DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011

Extending DCAM for Metadata Provenance

Kai Eckert
Mannheim University Library, Germany
eckert@bib.uni-mannheim.de

Daniel Garijo
Universidad Politécnica de Madrid, Spain
dgarijo@delicias.dia.fi.upm.es

Michael Panzer
OCLC Online Computer Library Center, Inc., USA
panzerm@oclc.org

Abstract

The Metadata Provenance Task Group aims to define a data model that allows for making assertions about description sets. Creating a shared model of the data elements required to describe an aggregation of metadata statements allows to collectively import, access, use and publish facts about the quality, rights, timeliness, data source type, trust situation, etc. of the described statements. In this paper we outline the preliminary model created by the task group, together with first examples that demonstrate how the model is to be used.

Keywords: metadata; provenance; DCAM; model

1. Introduction

The rise of the Web of Data during the last few years has increased the amount of information available for users in a wide range of domains: digital libraries, scientific workflows, or social networks among others. In order to provide high quality content to users, content providers have started to pay more attention to the provenance of their content: where does it come from, who created it, or how was it modified by other sources to produce its current version?

1.1. Motivation

Many vocabularies and specifications have been developed for representing provenance, but no commonly accepted standard or official recommendation has emerged yet. The W3C Provenance Incubator Group[1] was launched in September 2009 with the objective of creating a roadmap and a state of the art report of the current approaches, taking the first steps towards a domain-independent standard. The group also analyzed the "gaps" when trying to provide provenance solutions in different domains (news aggregation[2], scientific workflows[3] and business contracts[4]) but didn't analyze deeply the topic of the provenance of metadata itself, as well as the representation of provenance information as metadata together with the described resources.

The latter is already practiced in some areas like digital libraries or scientific workflows, but these approaches are independent of each other and thus use different models and vocabularies. Therefore, our motivation for a Dublin Core application profile for metadata provenance is twofold: Firstly, we want to represent existing metadata provenance information in a simple and unified way that is well suited for an application of Dublin Core. Secondly, we want to enable the provision of provenance information for Dublin Core metadata in a Dublin Core compatible way.

1.2. Related Work

An early initiative to define a vocabulary and usage guidelines for the provenance of metadata was the ACore (Iannella & Campbell, 1999) and, based on it, the proposal (Hansen & Andresen,

1 http://www.w3.org/2005/Incubator/prov/wiki/Main_Page
2 http://www.w3.org/2005/Incubator/prov/wiki/Analysis_of_News_Aggregator_Scenario
3 http://www.w3.org/2005/Incubator/prov/wiki/Analysis_of_Disease_Outbreak_Scenario
4 http://www.w3.org/2005/Incubator/prov/wiki/Analysis_of_Business_Contract_Scenario

https://doi.org/10.23106/dcmi.952135636

2001) for the DCMI Administrative Metadata Working Group[5]. The working group finished in 2003 and presented the Administrative Components (AC), addressing metadata for the entire record, for update and change, and for batch interchange of records (Hansen & Andresen, 2003). Both initiatives focused more on the definition of specific vocabularies to describe the provenance of metadata. There was not yet a concise model to relate the provenance information with the metadata.

Later initiatives have focused more on domain-independent provenance representation. Vocabularies like OPM (Moreau et. al., 2010), Provenir (Sahoo et. al., 2010) and, more domain-specific, the Provenance Vocabulary (Hartig, 2009) allow for representing various levels of provenance as a hierarchy, but they are agnostic about the resource they are providing provenance about. So in the context of metadata, they leave the implementer alone to decide how to identify metadata as a resource.

Other initiatives like OAI-ORE[6] or OAI-PMH[7] integrate the provenance information with the metadata, but are either too generic (ORE is not specifically designed to represent provenance information) or too specific (PMH only provides provenance for aggregations of metadata for the purpose of metadata harvesting).

The alternative domain model we are presenting here has some structural resemblance with the notion of a "nano-publication" as described by Groth, Gibson, and Velterop (2010). A nano-publication consists of a single scientific statement combined with a set of annotations describing the statement's publication context (i.e., its descriptive metadata), essentially providing a minute element of the publication in which this statement originally appeared. There are, however, a couple of key differences, which cause this model to not be directly applicable to the problem of metadata provenance. While the focus on a "statement" as the annotated resource allows for the use of some data-related properties in annotations, the focus of the model is still not limited to data or even metadata, but to all "research statements" as defined in the SWAN ontology[8].

In addition, the requirement of annotating single statements only (as one-statement named graphs) raises the question of scalability in large triple stores (from both performance and data management standpoints). While the nano-publication model conceptually allows for the aggregation of all nano-publications about the same statement as "S-Evidence," this still does not satisfy the flexibility requirements of the Dublin Core Abstract Model, where a description set (i.e., metadata) might consist of several different descriptions or statements. Handling the annotation of an entire description set with the nano-publication model would require unnecessary redundancy and cause triple explosion, something that was sought to be avoided by using named graphs.

1.3. Problem definition.

The main objective of the Dublin Core Metadata Provenance Task Group[9] is to provide the means and guidelines to model and handle metadata provenance as a type of metadata. The approach taken for this task has been to create a model as simple as possible, providing real world examples and mappings to other provenance approaches and comparing the complexity of the outcomes.

