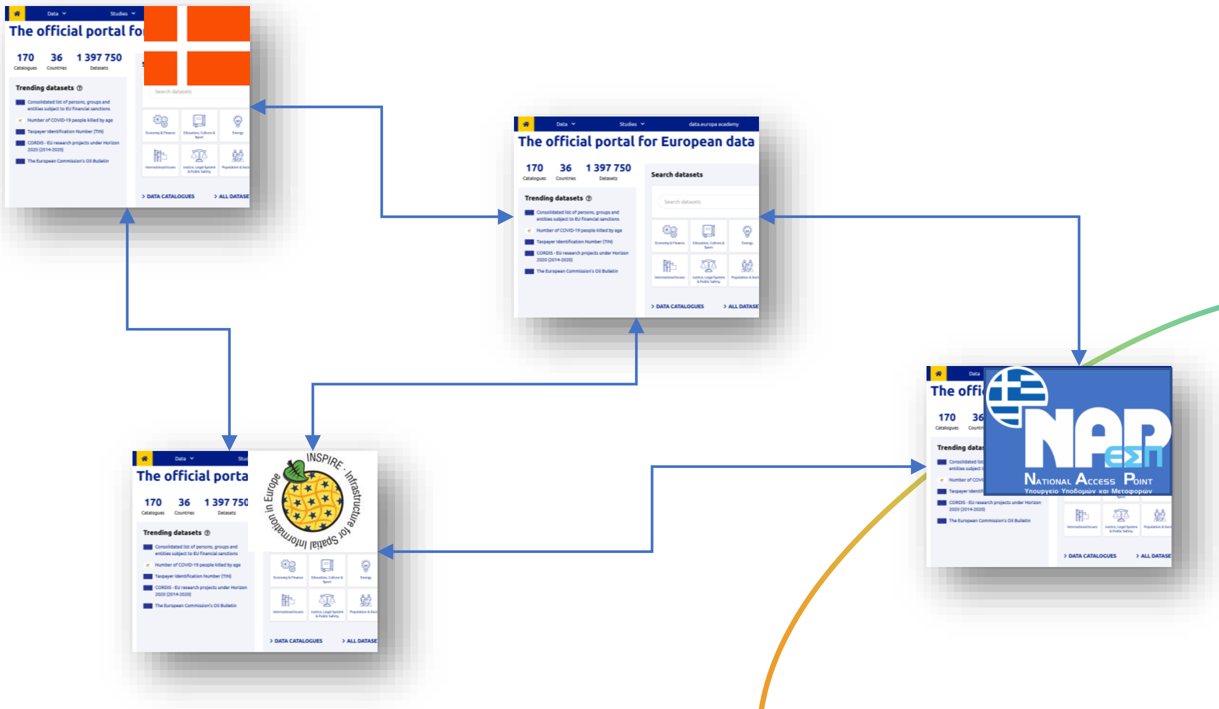
Identifiers design principles

Around the following concepts:

- Responsibility
- Persistency
- Dereferenceability

Use cases for identifiers

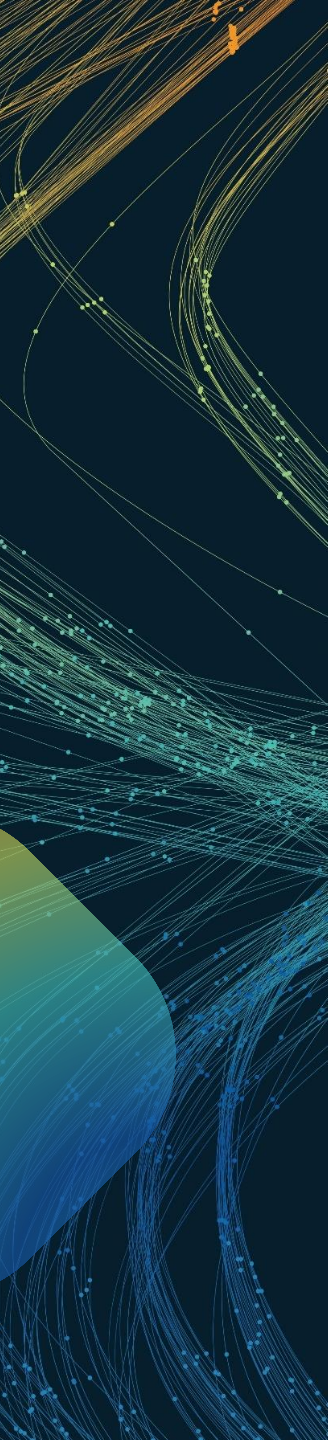
- facilitating processing
- facilitating networking
- facilitating portal development
- facilitating harvesting


The official portal for European data
170 Catalogues
36 Countries
1 397 750 Datasets
Trending datasets
Search datasets
DATA CATALOGUES
ALL DATASETS
INSPIRE · Infrastructure for Spatial Information in Europe
NAP
National Access Point
Υπουργείο Υποδομών και Μεταφορών

Existing work

- Existing guidelines are not sufficient
- DCAT-AP has the following relevant properties
  - 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

Proposal



Proposal presentation

- The proposal is based on the discussions and decisions taken in webinar of 10 March
- Reorganised presentation to highlight our main objective behind this initiative:
  - supporting the DCAT-AP community
  - Strengthening the catalogue network
  - Keep catalogues as independent from each other
  - Provide catalogue implementable solution
- Available on

  https://github.com/SEMICeu/DCAT-AP/blob/2.1.1-draft/releases/2.1.1/usageguide-identifiers.md

Approach

- Introduction of a discussion topic
  - As concrete as possible.
- Discussion with the WG to understand their opinion

DISCUSSION

Flow of discussion

Share metadata on identifiers

adms:identifier

Example scenarios

Illustration proposal

Harvesting

Main identifier

dct:identifier

RDF format guidelines

Proposal guidelines

Application

The extend of the application of the rules

Implementation of the proposal

- adms:identifier
    - immediate benefit
    - Suggestion: adoption via bug fix release

- dct:identifier
    - requires possibly internal catalogue adaptations
    - Suggestion: adoption via minor release

Usage Proposal
Share metadata on identifier

Proposal

Use adms:identifier to describe metadata about identifiers.

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 becomes an ever growing collection of identifiers assigned

```
<D1> dct:identifier “D1”.

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

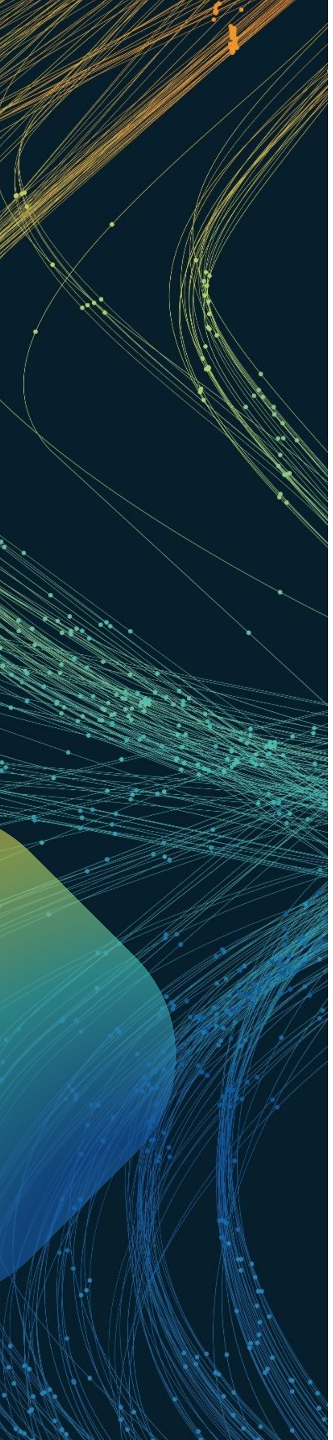
Proposal

Property label	URI	Range	Cardinality
identifier	adms:identifier	adms:Identifier	0..n

Definition	Usage Note
described identifier for the Dataset	Each identifier a catalogue or a process assigns and which is publicly accessible (e.g. via a data portal) should be included.

Changelog

- label change: "other identifier" -> "identifier"
- usage note change: "This property refers to a secondary identifier of the Dataset" -> "described identifier for the Dataset"



Requirements on the range

Impose minimal information for an adms:Identifier

Property	URI	Range	Cardinality	Definition
notation	skos:notation	Literal	1..1	content string which is the identifier
schema manager name	adms:schemaAgency	Literal	1..1	the name of the agency that manages the identifier scheme
schema manager agent	dct:creator	foaf:Agent	1..1	the agency that manages the identifier scheme

Already the providing source of the identifier is aiding decision making. At least one should be provided.

Additional components

Extend adms:Identifier with additional 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

Is there interest in such additional components?

```
<D> adms:identifier [
        skos:notation “{context}:{uuid}”;
        dct:creator  <publisher> ;
        m8g:namespace “{context}”;
        m8g:localIdentifier “{uuid}” ;
       m8g:versionIdentifier “<idcreationtime>”
]
```

DISCUSSION

Proposal impact

All identifiers should be **included**:

- the value of dct:identifier should be included
- the RDF URI should be included

All identifiers should be **shared**:

- on harvesting, the aggregator should never lose identifier information so if the RDF URI is changed, or dct:identifier value is changed during harvesting the original identifiers along the new identifiers should be part of the adms:identifier list.

Enforcement

SHACL validation rules **can** check

- if the value of dct:identifier is part of adms:identifier notation

and to be investigated

- The URI is a part of adms:identifier

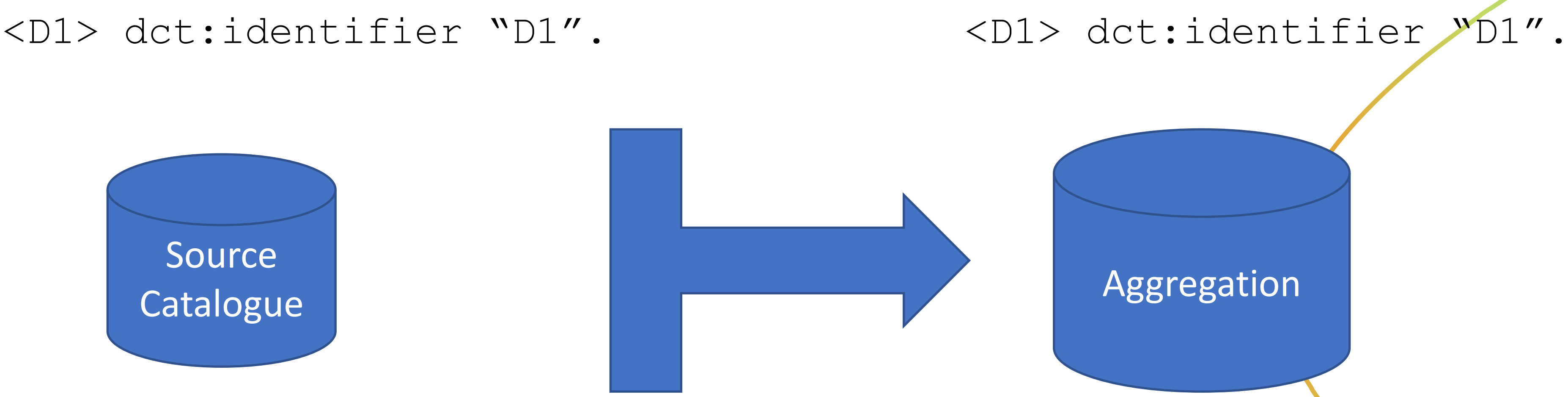
But **cannot** check the propagation/sharing aspect.

Example scenarios

Illustration of 1 situation

Harvesting a catalogue

Simple copying is **not** the advice.



```
<D1> dct:identifier "D1".
```

```
<D1> dct:identifier "D1".
```

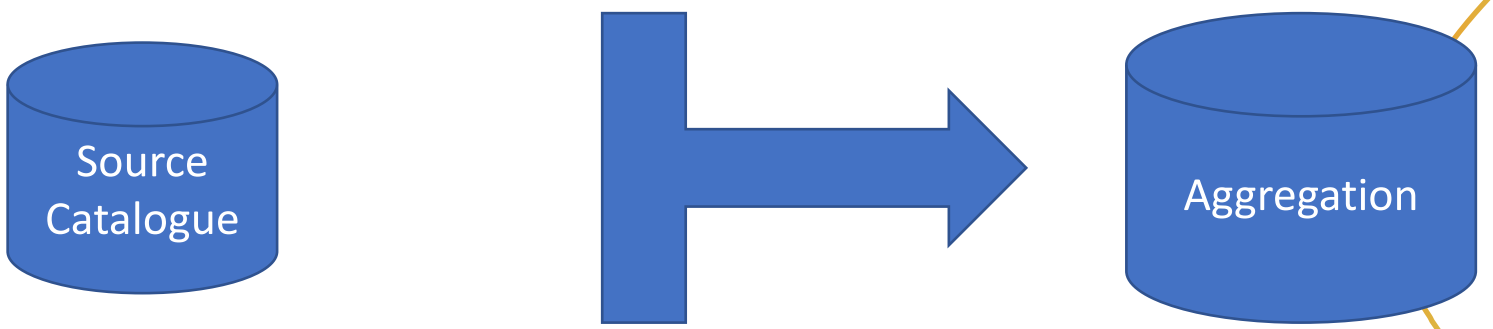
Harvesting a catalogue

Harvesters are advised to complete the adms:identifier list for the harvested datasets.

```
<D1> dct:identifier “D1”.
```

```
<D1> dct:identifier “D1”.


<D1> adms:identifier [
       skos:notation “D1” ;
       dct:creator  <Source-Catalogue>
]
```



Aggregation catalogues
share your identifier

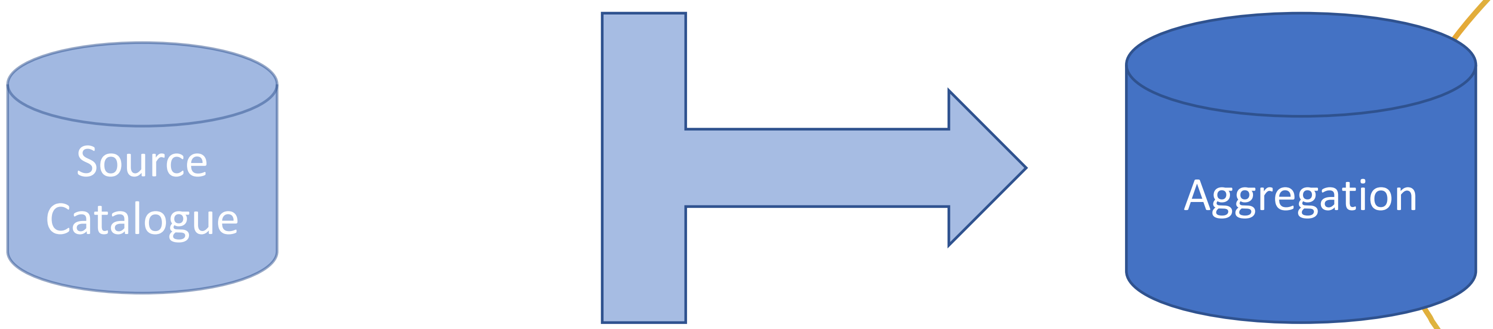
Harvesters are advised to add adms:identifier for newly created identifiers

```
<D1> dct:identifier “D1”.

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

<D1> adms:identifier [
          skos:notation “D1” ;
          dct:creator  <Source-Catalogue>
]
```

```
<D1> dct:identifier “D1”.
```



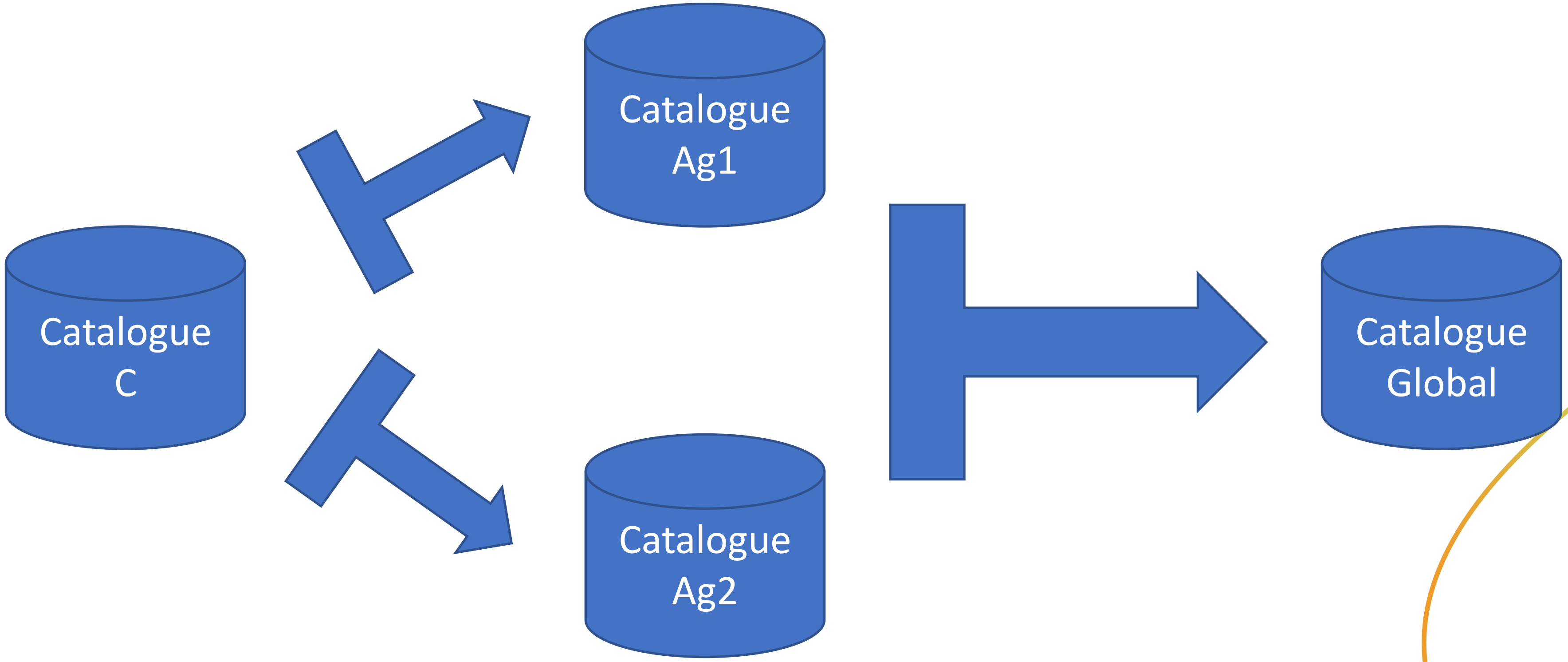
Aggregation catalogues
share your identifier

Aggregating catalogues want to have uniform identifiers in their catalogue in order to support querying.

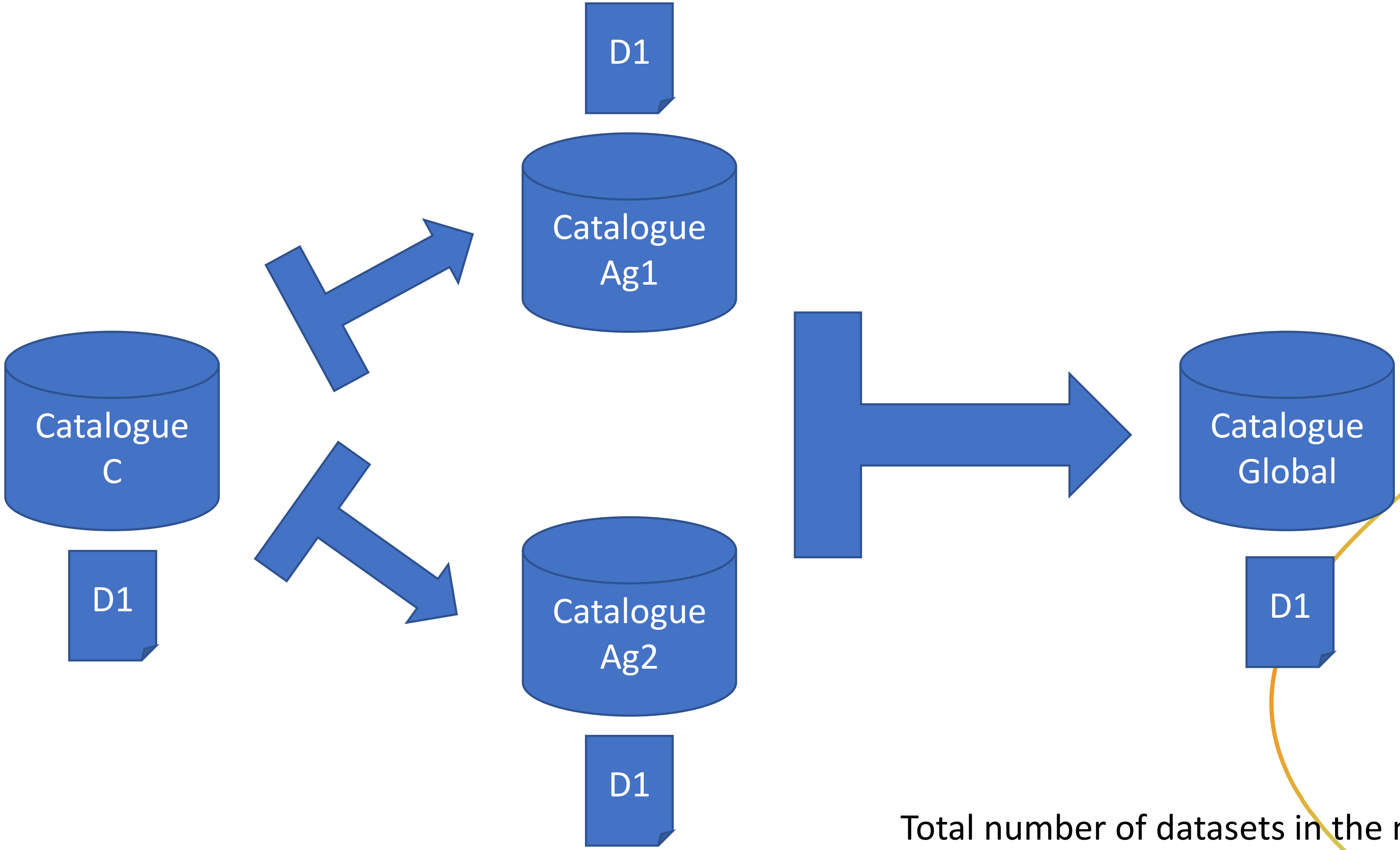
These local specific identifiers are usually also published (through the data portal API and UI).

Let’s consider these as other names/identifiers given by the aggregator and share them through the catalogue network.

Harvesting network

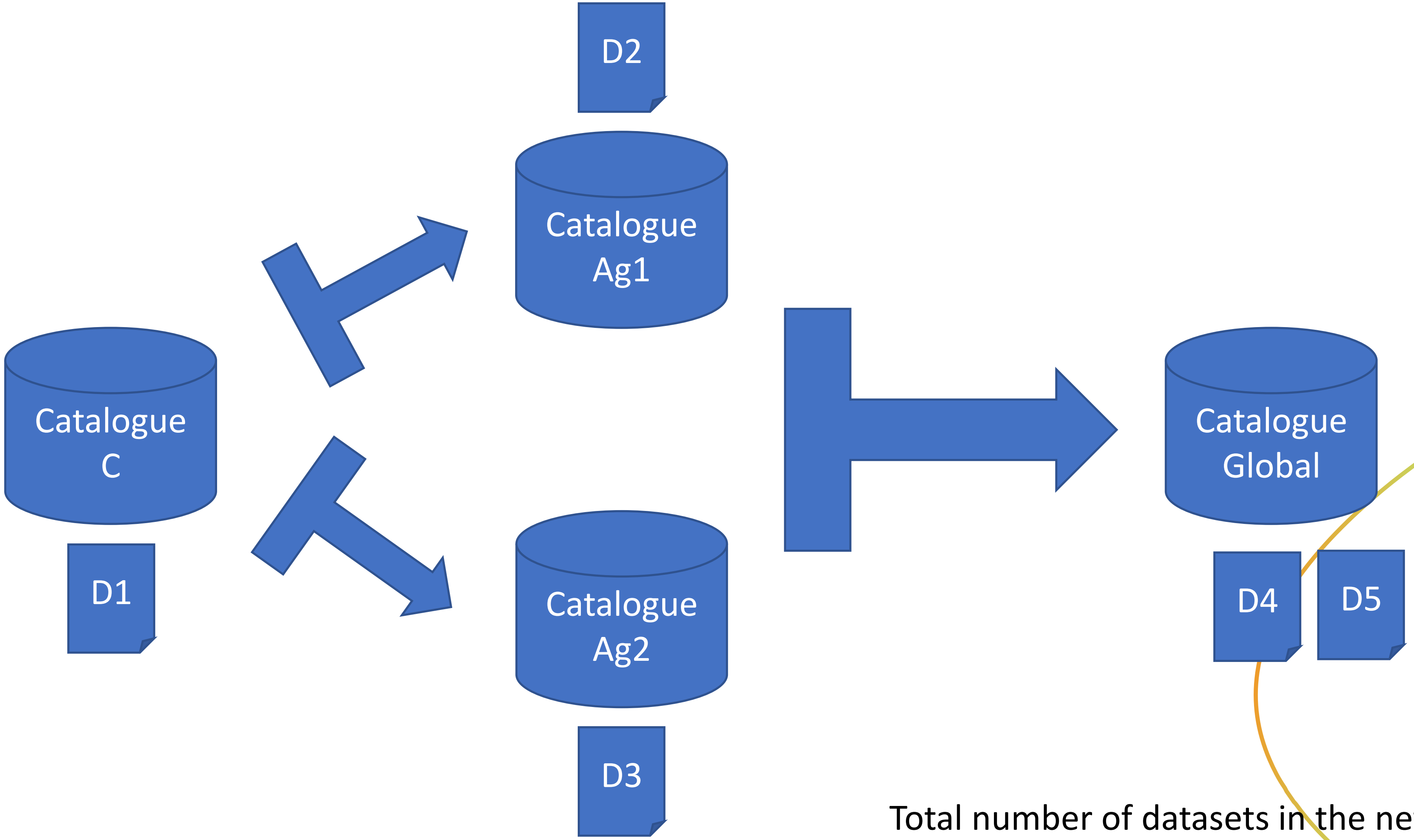


Harvesting network: expected behavior



Total number of datasets in the network = 1

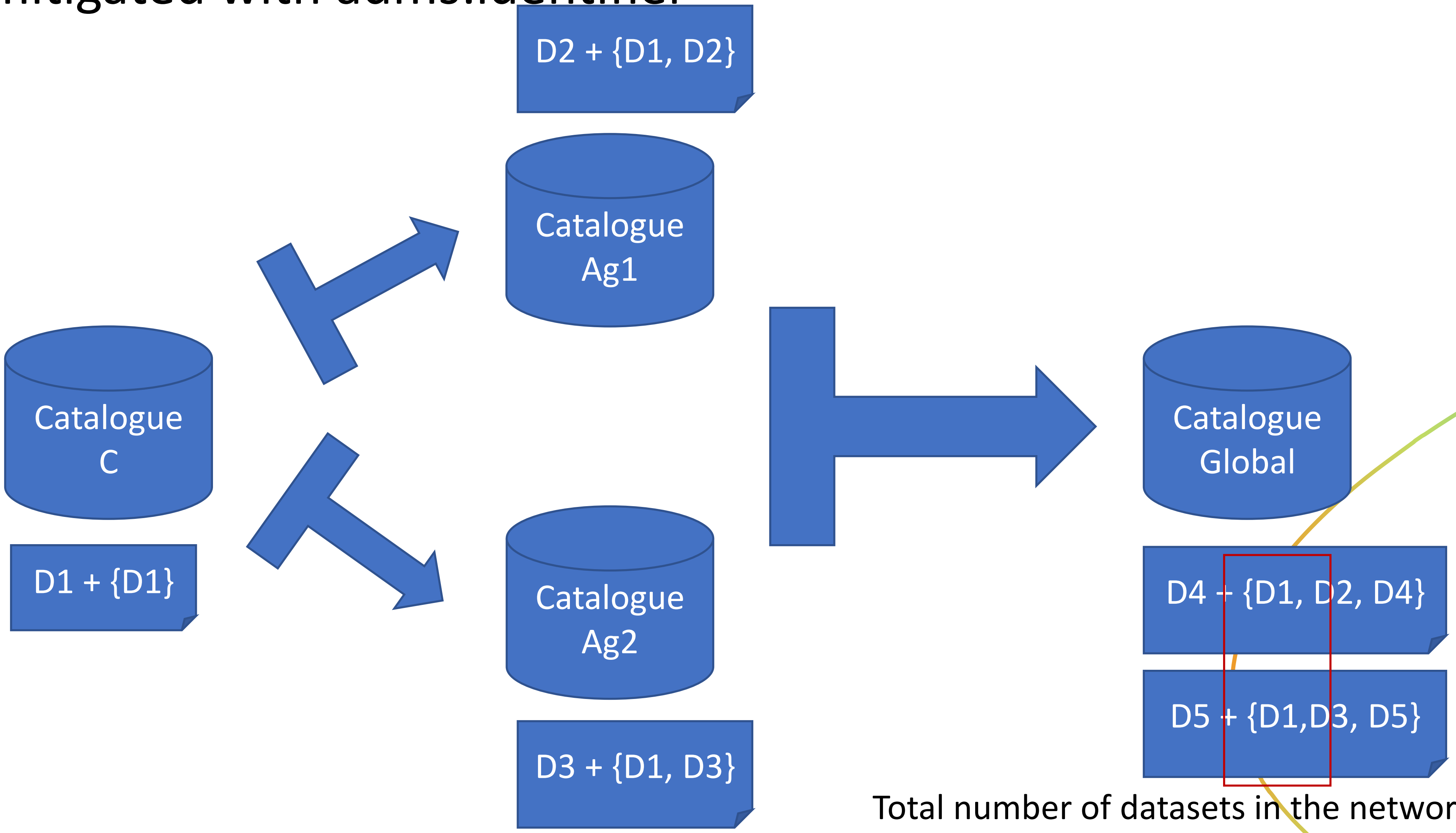
Harvesting network: undesired behavior



Total number of datasets in the network = 5

Harvesting network: undesired behavior

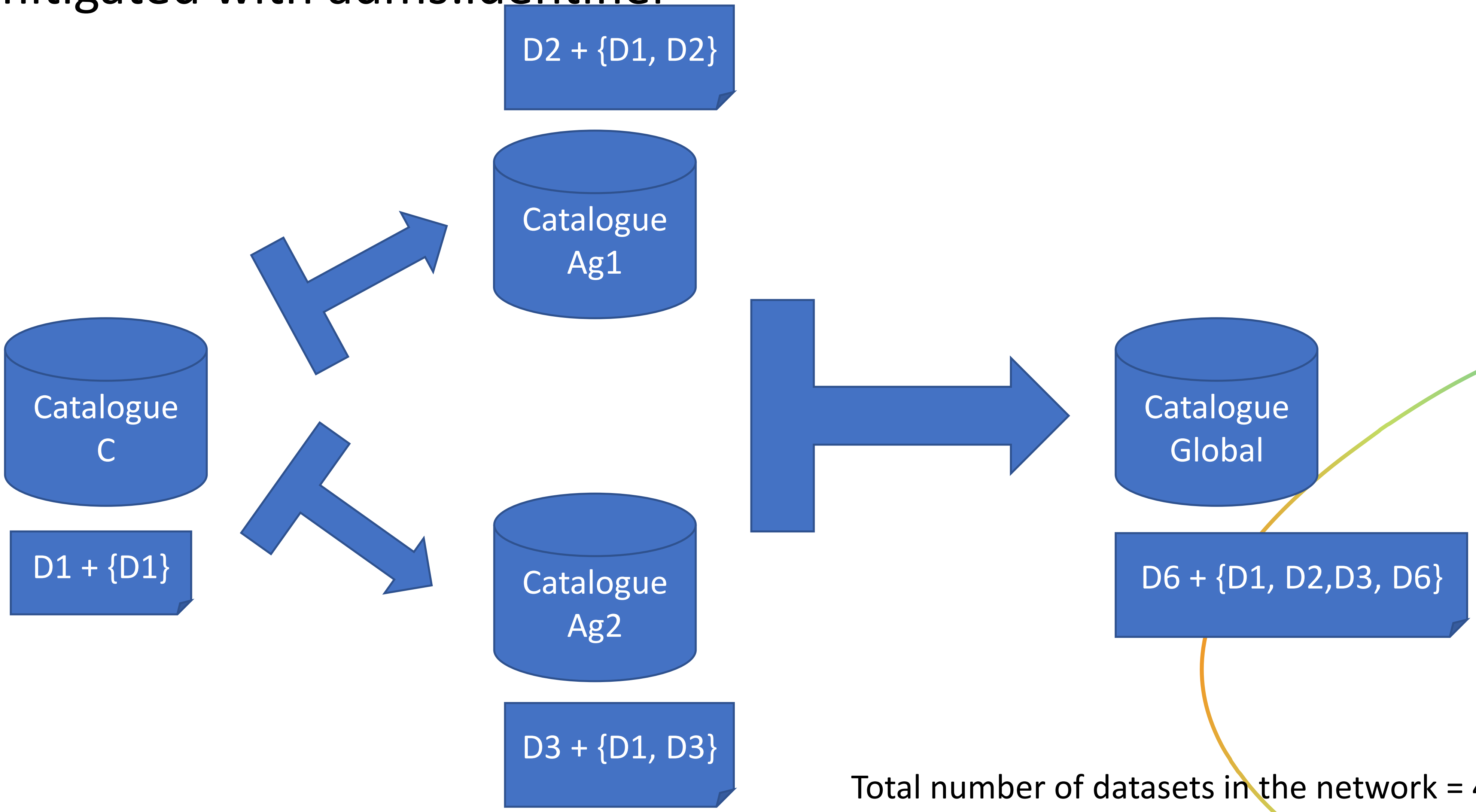
mitigated with adms:identifier



Total number of datasets in the network = 5

Harvesting network: undesired behavior

mitigated with adms:identifier



Total number of datasets in the network = 4 -> 1

Usage Proposal main identifier

Proposal

