PREMIS PREservation Metadata Implementation Strategies

PREMIS Ontology and controlled vocabularies

The PREMIS ontology working group

Sam Coppens (University of Ghent)
Rebecca Guenther (Library of Congress)
Kevin Ford (Library of Congress)
Sébastien Peyrard (National Library of France)
Tom Creighton (Family Search)

Background and purpose

- Background
  - PREMIS used at UGhent in RDF format – full draft ontology
  - Taken on a global level by a Working Group
    - Oct. 2010 – Oct. 2011
  - Update to add PREMIS 2.2 changes; refactoring; documentation
    - Aug. 2012 – May 2013

- Purpose
  - Setting up an RDF serialization of the PREMIS data model and dictionary
  - Take advantage of RDF specificities
  - Remain as close as possible to the data dictionary's clearly defined semantics
  - Propose a framework where existing controlled vocabularies at id.loc.gov can be reused

Reminder: RDF, OWL and ontologies

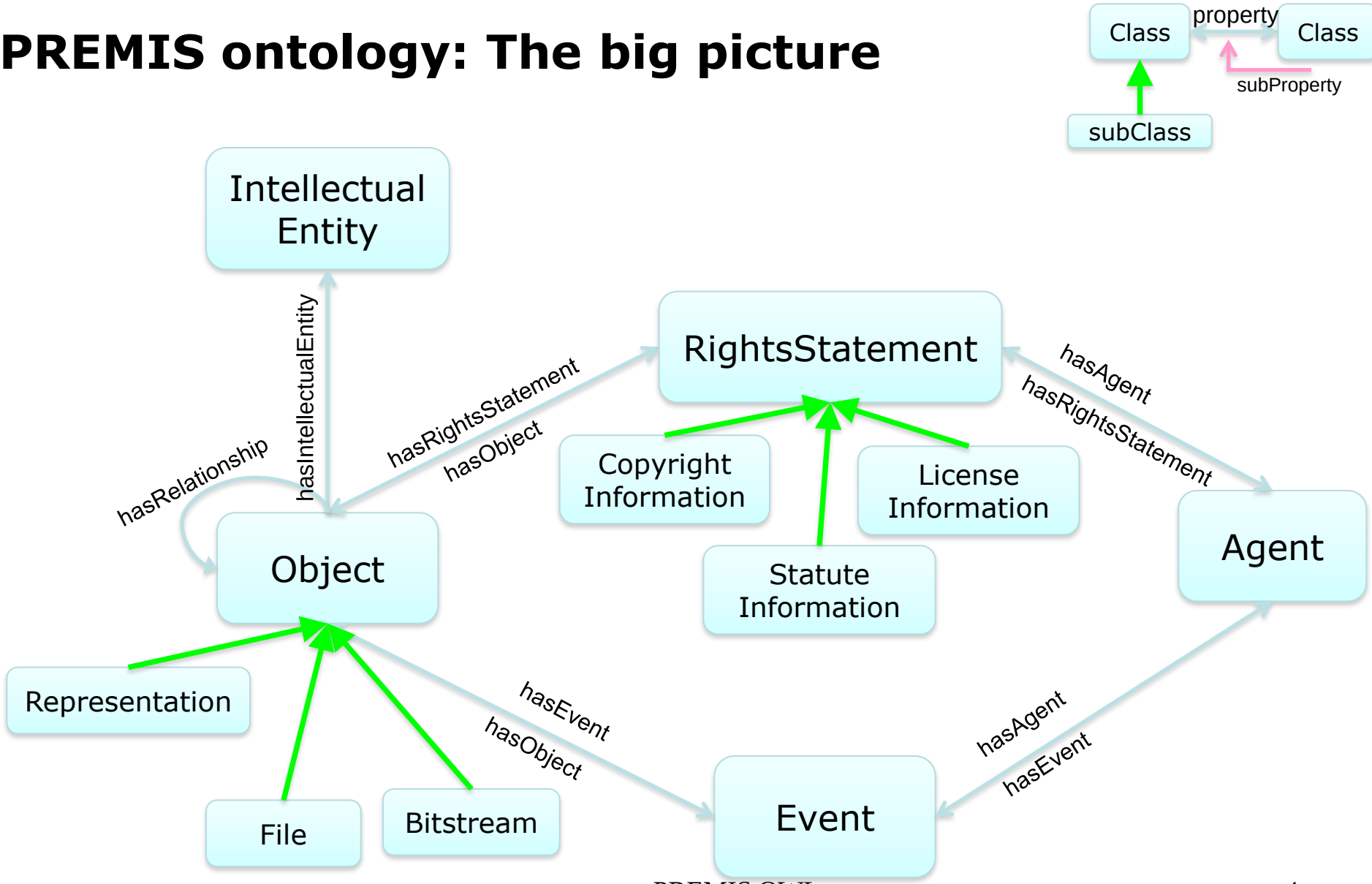
- RDF: a formalized way to describe things

```
subject verb  object
<URI>   <URI> <URI>/"string"
```