Dublin Core provides a domain model – the Dublin Core Abstract Model[10] (DCAM) –, which tries to abstract from actual data models used in metadata implementations. The currently most prominent domain-independent data model for metadata is probably RDF. While RDF and DCAM look quite similar, there are enough differences that can lead to implementation problems,

5 http://dublincore.org/groups/admin/
6 http://www.openarchives.org/ore/
7 http://www.openarchives.org/pmh/
8 http://www.w3.org/TR/hcls-swan/
9 http://wiki.bib.uni-mannheim.de/dc-provenance/doku.php?id=dc-provenance
10 http://dublincore.org/documents/abstract-model/

https://doi.org/10.23106/dcmi.952135636

e.g., the missing ability of RDF to represent a description set. In the DC community, the very need for a DCAM is discussed, with the option to deprecate it completely in favor of RDF (Baker & Johnston, 2010). However, RDF is not the only data model for metadata out there, and it makes sense to introduce new metadata concepts in an implementation-independent manner. For this reason, we built our first proposal for a metadata provenance domain model on DCAM concepts.

The remainder of this paper is structured as follows: In section two, we describe the metadata provenance domain model in terms of the DCAM, then discuss first thoughts about an element vocabulary in section three, showing how the model can be implemented in RDF in section four, and providing a complete use case example of a more complex vocabulary in section five. Finally, section six presents our conclusions and future lines of work.

2. The basic domain model

In this section we introduce and explain the current proposal of a domain model for managing metadata provenance. The domain model is independent of an employed element vocabulary that would be used in statements to represent the actual provenance information. Instead, it forms the abstract framework that relates the provenance information to existing metadata and especially relates the classes that are introduced in the model to the existing classes in the DCAM.

2.1. Domain model

The proposed model extends the Dublin Core Abstract Model. In particular, it uses the following classes:

- Description Set (from DCAM terminology[11]): A set of one or more Descriptions, each of which describes a single resource.
- Description (from DCAM terminology): One or more Statements about one, and only one, resource.
- Statement[12] (from DCAM terminology): An instantiation of a property-value pair made up of a property URI (a URI that identifies a property) and a value surrogate.
- Annotation: One or more Statements about one Description Set. Subclass of Description.
- Annotation Set: A set of one or more Annotations. Subclass of Description Set.

Figure 1 illustrates the relationships between the new classes and the existing DCAM classes in the form of an UML diagram. As a basis of the aforementioned application model for metadata provenance, the main purpose of the UML diagram is to show (1) ways in which the new entities *Annotation* and *Annotation Set* relate to and extend the existing Dublin Core Abstract Model (DCAM) entities, (2) how an annotation should be associated with the metadata it provides provenance information about, and (3) how annotations are gathered into annotation sets. Note that the domain model (as an extension of DCAM) is an abstract model that is independent of actual implementations like XML Schema or RDF. It is also independent of the employed vocabulary that is used to create the annotations, i.e., the provenance statements.

[11] http://dublincore.org/documents/abstract-model/#sect-7

[12] Keep in mind that a DCAM statement differs from an RDF statement as it only represents a value pair, while an RDF statement already contains the connection of a value pair with a resource. Consequently, descriptions would be defined in RDF only implicitly as triples describing the same resource.

https://doi.org/10.23106/dcmi.952135636

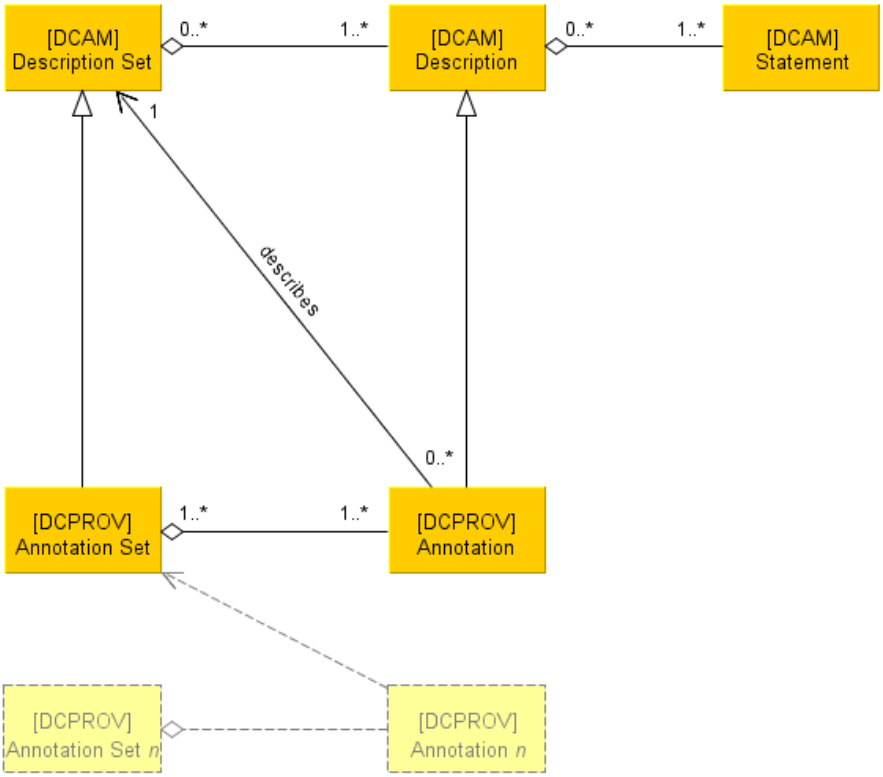

FIG. 1. UML class diagram of the domain model.

