The Matterhorn RDF Data Model

Formalizing Archival Metadata With SHACL

Tobias Wildi
docuteam GmbH
Baden, Switzerland
t.wildi@docuteam.ch

Alain Dubois
Archives de l'Etat du Valais
Sion, Switzerland
alain.dubois@admin.vs.ch

Matterhorn RDF is a linked data-based model for archival metadata with the goal of improving the contextualization of archival records. It covers the three standards ISAD(G), ISAAR(CPF) and ISDF, as well as the areas "Preservation Description Information" and "Representation Information" of the OAIS information model. For the implementation of Matterhorn RDF, classes and properties of existing ontologies are used. The formalization of the model is realized with the help of SHACL shapes. [1]

Keywords: Archival metadata model, linked data, ontology, SHACL, RiC, contextualization

Conference Topics: Exploring New Horizons.

I. Introduction

This paper describes a model for archival metadata based on semantic technologies. The model represents both descriptive and technical metadata, specifically the standards ISAD(G), ISAAR (CPF) and ISDF of the International Council on Archives (ICA), as well as "Representation Information" and "Preservation Description Information" from the OAIS information model. The model also takes into account the current work of the ICA's Expert Group on Archival Description (EGAD), but chooses a different design approach than their conceptual model Records in Context (RiC).

The first part of this document defines the goal and scope of Matterhorn RDF. The second part substantiates why semantic technologies are used for the model and how they eliminate the disadvantages of today's XML-based data models. The third part outlines the design principles of Matterhorn RDF. This includes the decision not to develop an new ontology but rather exclusively use classes and properties of existing ontologies. The Shapes Constraint Language (SHACL) is used to formalize and validate Matterhorn RDF. The fourth and fifth parts explain the concept model and the class model of Matterhorn RDF. The most important and at the same time unspectacular finding of both these parts is the realisation that the innovation of Matterhorn RDF lies in the adaptation of existing models and ontologies for use in archives. The last part provides an outlook on the potential of Matterhorn RDF in terms of its technical implementation.

II. Improved contextualization as a goal

Archival metadata have the function of keeping the context in which documents were created comprehensible over a long period of time. Archival material has to be placed in a context to have any value. Thus, documents are contextualised through the description of their content (What?), the actors involved (Who?) and the process of creation (How?). The triangle of what, who and how has been covered to date by the three standards ISAD(G), ISAAR (CPF) and ISDF. While EAD and EAC can be coded in XML, the same is not true for ISDF. The three standards were developed by ICA over several years, with the result that they partly overlap and it is now unclear as to how relationships between them are to be mapped. The aim of Matterhorn RDF is firstly to ensure the encoding of the three standards and secondly to show how relationships between them can be modelled.

The need to revise, standardize and improve the relationship between the existing standards also manifested itself within the ICA. The Expert Group on Archival Description (EGAD) was founded in 2012 with the task of developing a new model under the title "Records in Context". Matterhorn RDF is not to be seen as an alternative to RiC, but rather seeks to

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

elaborate the RiC concept model in a future version, taking into account, however, different design considerations to those which EGAD currently implements.

The perimeter of Matterhorn RDF goes beyond descriptive metadata: the model also includes technical metadata necessary for the long-term preservation of digital objects. These are "Preservation Description Information" and "Representation Information" from the OAIS information model. Matterhorn RDF thus lays the foundation for a model that contains both the content and the technical contextualization of a record.

III. Semantic technologies instead of XML

Matterhorn METS, the predecessor of Matterhorn RDF, was registered with the Library of Congress in 2012 in the form of a METS profile. [2] Today, Matterhorn METS is used by around 25 institutions in Switzerland, Germany and France. This XML-based model is based on the standards METS, PREMIS, EAD and EAC. [3]

The modelling of archival metadata in XML leads to problems in the technical implementation for several reasons. Firstly, the typical hierarchies for archives (tectonics) generate deeply nested structures in XML. Secondly, the two standards EAD and PREMIS require elaborate XML constructs compared with the information actually transported. Thirdly, the use of persistent identifiers in XML is by no means self-evident and must be explicitly specified.