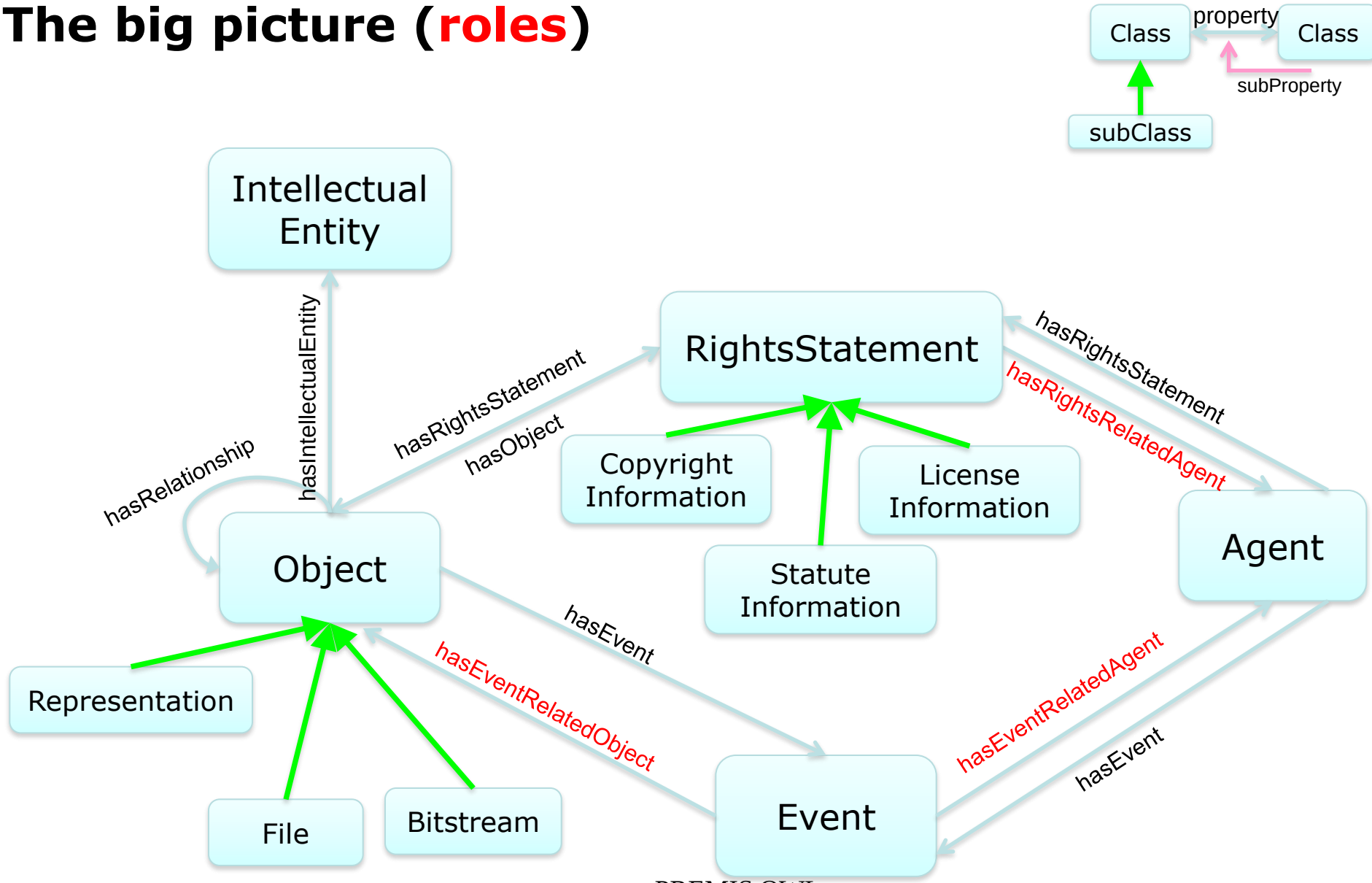
- URIs:
  - web addresses beginning with http:, info:, urn:...
  - **identifies** the stuff we want to describe

- OWL, RDFS: express vocabularies to describe stuff
  - Classes — **categories** for the things we describe
  - Properties — **verbs** to describe and relate things
  - They can be hierarchized

PREMIS ontology: The big picture



The big picture (roles)



PREMIS OWL: what for?

- Possible uses:
  - Having "RDF-ready-to-use" preservation concepts available as an ontology
  - RDF for Data management function
  - The metadata face of DDP (Distributed Digital Preservation)?
    - Reference format, software, hardware information
  - Link to other non-preservation databases (library catalogues, wikidata…)
    - Allowing bridges between curation, preservation, cataloguing…

PREMIS controlled vocabularies

Semantic unit	2.2 eventType
Semantic components	None
Definition	A categorization of the nature of the event.
Rationale	Categorizing events will aid the preservation repository in machine processing of event information, particularly in reporting.
Data constraint	Value should be taken from a controlled vocabulary.
Examples	E77 [a code used within a repository for a particular event type] Ingest
Repeatability	Not repeatable
Obligation	Mandatory
Usage notes	Each repository should define its own controlled vocabulary of *eventType* values. A suggested starter list for consideration (see also the Glossary for more detailed definitions):

PREMIS vocabularies

- June 2013: publication of 24 preservation vocabularies on id.loc.gov
  - 3 old ones updated
  - 21 new vocabularies added
- Not only applicable for RDF!
- See http://id.loc.gov/vocabulary/preservation

Let's look at it

Ontology:
http://www.loc.gov/premis/rdf/v1#

Vocabularies:
http://id.loc.gov/vocabulary/preservation

From the PREMIS Data Dictionary to RDF: Specific choices

1. Identifiers
2. Controlled vocabularies
3. Extensions
4. Format registry keys
5. Rights entity modelling

1. Identifiers

- In the Data Dictionary, the identifier *qualifies* the object
- In RDF, a URL/URI identifier *is* the Object

Object
    ObjectIdentifier
        ObjectIdentifierType: URI
        ObjectIdentifierValue:
        info:ark/12148/bpt6k102002g
    ObjectCategory: representation

<info:ark/9999/c1234>
    rdf:type
    premis:Representation.

1. Identifiers

- managing identifiers when they are **not** URIs

subject verb object

file description

```
<file1> premis:identifier <file1-ID>.
<file1-ID> rdf:type premis:Identifier;
           premis:identifierType
  "someUniversityIdentifierType";
           premis:identifierValue "12345678".
```

Easy to use
No controlled vocab to define

OR

local vocabulary

```
<http://university.edu/local#someIdentifierType>
  rdfs:subPropertyOf premis:identifier.
```

Far more concise
Takes advantage of RDF features

file description

```
<object1> <http://university.edu/local#someIdentifierType>
  "12345678".
```

2. Controlled vocabularies: the mechanism

- An example

<http://id.loc.gov/vocabulary/preservation/eventType>

PREMIS standard description:

event

eventIdentifier

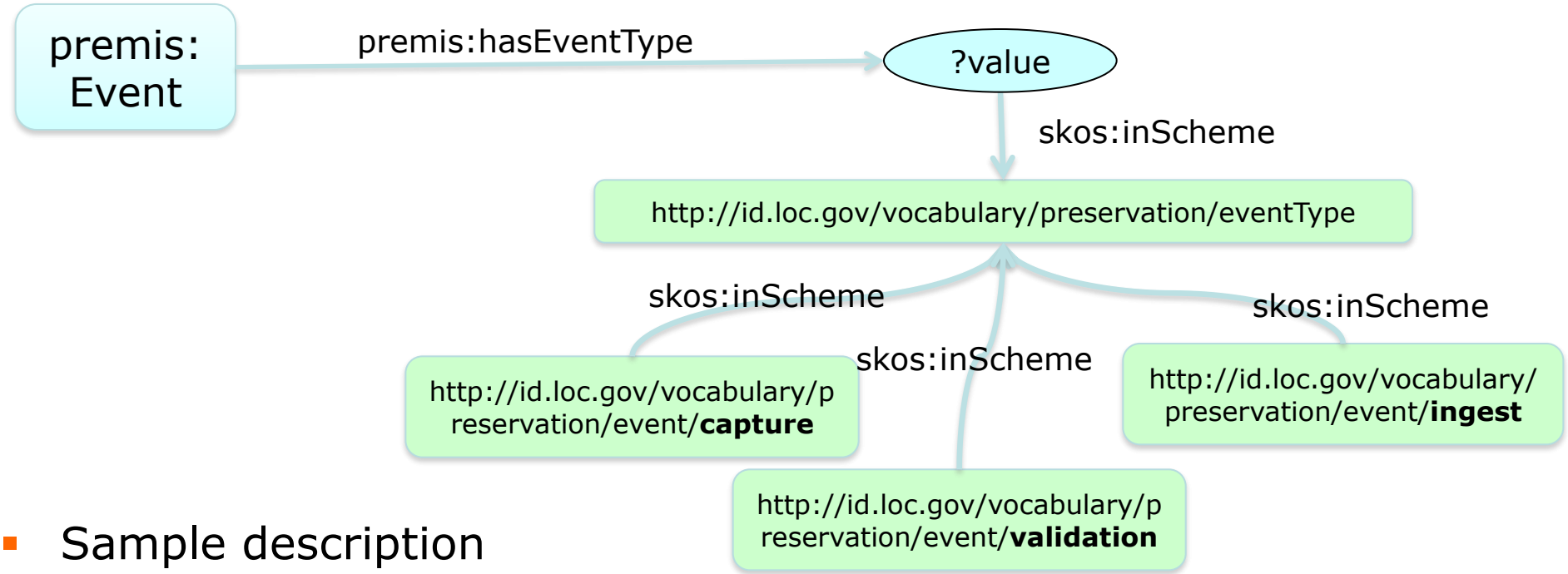
eventIdentifierType: UUID

eventIdentifierValue: 87c7a310-0fbd-11e3-8ffd-0800200c9a66

eventType: ingest

- In RDF, the controlled value is registered as a SKOS vocabulary at id.loc.gov
- **This vocabulary can be expanded with local values**

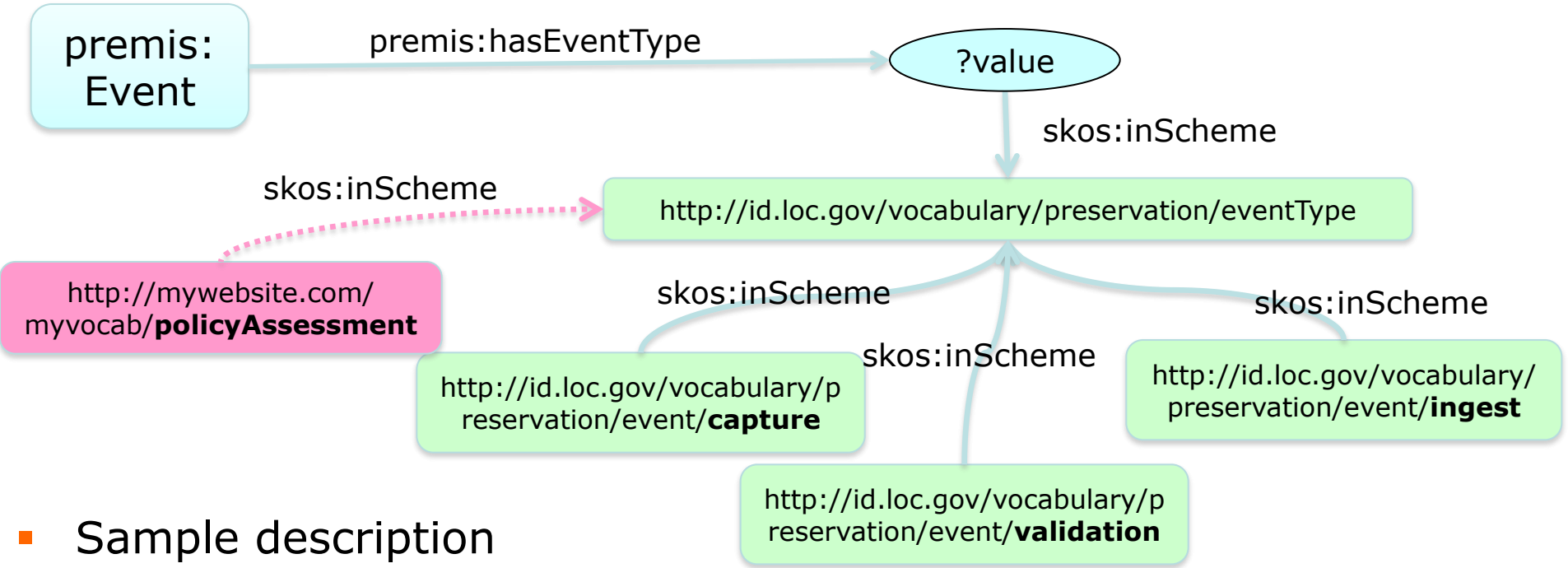
2. PREMIS OWL: alignment with other vocabularies



- Sample description

```
subject verb object
<http://example.com/id/87c7a310-0fbd-11e3-8ffd-0800200c9a66> rdf:type
   premis:Event.
<http://example.com/id/87c7a310-0fbd-11e3-8ffd-0800200c9a66> premis:eventType
   <http://id.loc.gov/vocabulary/eventType/ingest>.
```

2. PREMIS OWL: alignment with other vocabularies



- Sample description

```
subject verb object
<http://example.com/id/87c7a310-0fbd-11e3-8ffd-0800200c9a66> rdf:type
   premis:Event.
<http://example.com/87c7a310-0fbd-11e3-8ffd-0800200c9a66> premis:eventType
   <http://mywebsite.com/myVocab/policyAssessment>.
```