2.2. The metadata provenance annotation

According to the domain model, annotations and annotation sets are specifications of their DCAM counterparts, i.e., subclasses in an RDF model. Just like a description set is an aggregation of descriptions (statements about a single resource), an annotation set is an aggregation of annotations (statements about a single description set) – one difference being a change in cardinality of this relationship, the motivation of which will be explained below.

This means that every annotation set is also a description set in the sense of the DCAM, and can be treated as such. If that is the case, however, why not just stick with the DCAM entities to deal with metadata provenance instead of introducing two new key entities?

With the derivation of subclasses from DCAM we want to reflect the fact that annotations are special kinds of descriptions, because they are *only* concerned with description sets, not arbitrary resources. With this distinction of annotations and the grouping in annotation sets, we make the (provenance) annotations identifiable and also easily retrievable given a known description set.

2.3. Connecting annotations and description sets

Annotations are associated only with description sets, which in turn contain one or more descriptions. The relationship between annotations and description sets (the "role" of annotations in UML terms) is generically stated in the model as being *descriptive*. The concrete mechanism or vocabulary element employed here to further specify this relationship will depend on the metadata or resource description model used in a specific metadata application or use case (e.g., RDF). The *"describes"* relationship in the diagram must not be confused with a specific property in RDF. In an RDF implementation, the *"describes"* relationship would manifest itself merely by the fact that the description set is used as a subject for the triples that form the annotations, independent of the *specific* relationships or properties used for these triples.

https://doi.org/10.23106/dcmi.952135636

The cardinality of 1 of the association on the side of the description set indicates that an annotation must only be related to a single description set. The same annotation cannot be associated with more than one description set for two reasons. On the one hand, it has to be compliant with the DCAM definition of description ("*statements about one, and only one, resource*"), from which annotation is derived, on the other hand, it makes expressions of the domain model in metadata frameworks like RDF easier, where one annotation about two different description sets would result in two completely different triples.

Annotations are aggregated in annotation sets, just as descriptions are generally aggregated in description sets. The main difference between these can be found, once more, in cardinality contraints. Whereas the association of a description with a description set is optional, this does not hold for the association between an annotation and an annotation set. An annotation has to be part of at least one annotation set; conversely, every annotation set aggregates at least one annotation.

The rationale for this cardinality constraint is mainly to facilitate basic discoverability of annotations. Since (1) a variety of relationships can be used for annotating (i.e., *describing*) description sets, and (2) not all entities associated with a description set in that manner may be metadata provenance related, the annotation set as a container or wrapper has to provide a reliable means of retrieving metadata provenance information.

Also, this constraint ensures that metadata provenance information can be further annotated by associating higher-level annotations with a lower-level annotation set, as seen in the lower row of Figure 1. Since an annotation set is a description set, it can itself be annotated by associating a further annotation set, i.e., it can as well capture provenance information about annotation sets. In this way, the model is able to handle an arbitrary number of levels of annotations.

3. Towards an Element Vocabulary

While the domain model outlines a mechanism that enables connecting an annotation with the annotated data, it does not describe the makeup of an annotation set for the specific context of metadata provenance, i.e., it does not provide an element vocabulary needed to put together and validate a concrete metadata provenance annotation set, but rather the generic scaffolding to accommodate such an element vocabulary.

As the work on the metadata provenance application profile progresses, the task group will continue analyzing use cases and requirements in order to derive an element vocabulary that will then be used to define necessary and sufficient conditions for compliant annotation sets. As is common practice in other application profiles, the resulting element vocabulary for creating actual annotations will most likely consist of a mix of common Dublin Core terms to state basic provenance information like creator, creation date, sources, contributors, etc., mixed with terms from experimental or established provenance vocabularies like OPM, while at the same time defining a migration path to new standardizing efforts like the Provenance Interchange Language (PIL) which will be defined as one of the deliverables of the recently founded W3C Provenance Interchange Working Group[13]. For this task we will partially rely on existing mappings between common provenance models[14], which translate some concepts of the most popular provenance models (including Dublin Core) to OPM and have served as reference for the initial set of concepts to be represented in the PIL.

4. RDF Implementation

We have already justified the use of the implementation-independent Dublin Core Abstract Model as a basis for our proposed domain model. Using this approach, we believe having a clean starting point for actual implementations, as there are already concrete recommendations for

[13] http://www.w3.org/2011/01/prov-wg-charter.html
[14] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings

https://doi.org/10.23106/dcmi.952135636

DCAM implementations, like DC-RDF or DC-XML. The main demand that is placed on the underlying model or format (e.g., RDF or XML) is the possibility to represent a description set (including recognizing it as a resource in its own right).

