Learning PREMIS Knowledge Base: a Tool for Humans and Machines

Integrating Original Documentation for Annotating Ontology

Angela Di Iorio, Marco Schaerf

DIAG - Department of Computer, Control, and Management Engineering Antonio Ruberti

SAPIENZA
UNIVERSITÀ DI ROMA

HARVARD
LIBRARY

MIT
Libraries

PREMIS tools: Data Dictionary and Ontology

XML

W3C OWL

PREMIS Ontology

VERSION 2

VERSION 2

VERSION 3

VERSION 3

Modelling PREMIS as Linked Data Vocabulary

VERSION 2

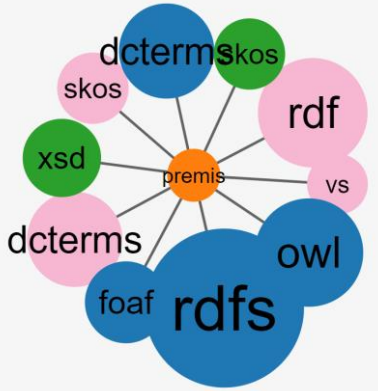
VERSION 3

PREMIS Ontology

- significant connections

+ codified knowledge

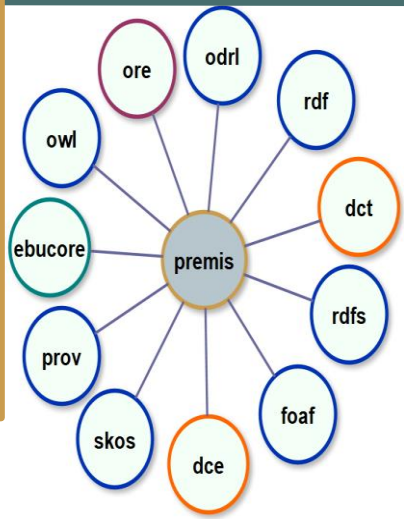
VERSION 2



+ significant connections

- codified knowledge

VERSION 3



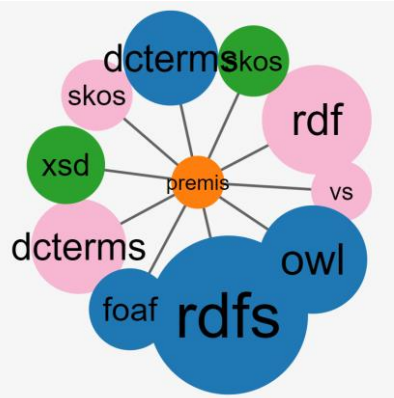
C. W. Choo. The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. International journal of information management, 16(5):329340, 1996.

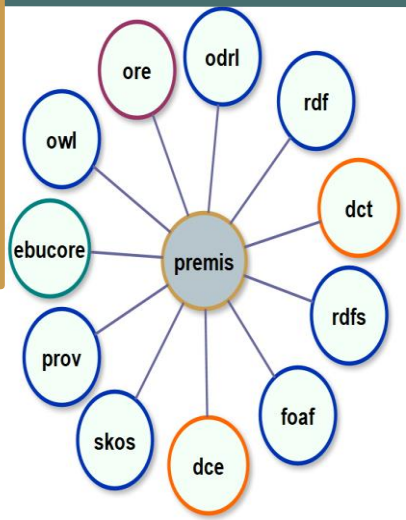
Modeling PREMIS as a Linked Data Vocabulary

PREMIS Ontology

VERSION 2 → **VERSION 2**

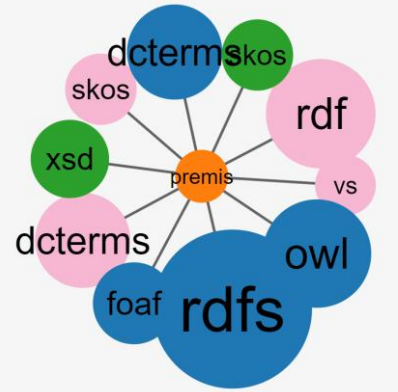
VERSION 3 → **VERSION 3**





Reciprocal influence in modeling

PREMIS Ontology

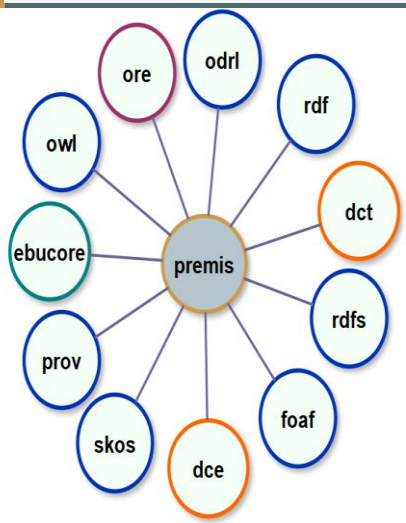


VERSION 2

VERSION 2

VERSION 3

VERSION 3



Example: the knowledge about Fixity

PREMIS Ontology

VERSION 2

```
<file1> a premis:File ;
premis:hasFixity <file1fixity> .
<file1fixity> a premis:Fixity ;
premis:hasMessageDigestAlgorithm crypHashFunc:md5 ;
premis:hasMessageDigest
"37ba62655d93e540a8195a6f02ec8bdc"^^xsd:string ;
premis:hasMessageDigestOriginator
"PHP 5.2.10 function"^^xsd:string .
```

VERSION 2

Semantic units pages 59-63
Special Topics page 258
Glossary page 270

VERSION 3

```
<file1> a premis:File ;
premis:fixity <file1fixity> .
<file1fixity> a crypHashFunc:md5 ;
rdf:value "258622b1688250cb619f3c9ccaefb7eb" ;
dct:creator <GNUgperf31> .
```

VERSION 3

Functional information about ontological entities

PREMIS Ontology

VERSION 2

Semantic units pages 59-63
Special Topics page 258
Glossary page 270

VERSION 3

<file1> a premis:File ;
premis:hasFixity
<file1fixity> a
premis:hasM
premis:hasM
"37ba62
premis:
"PHP 5.2

How functional information about premis:Fixity should be captured, structured in a computable form, and made accessible to the digital preservation community?