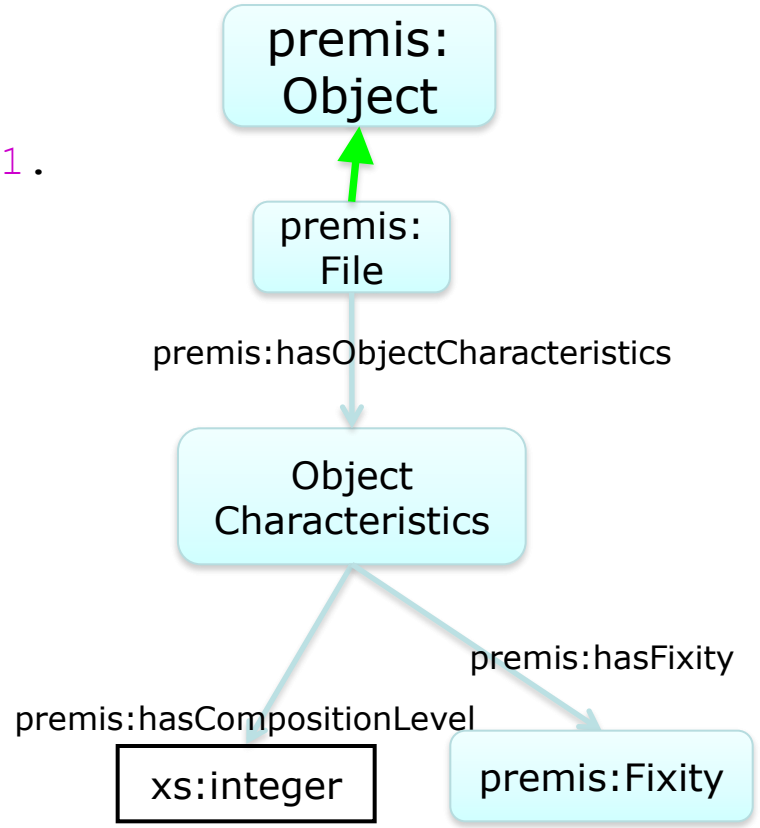
3. Extensions

- Extensions not explicitly stated: built-in RDF mechanism

Sample description

```
   subject verb object
<info:ark/9999/b1234> rdf:type premis:file;
   premis:hasObjectCharacteristics ?objChar1.
?objChar1 rdf:type
   premis:ObjectCharacteristics;
   premis:hasCompositionLevel "0";
   textmd:charset "UTF-8";
   textmd:byte_size "8";
   textmd:markup_basis "XML";
   textmd:markup_version "1.0";
   textmd:markup_language
   <http://www.loc.gov/standards/alto/ns-
   v2#>.
```

Ontology



4. Format Registry Keys

- Ability to directly link to a format URI
- E.g. in UDFR: <http://udfr.org/udfr/u1r2617>

- **Data Dictionary:**

objectIdentifier
objectCategory "file"
objectCharacteristics
format
formatDesignation
formatName: image/tiff
formatVersion: 6.0
formatRegistry
formatRegistryName: UDFR
formatRegistryKey: u1r2617
formatRegistryRole: specification

Sample RDF description

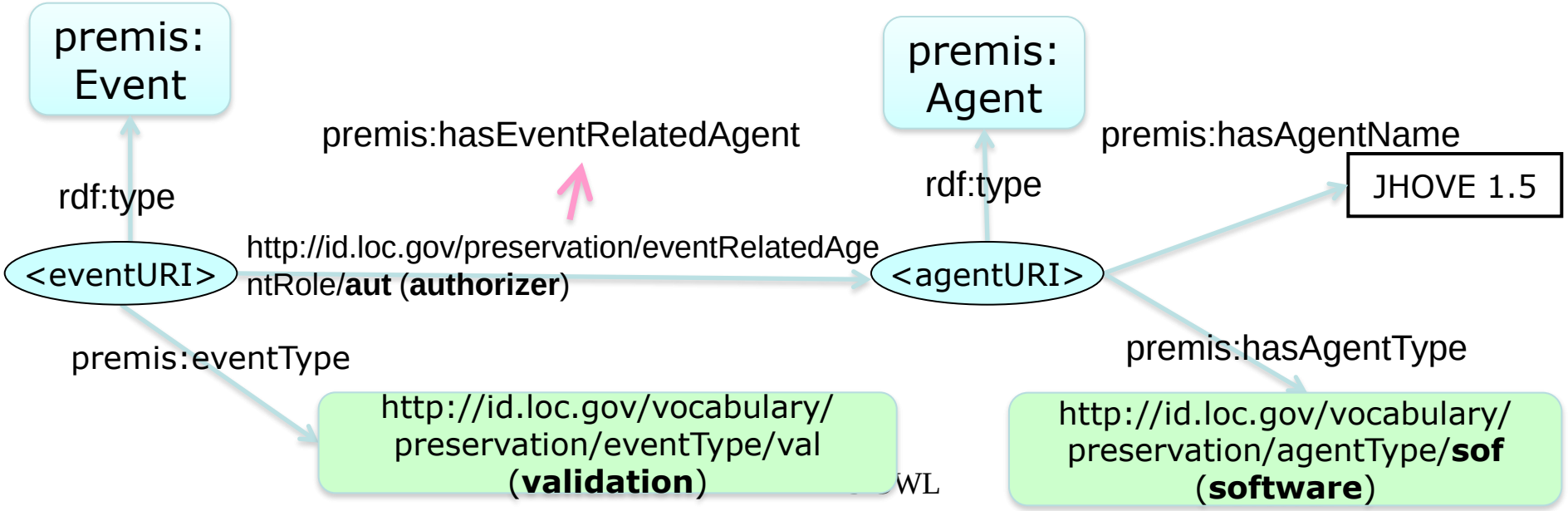
```
subject verb object

info:ark/9999/c1234> rdf:type
  premis:file;
  premis:hasObjectCharacteristics
  ?objChar1.
objChar1 rdf:type
  premis:ObjectCharacteristics;
  premis:hasFormat
  <http://udfr.org/udfr/u1r2617>.
```