For a successor model, semantic technologies were the obvious choice in order to simplify structures and better model relationships between individual resources. There were three reasons for using Linked Data. Firstly, each resource can be uniquely identified using a URI. This is an advantage over the original XML-based approach, where identifiers were unique only within a single METS file. Secondly, the relationships between resources can be qualified. For example, not only is a relationship between two people propagated, the relationship is additionally qualified with the help of so called predicates like "child of" or "married to". The third and most important reason is that the use of external resources and knowledge sources for cataloguing is greatly simplified. Archival cataloguing today largely consists of filling in free text fields in database applications. In contrast to library cataloguing, this procedure is less systematic. With Linked Data, the full text description is at least partially replaced by linking to already existing knowledge sources. These can be entries in Wikidata, GND or VIAF, for example, each of which can be uniquely referenced via a URI. The reference to long-term stable external resources promotes the efficiency and accuracy of archive cataloguing. And vice versa, resources in one's own archive can be used much more easily by third parties.

IV. Design Principles of Matterhorn

The central design principle of Matterhorn RDF is that, as a linked data-based model, it does not have its own ontology. The model is based exclusively on classes and properties of existing ontologies. It regroups and correlates them with each other using a conceptual model. This design principle is derived from the Best Practices for Publishing Linked Data of the W3C, which state: "Standardized vocabularies should be reused as much as possible". [4] State actors, including many archives, are especially called to account: "Government publishers are encouraged to use standardized vocabularies rather than reinventing the wheel, wherever possible."

The decision not to create a domain-specific ontology for archival metadata allowed for the development of a data model in a relatively short period of time and resource-saving manner. The fact that no data dictionary had to be written in order to precisely execute the semantic meaning of each property, was especially time-saving. It was sufficient to refer to the descriptions of the respective ontologies.

V. Overview and most important elements

The Matterhorn RDF model is conceptually based on the three standards ISAD(G), ISAAR (CPF) and ISDF,[1] as well as the specifications for Preservation Description Information and Representation Information from the OAIS information model. The model is very similar to the PREMIS3 ontology and

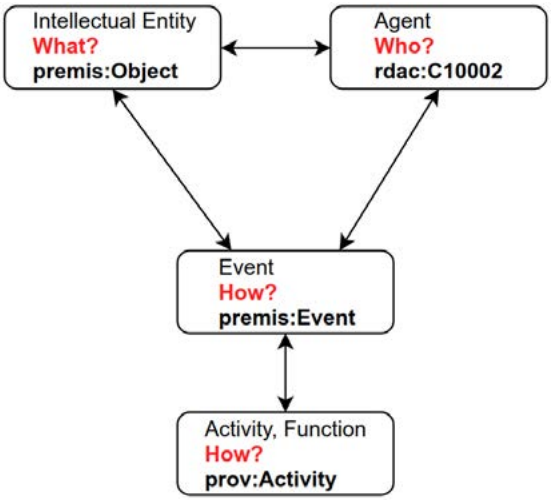
[1] As soon as RiC is consolidated, the RiC concept model will be implemented in the next version of Matterhorn RDF.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

works with the following three core classes:

- **Intellectual Entities (Records):** premis:object from PREMIS3 ontology
- **Agents:** rdac:C10002 from RDA ontology
- **Functions and Events:** prov:Activity from PROV ontology of the W3C

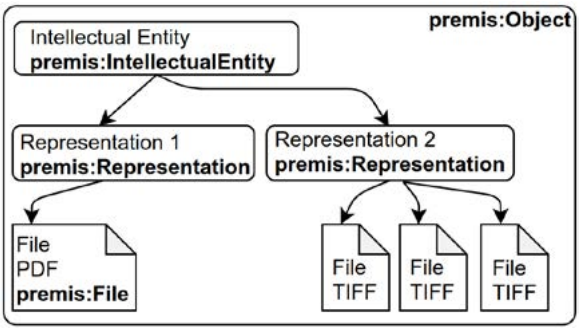
These classes are structured hierarchically into subclasses. The classes are related as follows:



Only some of the used properties in the Matterhorn RDF model come from the ontologies of the corresponding classes. In addition, properties from Dublin Core, Ebucore or the standardized preservation vocabularies of the Library of Congress are used. The PREMIS standard does not include properties for descriptive metadata, therefore, attributes from other ontologies must be used. Dublin Core, Dublin Core Terms and RDA (Resource Description and Access) contain attributes that semantically correspond to the respective ISAD(G) fields.

The description of archival content takes place in the premis:IntellectualEntity class, a subclass of premis:Object. Intellectual entities are brought into a hierarchical relationship to each other via "has part" relationships, thus modelling the ISAD(G) tectonic. Horizontal or associative relationships between intellectual entities can also be modelled. An important feature is that a record or a single intellectual entity can be displayed by several representations at the same time. For example, a text document (= Intellectual Entity) can be represented by a PDF file as well as several TIFF files. To model this, the two following premis:Object subclasses, premis: Representation and premis:File, are used. These subclasses do not contain any descriptive metadata, they do, however, contain technical metadata from the PREMIS ontology. Thus, descriptive and technical metadata are combined in a single data model.

The graphical representation is as follows:



The actors defined by ISAAR(CPF) are represented in the class rdac:C10002. RDA is a set of library rules for cataloguing and publishing. [5] With FRBR, RDA has its own data model, which we are not concerned with in this context, because it is based on the concept of the "work", which is relevant only to libraries and has no bearing on archives. The part of the RDA-Ontology concerningthe so-called "Agent Properties" is, however, semantically largely congruent with the ISAAR (CPF)-Standard. Therefore, the already existing and widespread RDA-Ontology can be used to encode ISAAR (CPF). The class rdac:C10002 includes "A person, family, or corporate body", i.e. exactly the same concepts as ISAAR (CPF). Corresponding properties to the ISAAR(CPF) fields can be found in RDA and for auxiliary fields (versioning, language etc.) Matterhorn RDF uses the Dublin Core Terms ontology.

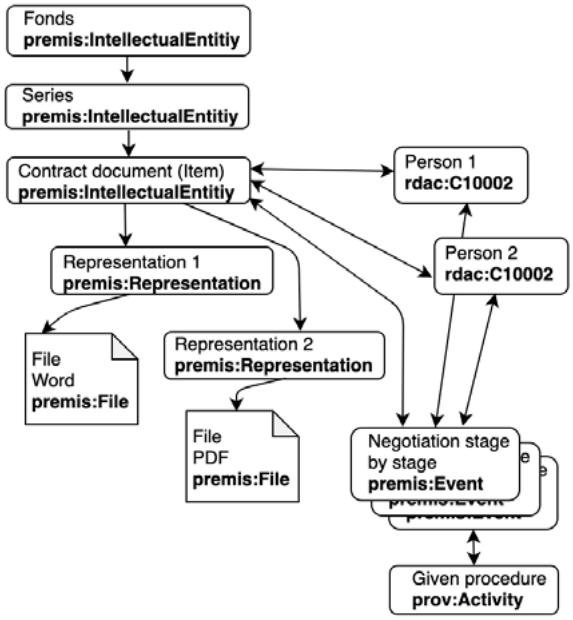
Functions, i.e. administrative tasks, processes and activities, are described with the help of the ISDF standard. These metadata form the basis for documenting the creation (and use) of records. The PROV data model and the PROV ontology of the W3C can be used to implement this. PROV is widely used and recommended by W3C for the modelling of "entities, activities and people". Matterhorn RDF, however, exclusively uses PROV's area of activities.

Two prov:Activity-subclasses model the process description on the one hand and the process documentation on the other hand. In prov:Activity the generic description of a business process or

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

administrative procedure can be found in the form of a sequence of various related activities. An activity is a generic concept for the work that a person or organization performs. It can stand alone or be composed of sub-activities. In the premis:Events class, a subclass of prov:Activity, the actual course of a business process is documented by means of individual events.