<file1> a premis
premis:fixity <file1fixity> .
<file1fixity> a crypHashFunc:md5 ;
rdf:value "258622b1688250cb619f3c9ccaefb7eb" ;
dct:creator <GNUgperf31> .

PREMIS Knowledge Base

PREMIS Ontology

VERSION 2

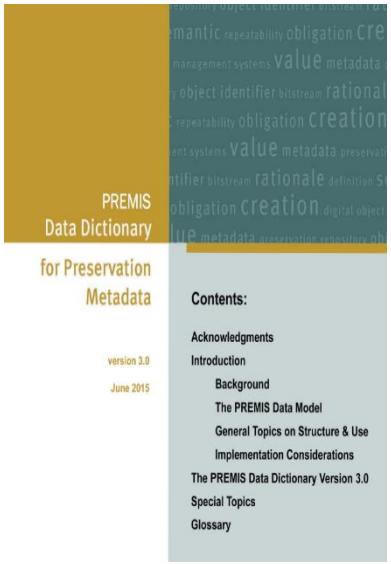
VERSION 2

PREMIS Knowledge Base

VERSION 3

VERSION 3

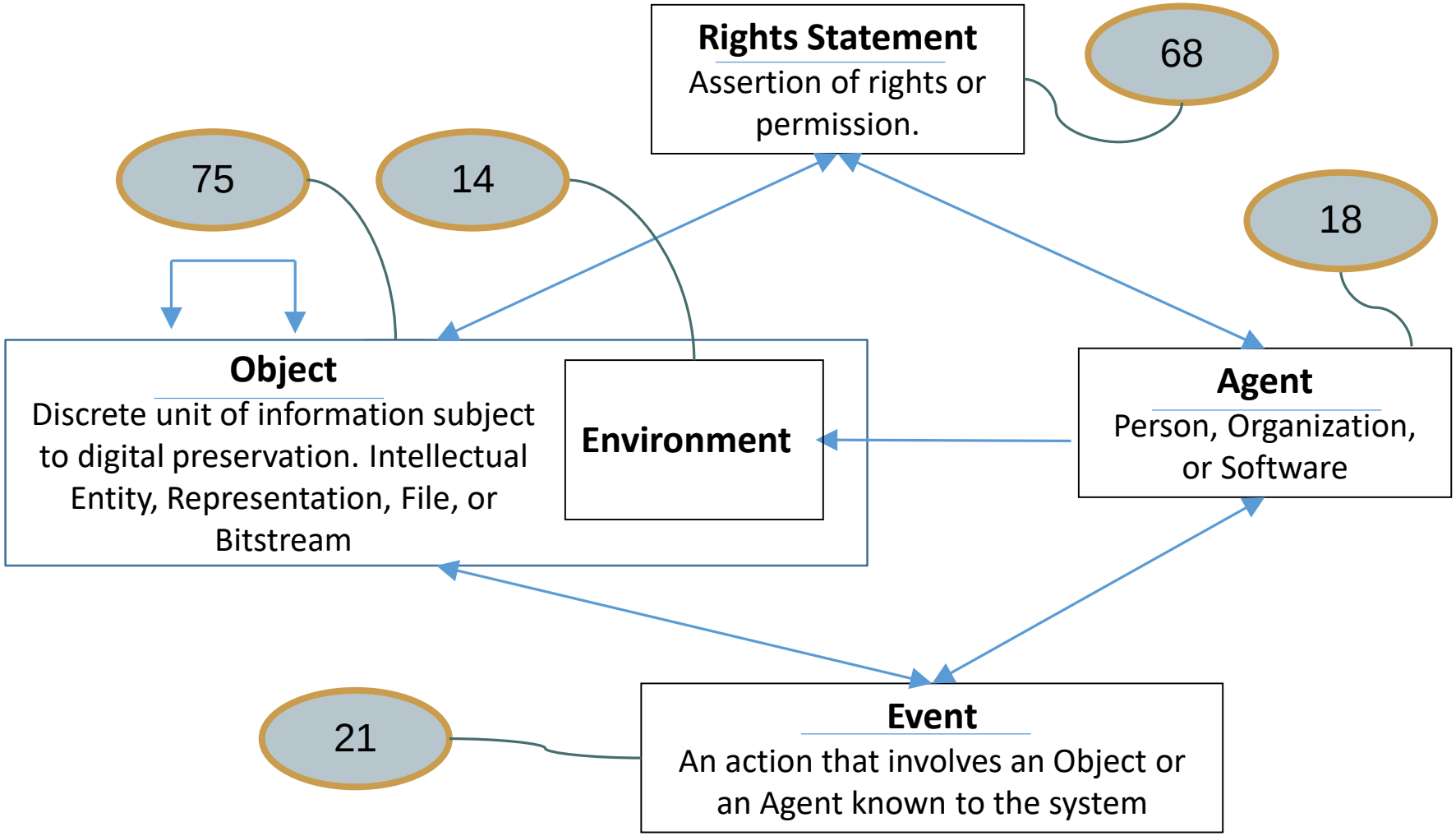
The attributes of PREMIS Semantic Units



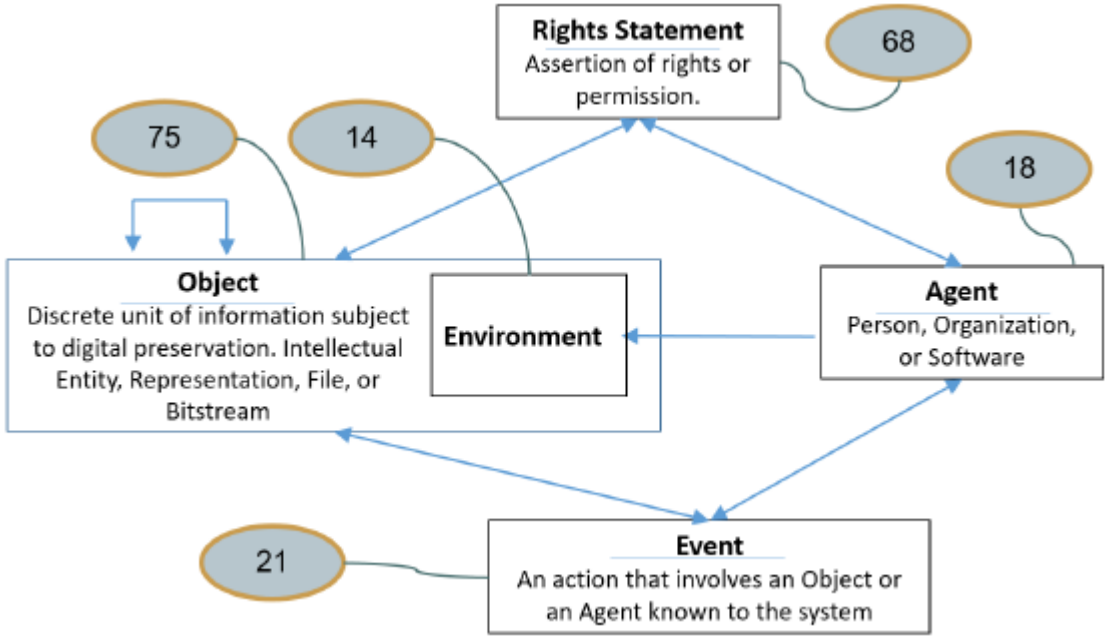
Semantic unit	1.5.2 fixity		
Semantic components	1.5.2.1 messageDigestAlgorithm 1.5.2.2 messageDigest 1.5.2.3 messageDigestOriginator		
Definition	Information used to verify whether an object has been altered in an undocumented or unauthorized way.		
Data constraint	Container		
Object category	Intellectual Entity / Representation	File	Bitstream
Applicability	Not applicable (see usage note)	Applicable	Applicable (see usage note)
Repeatability		Repeatable	Repeatable
Obligation		Optional	Optional
Creation / Maintenance notes	Automatically calculated and recorded by repository.		
Usage notes	To perform a fixity check, a message digest calculated at some earlier time is compared with a message digest calculated at a later time. If the digests are the same, the object was not altered in the interim. (Note that the terms "message digest" and "checksum" are commonly used interchangeably. However, the term "checksum" is more correctly used for the product of a cyclical redundancy check (CRC), whereas the term "message digest" refers to the result of a cryptographic hash function, which is what is referred to here.) The act of performing a fixity check and the date it occurred would be recorded as an Event. The result of the check would be recorded as the *eventOutcome*. Therefore, only the *messageDigestAlgorithm* and *messageDigest* need to be recorded as *objectCharacteristics* for future comparison. Representation