Metadata-Driven Approach for Keeping Interpretability of Digital Objects through Formal Provenance Description

Chunqiu Li
Graduate School of Library, Information and Media Studies,
University of Tsukuba
Japan
lichunqiuaa@126.com

Shigeo Sugimoto
Faculty of Library, Information and Media Science,
University of Tsukuba
Japan
sugimoto@slis.tsukuba.ac.jp

ABSTRACT

Metadata about digital objects help users find, understand, use and reuse those objects. Longevity of digital objects is a vital issue for digital preservation, which means that the metadata about digital objects must be maintained as well, so that their content and meaning should be maintained over time. Open Archival Information System (OAIS) defines three metadata components, which have to be maintained with Digital Object – Representation Information of Digital Object, Preservation Description Information (PDI) in an Information Package, and the Content Information given to every Information Package. Provenance of a digital object, which is one of the five categories of PDI, is a crucial record of the history of the object over its lifecycle. Since metadata are exchanged as digital objects on the Web, machine-readable and interoperable provenance description of metadata is required for the long-term maintenance of metadata. This paper presents issues in the longevity of metadata, especially the issue of metadata provenance based on the Singapore Framework for Dublin Core Application Profiles (DCAP), which is well known for metadata interoperability in the networked information environment. The paper first briefly discusses features of metadata as first class objects on the Web. Then, we address potential risks in affecting interpretability of digital objects and issues in the consistent maintenance of metadata. Next, the W3C PROV standard for general provenance description and Resource Description Framework (RDF) for metadata exchange on the Web are adopted as the base models for provenance description of metadata. We developed simple provenance description models for formal provenance description for both structural features and vocabularies of metadata. The models are designed based on Entities and Activities defined by the W3C PROV in correspondence with primitive changes of metadata application profiles and metadata vocabularies, respectively. We also provide formal provenance description examples corresponding to structural changes in a metadata application profile along with semantic changes in the use of its metadata vocabulary. We discuss limitations of our proposed models and review provenance related research. Finally, the main findings of this research are summarized in the conclusions.

KEYWORDS

Metadata, digital object, metadata schema, application profile, metadata vocabulary, provenance, long-term maintenance

1 INTRODUCTION

Long-term accessibility of digital collections requires keeping digital objects usable over time and across communities. Metadata plays an important role in the continued accessibility of digital collections and is used in a wide range of fields, such as computer science, library and information science, archival science, and so forth. Recent developments in research data sharing (e.g., Research Data Alliance [1]) and cultural resources aggregation (e.g., Europeana[2] and DPLA[3]) have also increased demands to keep research data and cultural collections alive over time. Maintaining metadata over time is important to keep research data and cultural contents reusable over time.

This paper discusses a metadata-centric study to keep digital objects interpretable, focusing on issues such as long-term maintenance of metadata schemas and metadata vocabularies. Metadata schema defines structural, syntactic and semantic features of metadata and uses metadata vocabulary that is a set of metadata terms to describe metadata record. In the long run, metadata provenance that records revision history of metadata schema and metadata vocabulary should be consistently recorded, since provenance is useful to auditing errors, justifying data authenticity, and identifying invalidated data, etc. In this paper, we developed simple provenance description models for metadata schemas and metadata vocabularies, respectively. The models enable maintainers to formally describe primitive revisions between two consecutive versions of a metadata schema and a metadata vocabulary in a machine-processable form, so that

[1] https://www.rd-alliance.org/
[2] http://www.europeana.eu/portal/en
[3] Digital Public Library of America, see https://dp.la/

their revision history can be consistently managed and effectively traced over time.

This paper provides a brief overview of the main results gained from our earlier research about metadata longevity conducted by the authors [1,2]. The rest of this paper is structured as follows. Section 2 analyzes features of metadata on the Web. Section 3 gives an overview of risks affecting metadata longevity. Section 4 discusses issues of metadata interpretability, long-term maintenance of metadata schema and metadata vocabulary, as well as metadata provenance. Section 5 presents our newly devised models for formal provenance description of metadata, and Section 6 discusses the formal provenance descriptions of metadata application profiles and metadata vocabularies based on our models using examples for illustration. Section 7 presents limitations of our models and briefly reviews related research on provenance description. Section 8 summarizes main findings of this study.

2 FEATURES OF METADATA ON THE WEB

Metadata (Greek: meta- + Latin: data "information") [3] as structured data is generally defined as "data about data". In library domain, a card catalog and its electronic counterpart are common examples of metadata. Metadata can be an object in databases or in systems. Metadata in the networked information environment has features different from conventional metadata that is primarily designed for use on a single database or a set of databases. An instance of metadata on the Web is no longer an object enclosed in a database, but is a digital object which is transferred from one site to another and shared among those sites. We call such metadata object as a "first class object". This paper discusses features of metadata on the Web as follows [4,5].

Structural Features: Metadata is typically structured according to a scheme. Structural features of metadata are assertions about data structure, mandatory levels, iteration constraints of description, and so forth. Such assertions represent attributes and values of resource in machine-readable form.

Syntactic Features: Metadata can be serialized in different syntaxes, e.g., HTML, XML, RDF/XML, Turtle, JSON, JSON-LD.

Semantic Features: The elementary semantics of metadata are specified and defined in a metadata vocabulary. The meaning of each metadata term and relationships between terms are identified as the semantic features of metadata. Uniform Resource Identifier (URI) is used as the base scheme to identify a term in the Linked Open Data (LOD) environment.

Metadata interoperability is crucial not only across communities but also over time. Metadata standards are the basis for interoperability of metadata [6]. However, schemes for metadata interoperability over time are still not well developed. In the long term, metadata need to be kept interpretable not only by humans but also by machines. We need to understand the potential risks and develop strategies to keep metadata instances consistently interpretable over time.

