

The Matterhorn RDF Data Model

Formalizing Archival Metadata With

SHACL

Alain Dubois (Archives de l'Etat du Valais (CH))
Tobias Wildi (docuteam GmbH (CH))

Introduction

- Matterhorn RDF is a **linked-data based model for archival metadata**
- Covers the three standards ISAD(G) / ISAAR(CPF) / ISDF
Will be based on the RiC CM as soon as V1.0 is available
- Based on the Matterhorn METS-Profile
<http://www.loc.gov/standards/mets/profiles/00000041.xml>
- Based on classes and properties from existing ontologies

Matterhorn RDF compared to EGAD / RiC: The same goals – two different approaches

- **EGAD's nongeneric approach:** Developing an RDF standard specific to archives, with gateways to library and museum standards
- **The Matterhorn RDF Data Model's generic approach:** Based upon RDF existing and consensual international standards, allowing to model Records in Context.
- In contrast to EGAD, the Matterhorn RDF Data Model is based on existing ontologies. It follows the best practices propagated by the W3C: *«It is best practice to use or extend an existing vocabulary before creating a new vocabulary.»*

The goal: improved contextualization

Contextualization of documents

- What? → Content (ISAD)
- Who? → Actors (ISAAR)
- How? → Process of creation (ISDF)

Contextualization goes further

- Preservation description information
- Representation Information

The goal is a model that contains **both the content and the technical contextualization** of a document or a record.

Semantic technologies instead of XML

Matterhorn METS

- is today used in > 25 institutions in Switzerland, Germany, France.
- Based on METS, PREMIS 2, EAD

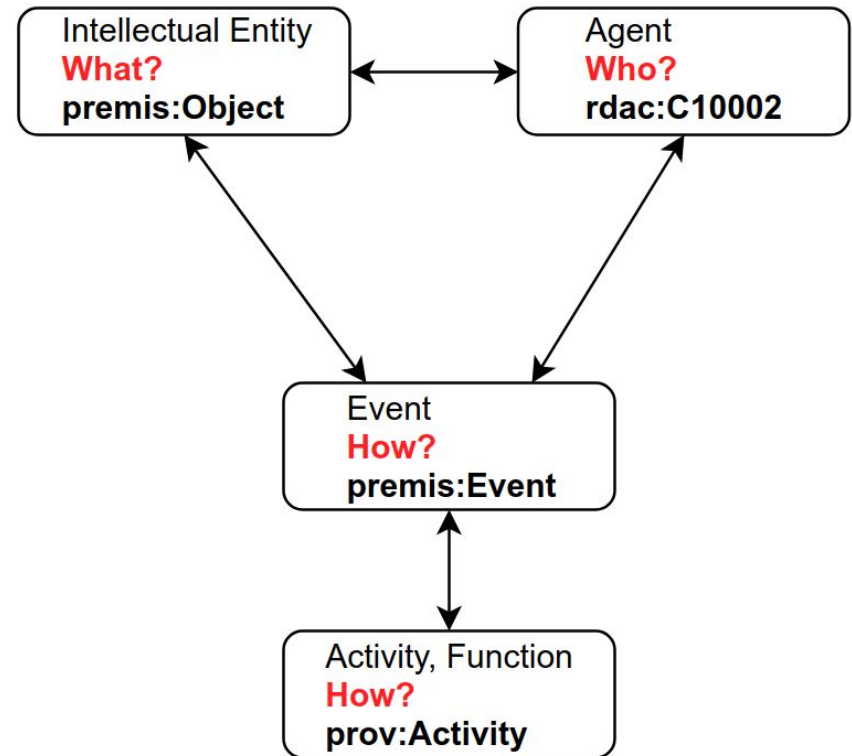
Change to semantic technologies because

- Resource can be uniquely identified using a URI
- Relationships between resources can be qualified
- The use of external resources and knowledge sources for cataloguing is greatly simplified (eg. Wikidata, GND, VIAF).

The core of Matterhorn RDF

Very similar to the PREMIS 3-model with three core classes

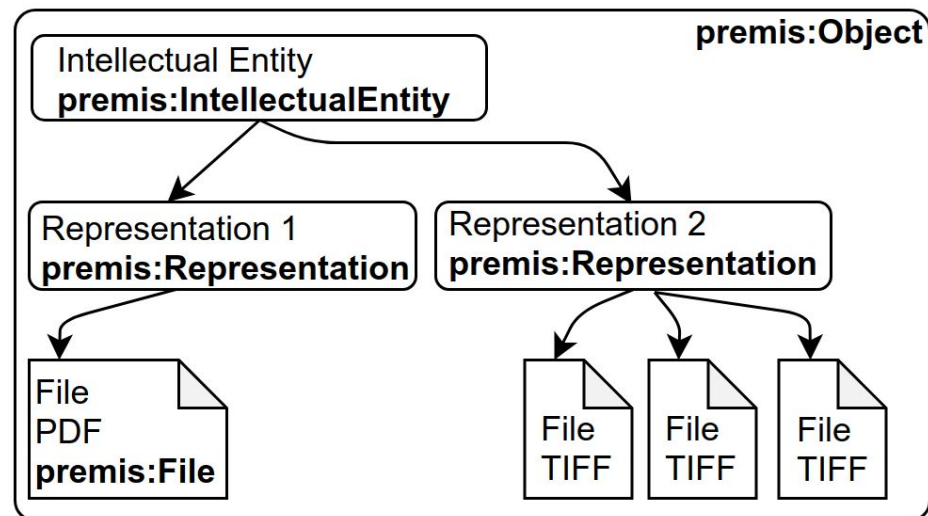
- **Intellectual Entities (Records):**
premis:object from PREMIS 3 ontology
- **Agents:** rdac:C10002 from RDA ontology
- **Functions and Events:**
prov:Activity from PROV ontology of the W3C