level: it could be argued that if a representation consists of a single file or if all the files comprised by a representation are combined (e.g., zipped) into a single file, then a fixity check could be performed on the representation. However, in both cases the fixity check is actually being performed on a file, which in this case happens to be coincidental with a representation. Bitstream level: message digests can be computed for bitstreams although they are not as common as with files. For example, the JPX format, which is a JPEG2000 format, supports the inclusion of MD5 or SHA-1 message digests in internal metadata that was calculated on any range of bytes of the file.		

PREMIS-SU attribute	Description	Obligation
Identifier	Identification number	*M*
Name	Descriptive name	*M*
Definition	The meaning of the PREMIS-SU	*M*
Rationale	why the PREMIS-SU is needed	*O*
Data constraints	how the value of the PREMIS-SU should be encoded (see Sect. 5.1)	*O*
Object category	specifies the type of the PREMIS Object	*O*
Applicability	whether the unit applies to a specific type of PREMIS Object	*O*
Examples	one or more examples of values the PREMIS-SU may take	*O*
Repeatability	whether a PREMIS-SU can take multiple values	*O*
Obligation	whether a value for the PREMIS-SU is mandatory or optional	*O*
Creation/Maintenance notes	notes about how the values for the PREMIS-SU may be obtained or updated	*O*
Usage notes	information about the intended use of the PREMIS-SU, or clarification of the definition	*O*

PREMIS-SUs for Data Model 3.0



Semantic units vs PREMIS-OWL ontological entities



196

PREMIS-OWL ontological entities	PREMIS-OWL ver. 2	PREMIS-OWL ver. 3
owl:class	37	32
owl:ObjectProperty	39	22
owl:DatatypeProperty	25	15
owl:FunctionalProperty	39	
owl:InverseFunctionalProperty	2	
owl:AnnotationProperty	1	
	143	**69**

