

University of Tsukuba, Japan

Oct. 10, 2014

DCMI 2014, Austin, Texas, U.S.A.

Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata

Chunqiu Li¹, Shigeo Sugimoto²

¹Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan

²Faculty of Library, Information and Media Science, University of Tsukuba, Japan

Sugimoto and Nagamori Laboratory

<http://mdlab.slis.tsukuba.ac.jp/en/>

Outline

Digital Preservation & Metadata Preservation

Research Goal: Provenance Description for Metadata Preservation

Metadata Provenance Model for Metadata Preservation Using PROV with PREMIS

Related Work & Future Research

Digital Preservation & Metadata Preservation

- Metadata Preservation vs. Digital Preservation
 - Digital preservation: to keep digital objects renderable over time
 - Metadata preservation: to keep metadata interpretable over time

Metadata Preservation

- **Semantics of metadata** has to be maintained over time
 - Preservation is worthless if the metadata meaning was lost
 - Preservation Description Information (PDI) of OAIS Information Package has to be interpretable over time
 - Metadata schemas have to be preserved as well as metadata instances

Metadata Preservation and Plain-text Preservation

- Metadata preservation is **not plain-text preservation** even if metadata instance may be serialized as a plain-text
- If a metadata instance is encoded in XML and stored in a plain-text file, the semantics of XML elements may be lost unless the meanings of the tags in the XML text are properly preserved

Preserving Metadata Schema

– Metadata Schema Changes –

- Metadata **term** definition changes
 - e.g., meaning may be broadened or narrowed, also a property may be added or deleted, domain or range of a property may be changed
- **Constraints** definition changes
 - e.g., the occurrence constraints for a property may be changed from “required” to “optional”, or from “repeatable” to “non-repeatable”

The Current State of Related Models

- Existing models are not specially developed for metadata preservation from both preservation and provenance viewpoints
 - Digital preservation models: OAIS, PREMIS
 - Provenance description models: PROV, OPM, and others

Research Goals

- To **propose a metadata provenance description model** for the long-term use of metadata via combining PROV and PREMIS
- To **clarify metadata provenance clearly** through the development of the metadata provenance description

What is provenance

- Provenance Definition in W3C PROV
“A record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing”
(<http://www.w3.org/TR/2013/REC-prov-dm-20130430/>)

Metadata Provenance

- In the research, we define metadata provenance as chronological information about metadata
- Metadata provenance typically describes responsible agents, influencing actions, associated events, and other related information about metadata over its lifecycle
- The provenance of metadata schema is also seen as metadata provenance

Metadata Schema Provenance

We classify metadata schema provenance to

- **Vocabulary** provenance
- **Structure** provenance
- Other provenance descriptions about encoding syntax guidelines, user guidelines, and functional requirements

Combining PROV and PREMIS

- **PROV** is designed generally for the **provenance description** but not primarily for preservation
- **PREMIS** is an international metadata standard for **digital preservation**
- Our main idea is to develop a novel model of provenance description for metadata longevity by combining PROV and PREMIS

Basic Concepts in PROV and PREMIS

PROV

Entity: any entities, e.g., physical, imaginary, digital, conceptual, intellectual or non-intellectual

Agent: person, organization, and software bearing responsibility to an entity

Activity: any activity on an entity and/or by an agent(s) not limited to preservation domain

PREMIS

Object: object in digital form

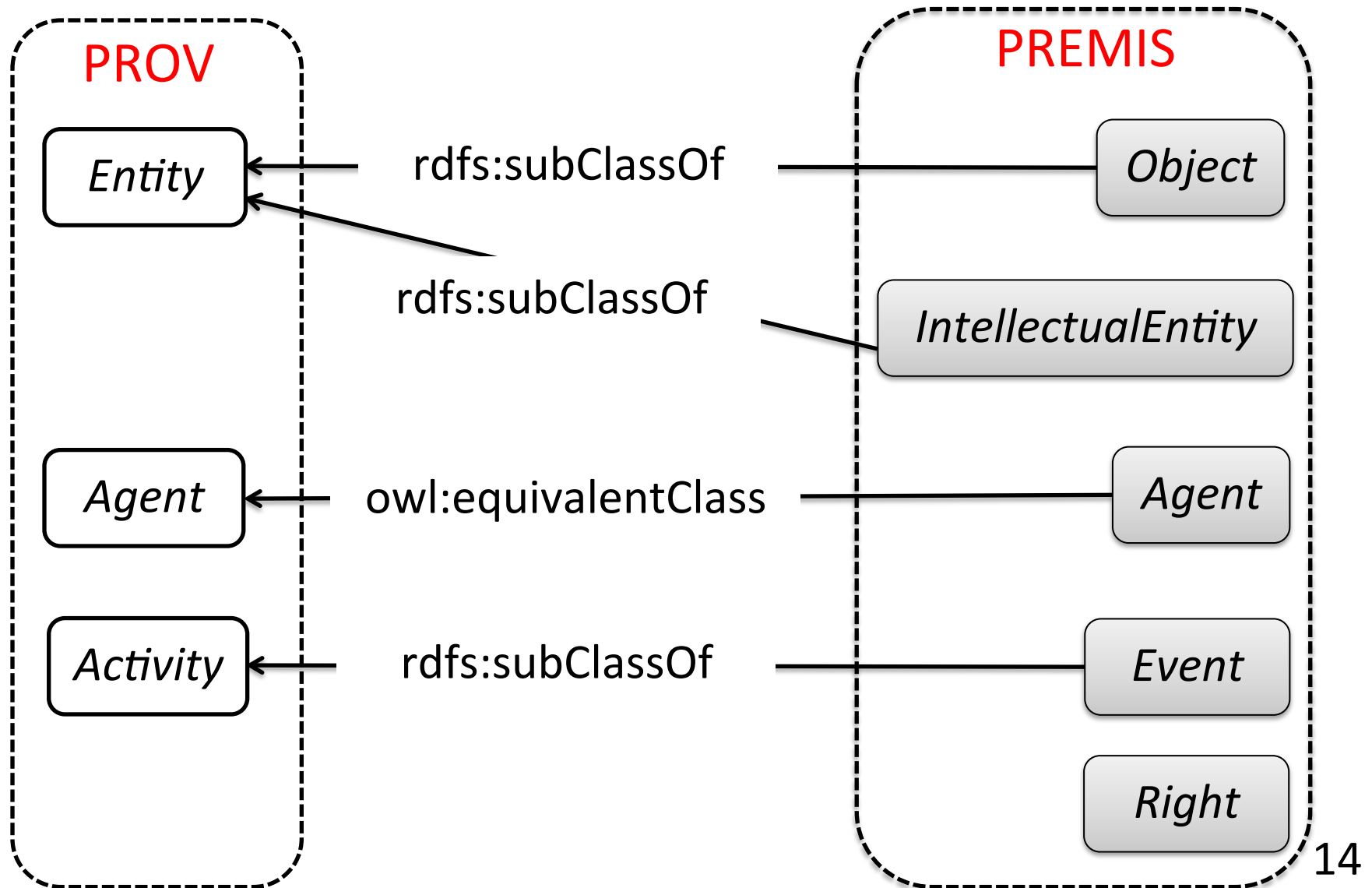
IntellectualEntity: *a single intellectual unit*

Agent: person, organization, software information

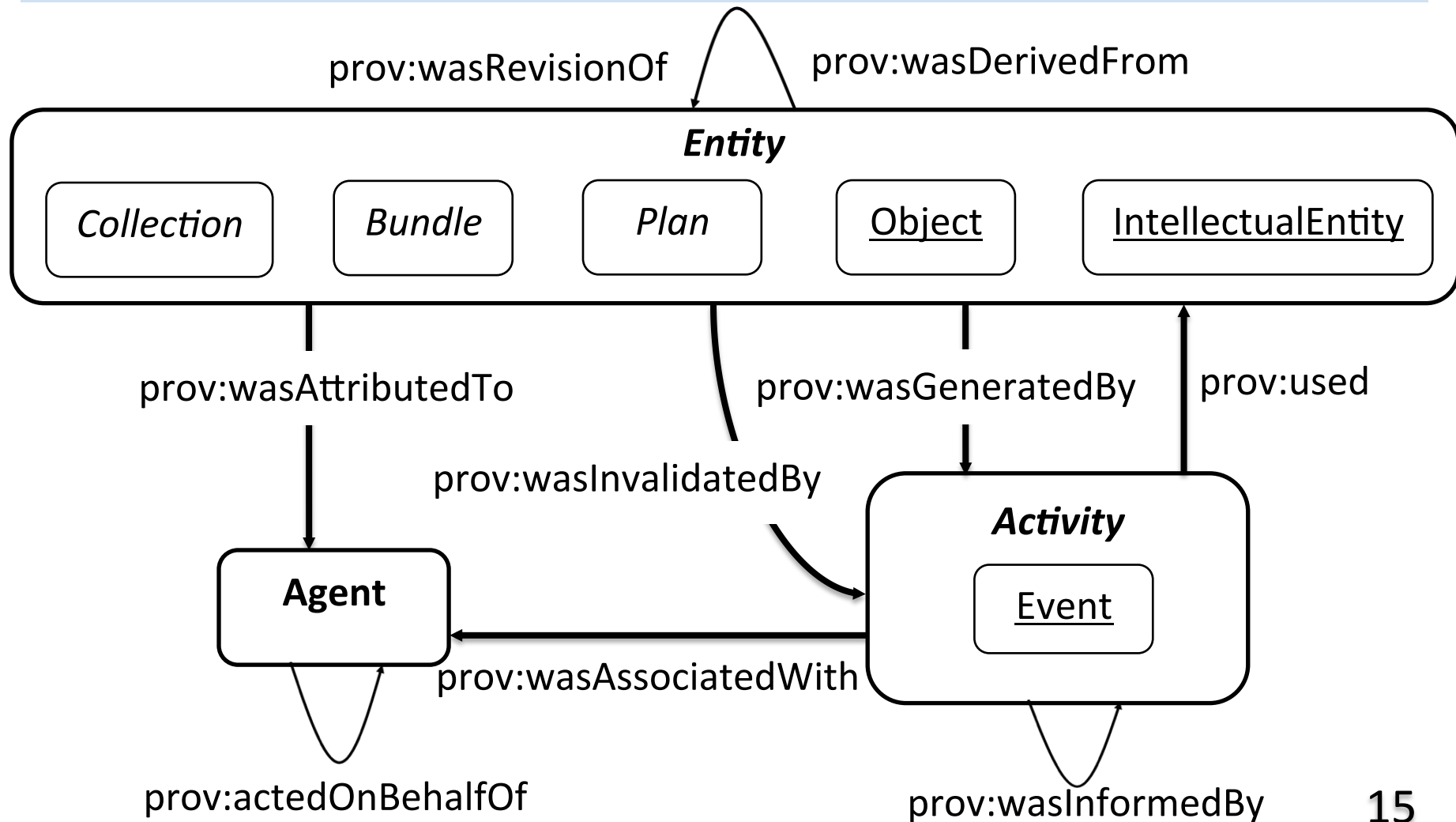
Event: major activity over preservation process

Right: *permission pertaining to object or/and agent*

Mapping between PROV and PREMIS

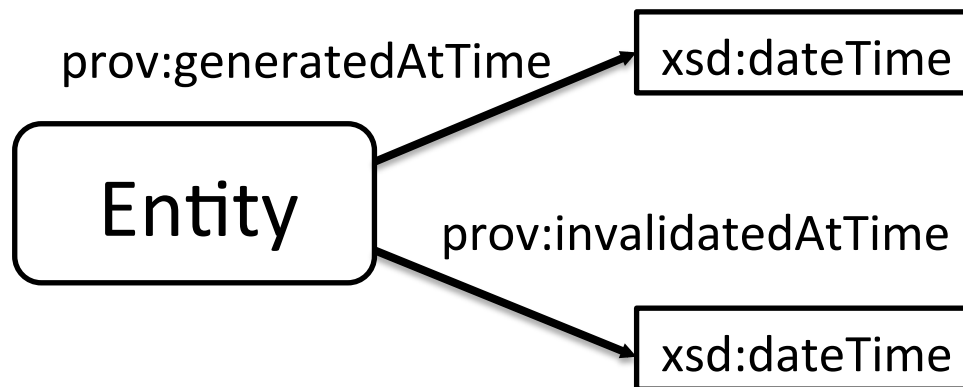


Proposal: Merged Model Combining PROV and PREMIS

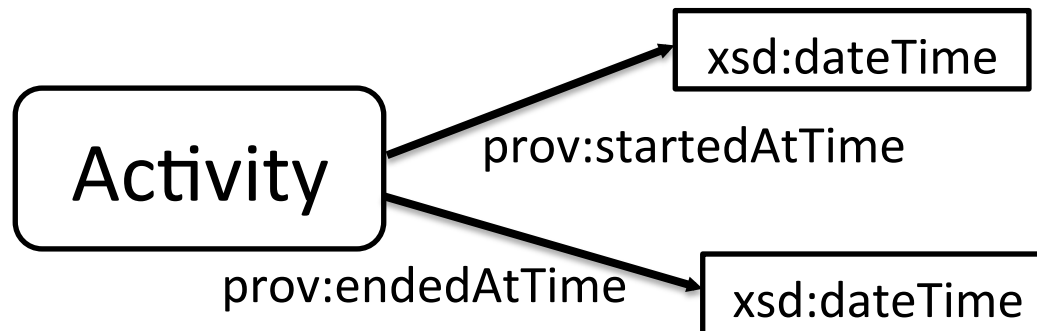


Chronological Information Description

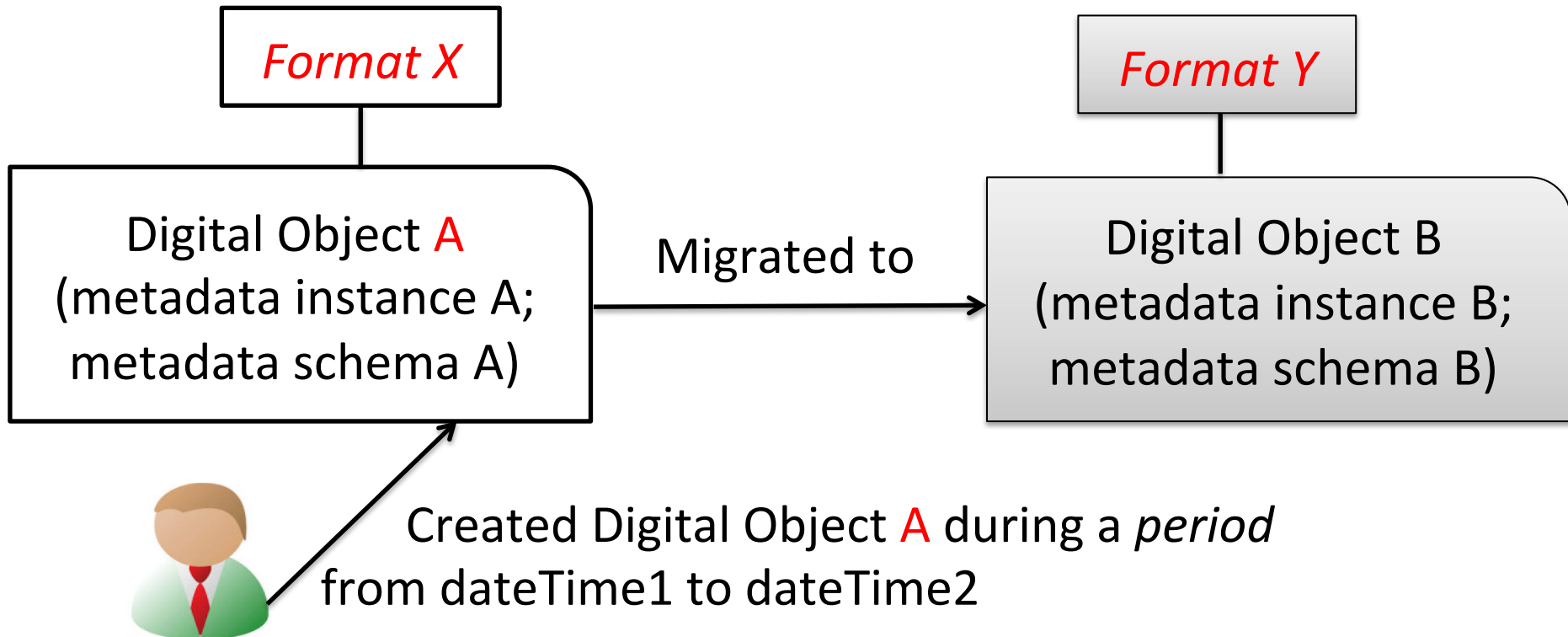
- Entity: generated time and invalidated time



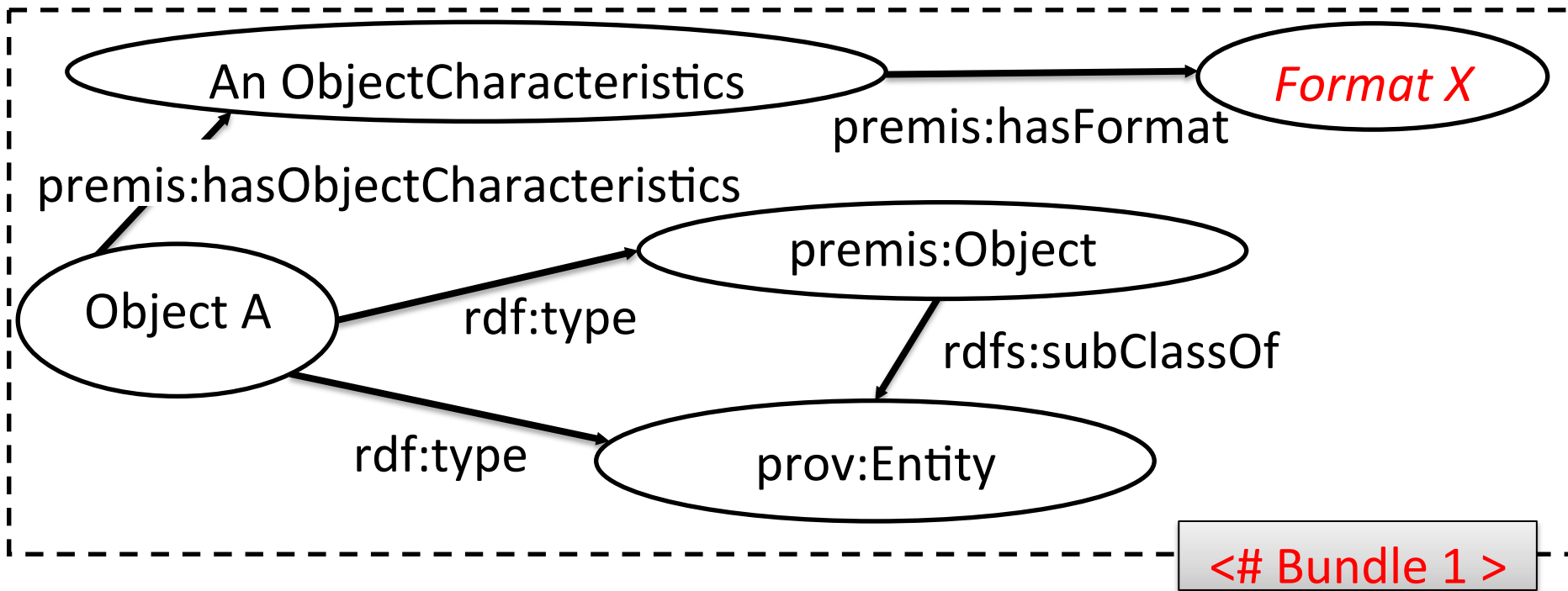
- Activity: started time and ended time



Example: Format Migration



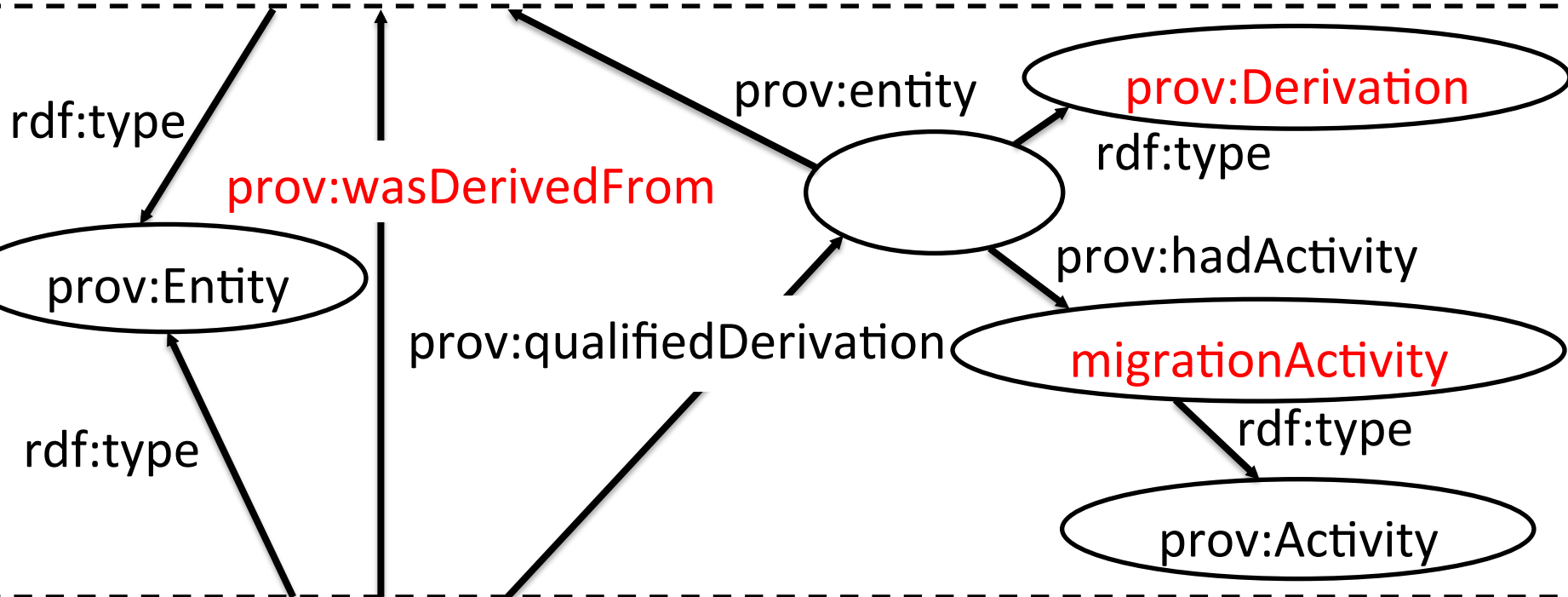
Format Characteristic Graph



Note: *Bundle* in PROV is defined as a **named set** of provenance descriptions

Provenance Graph about Format Derivation

Bundle 1: the format description graph of *Object A* in *Format X*



Bundle 2: the format description graph of *Object B* in *Format Y*

Related Work

- Europeana uses EDM to represent metadata provenance
- Project “Digitised Manuscripts to Europeana” (DM2E) defined a simple vocabulary for the datasets versioning
- Our research proposes a novel model combining PROV and PREMIS for the metadata provenance description and focuses on how metadata schema changes

References:

Eckert, Kai. (2012). Metadata Provenance in Europeana and the Semantic Web.

Eckert, Kai. (2013). Provenance and Annotations for Linked Data. Proceedings of the International Conference on Dublin Core and Metadata Applications.2013, 9-18.

Summary

- Metadata preservation as a function is distinct from digital preservation
- We proposed a merged model combining PROV and PREMIS for metadata provenance description in order to support metadata longevity

Future Research

- *Evaluating* the proposed model
 - Collecting versions of metadata schemas used in memory institutions
 - Creating metadata schema provenance data in RDF using the proposed model
- *Exploring* data service using the described metadata schema provenance data
 - To show metadata schema change history
 - To contribute to transformation among versioning metadata schemas

The End

THANKS FOR YOUR ATTENTION.

COMMENTS WELCOME.



University of Tsukuba, Japan

Assumed Questions

- Why the merger of PREMIS and PROV is necessary?
- Why did not merge OAIS and PROV?
- What is the possible use of metadata schema provenance?
- What is the relationship between provenance and preservation?
- What is the difficulty for ongoing work?

OAIS Archival Information Package

