Modeling cultural heritage data for online publication

Chris Dijkshoorn [a,b,*], Lora Aroyo [b], Jacco van Ossenbruggen [c] and Guus Schreiber [b]
[a] *Research Services, Rijksmuseum Amsterdam, the Netherlands*
E-mail: c.dijkshoorn@rijksmuseum.nl
[b] *Department of Computer Science, Vrije Universiteit Amsterdam, the Netherlands*
E-mails: lora.aroyo@vu.nl, guus.schreiber@vu.nl
[c] *Centrum Wiskunde & Informatica, the Netherlands*
E-mail: jacco.van.ossenbruggen@cwi.nl

Abstract. An increasing number of cultural heritage institutions publish data online. Ontologies can be used to structure published data, thereby increasing interoperability. To achieve widespread adoption of ontologies, institutions such as libraries, archives and museums have to be able to assess whether an ontology can adequately capture information about their artifacts. We identify six requirements that should be met by ontologies in the cultural heritage domain, based upon modeling challenges encountered while publishing data of the Rijksmuseum Amsterdam and challenges observed in related work. These challenges regard specialization, object- and event-centric approaches, temporality, representations, views and subject matter. For each challenge, we investigate common modeling approaches, by discussing two models regularly used in the museum sector: the CIDOC-Conceptual Reference Model and the Europeana Data Model. The outlined approaches and requirements provide insights for data modeling practices reaching beyond the cultural heritage sector.

Keywords: Ontologies, Data Models, Cultural Heritage, Museums, Semantic Web, Linked Data

1. Introduction

Cultural heritage institutions possess a wealth of information and there is a growing understanding that it would be beneficial to share this online. It is, however, a non-trivial step for institutions to move from traditional information dissemination methods, such as exhibition catalogs and research papers, to the publishing of data (Knoblock et al., 2017). To do so, cultural heritage institutions need to consider aspects that were previously not part of their core activities (for example using standardized data models and aligning data with other institutions). To aid this transition, the current paper will: 1) outline data models in the cultural heritage domain; 2) demonstrate using two example artifacts, that different modeling approaches impact the information that can be published; and 3) combine insights from modeling challenges to form requirements, in order to aid institutions in choosing an appropriate data model.

Most cultural heritage data is contained in collection management systems and catalogs that often function as data silos; systems that can not be accessed by the outside world (Hyvönen et al., 2005). Over the years, efforts to export this data have had many manifestations, resulting in different data formats and data models. This makes the reuse and integration of cultural heritage data a cumbersome

[*] Corresponding author. E-mail: c.dijkshoorn@rijksmuseum.nl.

task. Adoption of Linked Data practices improves the interoperability of cultural heritage data. Linked Data principles advocate systematic referring to resources and syntactically standardized publishing of data (Bizer et al., 2011). Ontologies make the semantics of published data explicit, by providing a shared conceptualization (Studer et al., 1998). The reuse of ontologies is encouraged and there are specialized ontologies available for many different domains (Janowicz et al., 2014). The cultural heritage domain is no exception, with models specifically tailored to libraries, archives and museums (Doerr, 2003).

A cultural heritage institution is confronted with many important choices when it decides to publish Linked Data and reuse an ontology (Obrst et al., 2014). One of the first decisions to make, is whether to invest in infrastructure to publish Linked Data or provide data to an aggregator of cultural heritage information. The former leaves the choice of which ontology to use up to the institution, while an aggregator might require data to be provided in a specific structure (Clayphan et al., 2016). The decision of which ontology to use has implications for the source data that can be included, as well as the shape of the resulting Linked Data. In this paper, we discuss the impact of such decisions, with respect to six modeling challenges. The challenges originate from related work, as well as our experience of publishing data about objects in the collection of the Rijksmuseum Amsterdam (Dijkshoorn et al., 2018). The mission of this museum is to provide an overview of Dutch art and history from the Middle Ages onwards. The museum is well known for its paintings by Old Masters, although the collection includes artifacts ranging from weapons to sculptures.

We introduce the modeling challenges using examples from the Rijksmuseum's collection and illustrate modeling approaches using two cultural heritage ontologies. The CIDOC Conceptual Reference Model (CIDOC-CRM) is specifically developed for the museum sector and is intended to be used to create interoperable data. The Europeana Data Model (EDM) is an ontology that enables cultural heritage institutions to structure collection data so that it can be used by the data aggregator Europeana. This model is designed to retain other more specific data models that are used by libraries, archives and museums. We chose these ontologies for illustrating modeling approaches since they are commonly considered by institutions for publishing data, but take a different approach in doing so. We draw requirements for cultural heritage ontologies from the challenges.

The remainder of this paper is structured as follows: in Section 2 we discuss related work on cultural heritage ontologies and datasets, to identify modeling challenges encountered while publishing cultural heritage data. We describe two exemplary artifacts of the Rijksmuseum in Section 3: a wedding portrait and a cased pair of pistols. These objects are used to illustrate modeling challenges, regarding specialization, object- and event-centric approaches, temporality, representations, views and subject matter. The challenges and modeling approaches addressing them, are discussed in Section 4. We discuss requirements distilled from these challenges in Section 5 and end with a conclusion and future work section.

2. Related work

In this paper we focus on how ontologies can be used to structure and represent information about artifacts in cultural heritage collections. In the context of computer science, Studer et al. (1998) defined an ontology as a "formal, explicit specification of a shared conceptualization". A conceptualization is an abstract view of the world we want to represent. Making the conceptualization explicit entails deciding on a language to use and constraining the interpretations of such language. "Formal" refers to the specification being machine readable (Guarino et al., 2009). In Section 4, we use the Europeana Data Model and the CIDOC Conceptual Reference Model to illustrate different approaches to modeling cultural heritage data.

The CIDOC Conceptual Reference Model (CIDOC-CRM) is an event-centric reference ontology for the cultural heritage sector, maintained by a special interest group of the ICOM international committee for documentation (Doerr, 2003). Constructs of the CIDOC-CRM are based on empirical studies of collection management systems (Crofts et al., 2011, p. i). The ontology aims to be a "discipline neutral", common semantic reference point, improving the semantic and structural interoperability of cultural heritage data. CIDOC-CRM has been accepted as an ISO standard for the interchange of cultural heritage information in 2006.

The Europeana Data Model (EDM) is developed to represent and structure cultural heritage data so that it can be delivered to the data aggregator Europeana (Doerr et al., 2010). EDM is used internally by Europeana to aggregate, process, enrich and disseminate data. The model reuses constructs from other data models, such as the Dublin Core metadata initiative[1] and the Open Archives Initiative Object Reuse and Exchange standard[2] (Isaac, 2014). EDM is a top-level ontology to which institutions can map their more specific data models (CIDOC-CRM can for example be embedded in EDM). This approach makes it useful beyond its original purpose of data delivery: the model is nowadays used by other aggregators, as well as institutions publishing their own data (Valentine and Isaac, 2015).