RDF provides at least two different ways to provide statements about statements: reification (Eckert et al., 2009) and named graphs (Carroll et al., 2005). Eckert et al. (2010) propose requirements for metadata provenance, and it is shown that the technical requirements are already met, but for true interoperability, further standardization of provenance mechanisms, especially the proper definition of sets of statements, e.g., in the form of named graphs, is needed. The same is emphasized by Zhao et al. (2010), who summarize the requirements established by the W3C Provenance Incubator Group. One result of the RDF Next Steps Workshop held in June 2010[15] was the likely introduction of some kind of graph identification – probably as named graphs or a mechanism similar to named graphs – into the next version of RDF. With the possibility of using named graphs in standard RDF, it seems almost self-evident that this would be the preferred way to work with provenance data in the future. The deprecation, even, of reification has been discussed[16] as, among other reasons, syntactical support is de facto limited to RDF/XML, semantical intricacies require careful usage conventions, and routine use causes the multiplication of stored triples. We want to demonstrate in the following an implementation based on named graphs without losing specificity compared to reification. However, as a stopgap measure until named graphs become fully available in RDF outside of SPARQL, the basics could also be accomplished, for example, with implicit graphs by means of reusing the URL that is used to provide and identify the actual RDF data set, as recommended by Bizer et al. (2007) and Sauermann and Cyganiak (2008).

Assume a metadata record for the "Mona Lisa," which was – a well known fact – created by Leonardo da Vinci. But of course, Leonardo da Vinci did not create the metadata record, which in our example was created by the *Bibliothèque nationale de France* (BnF). We use two graphs (or, alternatively, two RDF datasets with different URLs on the web) to define a description set and an annotation set according to our domain model (see also Figure 2):

```
# ---------------------------------------------------------------
# Named graph: http://example.org/data/ML-Desc
@prefix dc: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .

:MonaLisa dct:format dctype:StillImage ;
    dc:creator :LeonardoDaVinci .
# ---------------------------------------------------------------

# ---------------------------------------------------------------
# Named graph: http://example.org/data/ML-Anno
@prefix dc: <http://purl.org/dc/terms/> .

<http://example.org/data/ML-Desc> dc:creator :BnF .
<http://example.org/data/ML-Desc> a dcam:DescriptionSet .
<http://example.org/data/ML-Anno> a dcprov[17]:AnnotationSet .
# ---------------------------------------------------------------
```

The following table shows how some of the RDF resources map to their corresponding UML classes of the domain model.

[15] http://www.w3.org/2009/12/rdf-ws/Report.html
[16] http://www.w3.org/2011/rdf-wg/track/issues/25
[17] Here, `dcprov:` is used as a preliminary namespace prefix; currently, there has not yet been a persistent `dcprov` namespace defined.

https://doi.org/10.23106/dcmi.952135636

TABLE 1: Relations between RDF instances and the classes of the domain model.

RDF	UML
:MonaLisa dc:creator :LeonardoDaVinci .	Description
<http://example.org/data/ML-Desc> dc:creator :BnF .	Annotation
<http://example.org/data/ML-Desc>	Description Set
<http://example.org/data/ML-Anno>	Annotation Set

Our example consists of two statements about the resource `:MonaLisa`, one about the creator of the resource, the other about its format. The graph `<ML-Desc>` containing these statements forms a description set. Annotations about this metadata are contained in a second graph, `<ML-Anno>`, forming an annotation set.

Statements that are part of this graph are considered annotations, i.e., statements about the provenance of the *metadata* of the original resource `:MonaLisa`, not about the resource itself. The statement `<ML-Desc> dc:creator :BnF.` means that the *Bibliothèque Nationale de France* created the description of the `:MonaLisa` (i.e., its metadata) contained in the graph `:ML-Desc` as opposed to the creation of the `:MonaLisa` itself.