3 RISKS IN INTERPRETABILITY OF DIGITAL OBJECTS

Digital objects are preserved as a sequence of bits. It is of importance to ensure that the bits remain intact and correct over time. However, bit preservation alone is not sufficient for the long-term preservation of digital objects. Digital objects should be kept interpretable across the changes in many aspects over time. As OAIS reference model defines three metadata components, Representation Information, Preservation Description Information (PDI) and Content Information, metadata has to be preserved with digital objects. Those metadata may be stored in a database with the preserved digital objects as an archival information package (AIP). This means that metadata schemas and vocabularies used in those metadata have to be maintained over time as well as those AIPs. In the LOD environment, metadata schemas and vocabularies are digital objects, which should be kept usable for long-term as well as the preserved digital objects.

Preservation of digital objects requires preventing damages or loss of digital objects. We need to manage risks of damages or loss in the preservation process of digital objects. Risk management for keeping metadata and their schemas safe is a crucial research issue. In the OAIS reference model, risk management is an essential part of preservation planning [7]. However, OAIS does not discuss risk management for metadata and their schemas exchanged on the Web. Therefore, this paper analyzes the risks in effecting interpretability of digital objects from the viewpoint of metadata. The following risks might cause inconsistency in digital preservation: (1) Metadata schema and metadata vocabulary for a digital object may be unknown, improperly recorded, lost, changed, or obsolete, (2) Metadata schema and metadata vocabulary for a digital object may be improperly maintained and their revision history may not be consistently recorded, (3) Provenance information about the digital objects and their metadata schemas as well as metadata vocabularies may not be consistently recorded in machine-processable form, and (4) Resource identifiers of any of those instances may be inconsistent.

The following functions have to be studied in order to reduce the risks in metadata longevity: (1) preserving the documents of metadata schema and metadata vocabulary, (2) recording and maintaining metadata schema and metadata vocabulary along with their revision history, (3) recording necessary and interoperable provenance of metadata schema and metadata vocabulary, and (4) creating sustainable identification mechanisms and schemes for metadata instances.

4 LONG-TERM MAINTENANCE OF METADATA

This section discusses long-term maintenance of metadata schemas, long-term maintenance of metadata vocabularies, and metadata provenance as key aspects of metadata longevity.

4.1 Metadata Preservation: Keeping Temporal Interoperability of Metadata

Metadata plays crucial roles in digital collections, e.g., finding aids, rights management, etc. Therefore, metadata should be preserved as well as the digital resources in the collection.

PDI of OAIS has five categories which are Provenance, Reference, Fixity, Context, and Access Rights Information. Provenance documents the history of content information, and tells the origin or source of content information, and any changes that may have taken place since it was originated, who has had custody of it since it was originated, providing an audit trial for the content information [8]. Provenance provides the credibility information about a preserved resource to the users in the future as it contributes to evidence supporting authenticity. Provenance describes change history of a digital object over time and can be viewed as a special type of context information [9].

This paper focuses on the formal provenance description of the structural features of metadata and vocabularies defined for metadata. In this paper, we use the terms Metadata Schema and Metadata Vocabulary. We define metadata schema as a description of scheme that defines structural features of metadata and metadata vocabulary as a controlled set of terms defined for metadata. For example, Simple Dublin Core defines a metadata vocabulary composed of 15 metadata terms and very general structural features where any defined element is repeatable and optional. The metadata vocabulary of Simple Dublin Core is known as Dublin Core Metadata Element Set[4].

The Singapore Framework for Dublin Core Application Profile (DCAP) defines the components of metadata schemas for an application and related components, such as metadata vocabulary. A DCAP has five components, which are Usage Guidelines, Syntax Guidelines and Data Formats, Functional Requirements, Domain Model, and Description Set Profile [10]. The framework is developed based on the Web standards (e.g., RDF[5], RDF/S[6]) and is used as a basis for our research to discuss issues with the longevity of metadata. This is because the framework separates the structural and semantic features of metadata. In particular, we focus on provenance description of Description Set Profile in this paper.

In the LOD environment, RDF/S and OWL[7] are basic schemes to define metadata terms and vocabularies. Each term is defined as a resource identified by URI. We discuss longevity of metadata vocabulary from the viewpoint of provenance description of metadata terms. Term mapping between metadata vocabularies is often done to merge two or more sets of metadata described on different metadata vocabularies. Mapping itself is a crucial data resource to keep a record of the merger for future purposes. However, in this paper, maintaining mapping over time is not discussed although it is a crucial issue for long-term use of metadata.

4.2 Long-term Maintenance of Metadata Schemas

A metadata schema should be preserved as well as metadata instances created from the schema. Metadata schemas are preserved as a document for human readers in a conventional maintenance environment of metadata. The state of the art environment of the Web provides standards and models to formally describe metadata schemas in a machine-processable form, e.g., RDF, OWL and the Singapore Framework for Dublin Core Application Profiles. Metadata schemas are no longer simple document-like objects, but are complex objects transferrable over networks. Therefore, we need to develop technologies to maintain metadata schemas and vocabularies over time for the longevity of metadata.

In this paper, we propose a formal model for provenance description of structural constraints of metadata schema based on the Description Set Profile of Singapore Framework. The changes between two consecutive versions of a metadata schema should be recorded as provenance description of the metadata schema. The advantage of formal provenance description of metadata schemas over conventional change-logs is automated auditing to help find errors and inconsistencies between the different versions of the metadata schema.

4.3 Long-term Maintenance of Metadata Vocabularies

It is recommended to reuse existing and well-known metadata vocabularies to improve semantic interoperability of metadata. This means that application metadata schemas rely on standard metadata vocabularies, in particular maintenance of the definition of metadata terms. For example, changes in descriptions of meanings of a term may affect application metadata schemas that use the term. However, many metadata vocabularies have no clear policy regarding their change documentation [11], and hence monitoring metadata vocabularies is of importance. Maintenance agencies of standard vocabularies should have policies for maintaining the vocabularies for their sustainability.

[4] http://www.dublincore.org/documents/dces/
[5] http://www.w3.org/TR/rdf11-primer/
[6] http://www.w3.org/TR/rdf-schema/
[7] http://www.w3.org/TR/owl-features/

A metadata vocabulary is comprised of a set of terms and relationships between terms. When a newly defined version of a metadata vocabulary is published, there may be some changes from its previous version, e.g., the meaning of a term can be changed, relationship between terms can be revised, a composite term can be split to single terms, or single terms can be merged into one composite term. The changes to a metadata vocabulary should be also consistently recorded as provenance of the metadata vocabulary.

4.4 Metadata Provenance

Provenance (from the French provenir, "to come from") is the description of the history of an object. Provenance is used in several fields, such as identifying authorship of art works, justifying trustworthiness of data, reproducibility of scientific research, etc. Provenance of a digital object describes how the digital object came to the current state since its origination.

For long-term preservation of digital objects, we need to record provenance of the digital objects in a digital form, which we call digital provenance in this paper. Digital provenance is included in well-known preservation standards, such as the OAIS reference model and the PREMIS standard for preservation metadata. Digital provenance typically describes agents responsible for the custody and stewardship of digital objects, key events that occur over the course of the digital object's life cycle, and other information associated with the digital object's creation, management and preservation [8,9,12]. The OAIS reference model is a generalized package-based model but is not oriented to the Web environment. PREMIS is continually maintained and its OWL ontology as well as XML schema are being revised. However, metadata longevity is not well explored in either of these standards and is still a new research topic.

Metadata provenance is a record that typically describes responsible agents, influencing actions, associated events and other related information about metadata over its lifecycle. Both provenance of metadata schemas and provenance of metadata vocabularies are metadata provenance. W3C PROV standard[8] is a Web-oriented provenance standard for general provenance description and provenance interchange. In the Web environment, machine-processable, interoperable and traceable provenance is required. The following sections present formal provenance description of metadata proposed by the authors which has its basis in W3C PROV standard and in RDF.

5 FORMAL PROVENANCE DESCRIPTION OF METADATA

Provenance description of metadata can be in various forms, such as natural language, RDF, etc. This study records provenance description of metadata in RDF, which is widely used for metadata exchange on the Web. Formal provenance description in RDF holds advantages over semi-formal and informal provenance description.

5.1 W3C PROV for Metadata Provenance

The Provenance Working Group at W3C has published the PROV family of documents, including the PROV Data Model (PROV-DM), PROV Ontology (PROV-O), and so forth. The working group aims at the interoperable interchange of provenance information in heterogeneous environments such as the Web. PROV-DM is a conceptual data model, which defines a set of concepts and relations to represent provenance [13]. PROV-O defines a set of classes and properties as an OWL2 ontology allowing mapping PROV-DM to RDF [14].

An *Entity* is a physical, digital, conceptual, or other kinds of thing. An *Activity* is something that occurs over a period of time and acts upon or with *Entities*. *Activity* is used to represent how an *Entity* comes into existence and how attributes of an *Entity* change to become a new *Entity*. W3C PROV defines relationships between *Entities*, relationships between *Entities* and *Activities*, relationships between *Activities*, and other relation types. To describe metadata provenance, we classified *Entities* and *Activities* affecting revisions to metadata application profiles and metadata vocabularies, respectively. The following two sections will briefly address provenance description of metadata application profiles and metadata vocabularies.

5.2 Model for Formal Provenance Description of Metadata Application Profiles

This section shows the DSP-PROV model with functions to describe deletion, addition and revision of structural features of metadata schema. The classified *Activities* are generalized into three change types – *Addition*, *Deletion* and *Revision*. The classified *Entities* are Description Set Profile (DSP) and its components, which are named as structural schema instances in this paper.

DSP defines structural constraints of metadata. There are two levels of templates in a DSP, i.e., Description Template (DT) and Statement Template (ST). DTs contain "statement templates that apply to a single kind of description as well as constraints on the described resource". STs contain "all the constraints on the property, value strings, vocabulary encoding schemes, etc. that apply to a single kind of statement" [10]. Structural Constraints (SCs) contain mandatory levels, iteration constraints and other constraints on properties and property values defined in statement templates.

[8] http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Fig. 1 depicts the DSP-PROV model using UML Class diagram: (1) *Generalization* is represented with a hollow triangle on super-classes (i.e., *Entity* and *Activity*), (2) *Aggregation* is represented with a diamond on containing classes (for example, DSP, RevsionOnDSP), and (3) *Association* is represented by a line with an arrow that describes the relationship between *Entity* and *Activity*. The DSP-PROV model uses the properties from PROV-O when applicable. PROV *Invalidation* and PROV *Generation* respectively represent the deletion and addition of structural schema instances. PROV *Derivation*, PROV *Invalidation*, PROV *Generation* and PROV *Usage* together describe the revision of structural schema instances. If applicable, DSP-PROV can also describe relationships between *Activities* in the cases where an *Activity* used an *Entity* generated by another *Activity*.

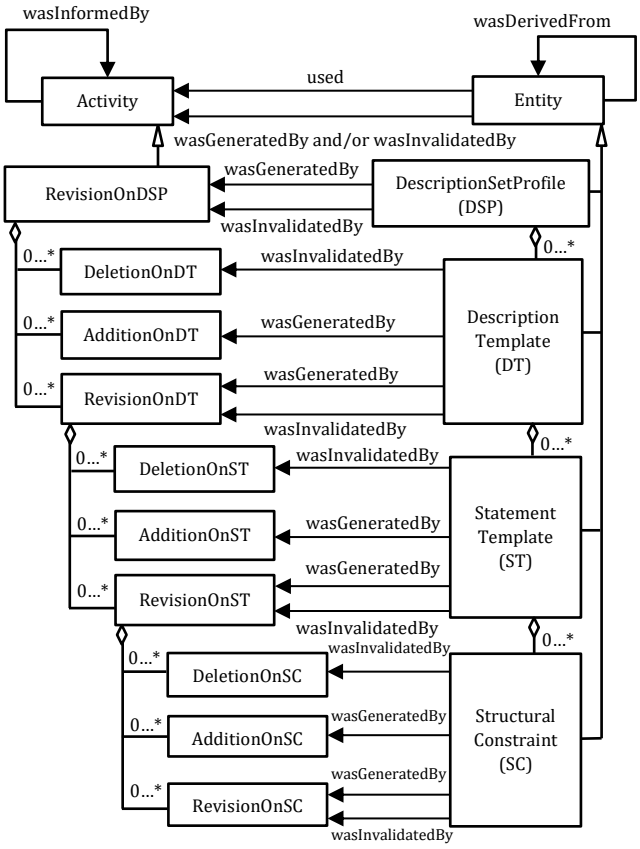

Figure 1: Overview of the DSP-PROV model.

Fig. 2 shows the classified *Activities* to describe structural changes of metadata schema. The naming convention of the *Activities* in this paper is "Activity Type + On + Abbreviation of structural schema instance". For instance, *Revision Activity* that acted upon a DT and led it to a new DT is named as an *Activity* instance of *RevisionOnDT*. Fig. 2 also indicates the relationships between classified *Activities*. The *Revision Activity* acted upon the containing *Entity* (e.g., a DSP) has sub-activities – *Deletion*, *Addition* and *Revision* acted upon its contained *Entity* (e.g., a DT of the DSP). DT changes caused by *DeletionOnDT*, *AdditionOnDT* and *RevisionOnDT* will result in DSP changes caused by *RevisionOnDSP*. Therefore, *RevisionOnDSP* has sub-activities, i.e., *DeletionOnDT*, *AdditionOnDT* and *RevisionOnDT*. Similarly, we can get the following two conclusions: *RevisionOnDT* has sub-activities, i.e., *DeletionOnST*, *AdditionOnST* and *RevisionOnST*; *RevisionOnST* has sub-activities, i.e., *DeletionOnSC*, *AdditionOnSC* and *RevisionOnSC*. This paper uses property *dcterms:hasPart* that is recommended for modeling sub-activities [15].

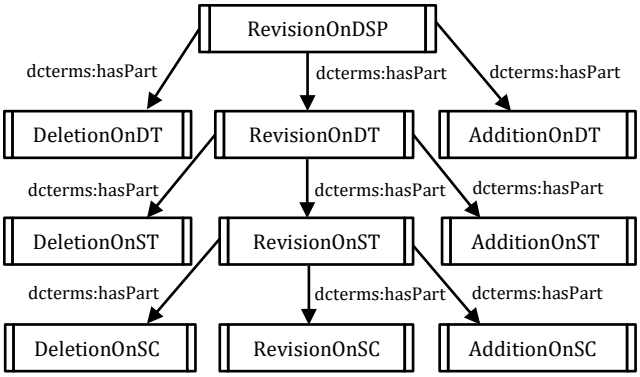

Figure 2: Activity relationships in the DSP-PROV model.

5.3 Model for Formal Provenance Description of Metadata Vocabularies

Fig. 3 shows Vocab-PROV model with functions to describe primitive changes of a metadata vocabulary and its metadata terms. The approach for building the Vocab-PROV model is similar to DSP-PROV model. We classified the *Activities* and *Entities* to describe the changes of metadata vocabularies.

Vocabulary, Term and Term Definition are classified as three subtypes of PROV *Entity* to describe the 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*. The *Replacement* is defined to describe the change cases such as term split and term merge.

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 the documentation for 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 the term definition.

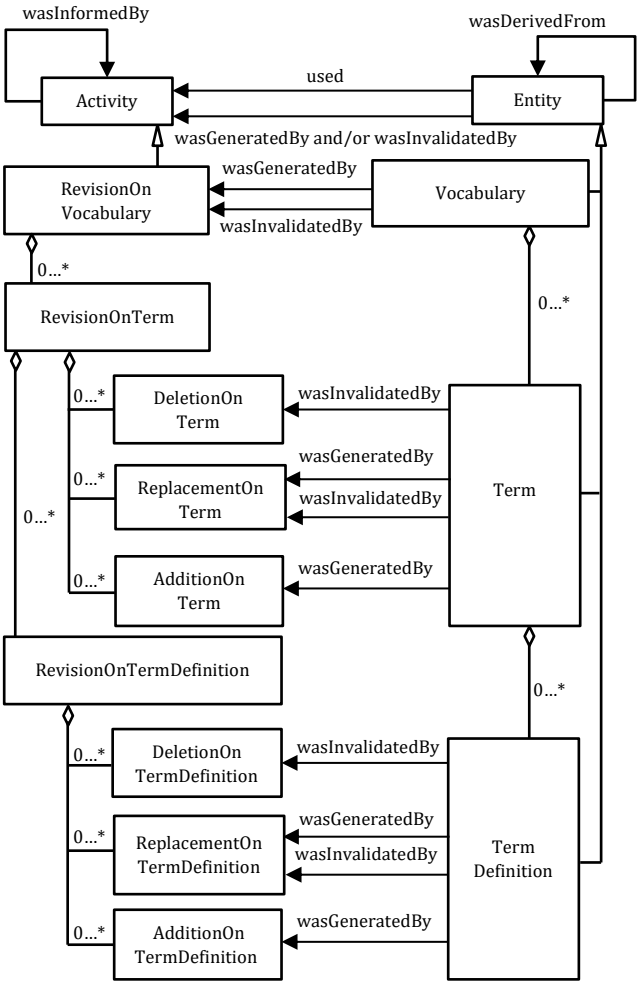

Figure 3: Overview of the Vocab-PROV model.

Fig. 4 shows the relationships between the classified *Activities* to describe provenance description of metadata vocabularies. A *RevisionOnVocabulary* is comprised of zero or more than one *RevisionOnTerm* and *RevisionOnTermDefinition*. Similarly, *RevisionOnTerm* has sub-activities that are *AdditionOnTerm*, *DeletionOnTerm* and *ReplacementOnTerm*; *RevisionOnTermDefinition* has sub-activities that are *AdditionOnTermDefinition*, *DeletionOnTermDefinition* and *ReplacementOnTermDefinition*.

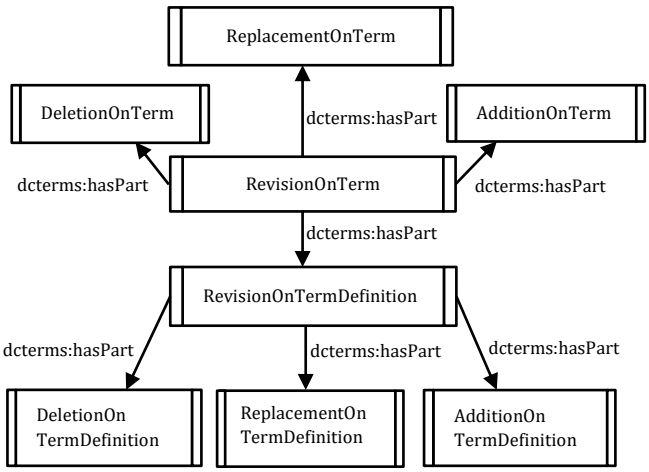

Figure 4: Activity relationships in the Vocab-PROV model.

6 FORMAL PROVENANCE DESCRIPTION EXAMPLES

A metadata application profile usually uses terms defined in existing metadata vocabularies. However, the metadata application profile may use the term meaning, which may be narrowed from the original meaning for better fit of the meaning to the application. The terms included in an existing vocabulary are usually defined within the namespace of the vocabulary without version information. Therefore, in this section, we do not take into account the versions of the terms and we focus on the changes of term meaning defined in the metadata application profiles.

Section 6.1 provides an example of semantic change and structural change that we found from the documents of DPLA MAPs. Section 6.2 and 6.3 discuss the provenance description about the changes using RDF graphs based on our proposed Vocab-PROV model and DSP-PROV model, respectively. In Section 6.4, we briefly present the relationships between the semantic change and structural change in the given change examples.

6.1 Example for Semantic Change along with Structural Change

Digital Public Library of America Metadata Application Profile (DPLA MAP) [9] defines structural constraints of metadata, which include property, usage, obligation, range and others information in tabular form. DPLA MAP uses classes and properties from existing vocabularies, such as EDM, ORE, DC, DCTERMS, DCMITYPE, Geo vocabulary, etc. Three versions of DPLA MAP have been released, i.e., V3, V3.1, V4. DPLA MAP does not provide exact meaning and definition for its classes and properties. The value of the "Usage" column

[9] https://dp.la/info/developers/map/

provides the kind of information related to meaning and definition of a term (i.e., class and property), which is written as the value of "Term Meaning" in the table titled "comparison between V3.1 and V4" of Fig. 5. DPLA MAP V3.1 provides changes from V3 to V3.1 and DPLA MAP V4 provides changes from V3.1 to V4 in natural language.

Fig. 5 shows change examples from DPLA MAP V3.1 to V4 that includes both structural change in DPLA MAP and semantic change in its metadata vocabulary. Both DPLA MAP V3.1 and DPLA MAP V4 define property *edm:object* for class *ore:Aggregation* to describe "object". The meaning of *edm:object* in DPLA MAP V3.1 and V4 are "Unambiguous URL to the DPLA content preview" and "The URL of a suitable source object in the best resolution available on the website of the Data Provider from which *edm:preview* could be generated for use in a portal", respectively. Fig.5, Fig.6 and Fig. 8 use a short expression of the two definitions, i.e., "Unambiguous…" and "The URL…".

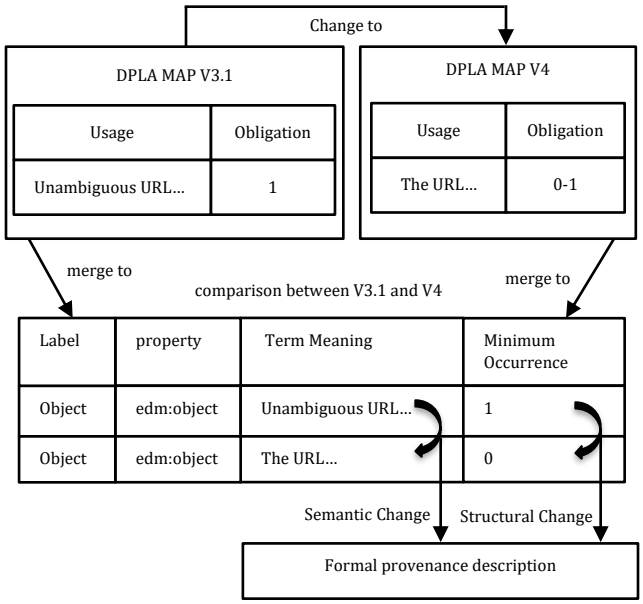

Figure 5: Example of semantic change along with structural change.