Both ontologies are used by museums that publish a Linked Data version of their collection. For instance, the Amsterdam museum ontology and VVV ontology specialize the EDM top-level ontology to structure collection data (de Boer et al., 2012; Dragoni et al., 2016). In addition, the Rijksmuseum dataset is published using a combination of the Dublin Core model and EDM (Dijkshoorn et al., 2018). A collaboration of 14 American art museums mapped collection data to CIDOC-CRM (Knoblock et al., 2017). This ontology is also used to publish collection data of the Yale center of British Art[3], British Museum[4] and Russian Museum (Mouromtsev et al., 2015).

There are many more cultural heritage collections available as Linked Data, through so-called "aggregators". Aggregators are organizations that host multiple collections, creating an integrated point of access to artifacts of different institutions. MuseumFinland is a collaboration of Finnish museums, while the LODAC Museum includes many Japanese museums (Hyvönen et al., 2005; Matsumura et al., 2012). Europeana is an aggregator of European cultural heritage data, which connected data of over 3,000 institutions in 2017 (Dunning and Verspille, 2017).

3. Two examples: a portrait and a pair of pistols

In this section, we introduce two artifacts from the Rijksmuseum Amsterdam. The first artifact is part of a pair of wedding portraits and the second is a set of pistols. Based on related work and our own experience modeling the Rijksmuseum data, we illustrate typical modeling challenges of the cultural heritage domain with these artifacts in the next section.

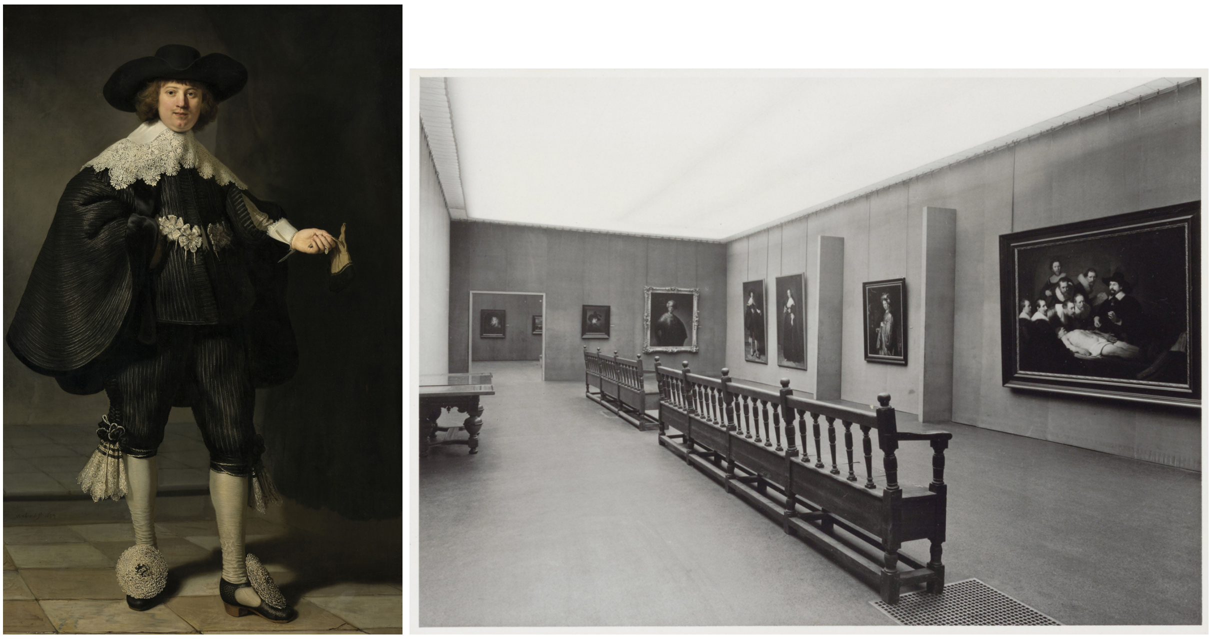
Portrait of Marten Soolmans. In 1634 Rembrandt van Rijn painted a pair of portraits in honor of the wedding of Marten Soolmans and Oopjen Coppit. The paintings show the young married couple in exuberant detail, dressed in black and adorned with many lace details. Figure 1a shows the portrait of Marten Soolmans. While part of private collections for ages, in an exceptional construction, the Dutch

[1] http://dublincore.org

[2] http://openarchives.org/ore

[3] http://collection.britishart.yale.edu

[4] http://collection.britishmuseum.org



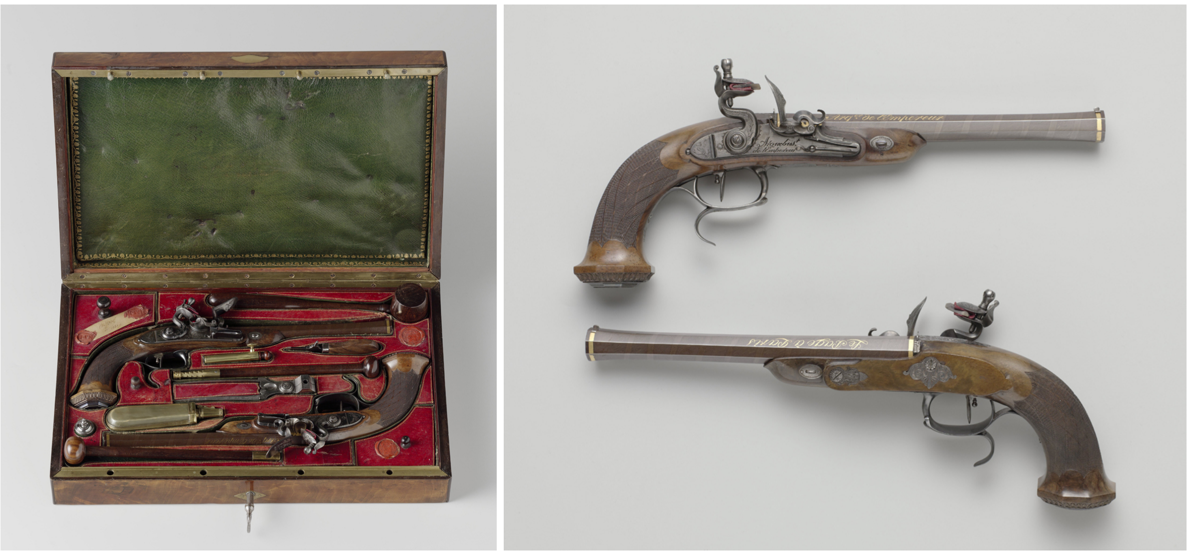
(a) Portrait of Marten Soolmans

(b) Rembrandt exhibition in 1956

Fig. 1. Portrait of Marten Soolman, created by Rembrandt in 1634 and included in an exhibition at the Rijksmuseum in 1956.

and French governments managed to acquire both paintings. The portraits will always be exhibited together and their location will alternate every five years between the Rijksmuseum and the Louvre. The Rijksmuseum maintains an archive, which includes files that document exhibitions. Figure 1b shows a picture of the two marital portraits exhibited at the Rijksmuseum in 1956, as part of an exhibition in honor of the 350th anniversary of the birth of Rembrandt. Two institutions with different views on the same artifact, related events and the availability of digital representations, makes modeling information about the portrait of Marten Soolmans an illustrative example of the challenges that can be encountered during the publication of cultural heritage data.

Cased pair of pistols. The second example is a cased pair of pistols, shown in Figure 2. These flintlock pistols were manufactured in the workshop of Jean Le Page, around the year 1808. The pistols lend their historical significance by reputedly being owned by Napoleon I Bonaparte, emperor of France. After the battle of Waterloo, the cassette containing the pistols was found in the traveling carriage of Napoleon and there are letters supporting the assessment that Napoleon once owned the cassette. Besides the pistols, the cassette also contains accessories, such as a powder horn, bullet mold, rammer and hammer. The pistols are made from multiple materials, such as walnut, steel and gold. The weapons are adorned with engravings both written, as well as figurative. For example, an eagle is depicted on the side of a pistol, while its barrel is engraved in gold, with the text “Arger de l’Empereur”. The pistols are well suited for illustrating the challenges of modeling cultural heritage data, because of their components that have been created at different moments in time, the detailed provenance information and the depicted subject matter. But the first challenge we discuss in Section 4 is how to differentiate between pistols and portraits, by specializing data models.



(a) Cased pair of pistols (b) Pair of pistols

Fig. 2. Cased pair of pistols, reputedly owned by Napoleon, created by Jean Le Page.

4. Modeling approaches in the cultural heritage domain

In this section we discuss modeling challenges regarding specialization, object- and event-centric approaches, temporality, representations, views and subject matter. For each challenge, we discuss the general issues in more depth and show current modeling approaches of the Europeana Data Model and the CIDOC-Conceptual Reference Model. From both ontologies we analyze encodings that are compatible with Linked Data, to which we will refer to as data models. We illustrate the discussion using the two examples introduced in the previous section.

4.1. How to specialize an interoperable data model

Libraries, museums and archives hold many different types of artifacts. The Rijksmuseum alone has ten different sub-collections, ranging from paintings to furniture. To achieve the desired level of interoperability across these sub-collections, some level of abstraction is needed to support descriptions on a more generic level. An overly generic data model, however, might “trivialize” descriptions of these distinct artifacts. This happens when important, but collection-specific information is systematically left out because it does not “fit” the general data model used. This has the additional risk that curators and domain specialists stop to support data publishing, if they are under the impression that this implies committing to generic models that do not fit their domain sufficiently. At the same time, there are use cases which require generic descriptions of artifacts, for example, to achieve interoperability with collections with slightly different characteristics. This interoperability is important from a managerial perspective, allowing the participation in vertical as well as horizontal integration projects. In relational or XML-oriented data models, users find it often hard to specialize data models without losing their interoperable generic structures. But even when Semantic Web technologies are used, there are still choices to be made on *how* to provide such collection-specific aspects.

Specialization by extending a top-level class hierarchy. One approach is to develop a commonly agreed upon, top-level class hierarchy, that provides the required level of abstraction and interoperability but allows more specific descriptions by refining the generic classes given. CIDOC-CRM defines 82 of such top-level classes, whereas EDM describes 18 classes. The documentation of EDM recommends to use the most specific construct available, thereby contributing to the precision of descriptions (Isaac, 2013, p. 11). However, the most specific EDM class that can be assigned to the pistols and painting is the fairly general *provided cultural heritage object*. CIDOC-CRM is a reference ontology with a similar approach and the most specific class that can be assigned to a cultural heritage object is *physical man-made thing*. Neither of these two general classes allows us to differentiate between a painting and a pistol.

This means that for many purposes, institutes may wish to add more specific classes, either by using a shared profile or by using an institute-specific set of extensions. An institution could, for example, introduce the class *weapon* and relate it to the CIDOC-CRM class *physical man-made thing*. Once the class *weapon* is assigned to one of the pistols, it can still be deduced that it is a *physical man-made thing*. At the moment a class *painting* is added as well, it is possible to differentiate between the two types of artifacts.

Typing using terminologies. Additional typing of instances with terms from hierarchically structured vocabularies is another common approach to achieve specialization without sacrificing generality. These structured vocabularies can take different forms, such as thesauri, classification schemes and gazetteers (Hooland and Verborgh, 2014, Chapter 4). An example is the Art and Architecture Thesaurus[5] (AAT), which is used to relate artifacts to materials and techniques. These structured organizations of concepts serve as shared vocabularies for data publishers and can improve artifact retrieval tasks (Baker et al., 2013; Wielinga et al., 2001).

Both CIDOC-CRM and EDM include the property *has type*. Using this property we can state that one of the pistols is of type "flintlock pistol". Such terms can often be reused, for example, from structured vocabularies such as the AAT. Many of these vocabularies are structured using the Simple Knowledge Organization System (SKOS)[6]. The terms in a SKOS vocabulary are connected using broader and narrower relations, forming a hierarchy. Using this hierarchy it is possible to deduce that a flintlock pistol is a more specific term than a weapon. This typing of instances using terms does not, however, impact the more formal RDF or OWL instance/class semantics, nor does it limit the connections that can be made between different instances.

Specialization by extending a property hierarchy. Properties are used to relate instances to other instances or literal values. CIDOC-CRM includes a total of 262 property definitions, where the EDM definition defines 35 properties and refers to 40 properties of other data models. Properties can also form hierarchies, which in turn can be extended. When extending the property hierarchy, it is important that the meaning of a sub-property is subsumed by that of properties higher up in the hierarchy. An institution could, for example, introduce the property *was painted for*, to relate the portrait of Marten to the wedding. This property should be a sub-property of *was made for* and not of *was used for*. While both properties relate things to activities, *was used for* is a sub-property of *was present at*, something which is not necessarily true for the wedding portrait.

Domains and ranges can be added to properties, thereby indicating which types of instances a property can relate. In Figures 3 and 4, the texts in ovals serve as indicators of these domains and ranges. The

[5] http://www.getty.edu/research/tools/vocabularies/aat/

[6] http://www.w3.org/TR/skos-reference/

EDM property *was present at*, for example, relates *information resources*, *things* and *agents* to *events*. Take, for example, the following statement: "the pistol was present at the battle of Waterloo". The range of the property *was present at* indicates that the instance *the battle of Waterloo* should be of class *event*. Inconsistencies can occur if the wrong properties are extended or aligned and reasoning is used to deduct additional information, something which we will discuss in more depth below.

Ontological commitment and the danger of alignment inconsistencies. A minimal ontological commitment assures maximum reusability (Studer et al., 1998). 63 Properties described in the EDM specification do not have a full domain and range specification, where every one of the 262 CIDOC-CRM properties has a domain and range specified. The omission of domain and range specifications makes the ontological commitment of EDM lower than CIDOC-CRM. While this allows EDM properties to connect multiple classes of instances, it refrains reasoners from automatically deducing the types of instances occurring as a domain or range of an EDM property. The ontological commitment of EDM is, however, impacted by alignments with constructs of CIDOC-CRM. As we will see from the example below, this can lead to undesired inconsistencies in the data when reasoning is used. Institutions that relate properties to EDM or CIDOC-CRM should consider the ramifications of their alignments carefully.

When new constructs are related to existing data models, care should be taken not to create inconsistencies. The specification of EDM aligns six classes and seven properties with constructs of CIDOC-CRM. An example of problems caused by an inconsistent alignment is the EDM property *is successor of*. This property has no defined domain or range, which allows someone to state that the book the Two Towers is successor of the Fellowship of the Ring, but also that Queen Elizabeth II is successor of King George VI. The EDM property *is successor of* is a sub-property of *is similar to*. EDM aligns this property with the CIDOC-CRM property *shows features of*. The latter has a domain and range of thing. Through reasoning, we can now deduct that the domain and range of the property *is successor of* is class thing. This is not problematic for books, but to categorize the Queen and former king as things is less appropriate. In Section 5 we further discuss the requirement for specializing data models without sacrificing interoperability, here we continue with outlining the differences between an object- and event-centric modeling approach.

4.2. Choosing between an event- and object-centric approach

Two major approaches can be used to describe cultural heritage data: the event-centric and object-centric approach. The latter puts the artifact at the center of the data model. In an object-centric data model, an artifact is directly connected to the data that describes its features. An artifact has, for example, a creator, creation date, owner and location. Event-centric data models describe artifacts using related events. A production leads to the creation of the artifact, an acquisition leads to a change of owner and a move leads to a change of location. The information that can be conveyed as well as the structure of data is impacted by a choice for one of the two approaches.

Measuring the appropriateness of a data model can be done by considering the balance between the amount of information that it is able to convey and the effort that is required to create the data structured according to the model (Lagoze and Hunter, 2006). Many cultural heritage institutions have either a collection management system or a library catalog system in place. These object-centric systems record which artifacts are part of a collection and are often the source of data published online. Attempting to convert this source data into data structured according to an event-centric data model requires much more effort than a conversion to an object-centric model. However, event-centric data could convey more

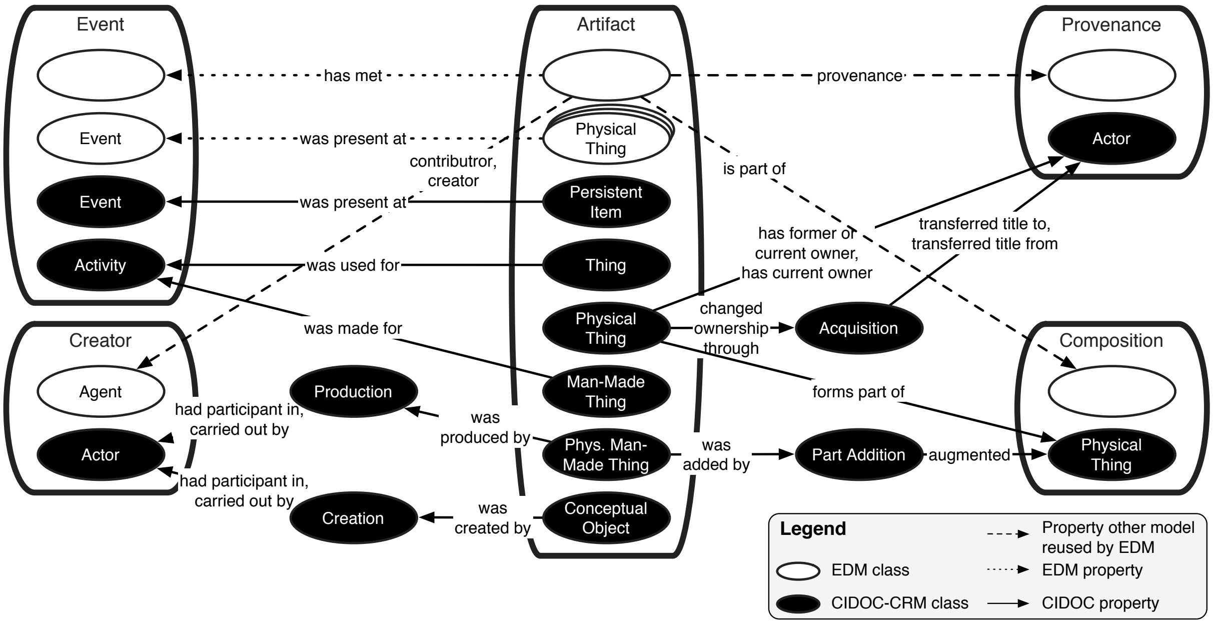


Fig. 3. Representations of four elements of an artifacts life cycle: event, creator, provenance and composition.

detailed information about the creation, evolution and transition of artifacts over time. We regard these differences by discussing the creation of an artifact in more depth.

Event-centric approach. CIDOC-CRM is an example of a data model which uses an event-centric approach. As can be seen in Figure 3, many features of an artifact are modeled using an intermediary event. An artifact is for example related to its creator, by creating a path from artifact to actor, with a production or creation event connecting the two. For the portrait of Marten, we can now state that it was produced by a production event, in which Rembrandt was involved. Attaching attributes to this event allows us to provide more details about the creation of an artifact. We can for example state that the event took place in the year 1634 and happened in Amsterdam.

The granularity of event descriptions can be increased using composition. To illustrate, multiple events can lead to the creation of one artifact. The pistols of Napoleon exist of multiple parts, such as the barrel and the grip, which are all the result of different production events. CIDOC-CRM caters for bundling the events leading to a creation, by decomposing an event into multiple related events using the property *consists of*. In practice, this can lead to long paths connecting an artifact to its creator: an artifact is produced by a production, which consists of a production carried out by an actor, who is identified by an appellation, which has label “Jean Le Page”. Object-centric approaches are less verbose since they allow an artifact to be directly connected to a string or an agent concept with a label.

The object- and event-centric approaches can exist side by side. Most information in CIDOC-CRM is conveyed using events, but so-called shortcuts can also be used to connect instances without the use of intermediary events. The property *has current owner* for example connects a physical thing directly to an actor, thereby using an object-centric approach. EDM supports both the object-centric as well as the event-centric approach. The expressiveness of the event-centric constructs in EDM is however limited since it only includes the property *was present at* and class *event*. Conveying detailed event-centric information requires making these constructs more specific, as discussed in Section 4.1

Object-centric approach. In EDM a stronger emphasis is given to the object-centric approach, due to the inclusion of many Dublin Core constructs. The reason for this emphasis is twofold: use of the object-centric approach is widespread and required constructs are readily available (Isaac, 2013, p. 17). The creator of an artifact can be indicated using the two properties *contributor* and *creator*, as shown in Figure 3. In contrast to the event-centric approach, the properties connect the artifact directly to the agent. The role of the contextual class is embedded in the properties semantics: they allow differentiating in the level of involvement of the creator. The property *creator* indicates the agent primarily responsible for the creation of the artifact, while the property *contributor* identifies someone who contributed to the artifact. Temporal information of the creation of the artifact can be conveyed using the property *created* and spatial information can be added using the property *coverage*.

Creation events bundle temporal and spatial information together with the actors involved. In contrast, the object-centric approach of Dublin Core uses three separate properties to relate an artifact to its creator, creation date and place of creation. This solution suffices if there is just one creator, but becomes problematic at the moment multiple actors with different roles are involved in the creation process. Say the barrel of the flintlock pistol was made by Jean Le Page, while Fleury Montagny engraved it. Since the date, place and creator are not connected, it is impossible to distinguish which agent was involved in what, where and when. For topics such as provenance, it is even more important to consider multiple related events. We discuss this in more detail in the next section.

4.3. *How to capture changes over time*

Capturing changes over time is relevant for cultural heritage data: artifacts are created, can be changed and might eventually be destroyed. There are also changes not directly affecting the artifact itself, but that for example regard ownership and location. We discuss why it is relevant to capture this temporal information using two examples: part addition and provenance. In some situations, it is useful to record if artifacts are augmented with new parts. An example of this is the changing of a frame of a painting. The portrait of Marten, for instance, has multiple fitting frames, but only one can be used at a given time. A museum needs to record which frame is in use and which other frames have been used before. This example shows that not only recording the current state of an artifact but also recording its changes is a worthwhile effort.

The provenance of an artifact is a series of events that regard the ownership of the artifact. The two pistols were for example reputedly owned by Napoleon, but after the battle of Waterloo bought by the tradesman Jean Sagermans, who gave them to his brother. Henry Sagermans, in turn, gave them to the State of the Netherlands. For many artifacts, not all provenance events are known. Tracing back owners can lead to new insights, sometimes showing that objects have been unrightfully obtained during some event, for example the Second World War. Provenance tells something about the history of an artifact and is often highly relevant information for researchers. The ability to capture changes over time is essential to support this type of research.

Textual descriptions of changes over time. EDM does allow recording changes of ownership using the property *provenance*. The range of this property is a provenance statement and adding this statement as plain text adheres to the EDM guidelines (Clayphan et al., 2016, p. 19). The following line is an excerpt of how the Rijksmuseum records the provenance statement of the painting *Jeremiah Lamenting the Destruction of Jerusalem* by Rembrandt: “Count Sergei Alexandrovich Stroganoff (1852-1923), St Petersburg and, after 1905, Paris; from whom, frs. 300,000, to Herman Rasch, Stockholm, 1922; from whom, fl. 150,000, to the museum, 1939”. This textual description is more extensive than that of the marital portrait and the cassette of pistols since the painting was acquired during the Second World War. Although the text includes rich information, it is difficult to interpret for machines and for example querying for previous owners is impossible without parsing it.

Embedding temporal information in properties. Temporal information can be embedded in the semantics of properties. The CIDOC-CRM properties *has former or current owner* and *has current owner* create a direct connection from artifact to actor. These properties embed temporal information in the properties using the words *former* and *current*. In the specification of CIDOC-CRM, it is advised to only use these properties whenever the date and place are unknown or only the current owner is known. Embedding more fine-grained temporal information into properties could be achieved by extending properties with an indicator for time, for example by creating the property *has owner in 2017*. This would, however, result in enumerating an impractical amount of properties.

As a result, object-centric approaches tend to describe one particular state of the world. In that state, the object has a specific shape, is owned by someone and is located at a place. Only considering this one state of the world refrains us from asking questions that for example regard changes in shape, previous owners and former locations. To illustrate, most properties of EDM originate from the object-centric Dublin Core data model. EDM has no constructs for capturing part additions or removal. As can be seen in Figure 3, there is the possibility to use the property *is part of* to record that an artifact consists out of other artifacts. This does, however, concern the current state of artifact and not how the artifact changed over time.

Events. Events can be used to record changes over time. CIDOC-CRM includes constructs that allow for fine-grained modeling of provenance and part addition and removal information. As shown in Figure 3, the model includes dedicated classes for part addition and part removal events. Adding a frame to a painting would be modeled using the following path: the frame was added by a part addition which augmented the portrait of Marten. Provenance is modeled using acquisition events. The property *changed ownership through* is used to connect an *acquisition* to the actors involved, with the properties *transferred title from* and *transferred title to*. Using this path, we can for example state that the two pistols changed ownership through an acquisition at the Oude Markt in Brussels in 1815 and thereby transferred title to Jean Sagermans.

CIDOC-CRM also includes shortcuts and extended paths for indicating the keeper of an artifact. After all, the owner and keeper of artifacts are not always the same actor. The acquisition of the marital portraits is an example: owned by the Rothschild family, both the Louvre as well as the Rijksmuseum were interested in acquiring the works. In the end, the paintings were bought with money from the Dutch as well as the French government, who now both own half of each painting. Thus, the portrait of Marten has two owners and its keeper alternates between the Rijksmuseum and the Louvre. As may have become clear from the examples above, events allow for conveying highly detailed temporal information. However, in some cases, this extra detail can make retrieval of information more cumbersome. Accessing the object-centric *has current keeper* property would lead directly to the current keeper of an artifact. At the moment this information would be modeled using events, the keepers would have to be sorted according to date, to obtain the current keeper.

4.4. How to describe representations of an artifact

Representations allow us to consider artifacts without being in the same physical space. A postcard of a statue, a poster of a painting or a recording of a concert all convey something about the represented

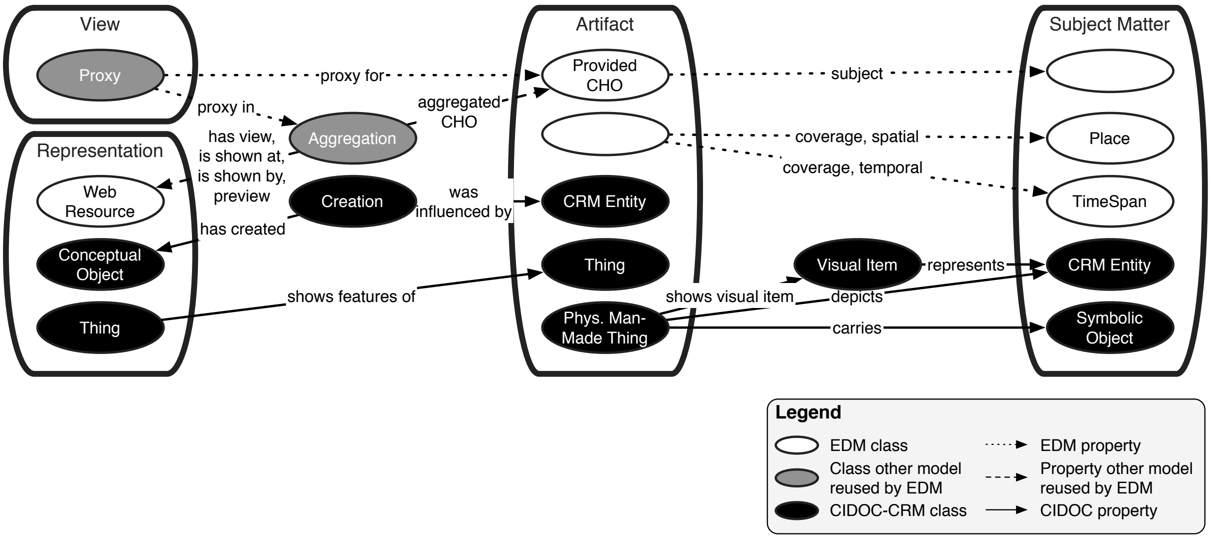


Fig. 4. A representation of three key modeling aspects of publishing cultural heritage data online: subject matter, view and some representation of the artifact.

artifact. It is important to note that by creating a representation, a new entity is created, which differs from the real world artifact. Say a photographer takes a picture of the portrait of Marten. If we would treat the picture as the same entity as the painting, the creator of the painting would be the photographer, as well as Rembrandt. At the moment the difference between representation and artifact is not made explicit in a data model, this either leads to conflicting information or refrains us from describing the representation in more detail.

Representations can take many shapes and forms. A difference, relevant for online publication, is whether a representation is analog or digital. Analog representations are for example posters, postcards and reproductions. To illustrate, a small reproduction of The Night Watch is exhibited next to the original. The original painting by Rembrandt used to be larger, but parts of it were removed in order to fit the city hall of Amsterdam. The reproduction provides insights into how the original composition must have looked like. The Rijksmuseum rarely keeps track of analog representations, as digital representations are more important since they can accompany online descriptions of artifacts.

The range of digital representations includes images, sounds, videos and 3D models. Many types of artifacts can be represented using an image, although for artifacts such as the cassette with pistols, multiple images are required to allow inspection of all sides. Different file encodings can lead to even more representations, for example introducing a lossless and a compressed version. For instance, the Rijksmuseum has 1083 images of the portrait of Marten alone. Many of these are close-ups of details, but also pictures taken with varying equipment, registering different light spectra, such as X-ray and infrared. This multitude of representations makes separate descriptions of representations all the more important.

Aggregations. EDM uses aggregations to connect data about artifacts to digital representations. An aggregation can only be connected to one artifact, but different properties can be used to connect it to multiple digital representations. The most generic property for doing so is *has view*, which does not have a range restriction beyond that the resource should be available on the web. Although this is not formally reflected in the range specification, three more specific properties limit the range of the web resource that functions as view. The range of *is shown by* is limited to digital representations in the best available quality. The property *is shown at* connects the aggregation to a website of the institution at which the artifact is shown. For the portrait of Marten this would be https://www.rijksmuseum.nl/en/collection/SK-A-5033. The range of *preview* is set to thumbnails that represent the artifact. Figure 4 provides an overview of these properties.

Similarity. CIDOC-CRM does not have properties dedicated to connecting artifacts to representations, yet more generic properties can be used to achieve the same effect. As can be seen in Figure 4, the first approach uses a creation event, relating the representation to the artifact using the property *was influenced by*, indicating the resemblance. A more direct connection is created with the use of the property *shows features of*, that indicates that the artifact is similar to the representation. The domain of this property should be the derivative, in this case, the digital representation. The property can be refined by adding the type of similarity. Where EDM properties can exclusively be used to refer to online representations, the properties of CIDOC-CRM can also refer to analog representations.

4.5. How to model multiple metadata sources with alternative views

Metadata is a point of view: it is created by someone who describes an artifact to the best of his or her knowledge, given a certain context. The context can be different, influenced by for example the institution or the intended use of data. This makes the metadata volatile, while the physical artifact that is described is not. These differences are not problematic in an environment that considers just one context, say a collection management system keeping track of the collection items of one museum. But at the moment data is published online, data created in different contexts starts to coexist.

This can lead to situations where data from different sources describes the same physical artifact, with potentially conflicting information. For example, the title used by the Rijksmuseum for the wedding portrait by Rembrandt is “Portret van Marten Soolmans”, while the Louvre uses “Portrait de Maerten Soolmans”. Besides the difference in language, the latter title is the result of the entanglement of names of Oopjen Coppits first and second husband. She married Maarten Soolmans, but after his death, she remarried Maerten Daey. For users it is essential to be aware of the context in which data is created, thereby allowing an informed decision of which information to use. It is, therefore, important that data models support capturing multiple sources describing the same artifact with possibly conflicting information.

Proxies. EDM caters for having multiple views on the same physical artifact using the class *proxy*. As shown in Figure 4, an instance of proxy is connected to two entities: it is a proxy for an aggregation and a proxy for an artifact. The aggregation bundles web resources provided by one of the institutions together. In case of the portrait, the aggregation could bundle the digital images provided by the Rijksmuseum. The artifact that is connected to the proxy is an instance of type *provided cultural heritage object*, represented by an identifier. At the moment proxies are used, data describing the artifact is not connected to the instance of *provided cultural heritage object*. Instead, the data is connected to the proxy instance. The proxy representing the data provided by the Rijksmuseum is for example connected to “Portret van Marten Soolmans”, while the proxy of the Louvre is connected to “Portrait de Maerten Soolmans”.

There are two cases in which proxy constructions are useful outside the context of Europeana. The first regards other aggregators that want to be able to convey different views on the same artifact. The second concerns cultural heritage institutions that are aware that data about an artifact is already ingested by an aggregator and who want to add an additional view. The latter case will be rare, most institutions will not check for the presence of a record. As a result, an institution will not use the proxy construction at all in the data sent to an aggregator. Thus, data ingested by an aggregator needs to be manipulated to cater for multiple views (Isaac, 2013, p. 10). Metadata has to be moved from instances of class *provided cultural heritage object* to entities of class *proxy*.

If two institutions are describing the same physical artifact, the object identifier that the proxy refers to is ideally the same or matched. Most current identifiers redirect to locations specific to one institution. As an example, the Rijksmuseum uses http://hdl.handle.net/10934/RM0001.COLLECT.612987 for the portrait of Marten, which points to a page of the Rijksmuseum website. It is unlikely that the Louvre will use the same identifier. So either alignment techniques have to be in place in order for the proxy construction to be useful, or a move has to be made towards institution agnostic identifiers for artifacts. The Cultural Objects Name Authority[7] which is currently under development by the Getty research institute might provide such identifiers in the future.

4.6. How to contextualize artifacts based on subject matter

Properties of artifacts can be divided into perceptual information about the artifact and information on a conceptual level (Hollink et al., 2004). To illustrate, an eagle has been engraved on one of the pistols of Napoleon. While this visual item is engraved by a person on a physical man-made thing, it represents a species of birds. The portrait by Rembrandt shows the person Marten Soolmans, wearing festive clothing, including a lace collar. The portrait is however not a person nor made of lace. Therefore a clear distinction should be made between perceptual properties of the artifact (e.g. materials, dimensions) and conceptual properties (e.g. who, what, where is depicted).

Many cultural heritage artifacts convey conceptual properties in the form of subject matter: a statue might represent a person, a painting might depict an event and a book might carry a belief. The range of topics is diverse, limited only by the imagination of the creator and the interpretation of the beholder of the artifact. The library field uses subject matter extensively to allow retrieval of relevant books. In contrast, it is not considered part of current museum documentation practices. One of the reasons is the lack of agreement regarding the terminology that should be used (Doerr, 2003).

For improving the accessibility of their online collection, the Rijksmuseum recognizes the value of subject matter descriptions. The museum uses the Iconclass vocabulary[8] to annotate subject matter. This vocabulary only covers a part of the topics encountered on artifacts and usage of additional or more general sources of terminology are considered. Usage of terminology that crosses over institutional boundaries can prove to be a valuable point for integration of online cultural heritage data. We continue by discussing different approaches of relating artifacts to subject matter.

Unbound subject matter. Both data models include a path that allows connecting an artifact to all other entities available, illustrating the vastness of available topics. The most generic EDM subject matter property is *subject*. The range of subject is not defined and therefore all sorts of topics can be related to the artifact. The most elaborate CIDOC-CRM subject matter path connects the artifact using the property *shows visual item* to entities of class *visual item*, which in turn is connected by the property *represents* to the root of the class hierarchy of CIDOC-CRM: *CRM entity*. We can say that one of Napoleons pistols shows the visual item of an eagle, which represents an entity of class *biological object*. The property *depicts* is a shortcut of this path, directly connecting the artifact to an entity, omitting the visual item. By connecting to the most general term, every other class in the CIDOC-CRM hierarchy can be shown on an artifact.

[7] http://www.getty.edu/research/tools/vocabularies/cona/

[8] http://www.iconclass.nl/

Subject matter limited by range. As shown in Figure 4, EDM includes three properties that limit the range of subject matter. The property *coverage* restricts the range to temporal or spatial topics. Coverage could, for example, relate the portrait by Rembrandt to Amsterdam and the second quarter of the 17th century, while subject could also relate the painting to the person Marten Soolmans. Temporal and spatial aspects can be further specified using the properties *spatial* and *temporal*. CIDOC-CRM includes a path that connects artifacts to symbolic objects using the property *carries*. This property can be used to connect non-visual works to symbolic objects, for example stating that a book carries a text. In the next section we distill requirements from the six challenges and approaches discussed in this section. Among these requirements is the possibility to contextualize artifacts, which will become more important with the increasing amount of cultural heritage data published online.

5. Discussion

Based upon the modeling challenges outlined in the previous section, we formulate six requirements for cultural heritage data models. By considering these requirements, institutions can make a more conscious choice when selecting a data model to publish data online. The gathered requirements, in addition to the modeling approaches, are relevant and applicable to many other domains.

(1) Possibility to specialize a data model without decreasing its interoperability
(2) Support for recording both attributes as well as events related to objects
(3) Ability to capture changes over time
(4) Ability to separate descriptions of artifacts and their representations
(5) Support for capturing multiple sources describing the same artifact, with possibly conflicting views
(6) Possibility to contextualize artifacts using subject matter

For domains where interoperability of data is desirable, providing methods that allow integration of data from heterogeneous sources is essential. The cultural heritage domain is diverse, with institutions ranging from archives to museums. Many data models have been created, that either have specific constructs for modeling a particular type of artifact, or have generic constructs for modeling different types of artifacts. A key insight has been that generic and specific modeling approaches can be combined, by creating ontologies which can be specialized **(requirement 1)**. At the moment certain patterns reoccur often, the ontology can be extended accordingly. To illustrate, CIDOC-CRM and EDM have been refined with collection specific constructs (de Boer et al., 2012; Szekely et al., 2013; Dragoni et al., 2016). Providing ontologies with a limited set of constructs, however, does put the responsibility of formulating specialized constructs at the side of institutions, while this might not be their strong suit. A danger is that institutions either solely rely on constructs provided by the ontologies, thereby losing a lot of detail contained within source data, or create flawed new constructs, that introduce inconsistencies.

The information that can be expressed is limited by the approach taken by an ontology. In the cultural heritage domain, the object- and event-centric approaches are common **(requirement 2)**. Data published online is often the result of a conversion from an existing data source, such as a catalog or a collection management system. These sources already take a particular stance. For instance, a collection management system primarily uses attributes to describe artifacts, thereby taking an object-centric approach. Making a transition from an object- to an event-centric approach requires effort and an institution needs to assess whether this is worth the investment. Choosing an event-centric approach provides a more natural way of conveying temporal data **(requirement 3)**. Although events add a layer of complexity, the ability to capture changes over time might be vital for some use cases.

Real-world artifacts can not be transferred over the internet, therefore we have to rely on descriptions of artifacts. A description can take the form of data, describing properties of the artifact. Representations such as images, sounds and videos can provide additional insights into the properties of an artifact. It is, however, essential to realize that a representation is a new artifact, which does not share all properties with the artifact and hence requires a separate description **(requirement 4)**. This subtle distinction becomes more apparent the moment the representation deviates more from the original. A video, like a recording of a painting from different angles, is obviously not a painting. But even "born-digital" artifacts can still have representations, such as other encodings. Both the CIDOC-CRM and EDM models include constructs for indicating how an artifact relates to a representation. These differences will become relevant in each domain that includes representations of real-world objects.

At the moment data is published online, it becomes part of an open world. This open world includes data from many sources, that possibly provide conflicting information. It is up to the user of data to decide which information to use, while understanding that data will never be one hundred percent correct and complete. Making an informed decision is enabled by the availability of the source and provenance of data. Where domain names can provide an indication of the origin of data, this context is lost when statements are made about resources outside the domain of the data publisher. For the aggregator Europeana this problem became important in an early stage since it had to manage data from many different sources, in addition to data resulting from its own enrichment strategies. EDM, therefore, provides constructs that allow for making the source of data explicit and supports having multiple descriptions of one resource **(requirement 5)**. Addressing this problem is relevant in other areas as well, for example in news stories. This will become even more important at the moment we move closer to making the internet one big data space.

Currently, the Rijksmuseum publishes Linked Data about over 350,000 objects, structured using a combination of EDM and Dublin Core. With the increase of available data, the need for contextualization rises. An important aspect of contextualization is subject matter **(requirement 6)**. Subject matter can be very diverse and often differs from the domain that cultural heritage ontologies intend to capture. Cultural heritage ontologies should, therefore, allow relating artifacts to contextual entities, even when these entities are structured differently.

6. Conclusion and future work

The cultural heritage domain is among the first domains to embrace the Semantic Web and now features mature ontologies that can be used to structure data. In Section 4 we discussed modeling challenges regarding specialization, object- and event-centric approaches, temporal changes, representations, multiple views and subject matter. Considering these challenges and abstracted requirements can help cultural heritage institutions to make an informed decision about the ramifications of choosing a particular ontology. The modeling practices in addition to the gathered requirements for ontologies are relevant and applicable to many other domains.

The Rijksmuseum currently publishes information about collection objects using a combination of the EDM and Dublin Core data models. Using an event-centric model instead of the current object-centric model, would overcome modeling limitations regarding changes over time and the recording of different roles of actors involved in a creation process. Although a new approach, that uses a combination of the EDM and CIDOC-CRM data models, requires a signification mapping effort, it would address all the requirements discussed in this paper. Additionally, the aim of the museum is to extend the data beyond the scope of collection management, by contextualizing objects with internal sources. These sources include bibliographic information from the art-historical library, documentation contained in the archives and research data. To convey this information adequately, ontologies from domains other than the museum sector will have to be considered.

Top-level and reference ontologies provide generic constructs, which can be refined by others. Requiring institutes to create their own specific extensions is error-prone and makes it more difficult to later align specific constructs. Creating standardized extensions for different domains and types of artifacts might help harmonize modeling efforts. This approach can already be observed within the CIDOC-CRM and EDM communities. Extensions and application profiles include models for ancient texts, fashion, archeology and scientific observations. Extending this list of topics will allow institutions to more easily publish interoperable, but detailed information.

Online publication of cultural heritage data enables the usage of data created outside the context of an institution. The increased use of controlled vocabularies in the cultural heritage sector is a first indicator of a more widespread acceptance of the usefulness of data created by others. Cross-institutional interlinking could greatly enhance the user experience, enabling a more thorough overview of the different facets of cultural heritage. At the moment more institutions are able to publish data on their own, aggregators could serve as discovery points for potential links. Increased usage of data from different sources will require data consumers to consider automated methods to validate data and assess trust in the obtained information.

The increasing amount of cultural heritage data published online will lead to new challenges. A major challenge will be ranking artifacts to adequately respond to information needs. Curators, librarians and archivist might have a natural feeling for doing this, but the required information is not always available online and with the rising number of available artifacts the need for contextualization will only grow. In the museum sector, recorded curation activities can serve as an additional source of context. However, this does not allow us to show, for example, the masterpiece of each artist in a collection. While we can provide ratings for many resources online, ranging from hotels to movies, this is rarely possible for cultural heritage artifacts. This type of subjective data is something not readily considered by cultural heritage institutions, although it might greatly improve the accessibility of information.

Acknowledgements

From the ResearchSpace team, we thank Dominic Oldman and Barry Norton for sharing valuable insights in the CIDOC-Conceptual Reference Model. We thank Lizzy Jongma and Trineke Kamerling from the Rijksmuseum for their support. Comments from Stefan Schlobach and Rinke Hoekstra helped shape this paper. We thank Eetu Mäkelä, Emmanuel Papadakis and Karl Grossner for their extremely useful reviews. This publication is supported by the Dutch national program COMMIT/ and the VRE4EIC project, funded from H2020 grant No. 676247.

References

Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G. & Summers, E. (2013). Key choices in the design of Simple Knowledge Organization System (SKOS). *Web Semantics: Science, Services and Agents on the World Wide Web*, *20*, 35–49. doi:http://dx.doi.org/10.1016/j.websem.2013.05.001.

Bizer, C., Heath, T. & Berners-Lee, T. (2011). Linked Data: the story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts* (pp. 205–227). IGI Global. Chapter 8. doi:10.4018/978-1-60960-593-3.ch008.

Clayphan, R., Charles, V. & Isaac, A. (2016). *Europeana Data Model – Mapping Guidelines* (2.3 ed.). Europeana.

Crofts, N., Doerr, M., Gill, T., Stead, S. & Stiff, M. (2011). *Definition of the CIDOC Conceptual Reference Model* (5.0.4 ed.). CIDOC CRM Special Interest Group.

de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J. & Schreiber, G. (2012). Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho and V. Presutti (Eds.), *Proceedings of the 9th Extended Semantic Web Conference*. ESWC '12 (Vol. 7295, pp. 733–747). Springer Berlin. doi:10.1007/978-3-642-30284-8_56.

Dijkshoorn, C., Jongma, L., Aroyo, L., Van Ossenbruggen, J., Schreiber, G., ter Weele, W. & Wielemaker, J. (2018). The Rijksmuseum collection as Linked Data. *Semantic Web Journal*, *9*(2), 221–230. doi:10.3233/SW-170257.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, *24*(3), 75–92.

Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C. & van de Sompel, H. (2010). The Europeana Data Model (EDM). In *World Library and Information Congress: 76th IFLA general conference and assembly* (pp. 10–15).

Dragoni, M., Cabrio, E., Tonelli, S. & Villata, S. (2016). Enriching a Small Artwork Collection Through Semantic Linking. In H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S.P. Ponzetto and C. Lange (Eds.), *Proceedings of the 13th Extended Semantic Web Conference*. ESWC '16 (pp. 724–740). Springer International Publishing. doi:10.1007/978-3-319-34129-3_44.

Dunning, A. & