DE GRUYTER
OPEN

Research Paper

Provenance Description of Metadata Vocabularies for the Long-term Maintenance of Metadata

Chunqiu Li[1†] & Shigeo Sugimoto[2]

[1]Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
[2]Faculty of Library, Information and Media Science, University of Tsukuba, Japan

Citation: Chunqiu Li & Shigeo Sugimoto (2017). Provenance Description of Metadata Vocabularies for the Long-term Maintenance of Metadata.
Vol. 2 No. 2, 2017
pp 41–55
DOI: 10.1515/jdis-2017-0007
Received: Dec. 4, 2016
Revised: Feb. 10, 2017
Accepted: Feb. 13, 2017

Abstract

Purpose: The purpose of this paper is to discuss provenance description of metadata terms and metadata vocabularies as a set of metadata terms. Provenance is crucial information to keep track of changes of metadata terms and metadata vocabularies for their consistent maintenance.

Design/methodology/approach: The W3C PROV standard for general provenance description and Resource Description Framework (RDF) are adopted as the base models to formally define provenance description for metadata vocabularies.

Findings: This paper defines a few primitive change types of metadata terms, and a provenance description model of the metadata terms based on the primitive change types. We also provide examples of provenance description in RDF graphs to show the proposed model.

Research limitations: The model proposed in this paper is defined based on a few primitive relationships (e.g. addition, deletion, and replacement) between pre-version and post-version of a metadata term. The model is simplified and the practical changes of metadata terms can be more complicated than the primitive relationships discussed in the model.

Practical implications: Formal provenance description of metadata vocabularies can improve maintainability of metadata vocabularies over time. Conventional maintenance of metadata terms is the maintenance of documents of terms. The proposed model enables effective and automated tracking of change history of metadata vocabularies using simple formal description scheme defined based on widely-used standards.

Originality/value: Changes in metadata vocabularies may cause inconsistencies in the long-term use of metadata. This paper proposes a simple and formal scheme of provenance description of metadata vocabularies. The proposed model works as the basis of automated maintenance of metadata terms and their vocabularies and is applicable to various types of changes.

† Corresponding author: Chunqiu Li (E-mail: lichunqiuaa@126.com).

Research Paper

Keywords Metadata maintenance; Metadata vocabulary; Metadata term; Metadata provenance; Metadata longevity

1 Introduction

Maintaining the accessibility of collections for future generations is a central mission of libraries and other memory institutions. Metadata longevity should be ensured to keep the long-term accessibility of data collections. However, we are facing the difficulties in metadata longevity, such as the consistent maintenance of metadata, maintenance of metadata vocabularies and metadata terms, structural and syntactic features of metadata, metadata description rules, and so forth. This paper focuses on consistent maintenance of metadata vocabularies and metadata terms. This is because the changes of definitions of a metadata term may not always be recorded appropriately. The definition of a metadata term may include meaning and usage of the term, relationships to other terms, human-readable labels, and so forth. Metadata terms are usually defined as a set of terms, which is called a metadata vocabulary. This paper aims to propose a metadata model designed to keep track of the changes to definitions of metadata terms and metadata vocabularies.

In digital preservation standards, e.g. Open Archival Information System (OAIS)[①] and PREMIS[②], provenance of digital objects is a required component that has to be recorded for longevity of digital objects. As provenance of metadata is crucial for metadata longevity of such digital objects, how to formally and consistently describe the provenance of metadata over time is an important issue. Provenance of metadata schemas and provenance of metadata vocabularies, as well as provenance of metadata terms have to be consistently recorded over time. This paper focuses on provenance of metadata vocabularies and metadata terms. Provenance description of a metadata term is a record that describes the revision history of the metadata term. Provenance description of a metadata vocabulary is crucial as well. This paper applies W3C PROV[③] to record provenance description of metadata vocabularies and their terms. The reason for adoption of W3C PROV is that it is developed as a standard for general provenance description and provenance interchange in a heterogeneous environment (Gil et al., 2013). W3C PROV has been commonly applied to specific domains, e.g. earth science and social sciences (Cuevas-Vicenttín et al., 2016; Lagoze, Willliams, & Vilhuber, 2013; Masó, Closa, & Gil, 2015; Missier & Chen, 2013; Tilmes et al., 2013).

① http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284
② http://www.loc.gov/standards/premis/
③ http://www.w3.org/TR/prov-overview/

The goal of this paper is to propose a model for formal provenance description of metadata vocabularies to keep track of primitive changes of their terms. The classified primitive change types can be applied to terms expressing either properties or classes of resources, i.e. both property vocabulary and value vocabulary.

The rest of this paper is organized as follows. Section 2 clarifies the meanings of Term and Term Definition in this paper. Section 3 presents requirements of provenance description of metadata vocabularies for metadata maintenance. Section 4 summarizes the related literature about metadata registries services and representation of changes. Section 5 applies W3C PROV to provenance description of metadata vocabularies. Section 6 provides a detailed description of the proposed model in this paper. The concluding remarks are given in Section 7.

2 Metadata Vocabulary and Terms

In the library community, commonly used metadata vocabularies are controlled vocabularies and metadata element sets (Hyland et al., 2013; Isaac et al., 2011), e.g. subject headings, authority files, Resource Description and Access (RDA)[④] element sets, and RDA value vocabularies. A metadata vocabulary is a set of metadata terms. In this paper, we use "metadata vocabulary" as a generic concept that includes two types, i.e. property vocabulary and value vocabulary. A property vocabulary is a set of terms expressing attributes of a resource and relationships between resources, which is often called metadata element set, e.g. Dublin Core metadata element set[⑤] and BIBFRAME vocabulary[⑥]. A value vocabulary is a set of terms expressing classes of resources and encoding schemes of property values, e.g. Library of Congress Subject Headings (LCSH)[⑦].

To propose general provenance description model for tracking primitive changes of metadata terms in metadata vocabularies, this study defines "Term" and "Term Definition" as follows.

Term in a metadata vocabulary is an individual entity, which represents a concept, a property, a class, and a metadata vocabulary. For example, a subject heading in LCSH, property "dct:title," class "dct:Agent," and vocabulary encoding scheme LCSH are examples of terms. In this study, we use "Term" in both meanings of property vocabulary term and value vocabulary term.

④ See http://www.rda-rsc.org and http://www.rda-jsc.org/archivedsite/rda.html for details.
⑤ http://dublincore.org/documents/dces/
⑥ http://bibframe.org/vocab/
⑦ See LCSH introduction at https://www.loc.gov/aba/publications/FreeLCSH/lcshintro.pdf.

Term Definition of a metadata term is a set of descriptions that defines features of the term. The features are the human-readable label(s) of the term, the meaning of the term, relationships between terms, usage of the term, and other information. *Term Definition* may be seen as a set of statements, each of which defines a feature of the term. For instance, "the broader term of Vehicles in LCSH is Transportation" is a *Term Definition* of Term "Vehicles;" "the label of term subject in Dublin Core metadata element set is Subject" is a *Term Definition* of Term "dc:subject." The two examples of *Term Definition* can be respectively represented as RDF triples, *lcsh:sh85142531 skos:broader lcsh:sh85137027* and *dc:subject rdfs:label "Subject"@en*. The *lcsh:sh85142531* stands for "*Vehicles*" while the *lcsh:sh85137027* stands for "*Transportation*."

3 Provenance of Metadata Vocabularies

3.1 Definition of Provenance of Metadata Vocabularies

Provenance comes from French verb "provenir." Provenance means source or history or derivation of an object, which can be work, data, etc. The provenance of a piece of data is the process that led to the piece of data in a computer system (Moreau, 2010). According to the W3C Provenance Working Group, provenance is a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing (Moreau et al., 2013). Provenance is used for many purposes, e.g. making judgments about information to determine whether to trust it, reproducing how something was generated (Gil et al., 2013).

Metadata vocabularies have to be maintained to keep metadata terms consistently interpretable. The definition of a metadata term may be changed, e.g. renaming of a term, revision of the meaning of the term, and revision of relationships to other related terms. It is crucial to trace changes of metadata terms in metadata vocabularies. Provenance description for long-term maintenance of metadata vocabularies is primarily the series of activities that have taken place on metadata vocabularies and their terms. This paper proposes a model to describe provenance description of metadata vocabularies based on W3C PROV. We classified entities and activities based on the relations defined in W3C PROV to describe primitive changes of metadata terms in metadata vocabularies. The recorded entities and activities are traceable to provide evidence for change tracking, which brings the benefits of provenance description of metadata vocabularies, e.g. preventing misinterpretation and auditing inconsistencies of metadata vocabularies. These benefits are valuable for the long-term maintenance of metadata vocabularies throughout their life cycle.

Provenance of metadata vocabularies is a record that describes the agents, activities, and entities involved in the lifecycle of metadata vocabularies. Provenance

of metadata vocabularies includes information about how metadata terms in a metadata vocabulary and its term definitions come to a specific state. The definitions of metadata terms can change over time. For instance, a term can be split into two related terms, or the semantic relationship between two terms can change over time. Those who are responsible for maintaining metadata vocabularies need to pay attention to the changes and also document the changes.

3.2 Requirements of Provenance Description of Metadata Vocabularies for Metadata Maintenance

Groth et al. (2012) illustrated requirements of provenance on the Web. The requirements refer to many dimensions, e.g. activities, records of changes, derivation, and interoperability. These requirements present the content of provenance and their use requirements. However, these requirements are not directly oriented to metadata maintenance. Keeping track of provenance of metadata vocabularies is beneficial to the consistent maintenance of metadata vocabularies. Provenance description of metadata vocabularies should be recorded in machine-readable, traceable and interoperable form to support the effective check of inconsistency caused by changes.

Machine-readability: to record provenance description of metadata vocabularies in machine-readable form for machine process, e.g. RDF/XML and RDF/JSON.

Traceability: to use provenance description of metadata vocabularies for tracking the changes among different versions of a metadata vocabulary, e.g. tracking provenance description in RDF using SPARQL®.

Interoperability: to keep provenance description of metadata vocabularies interoperable in the heterogeneous Web environments.

4 Literature Review

This section discusses related works from the two aspects that are closely related to this study – metadata registry and representation of changes.

4.1 Metadata Registry Services for Metadata Interoperability

The reuse of existing metadata terms is essential to improve metadata interoperability. Metadata registry plays an important role in collecting and sharing metadata vocabularies to achieve metadata interoperability. Although metadata interoperability is an important aspect for long-term maintenance of metadata, metadata registry does not ensure metadata longevity. Metadata registry typically

® The SPARQL Protocol and RDF Query Language (SPARQL) is a query language and protocol for RDF. Please see the details at http://www.w3.org/TR/sparql11-query.

holds the following functions, i.e. registration, management, storage and sharing of metadata elements sets, and controlled vocabularies and application profiles. For example, Open Metadata Registry (OMR)[9], RDA Registry[10] and Dublin Core Metadata Initiative (DCMI)[11] metadata registry[12] are typical examples of metadata registries, which provide search and browse services of their registered metadata vocabularies.

OMR also provides service to vocabulary owners and managers about the versioning and change tracking of their registered vocabularies. The information about changed time, action, and the vocabulary maintainer who made the change are accessible on OMR history page. RDA vocabularies (element sets and value vocabularies) are maintained in the RDA Registry based on OMR with a combination of Git and GitHub. RDA Registry supports the semantic versioning of RDA vocabularies. The version designations follow the general principles of semantic versioning. GitHub provides the changes list of released RDA vocabularies in natural language, e.g. lists of "Adds new RDA entities," "Adds new RDA elements," "Adds new constrained RDA elements," "Deprecates published RDA elements," "Adds value vocabularies," and "Renames value vocabularies" (Phipps, Dunsire, & Hillmann, 2015). However, these changes of RDA vocabularies are not kept interpretable to machines over time.

4.2 Representation of Changes

Javed, Abgaz, and Pahl (2014) proposed a layered change log model to record the changes of ontology using RDF triple-based representation. The changes are recorded using their own change metadata ontology and existing Provenance Vocabulary Core Ontology terms. Chawuthai et al. (2016) presented a logical model named Linked Taxonomic Knowledge (LTK) and LTK Ontology for preserving and representing changes in taxonomic knowledge for linked data. The changes in conception or in the relationship between taxa are preserved as events along with aspects of time, provenance, causes, and effects. A tool supporting version management of RDF vocabularies named SemVersion has been developed (Kendall et al., 2008). SemVersion provides structural and semantic versioning for RDF models and RDF-based ontology language like RDFS (Völkel & Groza, 2006).

Changeset vocabulary defines a set of terms (e.g. Addition, ChangeReason, and Removal) to describe changes between two versions of a resource description by using two sets of triples, i.e. additions and removals (Tunnicliffe & Davis, 2009).

[9] http://metadataregistry.org
[10] http://www.rdaregistry.info
[11] http://dublincore.org/
[12] http://dcmi.kc.tsukuba.ac.jp/dcregistry/

Changeset vocabulary represents changes to resource descriptions using RDF reification. An update is represented by a set of statements about statements and whether they are added or removed (Meinhardt, 2015). Changeset vocabulary is used by LCSH to describe the information of "Change Notes" of subject headings. The document-centric approved list of new headings and revisions to existing headings in LCSH are available on the Acquisitions and Bibliographic Access Web page[9]. The changes to the subject headings are provided together with the literal words like "ADD FIELD" or "DELETE FIELD." Although Changeset vocabulary is applicable to describe changes of metadata vocabularies, the use of RDF reification will make the description of changes of metadata vocabularies complex.

The W3C PROV standard for provenance description and provenance interchange is developed by W3C Provenance Working Group in 2013. The data model defined by W3C PROV, i.e. PROV-DM is used to encode the revision history of wiki pages (Missier & Chen, 2013). Getty Thesaurus of Geographic Names adopts W3C PROV to describe revision history of geographic names. W3C PROV is used to document the Activity information about the revision of geographic names, e.g. Activity type (Create, Modify) and temporal information associated with the Activity. Given to the extendibility of W3C PROV, this paper selects W3C PROV to record how metadata vocabularies change as provenance in RDF.

5 Application of W3C PROV to Metadata Vocabularies

5.1 Why Use W3C PROV

The W3C PROV standard includes a set of specifications which refers to many aspects of provenance, e.g. modeling, serialization, exchange, access, validation, semantics, and reasoning (Moreau et al., 2015). PROV-DM defines a conceptual data model along with relations to describe general provenance. PROV-O defines an OWL ontology consisting of a set of classes and properties for mapping PROV-DM to RDF. W3C PROV is for general provenance description and allows application to specific domains.

This paper applies W3C PROV to describe the provenance of metadata vocabularies. The main reason is that W3C PROV is a Web-oriented provenance standard for provenance description and provenance interchange. Entities and Activities are an important component to describe provenance in PROV-DM. An Entity is a physical, digital, conceptual, or other kind of thing (Gil et al., 2013). An "Activity" is something that occurs over a period of time and acts upon or with "Entities" (Moreau et al., 2013). An Activity can be used to represent how an Entity

[9] https://www.loc.gov/aba/cataloging/subject/weeklylists/

comes into existence, and how its attributes change to become a new Entity (Gil et al., 2013). To describe the provenance of metadata vocabularies based on W3C PROV, it is necessary to classify the Entities and Activities associated with changes among different versions of a metadata vocabulary. In other words, W3C PROV is used to describe the provenance of metadata vocabularies by defining what Entities have been changed and how the changes are caused by a series of Activities.

5.2 Entities and Activities for Provenance Description of Metadata Vocabularies

Vocabulary, *Term*, and *Term Definition* are classified as three subtypes of PROV Entity to describe provenance of metadata vocabularies. As illustrated above, a *Term* can be a concept or a class or a property. In the case of a concept, its definition may include its narrower term(s), broader term(s), association/related term(s), and other information. In the case of a class, its definition may include a description of its meaning, a label(s), a URI, super-class(es), sub-class(es), used property(ies), and other information. In the case of a property, its definition may include a description of its meaning, a label(s), a URI, super-property(ies), sub-property(ies), domain, range, expected value, and other information.

To describe the provenance of metadata vocabularies, Activities acting on the previously classified Entities are categorized into the following types, i.e. Revision, Addition, Deletion, and Replacement. Table 1 shows the correspondence of the classified Activities to the classified Entities. The mark "○" means "applicable" and "×" means "not-applicable." Table 2 illustrates the classified Activities with their names and definitions. It is notable that replacement of term can be the following cases, e.g. a composite term was split into more than one term; or more than one term was merged to a term; or a term was replaced by another term. Table 3 provides change types of metadata vocabularies as well as their terms with specific examples, which are mainly from the changes between BIBFRAME 2.0 vocabulary (BIBFRAME 2.0 vocabulary list view, 2016) and BIBFRAME 1.0 vocabulary (BIBFRAME 2.0 specifications notes, 2016). In this paper, the separation of a single term into two or more terms is called a split. An example of a split in a subject heading is given in Table 3.

Table 1. Activities acted on Entities for provenance of metadata vocabularies.

Subtypes of PROV Entity	Subtypes of PROV Activity			
	Revision	Addition	Deletion	Replacement
Vocabulary	○	×	×	×
Term	○	○	○	○
Term Definition	○	○	○	○

Table 2. Definitions of the classified Activities for provenance of metadata vocabularies.

Activity name	Definition
RevisionOnVocabulary	The revision of the contents or information of a metadata vocabulary
RevisionOnTerm	The revision of a term of the metadata vocabulary
AdditionOnTerm	The addition of a term
DeletionOnTerm	The deletion of a term
ReplacementOnTerm	The replacement of term(s) by other term(s)
RevisionOnTermDefinition	The revision of a term definition
AdditionOnTermDefinition	The addition of a term definition
DeletionOnTermDefinition	The deletion of a term definition
ReplacementOnTermDefinition	The replacement of a term definition by another term definition

Table 3. Primitive change types of metadata vocabularies and their terms with examples.

Change type	Example
Revision of a Vocabulary	BIBFRAME 1.0 vocabulary is revised to BIBFRAME 2.0 vocabulary.
Revision of a Term	
Addition of a Term	Class bf:Note is newly defined in BIBFRAME 2.0 vocabulary.
Deletion of a Term	Property bf:otherEditionOf that was defined in BIBFRAME 1.0 vocabulary is deleted in BIBFRAME 2.0 vocabulary.
Replacement of a Term	Property bf:credits in BIBFRAME 2.0 vocabulary essentially replaces bf:creditsNote in BIBFRAME 1.0 vocabulary; Subject heading "Folklore, Negro" is split into "Folklore, African" and "Folklore, Afro-American."
Revision of a Term Definition	
Addition of a Term Definition	The inverse property to property bf:absorbed is added in BIBFRAME 2.0 vocabulary.
Deletion of a Term Definition	The definitions of property bf:otherEditionOf that was defined in BIBFRAME 1.0 vocabulary is deleted in BIBFRAME 2.0 vocabulary.
Replacement of a Term Definition	The expected value of property bf:copyrightRegistration is corrected in BIBFRAME 2.0 vocabulary.

A revision of a vocabulary is caused by a revision of its terms. The revision of a term may be a revision of the term as an instance, or a revision of documentation of the term. For example, replacement of a single term by a set of terms is a revision of an instance, and replacement of a title text is a revision of term definition. Therefore, the relationships between the classified Activities are as follows. A *RevisionOnVocabulary* is comprised of *RevisionOnTerm* (zero or more than one) and *RevisionOnTermDefinition* (zero or more than one). Given to the practical change examples of revision of a term and revision of term definitions, *RevisionOnTerm* has three general types, i.e. *AdditionOnTerm, DeletionOnTerm*, and *ReplacementOnTerm; RevisionOnTermDefinition* has three general types, i.e. *AdditionOnTermDefinition, DeletionOnTermDefinition*, and *ReplacementOn TermDefinition*.

5.3 Relations Between the Classified Entities and Activities

The relations between Entities and Activities defined in W3C PROV include *Usage*, *Generation*, and *Invalidation*. *Usage* means utilization of an Entity by an

Activity. *Generation* means creation of a new Entity by an Activity. *Invalidation* means destruction, cessation or expiry of an existing Entity by an Activity (Lebo et al., 2013). The properties *prov:used*, *prov:wasGeneratedBy*, and *prov:wasInvalidatedBy* defined in PROV-O are used to respectively describe *Usage*, *Generation*, and *Invalidation*. W3C PROV also defines *Derivation* between Entities. A *Derivation* is a transformation of an Entity into another, an update of an Entity resulting in a new one, or the construction of a new Entity based on a pre-existing Entity (Lebo et al., 2013). The property *prov:wasDerivedFrom* is used to directionally connect the two Entities from the new Entity to the pre-existing Entity.

Figure 1(a) provides provenance description in RDF graphs defined for the example of term replacement in Table 3: Subject heading "Folklore, Negro" is split into "Folklore, African" and "Folklore, Afro-American" (Knowlton, 2005). The classes and properties with prefix "mv" are defined in this research. The property *mv:wasSplitTo* is to describe the split of a term to more than one term. The class *mv:Term* is to assert a term of a metadata vocabulary as an instance of *mv:Term* using the property *rdf:type*. The class *mv:ReplacementOnTerm* is to assert an Activity as an instance of *mv:ReplacementOnTerm* using the property *rdf:type*.

This paper assumes the following URIs to describe the headings: "Folklore, Negro" with "http://id.loc.gov/authorities/childrensSubjects/sj96004706," "Folklore, African" with "http://id.loc.gov/authorities/childrensSubjects/sj96004704," and "Folklore, Afro-American" with "http://id.loc.gov/authorities/childrensSubjects/sj96004705." An Activity instance of *mv:ReplacementOnTerm* made "Folklore, Negro" invalidated and generated two headings, i.e. "Folklore, African" and "Folklore, Afro-American." In the split of a LCSH term, the Library of Congress Subject Headings Supplemental Vocabularies: Children's Headings (LCSHAC) is a thesaurus that is used in conjunction with LCSH.

This paper identifies the thesaurus Entity before the split by URI "http://id.loc.gov/authorities/childrensSubjects/pv" and the thesaurus Entity after the split by URI "http://id.loc.gov/authorities/childrensSubjects/sv." These thesaurus Entities are named LCSHAC PV and LCSHAC SV, respectively. Figure 1(b) shows the derivation from LCSHAC PV to LCSHAC SV. LCSHAC SV was generated by an Activity instance of *mv:RevisionOnVocabulary* and LCSHAC PV became invalidated by the same Activity instance. The class *mv:Vocabulary* is defined to assert a metadata vocabulary as an instance of *mv:Vocabulary* using the property *rdf:type*. The class *mv:RevisionOnVocabulary* is defined to assert an Activity as an instance of *mv:RevisionOnVocabulary* using the property *rdf:type*. The Activity instance of *mv:RevisionOnVocabulary* connects with the Activity instance of *mv:ReplacementOnTerm* through the property *dcterms:hasPart*, which is used to describe the inclusion relationships between Activities in this study.

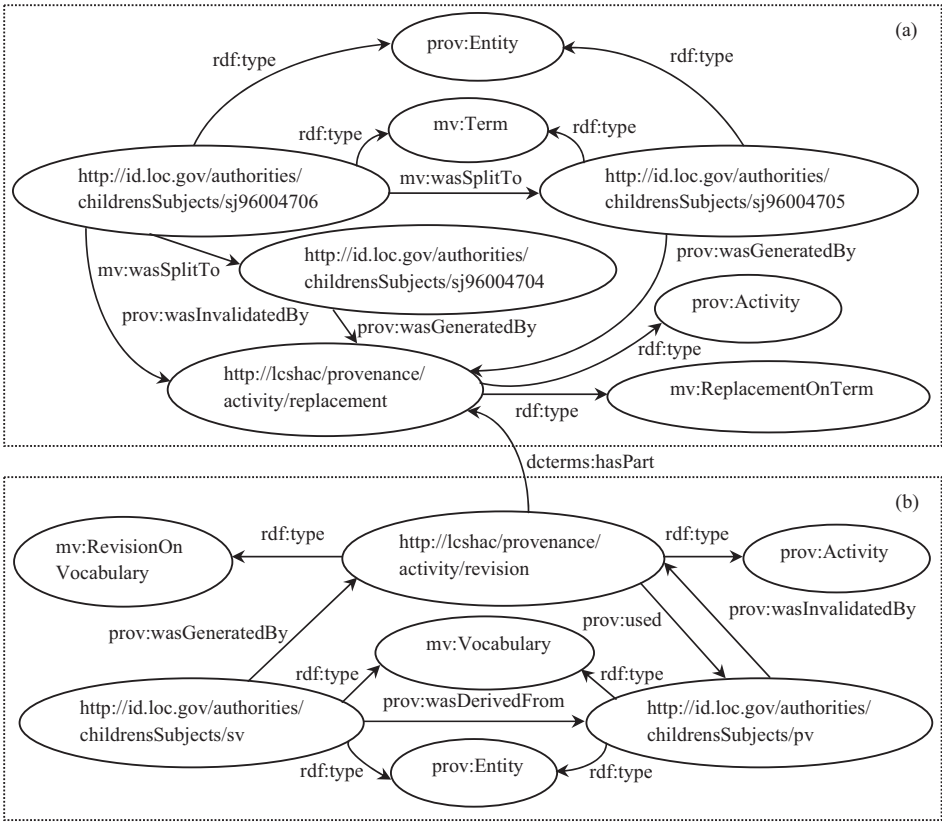

Figure 1. Example of provenance description of metadata vocabularies in RDF.

6 Discussion

The goal of this paper is to define a model for provenance description of metadata vocabularies based on W3C PROV and RDF. To achieve this, we defined primitive change types of metadata vocabularies and their metadata terms as shown in Tables 1, 2, and 3. Following the proposed model, the provenance description of metadata vocabularies and their metadata terms can be recorded in RDF, which is machine-readable and traceable using SPARQL. Keeping change history of metadata vocabularies traceable by machines is important to keep numerous metadata consistently interpretable.

The proposed model can describe the revision history of metadata terms. As shown in Figure 1(a), the subject heading “Folklore, Negro” (before the split) connects with “Folklore, African” and “Folklore, Afro-American” (after the split) through property *mv:wasSplitTo*. The proposed model can also describe the revision

history of documentation of metadata terms. For instance, the meaning of term "soundContent" in the RDA element set was changed from "Relates to an expression to a presence or absence of sound in a resource other than one that consists primarily of recorded sound" to "Relates to an expression to a presence or absence of sound in a resource" (RDA sound content, 2016).

Figure 2 defines the RDF model for the provenance description of a metadata term corresponding to the meaning revision example of term "soundContent." We use the URI "http://rdaregistry.info/Elements/e/P20225" from the RDA Registry to represent the term "soundContent" in an oval. The meaning of the term "soundContent" is supplied by the literal value of property *skos:definition* in a rectangle (solid line). The new meaning represented in lower dotted-rectangle was derived from the meaning represented in upper dotted-rectangle. The newly defined meaning was generated and the previously defined meaning became invalidated through the same Activity instance of *mv:ReplacementOnTermDefinition*.

Not only provenance description of metadata vocabularies but also provenance description of structural features of metadata is crucial for the long-term maintenance of metadata. Related to this paper, our previous papers present models for provenance description of metadata schemas (Li & Sugimoto, 2014; Li, Nagamori, & Sugimoto, 2015). The practical use and service development of metadata provenance to facilitate long-term maintenance of metadata is left as the future research.

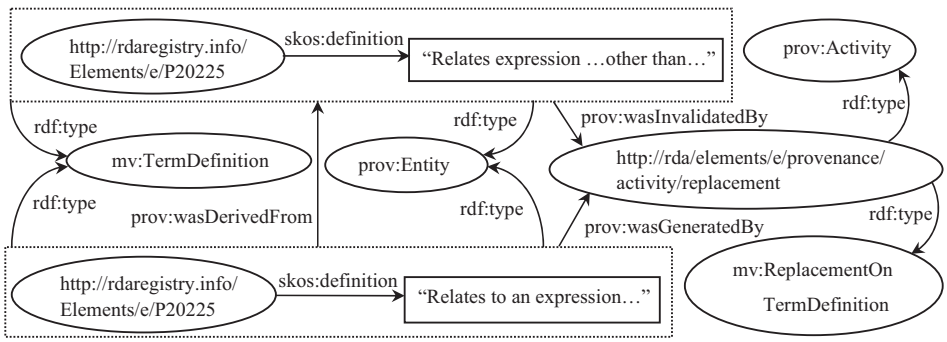

Figure 2. Example of provenance description of a metadata term in RDF.

7 Conclusion

Provenance tracking is an important issue for the long-term maintenance of metadata vocabularies. Evidence of such provenance of metadata vocabularies enables consistent maintenance of metadata vocabularies. This paper proposes a

model to formally describe provenance of metadata vocabularies, especially how metadata terms and term definitions (e.g. meaning and usage) change over time.

In this paper, the W3C PROV standard for general provenance description is applied to describe provenance of metadata vocabularies. We classified primitive change types of metadata terms in metadata vocabularies with specific examples. This study proposes a general model for provenance description of metadata vocabularies to track the primitive changes of metadata terms between different versions of a metadata vocabulary, e.g. split and merge of metadata terms and revision of meaning of metadata terms.

Acknowledgements

This study is supported in part by JSPS Kaken Grant-in-Aid for Scientific Research (A) (Grant No.: 16H01754). The study was initially presented at the 7th Asia-Pacific Conference on Library & Information Education and Practice (A-LIEP 2016) and selected for publication in *Journal of Data and Information Science* (*JDIS*) after peer review. We thank the two anonymous reviewers for their constructive comments, which helped us improve the manuscript.

Author Contributions

This paper is part of research results of the Ph.D. Project conducted by C.Q. Li (lichunqiuaa@126.com, corresponding author) under the guidance of Professor S. Sugimoto (sugimoto@slis.tsukuba.ac.jp). Both authors participated in the development of the provenance model and writing of the paper. Both authors proofread and approved the final manuscript.

References

BIBFRAME 2.0 vocabulary list view. (2016). Retrieved on June 20, 2016, from http://id.loc.gov/ontologies/bibframe.html.

BIBFRAME 2.0 specifications notes. (2016). Retrieved on June 20, 2016, from https://www.loc.gov/bibframe/docs/pdf/bf2-notes-june2016.pdf.

Chawuthai, R., Takeda, H., Wuwongse, V., & Jinbo, U. (2016). Presenting and preserving the change in taxonomic knowledge for linked data. Semantic Web, 7(6), 589–616.

Cuevas-Vicenttín, V., Ludäscher, B., Missier, P., Belhajjame, K., Chirigati, F., Wei, Y., … Cao, Y. (2016). ProvONE: A PROV extension data model for scientific workflow provenance. Retrived on June 20, 2016, from http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html.

Gil, Y., Miles, S., Belhajjame, K., Deus, H., Garijo, D., Klyne, G., … Zednik, S. (2013). PROV model primer. Retrieved on June 29, 2016, from http://www.inderscience.com/offer.php?id=63137.

Groth, P., Gil, Y., Cheney, J., & Miles, S. (2012). Requirements for provenance on the Web. The International Journal of Digital Curation, 7(1), 39–56.

Hyland, B., Atemezing, G., Pendleton, M., & Srivastava, B. (2013). Linked data glossary. Retrieved on June 20, 2016, from http://www.w3.org/TR/ld-glossary/.