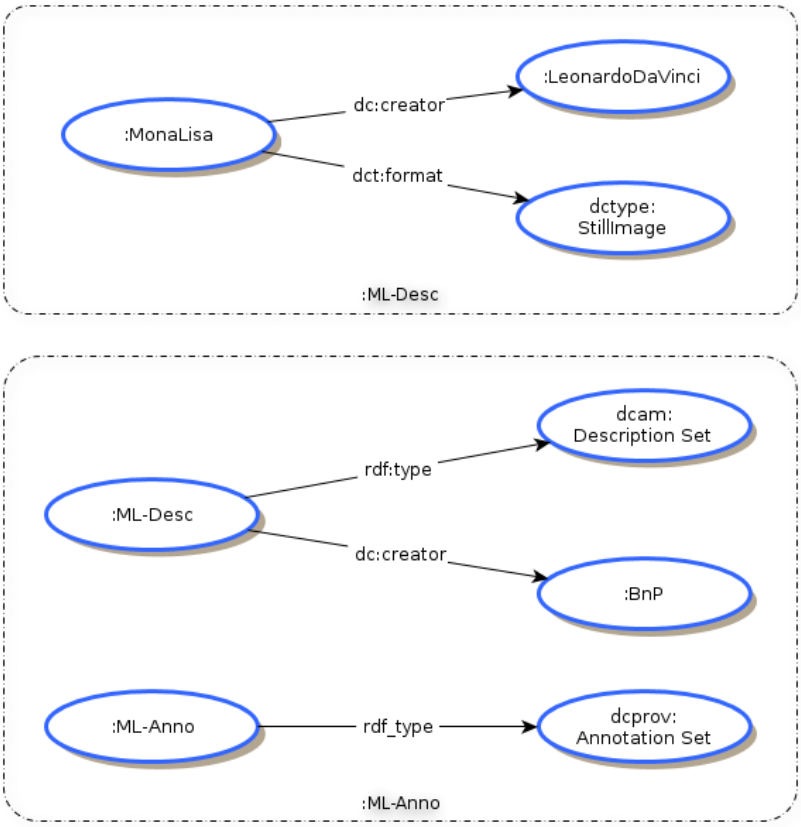

FIG. 2. Example of an RDF implementation.

As seen in this example and explicated above, the only straightforward way to express annotation sets in RDF appears to be by using named graph constructs. This is often a challenge because named graph support in the current version of the RDF standard is rudimentary. There are, however, indirect ways of associating triples with URIs that have been used in the Linked Data community, for example reusing the URL of an actual RDF web document (e.g., `:MonaLisa.rdf`) as a subject for provenance statements about RDF metadata that describes

https://doi.org/10.23106/dcmi.952135636

`:MonaLisa`. A drawback of this method is the inability to explicitly express which statements about `:MonaLisa` are elements of the `:MonaLisa.rdf` aggregation; dereferencing this URI only provides an informal indication based on the HTTP response message.

4.1. Discoverability of metadata provenance

Given a metadata statement *a*, the domain model provides a path to discover whether and which provenance related statement have been asserted for *a*. In RDF however, even a known individual triple may be part of several graphs (e.g., description sets), only some of which might have been annotated. Discovery in RDF is, therefore, a two-stage process. Firstly, it has to be determined of which description sets the triple is part, then it has to be established whether an annotation set exists for any one of these instances. To assert if some provenance information exists for some interpretation of a triple, the following SPARQL query can be used:

```
ASK {
      GRAPH ?ds { :MonaLisa dc:creator :LeonardoDaVinci . }
      GRAPH ?as { ?ds ?p ?o .
                  ?as a dcprov:AnnotationSet . }
}
```

The query will return "true" if some provenance metadata is available. To then gather more information, the query can be expanded.

```
SELECT ?ds ?p ?o WHERE {
      GRAPH ?ds { :MonaLisa dc:creator :LeonardoDaVinci . }
      GRAPH ?as { ?ds ?p ?o .
                  ?as a dcprov:AnnotationSet . }
}
```

This query finds all available provenance statements about the triple. The result shows that the metadata was created by *Bibliothèque Nationale de France*:

TABLE 2: Results of a query on provenance triples regarding a specific triple.

?ds	?p	?o
<http://example.org/data/ML-Desc>	dc:creator	:BnF

4.2. Work in progress: modeling the provenance metadata of travel guides

At the Universidad Politécnica de Madrid the project Web N+1[18] is currently underway, which aims to create a repository of metadata about tourist resources (i.e., guides, images, and videos). Each resource is assigned a different URI, which is used to associate it with its provenance information (creator, date of creation, references used, etc.) as well as additional descriptive metadata about the resource (size, title, subtitle, etc.). A reduced example for a travel guide can be seen in the following RDF code:

[18] http://webenemasuno.linkeddata.es/index_en.html

https://doi.org/10.23106/dcmi.952135636

```
<http://webenemasuno.linkeddata.es/elviajero/resource/Guide/20040117ELP
VIALBV_6.TES>
      rdf:type opmopviajero:Guide ;
      dcterms:rightsHolder
            <http://webenemasuno.linkeddata.es/elviajero/resource/Agent
            /DIARIO%20EL%20PA%C3%8DS%20S.L.>;
      dcterms:date "20040117" ;
      geo:location
            <http://webenemasuno.linkeddata.es/elviajero/resource/Point
            /POINT40.279228_-5.50261>;
      sioc:title "Descanso al calor de una chimenea encendida" ;
      opmopviajero:IPTCMediaType "text".
```

The metadata was created by a Spanish newspaper[19], but it was made public in RDF by the UPM at a certain date under a certain license, which should also be reflected in the created RDF. The RDF is exposed as Linked Data in a repository accessed via Pubby[20], a linked data frontend for SPARQL endpoints which allows exploring and navigating through the links of the endpoint. Pubby allows us to define an additional level of metadata, since it provides information about the RDF shown to the final user (e.g., the query used to retrieve the RDF from the server, the date of retrieval, the web service used to perform the query, etc.), describing it using the Provenance Vocabulary[21].

Therefore, we can organize the metadata in three different levels or groups: the first one groups the descriptions about the resource, the second one gathers the descriptions about the previous statements, and the last one refers to the RDF serialization of the first two groups, which is what is shown to the users.

The modeling of this paradigm with our domain model is done by adapting the different levels to the DescriptionSets and AnnotationSets entities. Figure 3 shows the relationships of the first two groups. A guide with URI ex:guideIdentifier is described in DescriptionSet1 by three triples (date, creator and rightsHolder), which have been created by the newspaper (Prisa Digital) and published by the UPM at a certain date (2011-06-01). These three statements form AnnotationSet1.

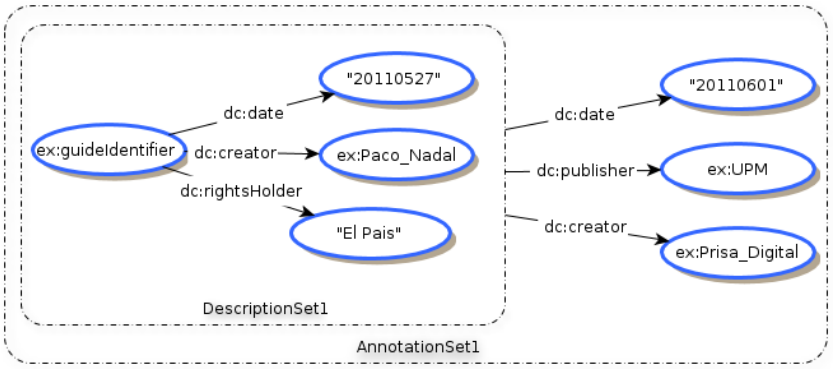

FIG. 3. Provenance information for the UPM guide project

The RDF code of this part can be obtained straightforwardly with the use of named graphs (in TriG syntax[22]), as follows:

19 http://elviajero.elpais.com/
20 http://www4.wiwiss.fu-berlin.de/pubby/
21 http://trdf.sourceforge.net/provenance/ns.html
22 http://www4.wiwiss.fu-berlin.de/bizer/TriG/

https://doi.org/10.23106/dcmi.952135636

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix ex: <http://example.org/data/>.

# DescriptionSet1: descriptive statements about a resource.
<http://example.org/data/guideIdentifier/desc1>
{
      ex:guideIdentifier dc:date "2011-05-27"^^xsd:date.
      ex:guideIdentifier dc:creator ex:Paco_Nadal.
      ex:guideIdentifier dc:rights "El País" .
      <http://example.org/data/guideIdentifier/desc1>
            a dcprov:DescriptionSet.
}
# AnnotationSet1: creator, date and publisher of DescriptionSet1
<http://example.org/data/AnnotationSet/annSet1>
{
      <http://example.org/data/guideIdentifier/desc1>
            dc:date "2011-05-28"^^xsd:date.
      <http://example.org/data/guideIdentifier/desc1>
            dc:creator ex:Prisa_Digital.
      <http://example.org/data/guideIdentifier/desc1>
            dc:publisher ex:UPM.
      <http://example.org/data/AnnotationSet/annSet1>
            a dcprov:AnnotationSet .
}
```

The third group is represented in our domain model with an additional AnnotationSet (AnnotationSet2), which describes the AnnotationSet1 using prv:createdBby (Figure 4).

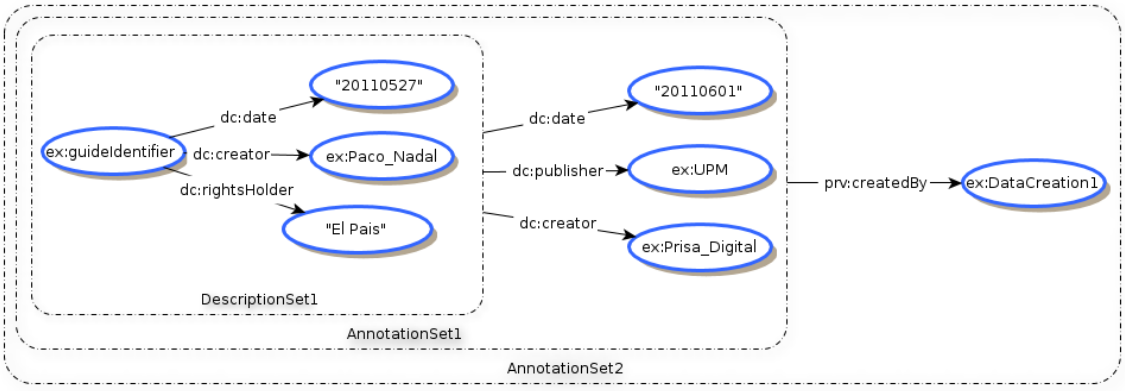

FIG. 4. Provenance information about the actual data creation in the next level.

With the addition of the next code snippet, the RDF of the example is complete:

https://doi.org/10.23106/dcmi.952135636