PREMIS OWL and Annotation Properties

PREMIS-OWL ontological entities	PREMIS-OWL ver. 2	PREMIS-OWL ver. 3
owl:class	37	32
owl:ObjectProperty	39	22
owl:DatatypeProperty	25	15
owl:FunctionalProperty	39	
owl:InverseFunctionalProperty	2	
owl:AnnotationProperty	1	

PREMIS-SU attributes	PREMIS-DD ver. 2	PREMIS-DD ver. 3	PREMIS-OWL2 Attribute Name	PREMIS-OWL annotation property	PREMIS-OWL ver. 2	PREMIS-OWL ver. 3
Data constraint	195	196	*Data Constraint/s:*		30/1	-
	-	-	*Extensions:*		23	-
Examples	98	110	*Example/s:*	rdfs:comment	15/28	-
Creation/Maintenance notes	24	22	*Creation/Maintenance Notes:*	skos:historyNote	16	-
Definition	195	196	*Definition:*	skos:definition	111	-
Rationale	85	93	*Rationale:*		55	-
	-	-	*Entity properties:*	skos:editorialNote	3	-
	-	-	*Entity types:*		1	-
Usage Notes	126	120	*Usage Note/s:*	skos:scopeNote	2/56	-

PREMIS-OWL3 annotation property	Value/Type	#
rdfs:isDefinedBy	http://www.loc.gov/premis/rdf/v3/	69
rdfs:label	Name of the ontological entity	69
rdfs:comment	Definition of the entity	28
rdfs:seeAlso	Link to controlled vocabularies	17

Linked Data and SKOS Vocabularies

Linked Data principles

1) Use URIs as names for things
"not just Web documents and digital content, but also real world objects and abstract concepts".

2) Use HTTP URIs so that people can look up those names
"to identify objects and abstract concepts".

3) When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
"use of a single data model for publishing structured data on the Web a simple graph-based data model that has been designed for use in the context of the Web".

4) Include links to other URIs, so that they can discover more things
"not only Web documents, but any type of thing".

SKOS Simple Knowledge Organization System

1) thesauri, taxonomies, classification schemes, etc., expressed in SKOS RDF are machine-readable and, software applications can use them as a Semantic Web knowledge;

2) lightweight, intuitive conceptual modeling language for developing and sharing controlled vocabularies in the Semantic Web;

3) a bridging technology between the rigorous logical formalism of OWL, and the informal and weakly-structured data.

Vocabularies and Annotation Properties

Property name	Description	Domain	Range
provide a human-readable description of a resource			
rdfs:label	A human-readable name for the subject.	rdfs:Resource	rdfs:Literal
rdfs:comment	A description of the subject resource.	rdfs:Resource	rdfs:Literal
provide a machine-actionable URI as description of a resource			
rdfs:seeAlso	Further information about the subject resource.	rdfs:Resource	rdfs:Resource
rdfs:isDefinedBy	The definition of the subject resource.	rdfs:Resource	rdfs:Resource

Can we use machine-readable URIs for connecting knowledge context for PREMIS Linked Datasets?

PREMIS-OWL3 annotation property	Value/Type	#
`rdfs:isDefinedBy`	http://www.loc.gov/premis/rdf/v3/	69
`rdfs:label`	Name of the ontological entity	69
`rdfs:comment`	Definition of the entity	28
`rdfs:seeAlso`	Link to controlled vocabularies	17

PREMIS-SU – Identification method

PREMIS-KB URIs are "talking segmented" identifiers

dlcmrc-vocs_2016-PREMISDD_2017_Object_fixity

Agent - dlcmrc is the identifier of the Library of Congress
the main responsibility for the maintenance of the resource content

Activity – vocs_2016 is the URI of provenance information about the PREMIS-KB generation activity

Entity – current documentation resource:
a) **PREMISDD_2017**, chaining the dataset name the year of the Linked Data generation;
b) **Object** the relevant container for belonging PREMIS-SUs;
c) **fixity** the PREMIS-SU name.

PREMIS-SU – Identification challenges

linkingEventIdentifier for **Object** and **Agent**;
linkingRightsStatementIdentifier for **Object** and **Agent**;
linkingObjectIdentifier for **Event** and **Rights**;
linkingAgentIdentifier for **Event** and **Rights**.

PREMIS-SU's name are not unique in the PREMIS-DD

Chaining names of the main entity & the PREMIS-SU works:
[…prefix…]_Object_linkingEventIdentifier

startDate and *endDate* for:

- copyrightApplicableDates,
- licenseApplicableDates,
- statuteApplicableDates,
- otherRightsApplicableDates,
- termOfGrant,
- termOfRestriction.