Isaac, A., Waites, W., Young, J., & Zeng, M. (2011). Library linked data incubator group: Datasets, value vocabularies, and metadata element sets. Retrieved on June 20, 2016, from https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/.

Javed, M., Abgaz, Y.M., & Pahl, C. (2014). Layered change log model: Bridging between ontology change representation and pattern mining. International Journal of Metadata Semantics and Ontologies, 9(3). Retrieved on June 20, 2016, from http://doi.org/10.1504/IJMSO.2014.063137.

Kendall, E., Novacek, V., Baker, T., & Miles, A. (2008). Principles of good practice for managing RDF vocabularies and OWL ontologies. Retrieved on June 20, 2016, from https://www.w3.org/2006/07/SWD/Vocab/principles.

Knowlton, S.A. (2005). Three decades since prejudices and antipathies: A study of changes in the Library of Congress Subject Headings. Cataloging & Classification Quarterly, 40(2), 123–145.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., … Zhao, J. (2013). PROV-O: The PROV ontology. Retrieved on June 25, 2016, from http://www.w3.org/TR/prov-o/.

Lagoze, C., Willliams, J., & Vilhuber, L. (2013). Encoding provenance metadata for social science datasets. In E. Garoufallou, & J. Greeberg (Eds.), Metadata and Semantics Research: 7th Metadata and Semantics Research Conference, MTSR 2013, Thessaloniki, Greece (pp.123–134). Berlin: Springer-Verlag.

Li, C.Q., Nagamori, M., & Sugimoto, S. (2015). Temporal interoperability of metadata: An interoperablity-based view for longevity of metadata. In Proceedings of 6th International Conference on Asia-Pacific Library and Information Education and Practice (pp. 212–222). Manila, Philippines. Retrieved on June 25, 2016, from http://www.vub.ac.be/BIBLIO/nieuwenhuysen/presentations/2015-10-aliep-manila/2015%20ALIEP%20Proceedings%20with%20ISSN.pdf.

Li, C.Q., & Sugimoto, S. (2014). Provenance description of metadata using PROV with PREMIS for long-term use of metadata. In 2014 Proceedings of the International Conference on Dublin Core and Metadata Applications (pp. 147–156). Austin, Texas, USA. Retrieved on June 25, 2016, from http://dcpapers.dublincore.org/pubs/article/view/3709.

Meinhardt, P. (2015). Versioning linked datasets: Towards preserving history on the semantic web. Potsdam. (University of Potsdam master's thesis) Retrieved on June 25, 2016, from https://hpi.de/fileadmin/user_upload/fachgebiete/meinel/Semantic-Technologies/theses/Masterthesis-Meinhardt-2015.pdf.

Missier, P., & Chen, Z. (2013). Extracting PROV provenance traces from Wikipedia hstory pages. In Proceedings of the Joint EDBT/ICDT 2013 Workshops (pp. 327–330). New York: ACM.

Moreau, L. (2010). The foundations for provenance on the web. Foundations and Trends in Web Science, 2(2–3), 99–241.

Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., … Tilmes, C. (2013). PROV-DM: The PROV data model. Retrieved on June 29, 2016, from https://www.w3.org/TR/prov-dm/.

Moreau, L., Groth, P., Cheney, J., Lebo, T., & Miles, S. (2015). The rationale of PROV. Web Semantics: Science, Services and Agents on the World Wide Web, 35(Part 4), 235–257.

Masó, J., Closa, G., & Gil, Y. (2015). Applying W3C PROV to express geospatial provenance at feature and attribute level. In B. Ludäscher, & B. Plale (Eds.), Provenance and Anonotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9–13, 2014 (pp.271–274). Berlin: Springer-Verlag.

Phipps, J., Dunsire, G., & Hillmann, D. (2015). Building a platform to manage RDA vocabularies and data for an international linked data world. Journal of Library Metadata, 15(3–4), 252–264.

RDA sound content. (2016). Retrieved on June 30, 2016, from http://metadataregistry.org/history/list/vocabulary_id/93.html.

Tunnicliffe, S., & Davis, I. (2009). Changeset. Retrieved on June 30, 2016, from http://vocab.org/changeset/.

Tilmes, C., Fox, P., Ma, Xi., McGuinness, D., Privette, A.P., Smith, A., … Zheng, J. (2013). Provenance representation for the national climate assessment in the global change information system. IEEE Transactions on Geoscience and Remote Sensing, 51(11), 5160–5168.

Völkel, M., & Groza, T. (2006). Sem version: RDF-based ontology versioning system. In Proceedings of the IADIS International Conference on WWW/Internet 2006 (ICWI 2006, pp.195–202). Murcia, Spain. Retrieved on June 30, 2016, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.84&rep=rep1&type=pdf.