Meeting Minutes – Webinar 10/03/2021

DCAT-AP Webinar Identifiers

Project	Action 2016-07 Promoting semantic interoperability amongst the EU Member States	Meeting Date/Time	26/04/2022 15:00-17:00 PM (GMT+1)
Meeting Type	Webinar	Meeting Location	Cisco Webex Meetings
Meeting Coordinator	Makx Dekkers	Issue Date	11/05/2022

Meeting Agenda
1. Welcome 2. Propose a consolidated overview of the blue track guidelines 3. Exchange of ideas on those guidelines to improve the knowledge graph on datasets and data services 4. Next steps

Attendee Name		**Organisation/Country**
Alberto Abella	AA	Spain
Aleksandra Lavreneva	AL	PwC
Alexey Lukashov	ALu	Cogni.zone
Anders Friis-Christensen	AFC	European Commission
Andrea Perego	AP	European Parliament
Anja Litka	ALi	Germany
Anja Loddenkemper	ALo	Germany
Bert Van Nuffelen	BVN	TenForce
Casper le Gras	CLG	Netherlands
Costas Simatos	CS	European Union

Dirk De Baere	DDB	Belgium - Flanders
Fabian Kirstein	FK	Germany
Fabrice Gouzi	FG	Switzerland
Fidel Santiago	FS	European Commission
Hagar Lowenthal	HL	Publications Office
Honza Förster	HF	Cogni.zone
Jesper Zedlitz	JZ	Germany
Jill Saligoe	JS	ArcGIS
Judie Attard	JA	Malta
Kees Trautwein	KT	Netherlands
Kestutis Andrijauskas	KA	Lithuania
Konstantins Bogucarskis	KB	Greece
Kuldar Aas	KAa	Estonia
Makx Dekkers	MD	Independent consultant
Lin Zhang	LZ	?
Ludger Rinsche	LR	Germany
Marco Combetto	MC	Italy
Matthias Palmér	MP	Sweden
Moritz Herter	MH	Germany
Nina Georgieva	NG	Bulgaria
Oystein Asnes	OA	Norway

Pascal Hurni	PH	Sweden
Pavlina Fragkou	PF	European Commission
Sander Van Dooren	SVD	Belgium
Stig B. Dormaenen	SBD	Norway
Thomas Weber	TW	Germany
Tod Dabolt	TD	U.S.

Summary of the meeting

Topic	Summary
Welcome	Pavlina Fragkou (PF) opened the webinar by welcoming the participants, and presented the motivation of the webinar, based on the decisions made in the previous webinar, by stating that alignment on the expectations on the usage of identifiers has been ● Identified as an implementation issue (already for long term), ● Set by DCAT-AP focus groups as an important priority, ● Enforced by emerging dataspaces.
Introduction	Bert Van Nuffelen (BVN) presented the agenda for the webinar: ● State of play: recap of previous webinar ● Proposals by the editorial team ● Next steps
State of play	BVN gave a short recap of the previous webinar, and presented the proposal of the editorial team, based on the decision made during the previous webinar. BVN highlighted that identifiers design principles are around the concepts responsibility, persistence and dereferenceability. BVN presented the use cases for identifiers, being that identifiers are useful for processing, networking, portal development and harvesting. BVN highlighted that there are already existing guidelines, but they are not sufficient and do not meet the expectations and challenges experienced in the past. He added that DCAT-AP has two properties

	for identifiers, dct:identifier and adms:identifier, based on which the editorial team will make proposals.
Proposals	BVN presented the proposal and the approach, based on the discussion during the previous webinar. It is written out on Github: https://github.com/SEMICeu/DCAT-AP/blob/2.1.1-draft/releases/2.1.1/usageguide-identifiers.md BVN presented the flow of discussion: • Share metadata on identifiers: adms:identifier • Examples scenarios • Main identifier • RDF format guidelines • Application **Share metadata on identifiers: adms:identifier** BVN presented the proposal of the editorial team, which is using adms:identifier to describe metadata about identifiers. The reason for this is that dct:identifier is just a literal without any context or ownership, adms:identifier allows us to provide this information. Furthermore it can become an ever growing collection of identifiers assigned. Concretely the property adms:identifier already exists. The proposition is to change the label "other identifier" to "identifier" in order to make clear that this would be the primary used identifier. Adrea Perego (AP) asked for clarification about the proposition for the text of the revised usage note. • Makx Dekkers (MD) replied that the new usage note will be "This property refers to each identifier that a catalogue or a process assigns. Alberto Abella (AA) asked whether the editors will talk about uniqueness later on. • BVN responded that here it is not about uniqueness or persistence, but the origin of the identifier. • AA requested to clarify whether the purpose is to have some grammar for building the identifier, by having its origin. • MD noted that this is not what the editors mean, as this is left up to the creator of the identifier. The editors just propose to include metadata of the identifier.

BVN presented the 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.

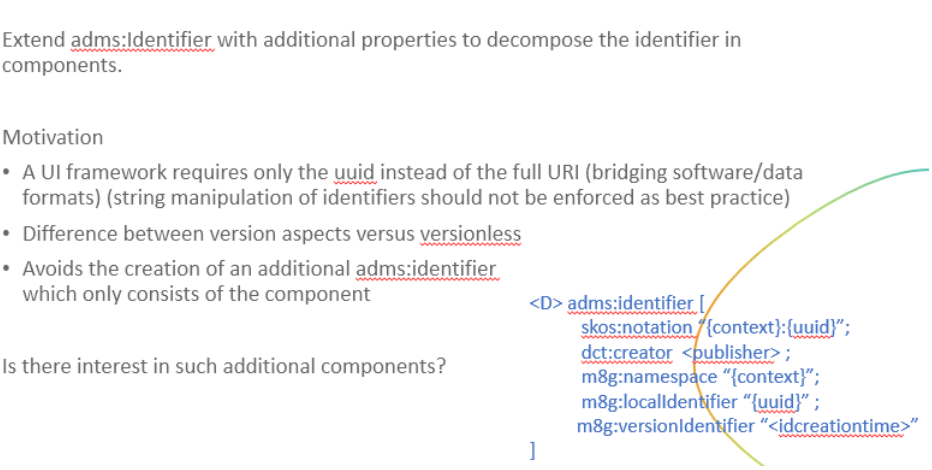
Ludger Rinsche (LR) commented that setting the attribute to 1..1 seems too strict. Even though it might be useful, it will take a lot of time to implement that by all data providers.

- MD responded that if metadata needs to be available, this should be provided.
- BVN added that if it is not provided, there is no difference between adms:identifier and dct:identifier.

Sander Van Dooren (SVD) commented that it could be a very personal thing, but that he is not comfortable with altering data that is federated. He would prefer to keep the record intact, and 'attach' everything that is added along the way external to the entity itself.

- MD responded that Sander's comments will be noted down.

BVN presented the additional components:



BVN presented the impact of the proposal, which is that all identifiers should be included, meaning both the value of dct:identifier as well as the RDF URI.

Matthias Palmer (MP) asked why there is a need for all this complexity. "*We already refer to datasets using their URIs, e.g. between datasets via properties like dcterms:relation, dcterms:hasVersion, dcterms:isVersionOf etc. The URI provides dereferencablitiy, ownership is implicit via the hostname in the URI being used, and uniqueness is a design issue in minting the URIs. He added that URI's of the datasets can be used when referring to other datasets. There is a well-established mechanism for this. This means that the URI will contain the owner of the dataset, being the owner of the URI.*"

- Makx Dekkers responded that this is not always the case unfortunately. Therefore with this proposal the editors try to make it clear what these identifiers are, because the ideal situation does not exist as people change data during federation.
- MP asked whether this is not something we should argue against, as this might cause more complications. He requested more information on how this will be used.
- MD responded that this will be presented later by BVN.
- Ludger Rinsche (LR) commented that as soon as the harvesting portal is adding its own meta-data, using a new URI for the copied and enhanced dataset can be useful.
- MP responded that it will break all links in RDF, unless they are updated in the harvesting process as well. In Sweden URI's are only changed if the dataset does not have a URI, e.g. if it is a blank node. In that case the harvester creates a

URI in a deterministic manner from a few of the available properties, which is basically "repairing".
- LR mentioned that data.europe assigns new URIs. It is a question of fate, and a question whether the copied dataset in the harvesting-portal is the same thing as the dataset in the harvested portal.
- MP replied that it should be the same dataset. He presumed that LR is treating the RDF as a mere container format along the lines of XML or JSON, which it is not. RDF is a language independent of the format that makes statements about things and changing the URI makes the semantics different.
- LR responded that if the concerned URI would be for a book or a human, he would completely agree with MP, however with a collection of metadata within a portal, it's not that straight forward.