5. linking[Entity]Role

- E.g. linkingAgentRole from an event
- Not about the agent nor the event, but about the **relationship** between an agent and an event.
- Designed as a subproperty
  - E.g. the performer of a file validation

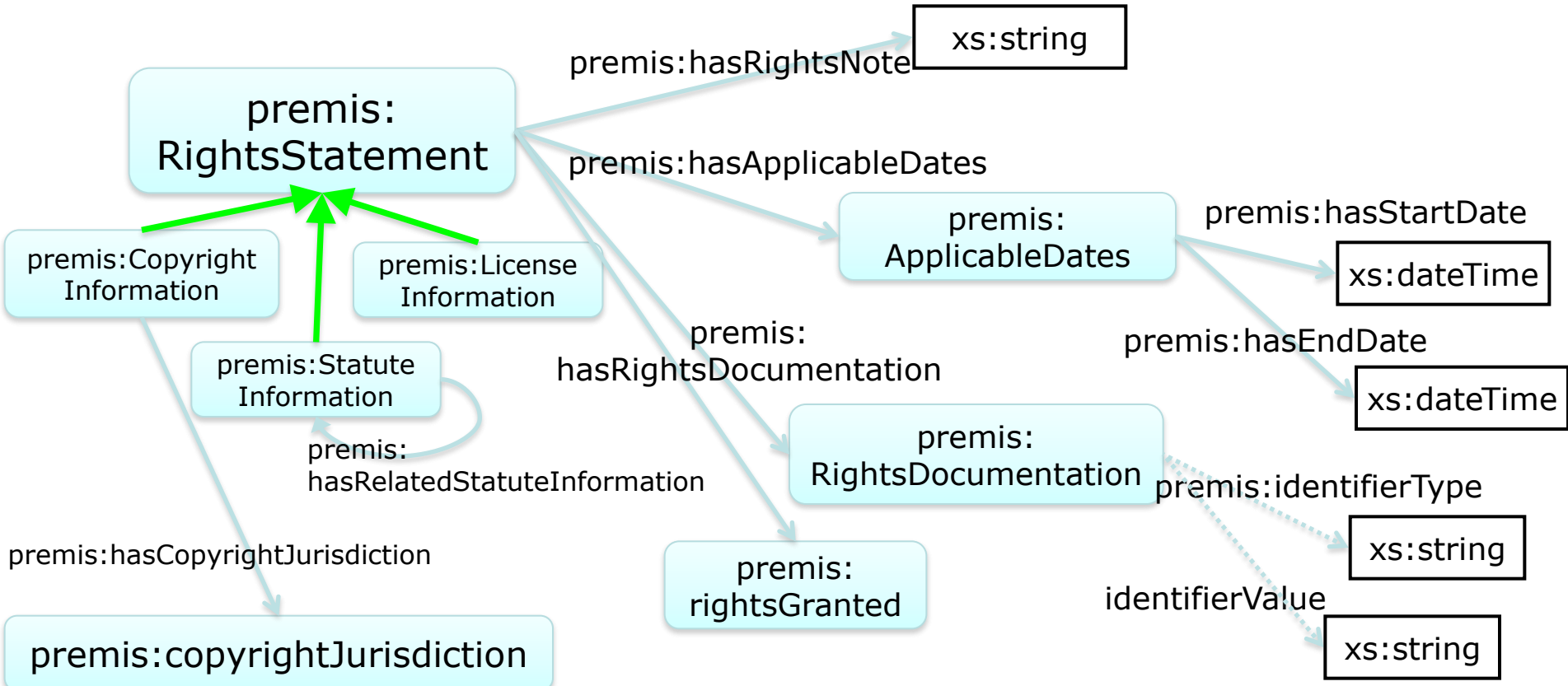
Data dictionary

Event
  […]
  linkingAgentIdentifier
    linkingAgentIdentifierType/Value
    linkingAgentRole: "authorizer"