Use dct:identifier to indicate the identifier assigned by the publisher/owner.

Motivation

- dct:identifier is a literal, without context and ownership
- dct:identifier then provides a clue to users of which identifier should be used preferably
- Its creates an incentive to publishers/owners to assign persistent identifiers

- Ambiguity in the definition is removed, so more reliable decisions can be taken.
- *Provides a shortcut to the original identifier*
- Provides a purpose to dct:identifier w.r.t. the adms:identifier proposal

(based on our discussion on 10 March 2022)

Proposal

Property label	URI	Range	Cardinality
Main identifier	dct:identifier	Literal	0..1

Definition	Usage Note
The main identifier for the Dataset	the value is assigned by the owner/publisher of the Dataset. *Use of a persistent identifier (e.g. DOI) is recommended.*

Changelog

- usage note change: "This property contains the main identifier for the Dataset, e.g. the URI or other unique identifier in the context of the Catalogue." -> "the value is assigned by the owner/publisher of the Dataset“
- Max cardinality: n -> 1.

DISCUSSION

Proposal impact

It is a **semantic change** because it eliminates one option.

In the catalogue network, the purpose and use of dct:identifier becomes uniform.

Catalogues which place their own uniform catalogue-specific identifier as unique value in dct:identifier are heavily impacted.

Grey zone are catalogues/publishers that do not provide dataset identifiers:

- Strict reading of usage note: dct:identifier must be provided
- Open reading of usage note: dct:identifier can be provided by the first catalogue that introduces the dataset in the catalogue network.

Without a universal enforcement of a specific dct:identifier representation, it is difficult to make it stricter.

RDF format guidelines

RDF expectations on identifiers

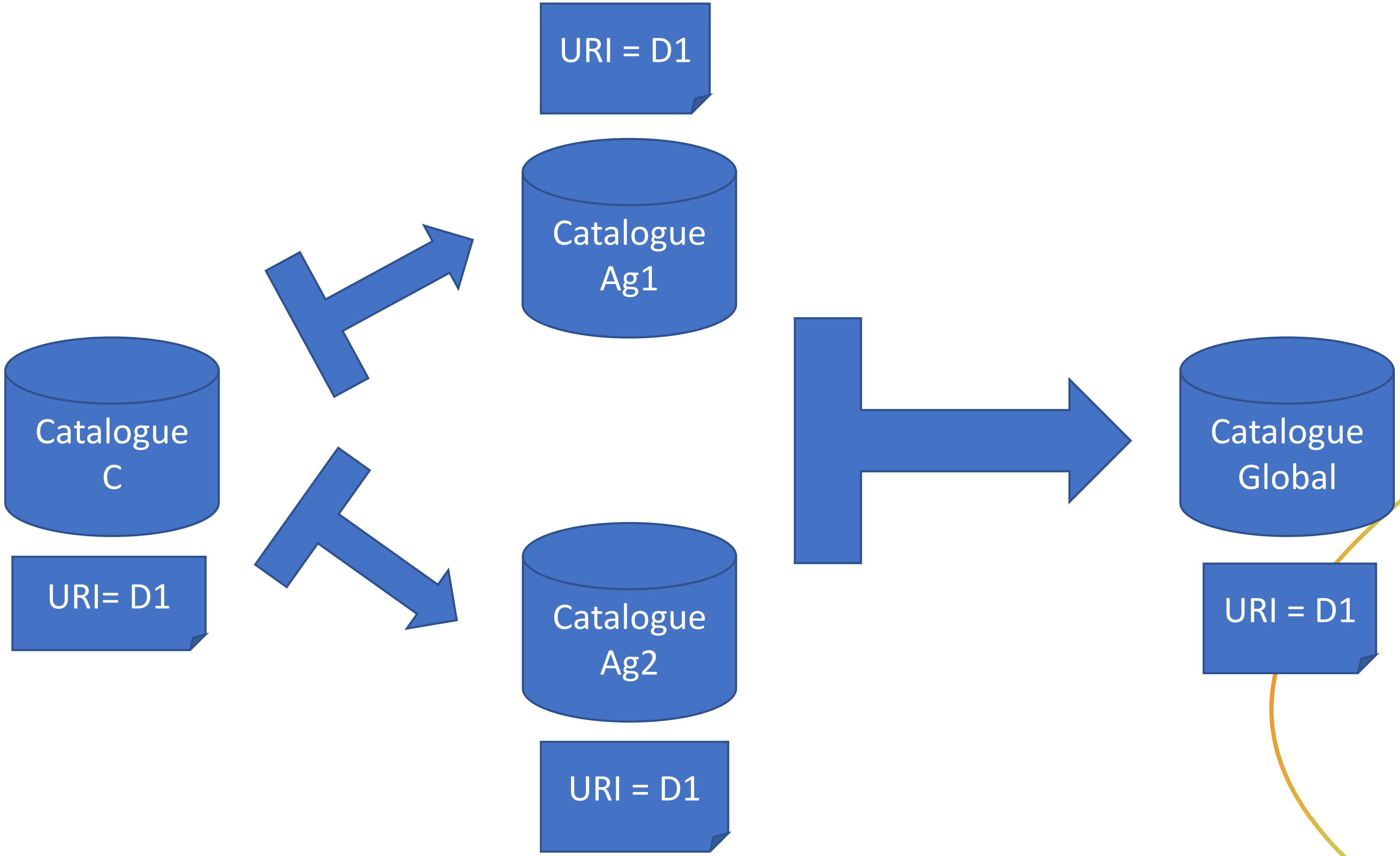
- (explicit) a node in an RDF graph is either named (in the form of a URI) or without an identifier (blank nodes).
- (implicit) URIs are preferable stable, persistent and dereferenceable
- (implicit) when processing RDF graphs a named node does not change name, but blank nodes do.

Merging two RDF files

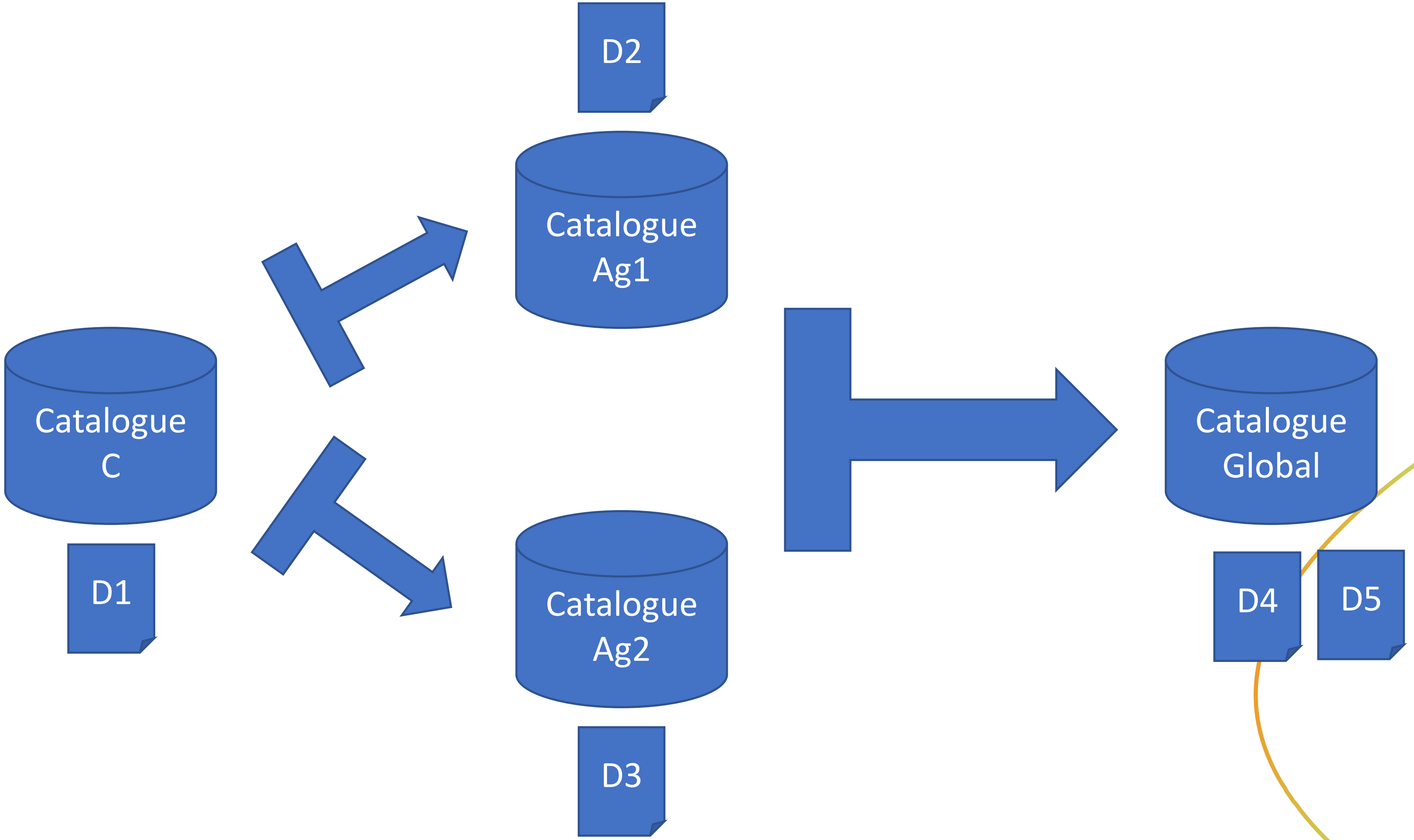
- The information associated with the same named node is fused together

- The information on a blank node is treated as independent entities and not fused together.

Harvesting with named nodes



Harvesting with blank nodes



Proposal RDF guidelines

- Use URIs for RDF nodes as name in case the catalogue also wants to share this as a persistent identifier.
  - Cfr dct:identifier proposal
  - Cfr advice to assign this as early as possible
  - Cfr do not treat URIs as blank nodes

- Postprocessing based on adms:identifier advised to reduce the injection of unintended copies in the network:
  - Sharing identifiers strengthens the catalogue network

Usage Proposal
Applicable to entities

change

Applicable to which entities (#141)

DISCUSSION

The previous discussed guidelines should apply to the following entities:

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

Possibly for

- Distribution
- Agent

- Catalogue
- Catalogue record

Next steps

Wrap up

- No discussion on the representation of an identifier
- No enforcement of the use of persistent identifiers, but highly recommended.
- Beneficial to the catalogue network, but not complicating the catalogue implementation itselves
- Works for exchange in RDF format as for non-RDF formats, even makes the RDF format exchange expectations clearer
- Broadly applicable

Next steps

Consider these changes as a **bug fix release 2.1.1,** as immediate impact is low.

Approach:

- Adapt the guidelines to the outcomes of this webinar.
- publish a draft release on GitHub for public review (short period)

Planning:

- Public review: during May 2022
- Release bug fix: June 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

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.

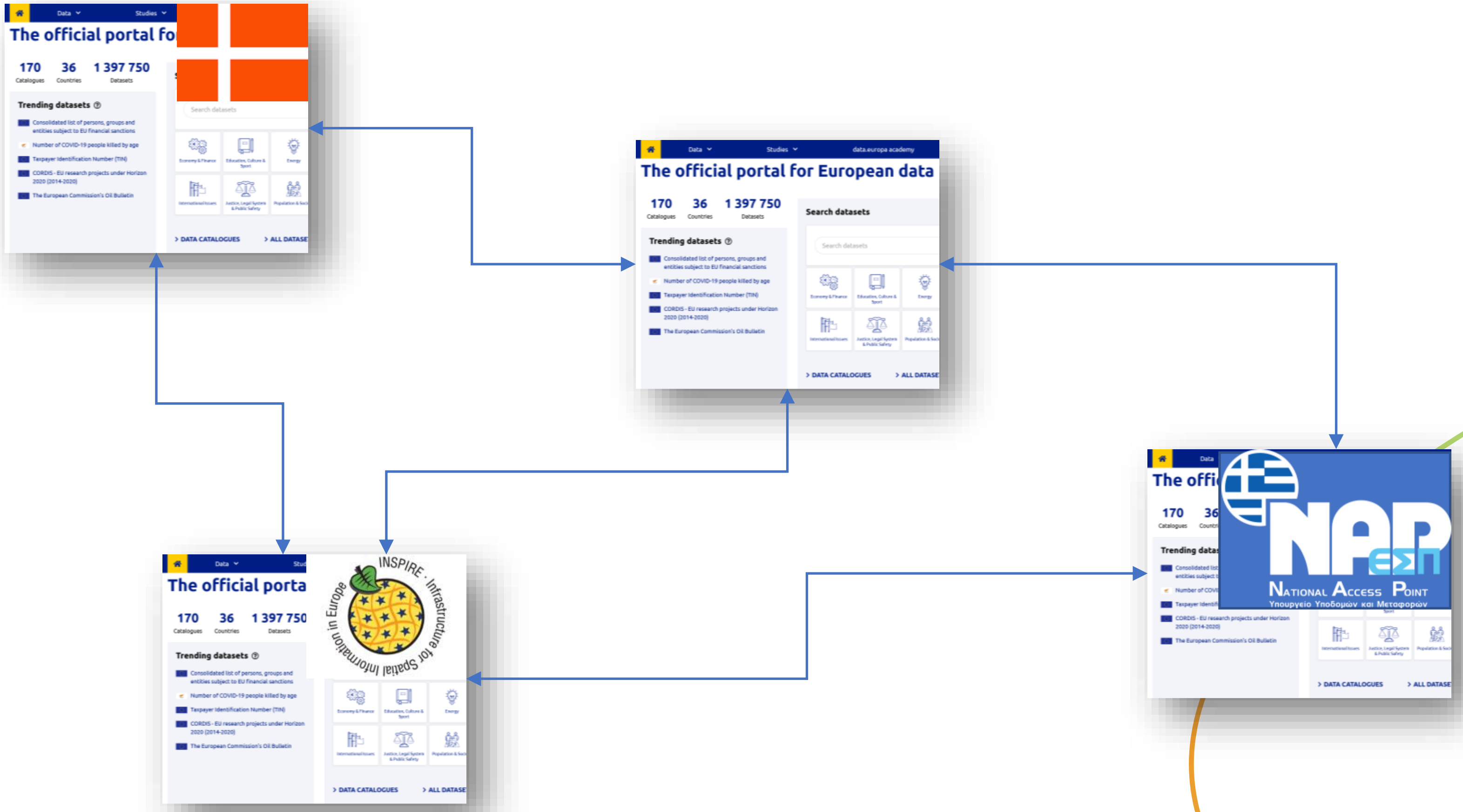
Use cases – harvesting

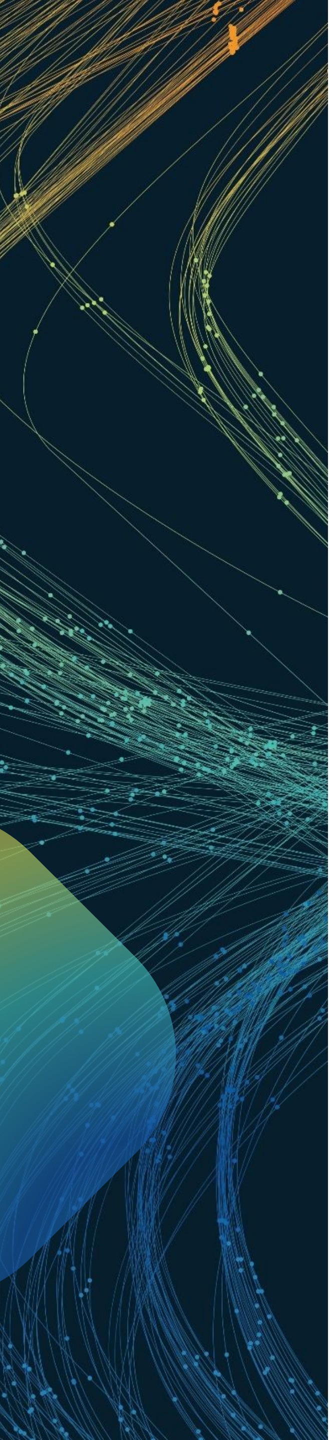
Harvesting is the process of aggregating source catalogues into a singe larger catalogue

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





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.

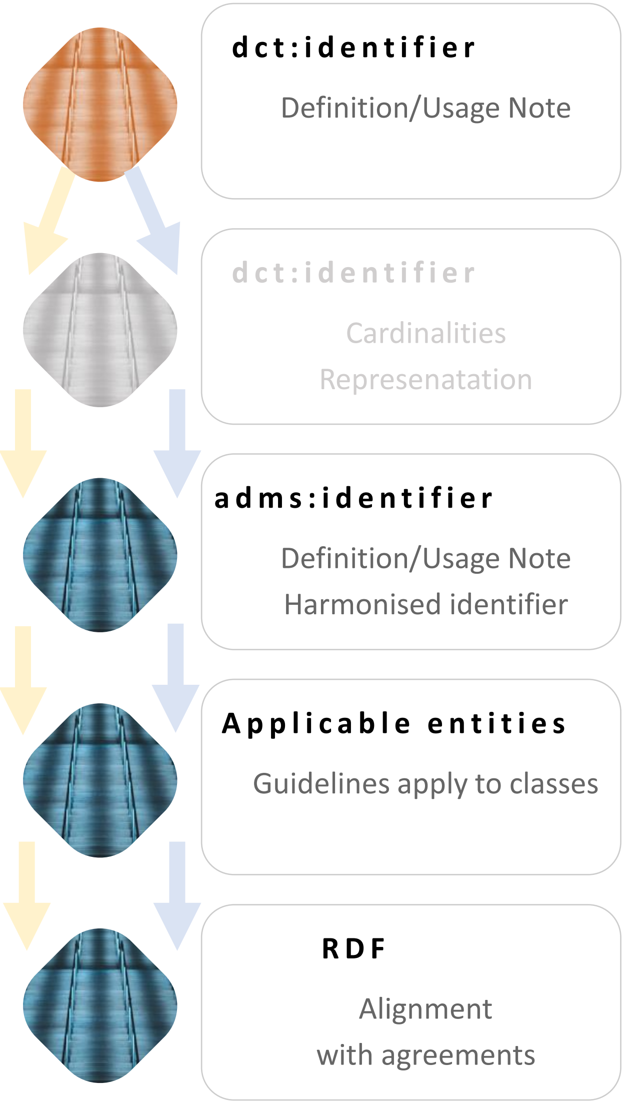
Identifier properties in DCAT(-AP)

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.

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