Jesper Zedlitz (JZ) agreed with the need for an additional identifier, because when harvesting a catalogue a new URL is assigned, and cannot use the original one as it will lead to another catalogue. There is a need to somehow store the new URL and the original URL, so in this case you will need additional identifiers.

- MD responded that this will be shown by BVN later on.

Kees Trautwein (KT) commented that UI addition should not be part of the standard. Because then decisions are made about the actual technical implementation of all parties using the standard. That is not correct. For every implementation it is always possible to add information for the UI. Could be included as a note, but not as a standard.

- BVN responded that if an UI framework says that the identifier must be all integers, you need to assign integers to the knowledge you have inside. When starting building public identifiers and sharing on this, this becomes part of the information cloud and makes it easier to share it and add it to a knowledge base.

MP commented that new URIs for landing pages in a portal should not necessarily be considered a new URI for the dataset, it is just another landing page for the dataset.

- Fabian Kirstein (FK) responded that relying on this is not possible as data providers give us datasets for which we

create our own URLs and therefore need another way to refer to these original datasets.

FK commented that the need to store the original identifiers is very complicated, because as SVD said we do not change the metadata itself. We would need to keep track of the changes and when harvesting again changes need to be tracked which is quite complicated. Therefore the DCAT dataset class should not be changed, but only updated.

- MD asked whether FK is proposing to put the information in the catalogue record instead of the catalogue itself.
- FK resplied that that is also not ideal, but it might be a better approach. It's always difficult to find the best way to do it.

JZ commented that data providers are required to change their own identifier. If dct:identifier as an adms:identifier needs to be added, then it would not need to be edited after harvesting data.

- BVN responded that that is the idea behind the proposal. It is what the editors hope the publisher will do closely to how the original publisher created it.

BVN concluded that there is a lot of concern about not changing the metadata. This raises the question how you would face any enrichment.

Makx Dekkers added that there are people who already said that they think no one should touch the metadata which is being harvested. There are also people here who say that it is good to add things as we go along.

AA commented that here only situations where harvesting is something that is done from time to time and not continuously are considered. Thinking about static datasets might not be the future.

- MD asked whether AA is suggesting that later on the editors will also look at the way tools for continuously changing need to be provided. He added that something is planned for that in the future.
- AA clarified that there is a need to design something that not only resolves current issues. He provided an example of a study in Spain: all the open data portals were visited to know how many datasets are published in Spain. This needed to be done manually. It resulted in about 58.000 datasets, because there is no way to know whether something is federated or

not. There is a need to be able to provide an identifier that really identifies the origin.

BVN presented the enforcement:

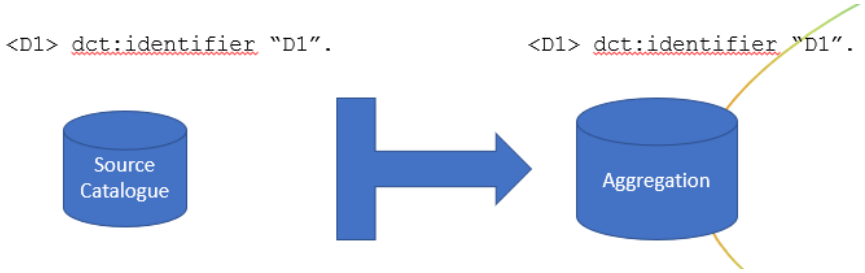
- SHACL validation rules can check if the value of dct:identifier is part of adms:identifier notation
- and to be investigated the URI is a part of adms:identifier
- but cannot check the propagation/sharing aspect.