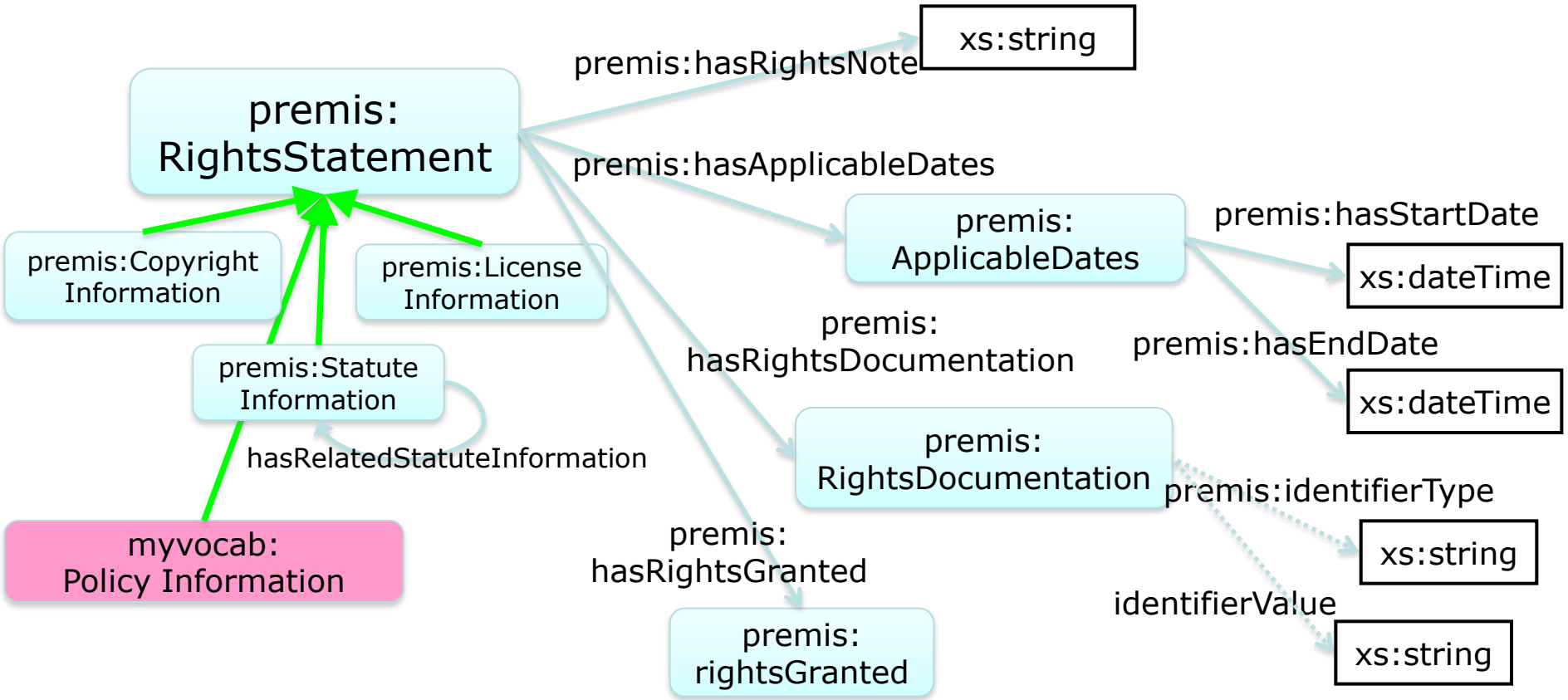
6. Implementation specificities: rights

- Take advantage of RDF features to make rights more compact



6. RightsStatement

- otherRights becomes an new subclass



Next steps?

- PREMIS 2.2 "OWL official version"
  - Alignment with PROV-O considered
- PREMIS 3.0 evolution
  - Update the ontology to reflect the PREMIS 3.0 changes

Thank you for your attention

Questions?

premis-ontology AT googlegroups.com