```
# AnnotationSet2: contains an annotation about how the data from
# AnnotationSet1 has been retrieved from the server.
<http://example.org/data/AnnotationSet/annSet2>
{
      <http://example.org/data/AnnotationSet/annSet1>
            prv:createdBy ex:DataCreation1.
      <http://example.org/data/AnnotationSet/annSet2>
            rdf:type  dcprov:AnnotationSet .
}
```

5. An illustrative example: OAI-PMH to DC-PROV

After the theoretical presentation of the proposed DC-PROV domain model and the RDF-based example implementation, we want to demonstrate the possible use by means of a real-world example: the translation of provenance information included in the metadata transported via OAI-PMH. The purpose of this example is twofold: On the one hand, it should help to understand the abstract classes presented in section two and show how they can be used independently of RDF. On the other hand, it hopefully supports the idea that real world data containing some metadata provenance information can be transformed into a unified data model that – albeit with some information loss – would enable true interoperability.

An OAI-PMH dataset may or may not include provenance related information. The provenance data – called origin description – contains the following elements (Lagoze et al., 2002):

- baseURL: the baseURL of the originating repository from which the metadata record was harvested
  - identifier: the unique identifier of the item in the originating repository from which the metadata record was disseminated
  - datestamp: the datestamp of the metadata record disseminated by the originating repository
  - metadataNamespace: the XML namespace URI of the metadata format of the record harvested from the originating repository
  - originDescription: an optional originDescription block which was obtained when the metadata record was harvested. A set of nested originDescription blocks will describe provenance over a sequence of harvests
  - harvestDate: the responseDate of the OAI-PMH response that resulted in the record being harvested from the originating repository
- altered: a boolean value which must be true if the harvested record was altered before being disseminated again

The metadata itself can be in an arbitrary format, the support of Dublin Core is obligatory for an OAI-PMH interface. But in this example, we don't want to deal with the translation of the metadata, we are concerned with the translation of the origin description.

The following example illustrates an origin description in OAI-PMH.

- originDescription
  - harvestDate="2002-02-08T08:55:46Z" altered="true"
  - baseURL = http://odd.oa.org
  - identifier = oai:odd.oa.org:z1x2y3
  - datestamp = 1999-08-07T06:05:04Z

https://doi.org/10.23106/dcmi.952135636

- metadataNamespace = http://odd.oa.org/odd_fmt

Figure 5 depicts the data transformed into the DC-PROV model. As the origin description refers to a source metadata set from which the information actually provided is derived, we are in fact dealing implicitly with two description sets, one containing the data in our PMH record, one representing the original data. The description sets are related by means of the `dc:source` property which is defined as "a related resource from which the described resource is derived". To avoid losing the information about whether the metadata was altered since the harvesting, we propose the definition of a new subproperty of `dc:source`, `dcprov:sourceModified`, which would be defined as "a related resource from which the described resource is derived by modifying it".

The identifier, according to OAI-PMH, is an identifier for the record, not the described resource. This implies that it can also be used as the URI for the description set. The contents of the description sets are completely arbitrary; we are not concerned with their representation in our model. As OAI-PMH always delivers Dublin Core, it can be used straightforwardly in this regard.

It is interesting that, with this approach, the provenance chain is intact if every party provides information in that way, i.e., we find it to be a quite natural fit between the OAI-PMH model and the proposed DC-PROV model.

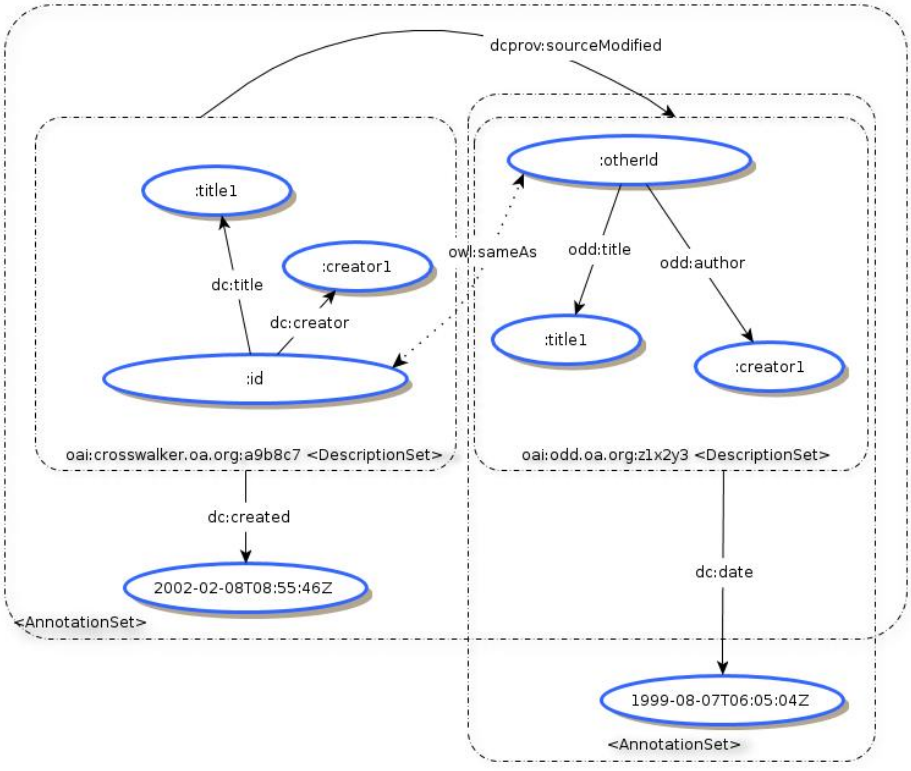

FIG. 5. OAI-PMH translated to DC-PROV

6. Conclusions and future work

In this paper we have introduced a domain model to handle metadata provenance annotations as an extension of DCAM, in order to (1) represent existing metadata provenance information in a simple and unified way that fits in the DCMI context, and (2) provide provenance information for DC metadata in a DCMI compatible way. We also have presented a possible implementation of this model in RDF using named graphs and shown how our domain model can be easily adopted by content providers in one real-world example modeled with OAI-PMH.