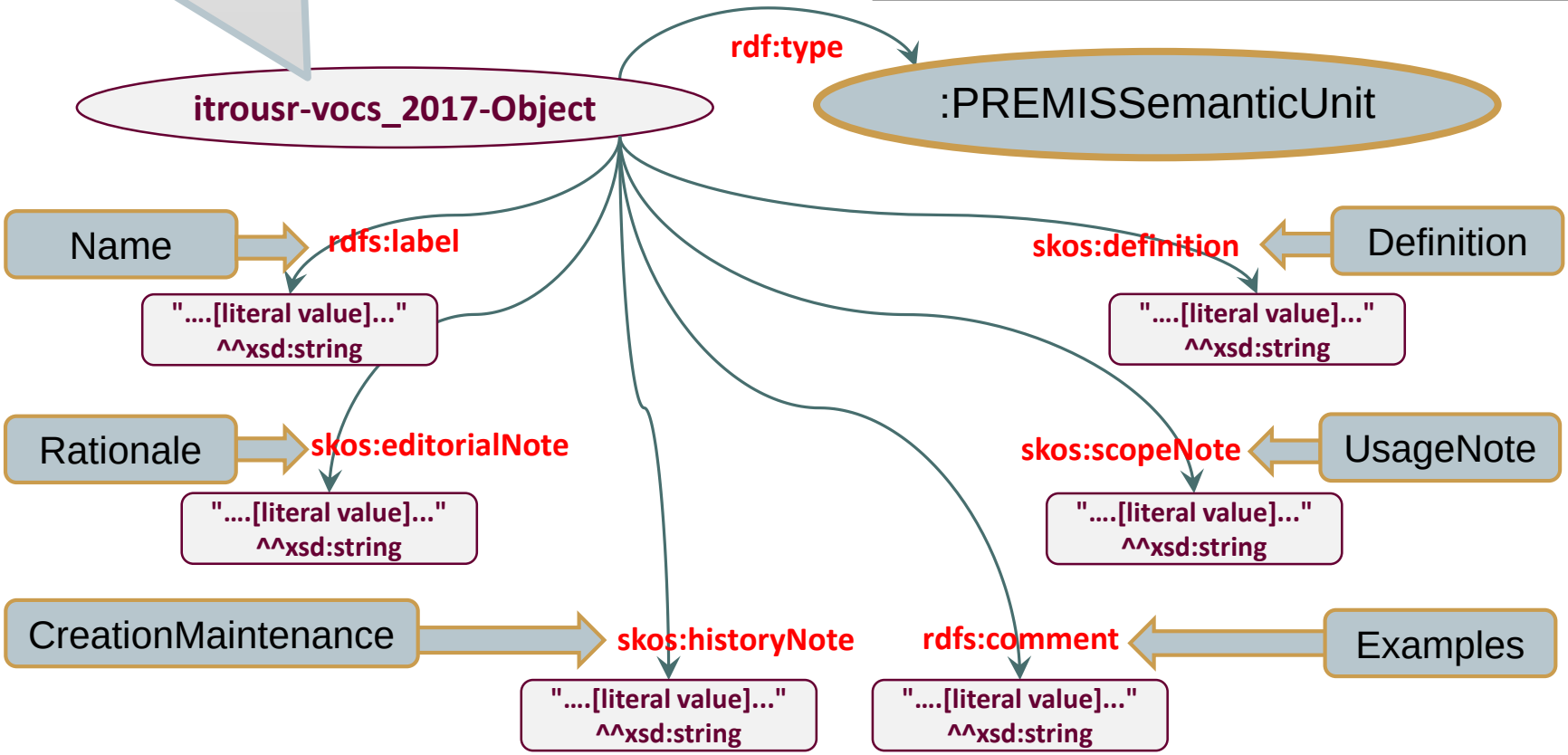
Identification method doesn't work, thus a deviation rule was applied:
dlcmrc-vocs_2016-PREMISDD_2017_Rights_termOfGrant_startDate

PREMIS-SU attributes -> SKOS mapping

Prototyping the local implementation:

- **itrousr**=identifier of the Sapienza Library System
- **vocs_2017**=provenance information URI of the PREMIS-KB generation activity
- **Object**=istance of the **:PREMISSemanticUnit** class

PREMIS-KB Properties	occurrences
rdfs:comment	105
rdfs:label	205
skos:definition	205
skos:editorialNote	91
skos:historyNote	18
skos:scopeNote	78



PREMIS-SU - Analysis

Container - is an umbrella for two or more semantic components, no value of its own.

an additional documentation resource, its hierarchical relationships can be described by skos semantic relationships (broader, narrower)

Datum - *None: can take any form of value*
- *Data type specification: integer, string, etc.*

The Range of a `DataProperty`

Authority - *value is taken from a controlled vocabulary.*

The Range of an `ObjectProperty`

PREMIS-DD ver. 3

Object			Agent			Event			Rights		
C	D	A	C	D	A	C	D	A	C	D	A
28	67	23	5	14	5	7	15	4	20	49	14