Examples scenarios

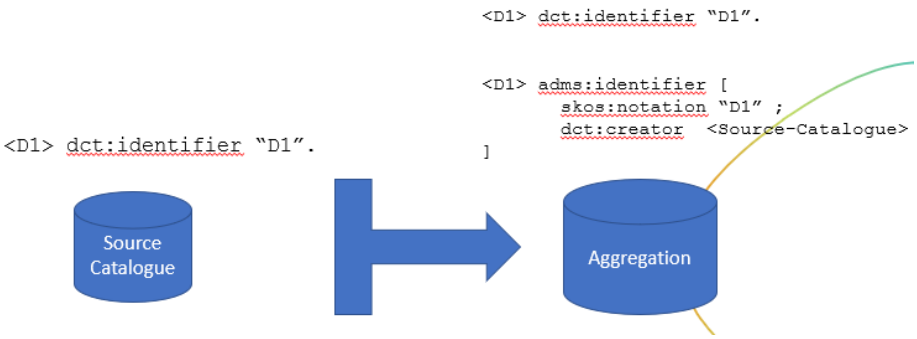
BVN presented example scenarios to make the proposal clearer.

1. Harvesting a catalogue:

   Simple copying is not the advice:

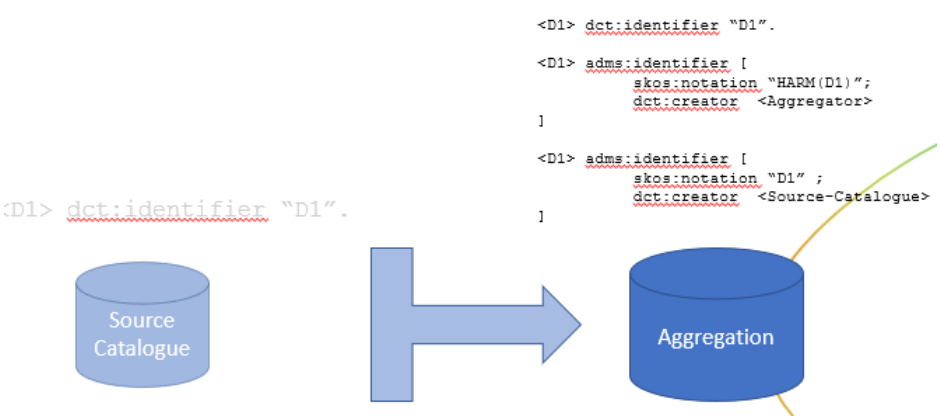


   Harvesters are advised to complete the adms:identifier list for the harvested datasets.



2. Aggregation catalogues:

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

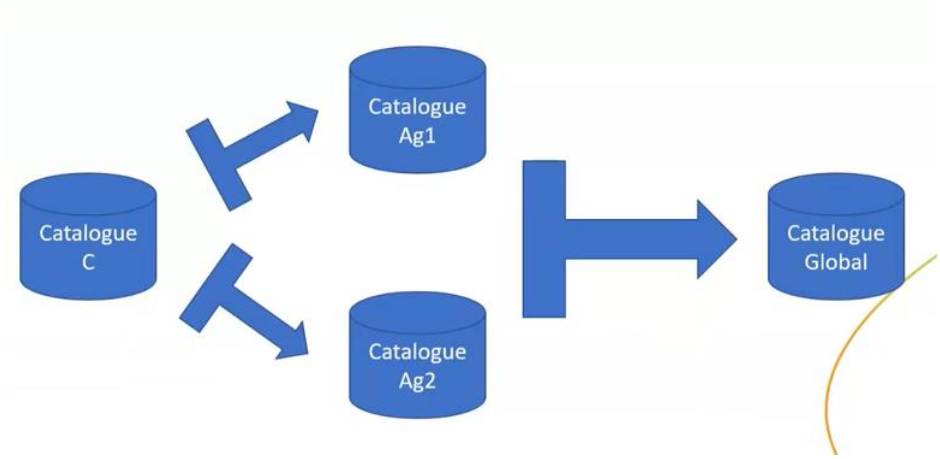


Aggregating catalogues want to have uniform identifiers in their catalogue in order to support querying.