The concept of representations

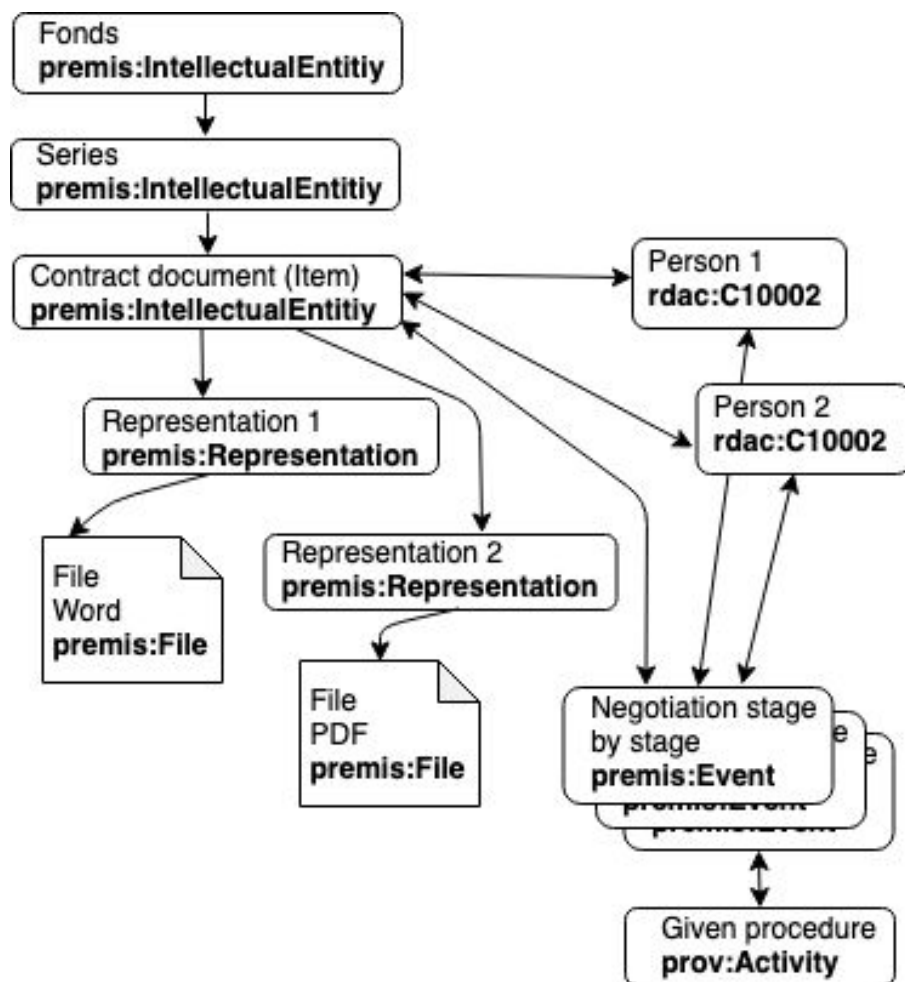
A document or record can be displayed by several representations at the same time. Example:

- Text document (= Intellectual Entity)
- Representation 1 consists of 1 PDF file
- Representation 2 consists of n TIFF files



Descriptive and technical metadata are combined in a single data model.

Example: contextualization of a contract



SHACL instead of an Ontology

- No ontology for Matterhorn RDF. Maintenance would cost too much time + effort.
- Formalization is done in Shapes Constraint Language (SHACL)
- For each property, restrictions regarding value ranges, minimum or maximum occurrence and data types are to be formulated.
- Statements made in triples can be validated

```
sh:property [  
    sh:path dc:title ;  
    rdfs:label "Title"@en ;  
    rdfs:label "Titel"@de ;  
    rdfs:label "Titre"@fr ;  
    rdfs:comment "ISAD 1.2" ;  
    owl:sameAs rico:title ;  
    sh:datatype xsd:string ;  
    sh:minCount 1 ;  
    sh:maxCount 1 ;  
    sh:nodeKind sh:Literal ;  
]
```

Conclusion

- Matterhorn RDF is a new approach to encoding and modeling archival metadata
- The innovation lies in the new combination of existing ontologies for the contextualization of records in archives
- Both descriptive and technical metadata are mapped with the model

Contacts

Alain Dubois

Archives de l'Etat du Valais,
Sion, Switzerland

alain.dubois@admin.vs.ch

@aladubois

Tobias Wildi

docuteam GmbH,
Baden, Switzerland

t.wildi@docuteam.ch

@wildit

<https://matterhorn.tools>