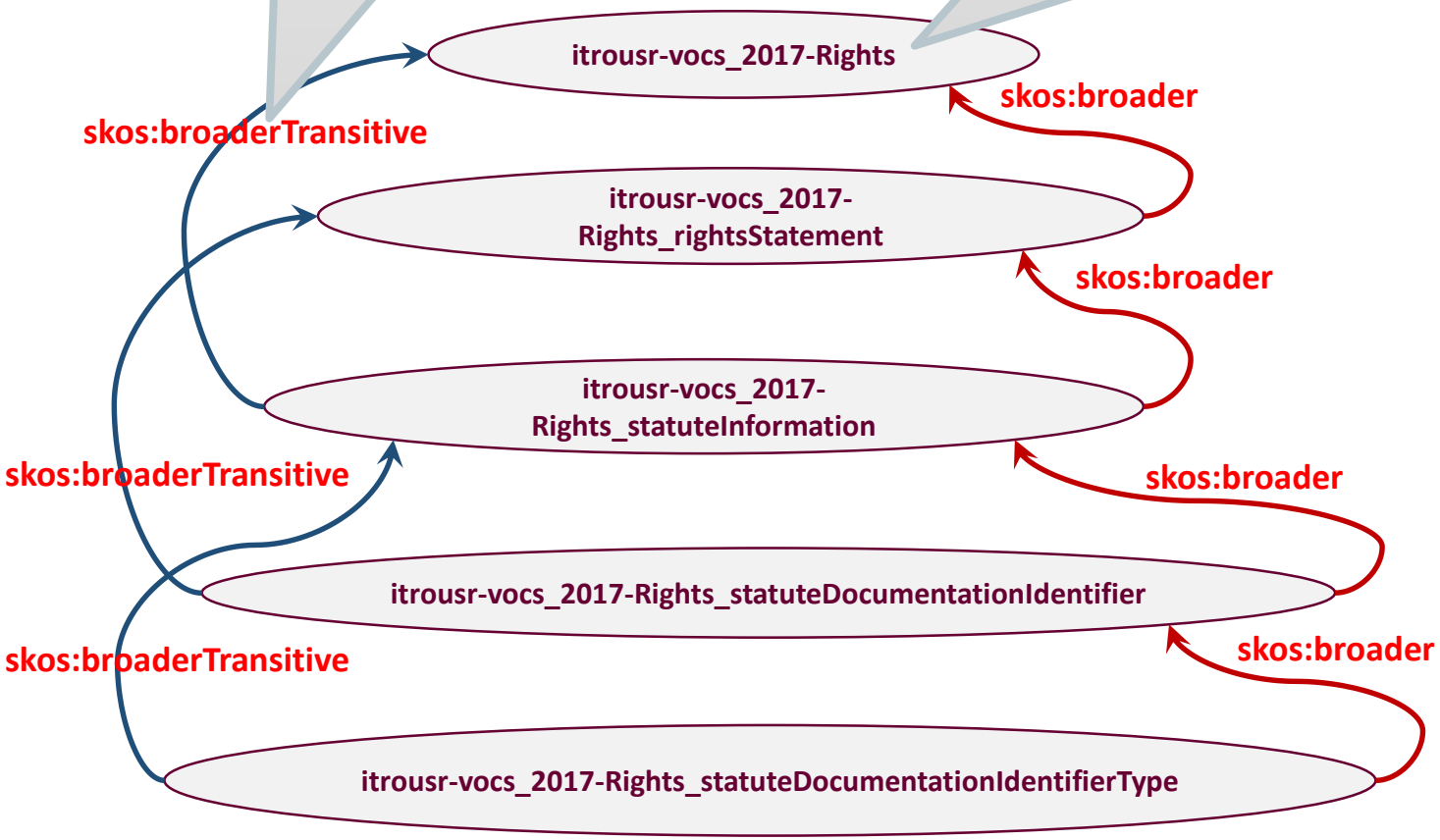
`eventIdentifierType`,
`eventType`,
`linkingAgentIdentifierType`
`linkingObjectIdentifierType`

SKOS semantic relations among PREMIS-SUs

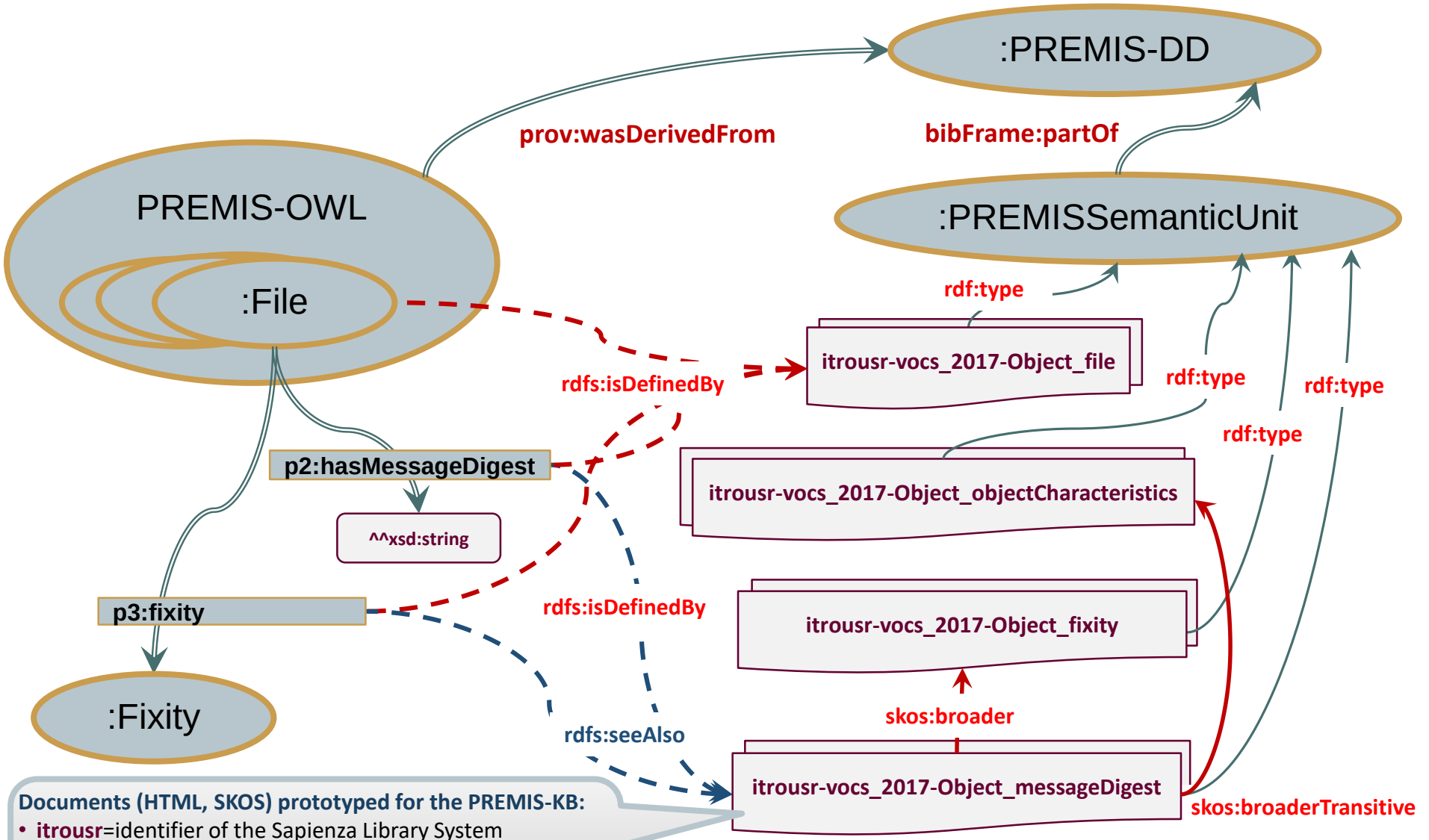
The PREMIS-DD hierarchical structure has driven the automatic generation of SKOS properties and related inverse properties