6.2 Formal Provenance Description for Semantic Change of Metadata Term

As shown in Fig. 5, the meaning of the term *edm:object* has been changed. The change was caused by a *replacement activity*, which is an instance of the class *mv:ReplacementOnTermDefinition* (note: "mv" is the prefix for the classes of the Vocab-PROV model [2]). According to our proposed Vocab-PROV model, we can formally describe the provenance description including the derivation of term definition as RDF graphs depicted in Fig. 6.

The meaning of the term *edm:object* is expressed as the literal value of property *skos:definition* in a rectangle (solid line). The new meaning represented in the lower dotted-rectangle was derived from the meaning represented in the upper dotted-rectangle. The newly defined meaning was generated and the previously defined meaning became invalidated through the same *Activity* instance of *mv:ReplacementOnTermDefinition*.

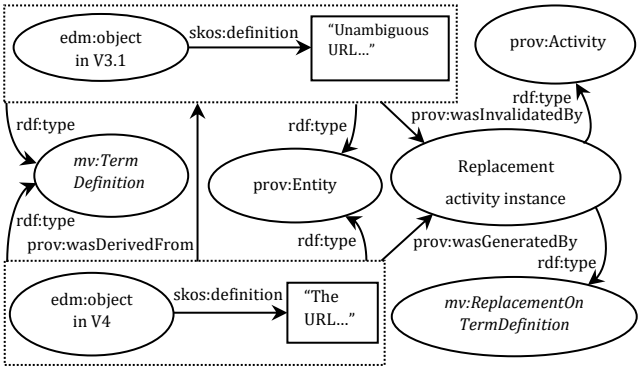

Figure 6: Provenance description for the above semantic change using RDF graphs.

6.3 Formal Provenance Description for Structural Change of Structural Constraint

As shown in Fig. 5, the minimum occurrence of *edm:object* has been changed from "1" to "0". The change was caused by a *revision activity*, which is an instance of the class *dspprov:RevisionOnSC* (note: "dspprov" is the prefix for the classes of the DSP-PROV model [1]). Fig. 7 shows RDF graphs of the provenance description about the structural constraint change.

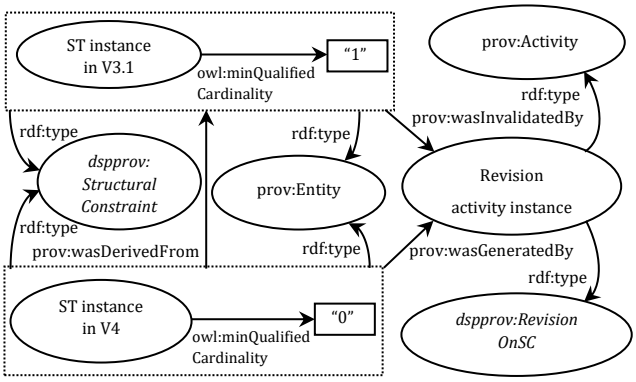

Figure 7: Provenance description for the above structural change using RDF graphs.

The minimum occurrence constraint defined in the Statement Template (ST) instance that defines all the structural constraints on the property *edm:object* is expressed as the literal value of property *owl:minQualifiedCardinality*. The new Structural Constraint (SC) represented in the lower

dotted-rectangle was derived from the previous constraint in the upper dotted-rectangle. The newly defined minimum occurrence constraint was generated and the previously defined minimum occurrence constraint became invalidated through the same *Activity* instance of *dspprov:RevisionOnSC*.

6.4 Connection between Semantic Change with Structural Change

In general, a metadata application profile uses terms from metadata vocabularies. Semantic changes of the terms used in a metadata application profile may be synchronized with structural changes of the metadata application profile. This section shows linkage of semantic change on a term and structural change in a metadata application profile. Figs. 6 and 7 show the provenance description about the semantic change and structural change in the examples given in Section 6.1. As shown in Fig. 8, the connection between Figs. 6 and 7 is the property constraint in the Statement Template (ST), which is expressed as the resource value of *owl:onProperty*.

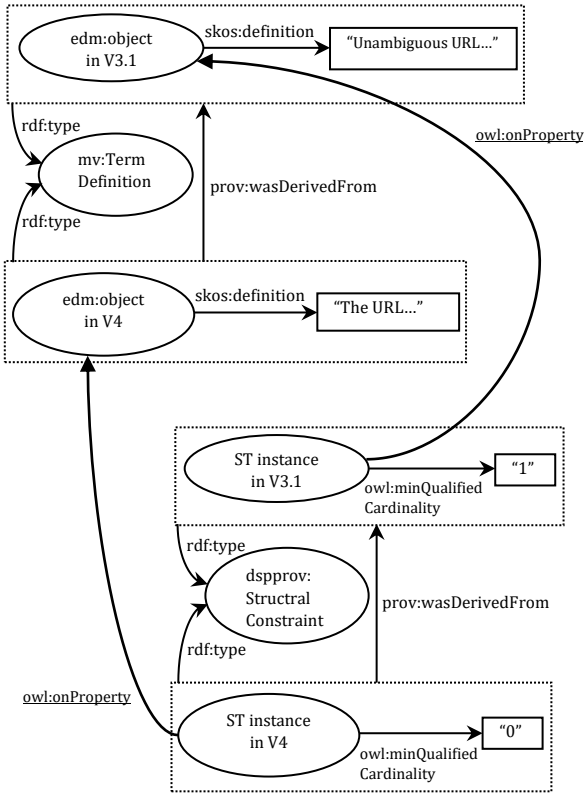

Figure 8: Connection between semantic change and structural change.