After months of discussions and feedback in the task group we can conclude that our domain

https://doi.org/10.23106/dcmi.952135636

model is stable and seems fit for its purpose (as illustrated by the examples), allowing the representation of as many (meta-)provenance levels as needed. It does so by having a simple specification following the style of Dublin Core, which is usable even if a small amount of information is lost depending on the models used in the source data.

As future work, we are currently developing several approaches to map our model to OAI-ORE and OPMV. By accomplishing these objectives, we will provide additional guidelines for publishing metadata provenance information (in the form of an application profile) and potential extensions to the DC element vocabulary for describing provenance in any domain.

Acknowledgement

We would like to thank the R&D project Web N+1 for providing the data for the use case in Section 4.2.

References

Baker, T. & Johnston, P. (October, 2010)A review of the DCMI Abstract Model with scenarios for its future. http://dublincore.org/architecturewiki/DcamInContext

Bizer, C., Cyganiak, R., & Heath, T. (2007). How to Publish Linked Data on the Web. Retrieved from http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Carroll, J. J., Bizer, C., Hayes, P., & Stickler, P. (2005). Named graphs, provenance and trust. *Proceedings of the 14th international conference on World Wide Web WWW 05*, *14*, 613. ACM Press. Retrieved from http://portal.acm.org/citation.cfm?doid=1060745.1060835

DCMI. (1998). Dublin Core Metadata Element Set, version 1.0: Reference description. Retrieved January 10, 2007, from http://www.dublincore.org/documents/1998/09/dces/.

Eckert, K., Pfeffer, M., & Stuckenschmidt, H. (2009). A Unified Approach for Representing Metametadata. *Proceedings of the 2009 Dublin Core Conference*. Retrieved from http://dcpapers.dublincore.org/ojs/pubs/article/view/973/948

Eckert, K., Pfeffer, M., & Völker, J. (2010). Towards Interoperable Metadata Provenance. *ISWC 2010*.

Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. *Inf. Serv. Use* 30, 1-2 (January 2010), 51-56.

Hansen, J., & Andresen, L. (2001) Administrative Dublin Core (A-Core) Element Retrieved from http://dublincore.org/groups/admin/proposal-20010910.shtml

Hansen, J., & Andresen, L. (2003). AC - Administrative Components: Dublin Core DCMI Administrative Metadata. Retrieved from http://www.bs.dk/standards/AdministrativeComponents.htm

Hartig, O. (2009). Provenance Information in the Web of Data. In *Proceedings of the Linked Data on the Web (LDOW) Workshop at WWW*, Madrid, Spain.

Heery, R. (2004). Metadata futures: Steps toward semantic interoperability. In Diane I. Hillmann & Elaine L. Westbrooks (Eds.), Metadata in practice (pp. 257-271). Chicago: American Library Association.

Hillmann, D. I., Sutton S. A., Phipps, J. & Laundry, R. J. (2006). A metadata registry from vocabularies up: The NSDL registry project. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2006, 65-75.

Iannella, R., & Campbell, D. (1999). The A-Core: Metadata about Content Metadata. Retrieved from http://metadata.net/admin/draft-iannella-admin-01.txt

Lagoze, C., Krafft D., Payette, S. & Jesuroga, S. (2005, November). What is a digital library anyway, anymore? Beyond search and access in the NSDL. D-Lib Magazine, 11(11). Retrieved, January 10, 2007, from http://www.dlib.org/dlib/november05/lagoze/11lagoze.html.

Lagoze, C., Van De Sompel, H., Nelson, M., & Warner, S. (2002). Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting - XML schema to hold provenance information in the "about" part of a record. Retrieved from http://www.openarchives.org/OAI/2.0/guidelines-provenance.htm

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers,J., Plale, B., Simmhan, Y., Stephan, E. & Van den Bussche , J.(July 2010). The open provenance model core specification (v1.1). *Future Generation Computer Systems*.

Sahoo, S.S., Barga, R., Sheth, A., Thirunarayan,K. & and Hitzler,P. (2010). *PrOM : A Semantic Web Framework for Provenance Management in Science*. 2010.

https://doi.org/10.23106/dcmi.952135636

Sauermann, L. & Cyganiak, R. (2008). Cool URIs for the Semantic Web: W3C Interest Group Note 03 December 2008. Retrieved from http://www.w3.org/TR/cooluris/

Zhao, J., Bizer, C., Gil, Y., Missier, P., & Sahoo, S. (2010). Provenance Requirements for the Next Version of RDF. *Proceedings of the W3C Workshop RDF Next Steps June 2627 2010 hosted by the National Center for Biomedical Ontology NCBO Stanford Palo Alto CA USA*. Retrieved from http://www.w3.org/2005/Incubator/prov/wiki/images/3/3f/RDFNextStep_ProvXG-submitted.pdf

https://doi.org/10.23106/dcmi.952135636