REPLY to nick

> If it's not feasible to provide standardized functionality for authenticity and integrity of DCAT files (or other distributions of the metadata) in the short term, then I think it would be reasonable to:
>
> 1. add a warning about the security implications of checksum properties when the metadata's authenticity has not been confirmed; and
> 2. list some ways to access DCAT metadata in an authenticated, secure way (downloaded over HTTPS from the expected origin, for example); and
> 3. mark it as an issue for a future version.
>
> Postponing features has to happen sometimes. But I would strongly recommend that there be a plan to address this in the future, rather than just postponing it as a way to avoid dealing with it.

Thanks, Nick, for your suggestions; we've included them in the Security and Privacy sections; check the second paragraph.

Please feel free to suggest improvements to the draft.

If you can live with the current draft, we will backlog this issue for further consideration in the next standardization round of DCAT (e.g., DCAT 4).

>I'm not convinced that it's wholly out of scope. One of the only features being added to this version is a checksum property, which is apparently intended to provide security protections, but doesn't provide the expected security protections if there's no way to provide integrity or authenticity of the DCAT metadata.

We have acknowledged that in the new paragraph.

>I'm not sure if the checksum property is fully defined enough that it can be generally interoperably used (is there implementation experience?), but that property assumes that there already exists a canonical way to refer to a distribution, if not a dataset.

This solution is adopted by DCAT-AP 2.1.0. The checksum property range in <code> spdx:Checksum</code> class, which specifies actual <code>spdx:checksumValue</code> and the <code>spdx:algorithm</code> used to produce the checksum.

DCAT distribution might be in many other formats than RDF. As for RDF, there is a Group working on the [RDF Dataset Canonicalization and Hash](), and we prefer to wait for their outcomes before recommending anything in that direction.

>Accessing datasets that could be tampered with, or not knowing the provenance or authorship or integrity of a dataset, is a real and significant threat; it affects far more than just the implementers of this spec. I don't think it can be our long-term plan that W3C Recommendations

don't provide any mechanism for basic, interoperable security properties and instead rely on the hope that every individual implementation or user will figure out its own way to provide security.

We agree that this is a pervasive and transversal issue that impacts every vocabulary the W3C recommends, and this is the main reason why the solution should be common to all vocabularies. RDF Dataset Canonicalization and Hash Working Group will likely provide a ground upon which RDF vocabularies will build. Anyway, any further input to consider in the next standardization round is more than welcome.

OLD STUFF

—-------------------------------

Suggestion from DXWG plenary

1 - state its not in scope of model,

2 - point to new community group

3 - thanks for feedback and note timeliness

4 - may rely on canonical serialisation and is a significant technical challenge

Thanks for the feedback. We discussed the issue of integrity and authenticity in the DXWG plenary [see https://www.w3.org/2022/10/11-dxwg-minutes].

The core of our work is DCAT as a metadata model, and integrity and authenticity seem to relate more to how DCAT is provided than the DCAT model itself. We are reluctant to address issues “Not at the core” of the group mandate. We want to avoid our DCAT-limited perspective can later conflict with more devoted solutions stemming

from new groups working on promoting transversal technology, which might be chosen to deliver DCAT metadata.

The RDF encoding as one of the most typical ways to serve DCAT provides an example of the above concerns. Typically, a DCAT encoding in RDF might end in an RDF store or file. In the case of an RDF store, it is the chosen software which needs to implement the caveat to ensure integrity and authenticity.
In the case of RDF files, other ongoing W3C groups such as the "canonicalizing and cryptographically hash of RDF Dataset" [1] deal with the integrity of RDF content. The existence of a dedicated effort shows the timeliness of your comments. If you think it might help, we can try to point at the ongoing initiatives in the DCAT document as the RDF Dataset Canonicalization and Hash Working Group. However, it seems reasonable first to wait for the RDF Dataset Canonicalization and Hash Working Group outcomes and check when their work consolidates if we can suggest adopting their recipes. Until then, we can move this issue to Future work - possible new requirements (i.e., https://github.com/w3c/dxwg/milestone/31).

Can you live with this strategy?

—---

Of course, If you have any other suggestions about remedies that can fall in the scope of the DXWG group, we are open to considering them. Nonetheless, suppose any new normative solution is required to be conceived, considering the stage in the standardization process we are now for DCAT 3. In that case, they likely need to be considered as requirements for the next standardization round (e.g., DCAT 4).