Prototyping the local implementation:

- **itrousr**=identifier of the Sapienza Library System
- **vocs_2017**=provenance information URI of the PREMIS-KB generation activity
- **Rights**=istance of the **:PREMISSemanticUnit** class



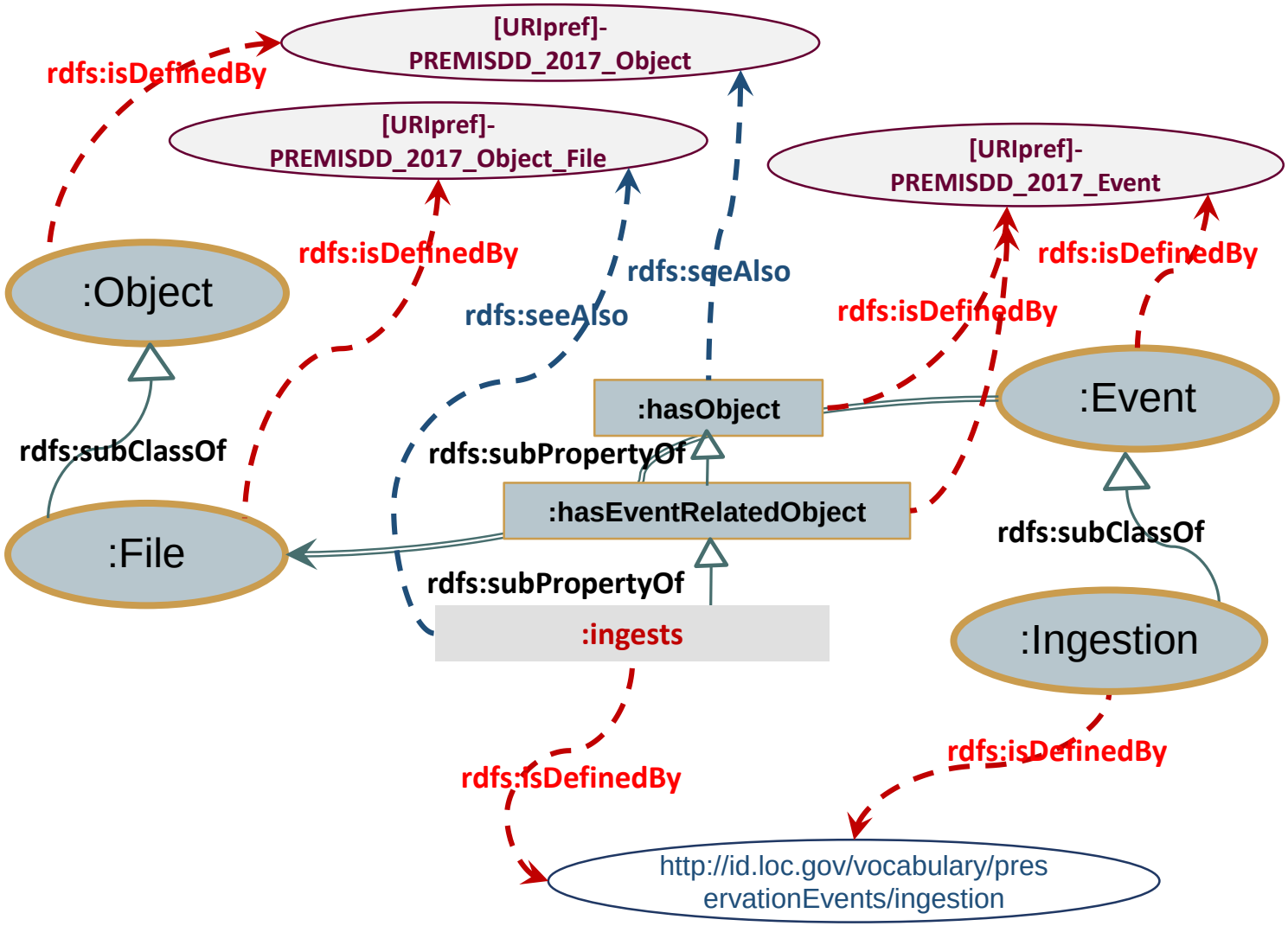
Ontology for the PREMIS-DD structure



Documents (HTML, SKOS) prototyped for the PREMIS-KB:

- **itrousr**=identifier of the Sapienza Library System
- **vocs_2017**=provenance information URI of the PREMIS-KB generation activity
- **Object_messageDigest**=Concept identifier

