Metadata Provenance: Dublin Core on the Next Level

Kai Eckert Mannheim University Library, Germany eckert@bib.uni-mannheim.de	Daniel Garijo Universidad Politécnica de Madrid, Spain dgarijo@delicias.fi.upm.es	Michael Panzer OCLC Online Computer Library Center, Inc., USA panzerm@oclc.org	Ömer Perçin[1] Mannheim University, Germany operin@mail.uni-mannheim.de

Keywords: metadata; provenance; DCAM; model

With this poster, we want to present the current state of the DCMI Metadata Provenance Task Group, which will wrap up its work at the time of DC-2011. The motivation for a Dublin Core extension for metadata provenance is twofold: Firstly, we want to represent existing metadata provenance information in a simple and unified way that is well suited as 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.

Thus, the main objective of the Dublin Core Metadata Provenance Task Group[2] is to provide the means and guidelines to model and handle metadata provenance. The approach followed 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.

The proposed model extends the Dublin Core Abstract Model[3]. In particular, it uses the following classes:

- Description Set (from DCAM terminology[4]): 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 (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 as a UML diagram. The UML diagram is independent of interpretations in specific vocabularies; it is intended to be expressed by using various metadata terms or schemes, although the final application profile will provide guidelines for implementation and validation of annotation sets.

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.

[1] Alphabetical order

[2] http://wiki.bib.uni-mannheim.de/dc-provenance/doku.php?id=dc-provenance

[3] http://dublincore.org/documents/abstract-model/

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

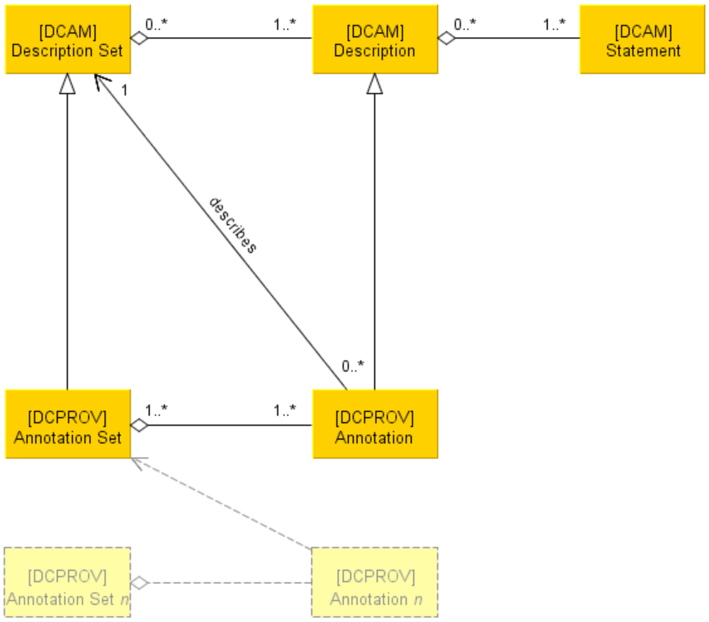

FIG. 1. UML class diagram of the domain model.

The domain model outlines a mechanism that enables connecting an annotation with the annotated data. It does not attempt to describe the makeup of an annotation set in the specific context of metadata provenance, i.e., it does not yet 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. Figure 1 also illustrates that annotations can also be used to describe annotation sets, as annotation sets are description sets as well. In this way, arbitrary levels of provenance information are possible.

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[5], while at the same time defining a migration path to new standardizing efforts like the Provenance Interchange Language (PIL).

Beside the presentation of the domain model in terms of DCAM, the poster will also be used to demonstrate (by means of real-world examples) how metadata provenance can be expressed in RDF. One example will be the representation of provenance information contained in an OAI-PMH[6] dataset in terms of the new model.

[5] http://openprovenance.org/

[6] http://www.openarchives.org/pmh/