—-
REplay to https://github.com/w3c/dxwg/issues/1526
Thanks for the feedback. We discussed the issue of integrity and authenticity in the DXWG plenary [see https://www.w3.org/2022/10/11-dxwg-minutes].

The core of our work is DCAT as a metadata model, and integrity and authenticity seem to relate more to how DCAT is provided than the DCAT model itself. We are quite wary to address issues “Not at the core” of the group mandate, as we want to avoid our DCAT-limited perspective can later conflict with more devoted solutions stemming from new groups working on promoting transversal technology which might be chosen to deliver DCAT metadata.

An example of the above dynamics can be found considering RDF encoding as one of the most typical ways to serve DCAT. Typically, a DCAT encoding in RDF might end in an RDF store or an RDF file. In the case of an RDF store, it is the chosen software which needs to implement the caveat to ensure integrity and authenticity.
In the case of RDF files, other ongoing W3C groups seem to specifically relate to the integrity of RDF metadata., e.g., the canonicalising and cryptographically hash of RDF

Dataset [1]. The existence of dedicated effort shows the timeliness of your comments. If you think it might help, we can try to point at the ongoing initiatives as the RDF Dataset Canonicalization and Hash Working Group, but it seems reasonable to wait for the RDF Dataset Canonicalization and Hash Working Group outcomes, and suggest adopting their recipes when their work consolidate.

Of course, If you have any other suggestions about remedies that can fall in the scope of the DXWG group, we are open to considering them. Nonetheless, especially if any new normative solution is required to be conceived, considering the stage in the standardization process we are now for DCAT 3, it is likely that they need to be considered as requirements for the next standardization round (e.g., DCAT 4).

[1] https://w3c.github.io/rch-wg-charter/explainer.html

——

For example, considering RDF encoding as one of the most typical ways to serve DCAT, the encoding might end in an RDF store or an RDF file.
In the case of an RDF store, it is the chosen software which needs to implement the caveat to ensure integrity and authenticity.
In the case of RDF files, there are other recent ongoing W3C efforts, which might relate more to the integrity of RDF metadata., e.g., the canonicalising and cryptographically hash of RDF Dataset, which is at the core of other W3C groups [1].

17. Security and Privacy

The DCAT vocabulary supports the attribution of data and metadata to various participants such as resource creators, publishers and other parties or agents via qualified relations, and as such defines terms that may be related to personal information. In addition, it also supports the association of rights and licenses with cataloged Resources and Distributions. These rights and licenses could potentially include or reference sensitive information such as user and asset identifiers as described in [ODRL-VOCAB]. Implementations that produce, maintain, publish or consume such vocabulary terms must take steps to ensure security and privacy considerations are addressed at the application level.

Issues pertaining to the Integrity and authenticity of DCAT might benefit from the result of other W3C activity. For example, DCAT metadata provided as RDF might benefit from progress delivered by the RDF Dataset Canonicalization and Hash Working Group.

—----

First attempt

Thanks for the feedback.

Yes, metadata integrity might be an issue, but doesn't this mainly depend on how DCAT is served? DCAT is a metadata model typically served on RDF for serialization.

We suspect it is not up to DCAT to provide a solution for ensuring RDF integrity, and that would need to be addressed by a dedicated group transversal to the different metadata models.

Apart from that, any suggestions about remedies that can fall in the scope of the DXWG group are more than welcome.

Considering the stage at which we are with DCAT 3 development, suggestions are likely to be registered as future requirements to be considered in the next round of standardization, e.g. DCAT 4.

____

- quite timely feedback https://w3c.github.io/rch-wg-charter/explainer.html
- people with gravity in the field

____

A second attempt (still to be finalized)

Thanks for the feedback.

DCAT is a metadata model, and integrity and authenticity seem to relate more to how DCAT is served than the DCAT model itself. So we feel wary that addressing these issues is a bit off-scope of the current group mandate, as it might conflict with the activity of other ongoing initiatives, and the assumption we can make at the level of the DCAT model can turn out in conflict with solutions elaborated by other groups in the meanwhile.

For example, considering RDF encoding as a typical way to serve DCAT, the encoding might end in an RDF store or an RDF file.

In the case of an RDF store, it is the chosen software which needs to implement the caveat to ensure integrity and authenticity. In the case of RDF files, other recent ongoing W3C efforts go in the direction of canonicalising and cryptographically hash of RDF Dataset might we think might relate to the integrity of DCAT metadata.