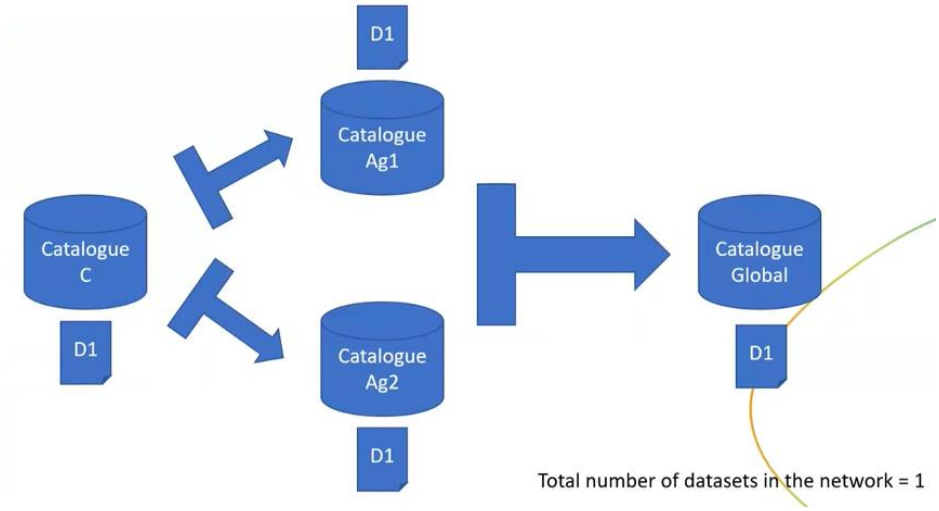
These local specific identifiers are usually also published (through the data portal API and UI).

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

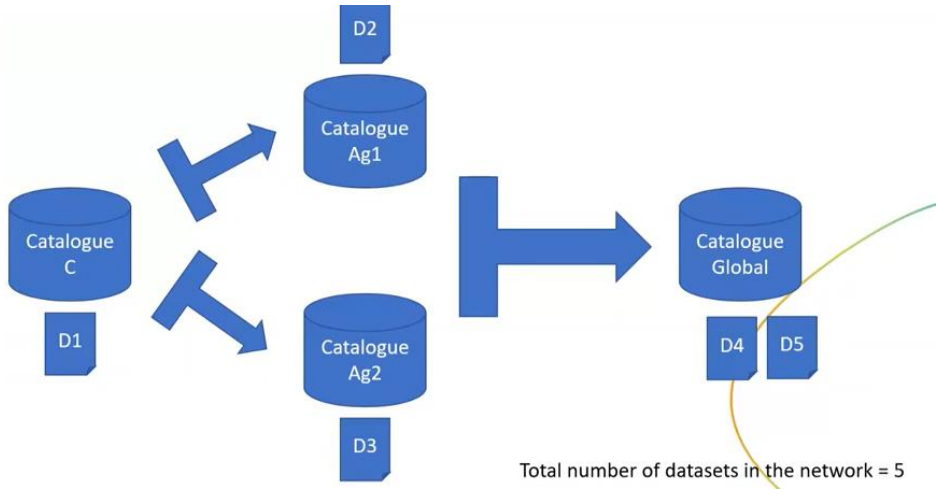
3. Harvesting network:



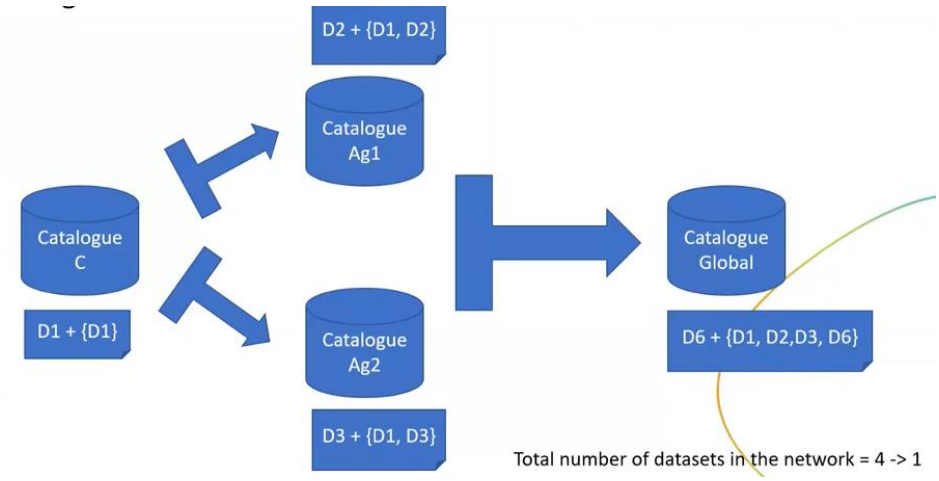
Desired behaviour:



Undesired behaviour:



Mitigated with adms:identifier:



Stig B. Dørmænen (SBD) mentioned that he is not familiar with the term "Aggregated catalogue". He asked whether it is this a different type of catalogue, and how you can tell if a catalogue is an aggregation, and another is not.

- AA responded that currently this is not possible.

FK commented that what is described is provenance information. We should not store this in the identifier. For building this kind of a harvesting graph, maybe we should think about another property or use provenance property to store this.

- MD agreed that a chain of provenance is being created. Unfortunately there is no possibility to do anything with provenance so the editors are trying to do this with identifiers, as currently identifiers do not work this way.
- BVN added that today with the identifiers it is not possible to do what was presented above. The proposal of the editorial team is to make this more formal, and if everyone follows it (it is little effort on your catalogue's side), but every catalogue that takes it can make this fusion decision.

Hagar Lowenthal (HL) shared her experience at the Publications Office as they have an extensive use of identifiers. It was decided to use rdf:about as identifier and explicitly map it to dct:identifier. The Publications Office use the adms:identifier for DOI's which allows them to assign a priority of which identifier to use for example for citation.

- MD added that there is a second proposal about dct:identifiers which will be presented later. The use of adms:identifier for DOI's is not in conflict with this

proposition. With the proposal here for each identifier it would be clear what the identifier is, and where it comes from. There is no reason that adms:identifier cannot also be used for persistent identifiers like DOI's.

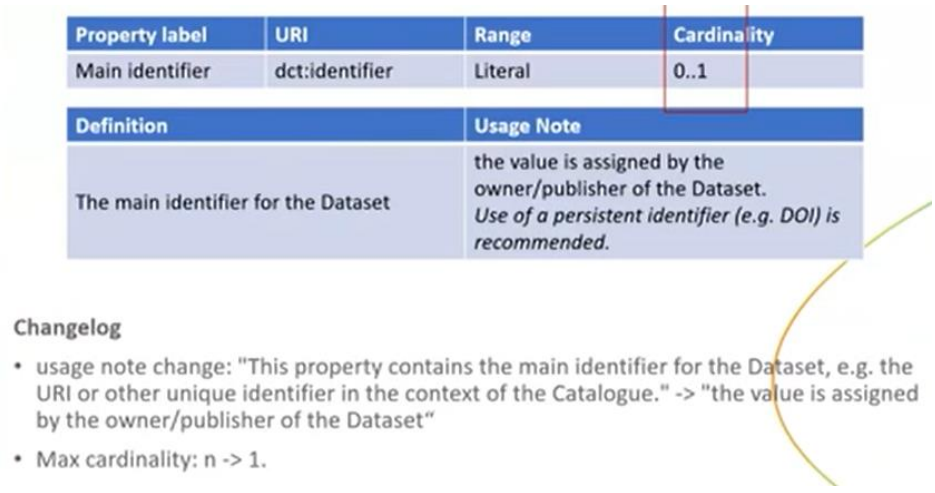
- BVN mentioned that the goal is that there should be no preference for the adms:identifier. There is no prescription that says what it should be, and in the past there was an assumption that dct:identifier was the one to be used. Now the editorial team only says that you can use the adms:identifier in the format you want, but report and share the agency. The benefit is that everyone can decide for their own how to do the referencing.

MD referred to MP's comment by highlighting that the editors are trying to repair what we might have done wrong in the beginning, and therefore try to have a clear way of using identifiers.

Usage proposal main identifier

BVN mentioned that the current definition and usage note provide two options. In the previous webinar it was decided to use dct:identifier to indicate the identifier assigned by the publisher/owner. In parallel to adms:identifier, you can have a dct:identifier which you can promote to your users by making it the preferred identifier to be used. This creates a shortcut to the original identifier.

BVN presented the usage proposal for the main identifier:



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.

Andrea Perego (AP) suggested that the current definition should be a bit clearer. Just reading this without any context, it is not clear that this should be assigned by the original publisher of the dataset, who is not necessarily the owner.

- MD clarified whether AP is proposing to remove the ambiguity in the usage note.
- AP confirmed that this is the suggestion.
- MD responded that the usage note will be changed correspondingly.

BVN presented the impact of this proposal:

- 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.

LR asked why the second change has more impact, as in the first change mandatory aspects are being added and it makes every existent use of adms:identifier invalid.

- MD responded that all the current usage of adms:identifier would remain completely valid.
- BVN added that the need for additional metadata should not be a problem, because if it is assigned and created as an adms:identifier, it should be known who the creator is.
- MD replied that it could be that if people assign an adms:identifier without metadata, then indeed implementation would not be in compliance. The editorial team needs to take this into account to see how metadata can be taken into account. He further added that in the first change elements are being added, when in the second case the change is more disruptive as people would need to change their approach.

- LR commented that he would agree with the approach if the first change would be back-log compatible.
- MD responded that providing this information is more useful, than not adding this information. He added that the editorial team will take LR's comment into account and make this clear.

MP asked whether dct:identifier would be unique within the catalogue instead of globally unique.

- BVN responded that in the current specification for dct:identifier it is allowed to be used as the identifier within the catalogue. This means that if two sources are harvested, dct:identifier would need to be replaced with the value used in that catalogue.
- MD asked whether it would be useful if the usage note would say that it should be globally unique.
- MP agreed with MD's proposition.

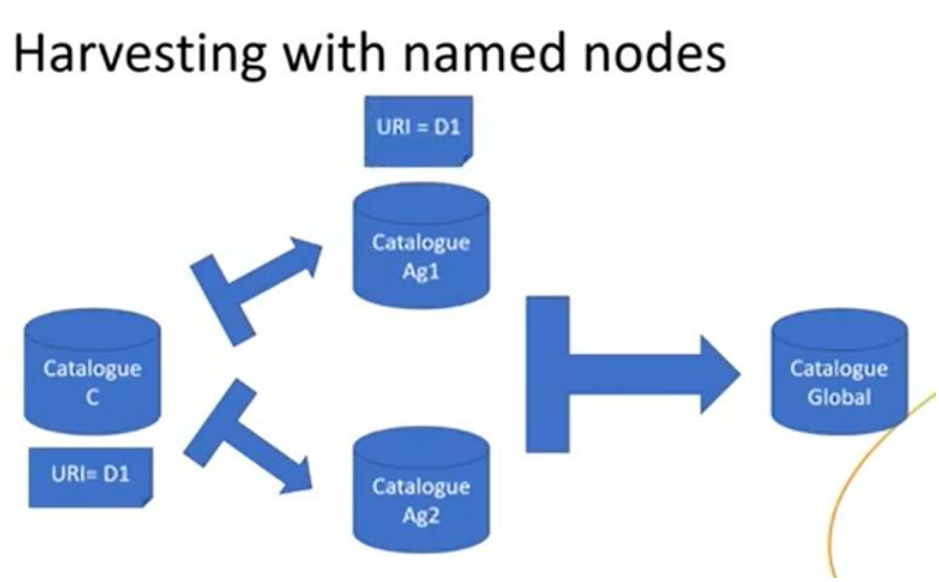
RDF format guidelines

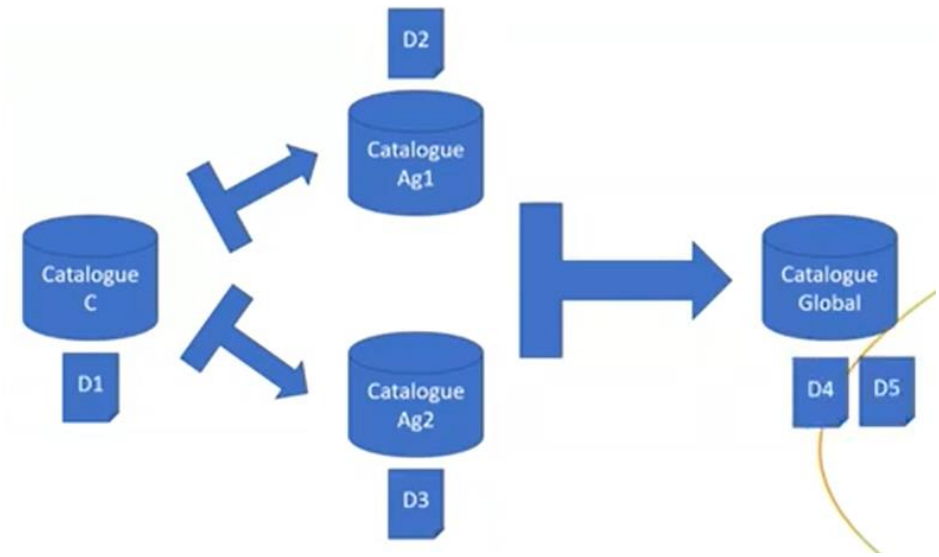
BVN presented the RDF expectations on identifiers:

- (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.

When merging two files, the impact would be that the information associated with the same named node is fused together, and The information on a blank node is treated as independent entities and not fused together.

BVN clarified with an example:





BVN presented the editorial team's proposal regarding using RDF guidelines for identifiers:

- 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

Matthias Palmer commented that blank nodes are good for some position in RDF-graph like time period etc., being only the main classes that should be main URI's regarding the DCAT2 recommendation.

- MD responded that DCAT-AP would have the same recommendation too, in cases where people use RDF.
- SVD agreed that blank nodes should not be used for datasets.

Application

BVN proposed that the previously discussed guidelines should apply to Datasets and Data Services, and asked the opinion of the working group regarding this.

MP suggested using it for URI's, Distributions, Agents and Catalogues. He added that it can create problems when there are duplicates of Agents. If a catalogue with blank nodes is being harvested, there are ways to detect these by using a stable alternative instead of a black node ID which is changing.

- MD responded that it is clear that there are indeed other ways to detect duplicates which are not based on identifiers but on other parts for the metadata.
- JZ agreed with using Distributions as blank nodes.
- SBD highlighted that scolonisation is also possible when creating URI's for blank nodes.
- SVD proposed to also apply it to Resource.
- MD responded that in DCAT-AP Resource is not a dedicated class.
- FK commented that for the data services that are part of distributions it might not be recommended to use URI's but to use blank nodes, because it has very limited properties and using a stand-alone URI might not make sense.

Maxk Dekkers summarised that the editorial team needs to look into conducting an implementation plan, with a timeline in mind to implement these proposals.

- BVN concluded that the change towards adms:identifier was experienced as more invasive than expected. It would require much more changes than expected beforehand.
- MD added that there are two groups of people:
    - One who thinks it would be good to repair what went wrong in the past.

	○ A second that would not bother with this repair and go for the proposal with dct:identifiers and rdf:identifier. ○ In general people think it is a good proposal, but it would need some effort to implement this. MP suggested to have a discussion on what happens when you change URI's and break links. There are references between datasets, and it can be that datasets are referenced across catalogues. If the URI's are changed, this link might be broken. ● MD clarified whether MP means that there is a need to look at what to do with identifiers in terms of relationships. ● MP confirmed. ● MD responded that this will be put on the list of discussions for further consideration. Makx Dekkers proposed to open a discussion on Github so that the working group can provide feedback on how and when to implement the proposals. Furthermore we need to think about the use of identifiers in referencing. HL proposed to look into using owl:sameAs. ● MD replied that the editorial team will take this suggestion into account and look into making a proposition regarding this. SBD shared an interesting link in this context: https://www.w3.org/2009/12/rdf-ws/papers/ws21.
Wrap up & Next steps	BVN wrapped up the webinar with the following summary of proposals by the editorial team: ● 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 BVN presented the next steps following on the two webinars regarding identifiers: ● 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 to be confirmed once agreement with working group on the approach is reached. Pavlina Fragkou thanked everyone for their contribution and provided comments, and added that the editorial team will make the necessary adaptations based on the provided comments by the working group. If an additional webinar would be needed to wrap up all discussions, it will be communicated to the working group.