7 DISCUSSION

In Section 7.1, we briefly discuss the limitation and implication of our proposed provenance models, i.e., DSP-PROV model and Vocab-PROV model. In Section 7.2 and 7.3, we review the current state of provenance description.

7.1 Limitation of the Proposed Models

The DSP-PROV model is designed based on Description Set Profile of Dublin Core Application Profile. However, there are no worldwide standards for the development of metadata application profile as not all metadata application profiles follow the same structure. This means the DSP-PROV model may not be applicable to all existing metadata application profiles.

The Vocab-PROV model generalizes primitive changes in a metadata vocabulary. In practice, the changes to a metadata vocabulary may be complex especially when considering temporal information over a long period. The Vocab-PROV model may not be applicable to complex changes. The maintenance of semantics of terms is a challenging issue and semantic interoperability of terms over time is difficult to achieve.

7.2 Provenance Description Models and Vocabularies

In our earlier research, we conducted survey of existing provenance description models and vocabularies [16]. There are already a wide range of models, ontologies and vocabularies that can be used for provenance description, such as Open Provenance Model (OPM), Open Provenance Model Vocabulary (OPMV), Open Provenance Model OWL Ontology (OPMO), Open Provenance Model (OPM) for Workflows (OPMW), Provenance Vocabulary (PRV), Vocabulary for Data and Dataset provenance (Voidp), Provenance, Authoring and Versioning Ontology (PAV), W7 Model, Provenir Ontology, BBC Provenance Ontology, W3C PROV standard and others. The CIDOC Conceptual Reference Model (CRM) in the museum community has also been extended to model provenance information of digital objects [17].

Provenance description is required in both conventional and Web environment. Recording provenance in a form interpretable by both computers and humans is required. However, existing technologies and standards are not specialized for metadata schema and metadata vocabulary. Specially, models for formal provenance description of metadata are not sufficiently explored. In the Web environment, there is a need to develop models for formal metadata provenance description. It is because that formal provenance description of metadata in machine-readable and interoperable form supports automated and effective metadata maintenance. In this study, we have developed models for provenance description of metadata application profiles and metadata vocabularies, respectively.

7.3 Provenance Description in Different Domains

Provenance related research has been conducted in a wide range of domains, such as museum, library and archive (MLA) community, computer science and e-science, etc. Provenance information can be used to identify authorship, ownership and authentication of objects, e.g., the Council of European Research Libraries (CERL) [10] website and the Getty Provenance Index Databases[11] provide search services for provenance information. International Research into the Preservation of Authentic Records in Electronic Systems (InterPARES) [12] project addresses the importance of provenance for keeping trustworthiness of digital records. Provenance in e-science can be used to reproduce research data.

The Bodleian libraries at the University of Oxford devised a data model to represent contextual information of research outputs in the Oxford University Research Archive (ORA) [13], which is a long-term data repository for scholarly research outputs. The model incorporates PROV-DM to describe activity related to research outputs, e.g., creation activity, funding activity, and publication activity. Activity-based description of relationships for a journal article using PROV-O is given as an example [18].

Capturing the provenance information of electronic records is a concern for archivists. Conventional provenance in the arrangement of archival records arises from their creators, for example, individuals, cooperated bodies or families [19]. The scope of provenance for archival records encompasses to creator history, records history, and custodial history. The archival standards such as, General International Standard Archival Description (ISAD(G)), Encoded Archival Description (EAD), International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR(CPF)), and Encoded Archival Context (EAC) define the description elements for provenance information. The recordkeeping metadata standard ISO 23081 provides us reference to capture audit trails in the records management process [20].

Getty Thesaurus of Geographic Names[14] 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.

As introduced above, many communities have paid attention to provenance description, especially the change history and activity related to objects. However, these provenance description elements are designed for specific domain requirements and not generalized for metadata provenance. That is, they cannot be directly applied to describe provenance of metadata application profiles and metadata vocabularies. Therefore, the aim of this research to propose general models for provenance description of metadata is novel.

[10] https://www.cerl.org/resources/provenance/main
[11] http://www.getty.edu/research/tools/provenance/search.html
[12] http://www.interpares.org/
[13] http://ora.ox.ac.uk/
[14] http://www.getty.edu/research/tools/vocabularies/tgn/

8 CONCLUSIONS

This paper addresses issues in longevity of metadata, especially temporal interoperability of metadata over time. The main contributions of this paper are: (1) development of the DSP-PROV model for provenance description of metadata application profiles, (2) development of the Vocab-PROV model for provenance description of metadata vocabularies, and (3) provision of examples of formal provenance description, especially considering both the structural constraint changes of metadata schemas along with the semantic changes of metadata vocabularies. Due to the space limitation of this paper, we will not specifically introduce the evaluation and description examples of the two proposed models. The details are presented in our other two papers [1,2].

The provenance description of metadata application profiles and metadata vocabularies together can reveal the revision history of structural features and semantic features of metadata instances, which can help users to interpret metadata instances. These descriptions in machine-processable form on the Web can be traced using Semantic Web technologies.

The main findings of this article are summarized into the following points: (1) Provenance information is crucial component to keep longevity of digital objects, (2) Provenance information should be consistently recorded in machine-processable form on the Web, (3) The devised DSP-PROV model and Vocab-PROV model enable us to keep track structural changes of metadata schemas and primitive changes of metadata terms, respectively, and (4) Formal provenance description holds advantages over provenance description in natural language. For instance, formal provenance description helps consistent maintenance of metadata over time; formal provenance description can be used to find errors in semi-provenance description that is recorded in natural language. In the future, we expect to explore practical services of provenance use cases in memory institutions.

Furthermore, implementation of existing provenance models with metadata standards (e.g., PREMIS dictionary; controlled vocabularies of Library of Congress) is also an applicable approach for provenance description of metadata. Our previous paper used this approach and briefly discussed provenance description of metadata schemas through combining the core of PROV data model with PREMIS data model [18]. In this paper, we adopted W3C PROV and RDF to propose provenance model for metadata longevity from the aspects of metadata application profile longevity and metadata vocabulary longevity.

Metadata provenance is still a new topic. We defined models for provenance descriptions. The models rely on several infrastructure issues in the practical Web environment, such as longevity of namespaces and URIs, and long-term maintenance of widely used vocabularies. This paper does not discuss these fundamental issues because they need to be discussed in the digital preservation communities and other wider communities due to their importance. We consider that this paper provides several fundamental issues for discussion for the future development of digital preservation.

ACKNOWLEDGMENTS

We express sincere thanks to professor Atsuyuki Morishima, associate professor Tetsuo Sakaguchi, and assistant professor Mitsuharu Nagamori for their contributions to this research through discussions. We thank Mr. Tsunagu Honma and all students at our laboratory for their kind help. We express our special thanks to Dr. Bhuva Narayan at University of Technology Sydney for her help to elaborate this paper. This work was partially supported by JSPS Kaken Grant-in-Aid for Scientific Research (A) (Grant No.: 16H01754).

REFERENCES

[1] Chunqiu Li and Shigeo Sugimoto. 2017. Provenance Description of Metadata Vocabularies for Long-term Maintenance of Metadata Schemas. Journal of Documentation. paper draft. Under revision based on peer-review.

[2] Chunqiu Li and Shigeo Sugimoto. 2017. Provenance Description of Metadata Vocabularies for the Long-term Maintenance of Metadata. Journal of Data and Information Science 2, 2, 41-55.

[3] Anne J. Gilliland, Tony Gill, Mary S. Woodley, and Maureen Whalen. 2008. Metadata and the Web. Introduction to Metadata (2nd. Ed.). Getty Research Institute, Los Angeles. 20-37.

[4] Jane Greenberg. 2003. Metadata and the World Wide Web. Encyclopedia of Library and Information Science 3, 1876-1888.

[5] Jane Greenberg. 2009. Understanding Metadata and Metadata Schemes. Cataloging & Classification Quarterly 40, 3-4, 17-36. DOI: http://dx.doi.org/10.1300/J104v40n03_02

[6] Lois Mai Chan and Marcia Lei Zeng. 2006. Metadata Interoperability and Standardization–A Study of Methodology Part I. D-Lib Magazine 12, 6. Retrieved from http://www.dlib.org/dlib/june06/chan/06chan.html

[7] Stefan Hein and Karlheinz Schmitt. 2013. Risk Management for Digital Long-Term Preservation Services. In *Proceedings of the 10th International Conference on Digital Preservation* (iPRES 2013). Lisbon. Retrieved from http://phaidra.univie.ac.at/o:378059

[8] Consultative Committee for Space Data System. 2012. CCSDS 650.0-M-2. Reference Model for Open Archival Information System (OAIS). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

[9] Catherine Lupovici and Masanès Julien. 2000. Metadata for Long-term Preservation. Biblioteque Nationale de France, NEDLIB Consortium.

[10] Mikael Nilsson, Thomas Baker, and Pete Johnston. 2008. The Singapore Framework for Dublin Core Application Profiles. Retrieved from http://dublincore.org/documents/singapore-framework/

[11] Baker Thomas, Vandenbussche Pierre-Yves, and Vatant Bernard. 2013. Requirements for Vocabulary Preservation and Governance. Library Hi Tech 31, 4, 657-668. DOI: http://dx.doi.org/10.1108/LHT-03-2013-0027

[12] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata. Version 3.0. Retrieved from http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

[13] Moreau Luc, Paolo Missier, Khalid Belhajjame,Reza B'Far,James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, Simon Miles, James Myers, Satya Sahoo, and Curt Tilmes. 2013. PROV-DM: The PROV Data Model. Retrieved from http://www.w3.org/TR/prov-dm/

[14] Lebo Timothy, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. PROV-O: The PROV Ontology. Retrieved from http://www.w3.org/TR/prov-o/

[15] W3C Semantic Web. PROV-FAQ. 2014. Retrieved from https://www.w3.org/2001/sw/wiki/PROV-FAQ

[16] Chunqiu Li and Shigeo Sugimoto. 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*. Dublin Core Metadata Initiative, USA, 147–156.

[17] Maria Theodoridou, Yannis Tzitzikas, Martin Doerr, Yannis Marketakis, and Valantis Melessanakis. 2010. Modeling and Querying Provenance by Extending CIDOC CRM. Distributed and Parallel Databases 27, 2, 169-210. DOI: http://dx.doi.org/ 10.1007/s10619-009-7059-2

[18] Tanya Gray Jones, Lucie Burgess, Neil Jefferies, Ansha Ranganathan, and Sally Rumsey. 2015. Contextual and Provenance Metadata in the Oxford University Research Archive (ORA). Metadata and Semantics Research. Communications in Computer and Information Science 544, Springer, Cham, 274–285. DOI: http://dx.doi.org/ 10.1007/978-3-319-24129-6_24

[19] Jinfang Niu. 2013. Provenance: Crossing Boundaries. Archives and Manuscripts 41, 2, 105-115. DOI: http://dx.doi.org/10.1080/01576895.2013.811426

[20] International Organization for Standardization. 2006. ISO 23081-1:2006 (en). Information and Documentation – Records Management Process – Metadata for Records – Part 1: Principles. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:23081:-1:ed-1:v1:en