The negotiation of a fictitious contract between two persons shall give an exemplary illustration of the entire model. The content of our contract document is described using the premis:IntellectualEntitiy class. There are two representations of the contract document (premis:Representation), a first premis:File in the form of a word file and a second premis:File in the form of a PDF. The premis:File class also stores technical metadata such as checksums and file format information. The contract was signed by two persons who are described using the rdac:C10002 class. The negotiation of the contract followed a given procedure, which is stored in prov:Activity. Each step in this process, including several rounds of negotiation, is documented in premis:Events. This provides us with metadata for our contract on all three questions What, Who and How, as well as technical metadata that form the basis for Preservation Planning. Thus, the contract is put into context and its creation is documented in a comprehensible way.



VI. Formalization and validation

Matterhorn RDF does not formulate its own ontology. The development and ongoing maintenance of a new ontology requires much time and effort. Nevertheless, it is possible to formalize the model. This should entail a description of the classes the model consists of as well as the definition of the necessary properties and their purpose. For each property, restrictions regarding value ranges, minimum or maximum occurence and data types are to be formulated. For XML-based data models the proven schema language is available for this purpose. For semantic models the equivalent Shapes Constraint Language (SHACL) has been available since 2017. [6] [7] SHACL is used to formulate so-called shapes, against which the statements made in the RDF triples are validated. The formulation of shapes is therefore an elegant way to describe an RDF-based data model built on existing classes.

The shapes are published online. [8] The development of the shapes for all elements of Matterhorn RDF should be completed by the end of 2019. The following example of the ISAD(G)-field “Title” of will show how such a shape looks like.

```
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 ;
] .
```

In this shape the property “dc:title” is specified in greater detail. The labels of the title field are defined in different languages, an important functionality for a multilingual country like Switzerland. A comment line refers to the ISAD(G) standard field 1.2. A further reference to the corresponding field in the RiC ontology is made with the help of owl:sameAs. The title field has to contain data of the type “string” and must appear exactly once. The entry of a value (“Literal”) is expected and not a reference to another node (“IRI” or “IRIOrLiteral”).

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

VII. Conclusion

The Expert Group on Archival Description (EGAD) is currently formulating its own ontology with RiC-O for the "Records in Context" concept model. With Matterhorn RDF we propose an alternative way to contextualize records. Our model is based on already existing and widely used ontologies, which brings an increase in efficiency not only in the development but especially in the maintenance of the model. The model can be formalized even without an ontology of one's own. SHACL is a suitable tool for this purpose. Matterhorn RDF and RiC-O should not be competing models. By using the SHACL-shapes to store the semantic equivalents of RiC-O, the matterhorn RDF-model ensures the necessary crosswalk between the two models.

The transition from encoding archival metadata in XML or relational databases to linked data-based solutions will fundamentally change the way archives are described. Today, the primary access to archival material takes place through a single hierarchy structured according to ISAD(G). In the future, access and entry points will also be possible via actors or business processes. The origin context of records is therefore no longer documented in rigid, non-adaptable XML schemas but in a flexibly extendable model.

The activity of archival description is shifting away from a barely systematized textual description in free text database fields towards linking archival content to already existing and clearly referenceable knowledge resources. The search and access to the archive will also change. Today's full text search for terms and character patterns is being replaced by structured access to clearly identifiable resources.

Matterhorn RDF is thus 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 and in the fact that both descriptive and technical metadata are mapped with the model.

References

[1] All information about Matterhorn RDF is available at http://matterhorn.tools

[2] Matterhorn METS profile: http://www.loc.gov/standards/mets/profiles/00000041.xml

[3] Wildi, Tobias. Spezifikation Matterhorn METS. Baden, 2017. http://matterhorn.tools/matterhorn-mets.pdf

[4] Best Practices for Publishing Linked Data. W3C Working Group Note 09 January 2014: https://www.w3.org/TR/ld-bp/#VOCABULARIES

[5] RDA Ressource Description & Access: http://www.rdaregistry.info/

[6] Shapes Constraint Language (SHACL). W3C Recommendation 20 July 2017: https://www.w3.org/TR/shacl/

[7] For validation of RDF see: Ying Ding, Paul Groth, "Validating RDF Data". Morgan & Claypool Publishers, 2017.

[8] https://bitbucket.org/docuteam/matterhorn

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.