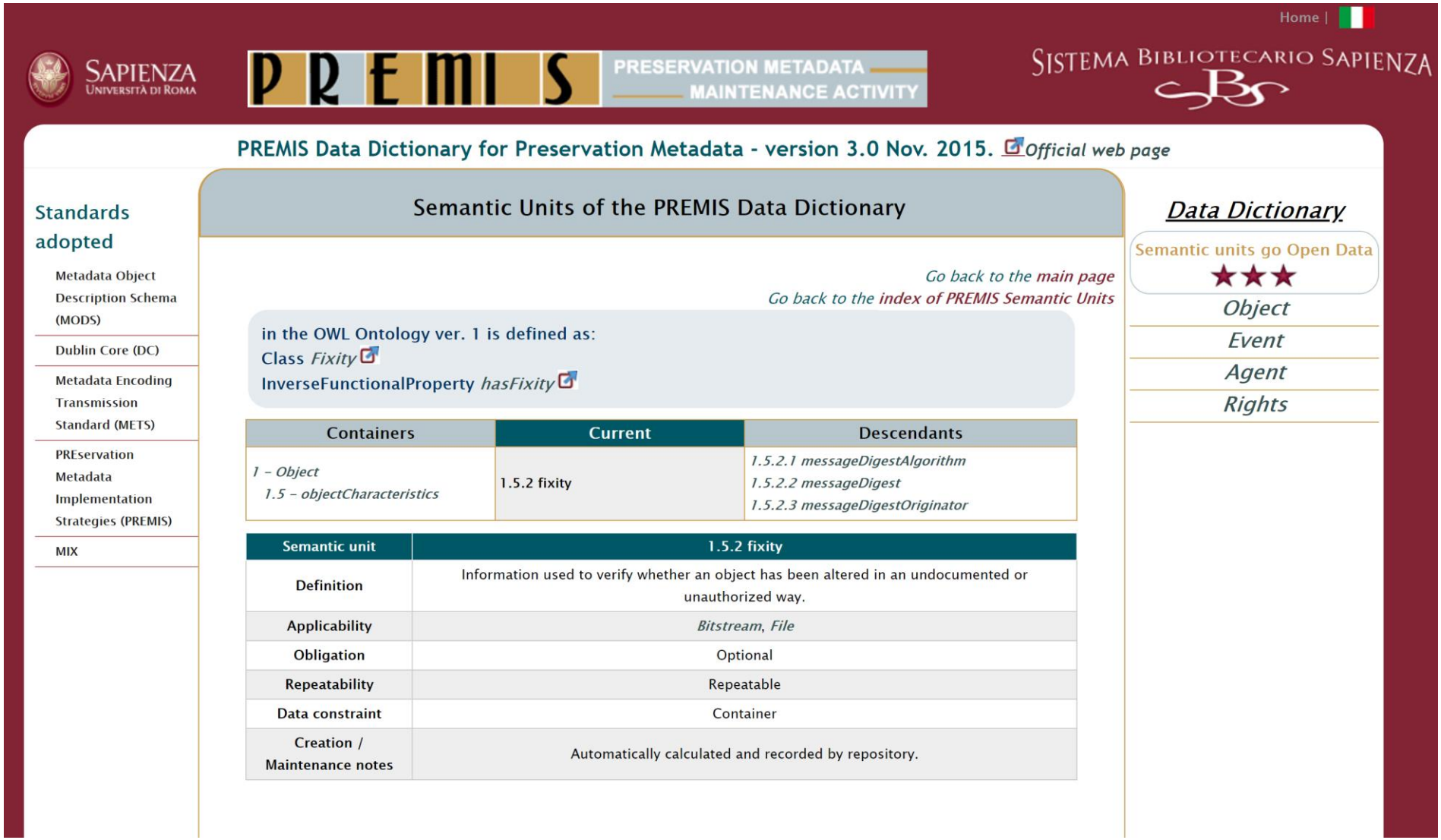
Integration of PREMIS-KB into PREMIS-OWL



Integration of PREMIS-KB into PREMIS-OWL

```
#---------->File class
premis:File
a owl:Class ;
rdfs:isDefinedBy
<premiskb/[URIpref]-PREMISDD_2017_Object_File> .
#---------->Event class
premis:Event
a owl:Class ;
rdfs:isDefinedBy
<premiskb/[URIpref]-PREMISDD_2017_Event> .
#---------->Ingestion subClassOf Event
premis:Ingestion
a owl:Class ;
rdfs:isDefinedBy
<http://id.loc.gov/vocabulary/preservation/eventType/ing>
#---------->ingests subPropertyOf hasEventRelatedObject
premis:ingests
a owl:ObjectProperty ;
rdfs:subPropertyOf premis:hasEventRelatedObject ;
rdfs:domain premis:Event ;
rdfs:range premis:Object ;
rdfs:isDefinedBy
<premiskb/[URIpref]-PREMISDD_2017_Event> .
rdfs:seeAlso
<premiskb/[URIpref]-PREMISDD_2017_Object> .
```

PREMIS-KB prototype



https://sbs.uniroma1.it/data/premisDataDictionary/dlcmrc-vocs_2016-PREMISDD_2017_Object_fixity

Impact of the PREMIS-KB

Computable PREMIS-KB will benefit:

- knowledge domain experts for PREMIS evolving changes
- community of developers for the implementation work;
- tasks of providing and consuming the PREMIS Linked Datasets;
- maintenance and design evolution of the PREMIS-OWL;
- long-life learning, the continuing education and the workers' professional development about the digital preservation metadata domain.

PREMIS-KB developments and challenges

- enrichment of PREMIS-KB with PREMIS-DD translations (i.e. current Japanese, Spanish) considering that SKOS is devoted to manage multilingualism: other work is needed for including other languages.
- where to host PREMIS-KB, LOC website and/or wikidata?
- tools for exploiting PREMIS-KB dataset and support the community

Question time

Thank you for your kind attention

Do you have any question?

PREMIS Knowledge Base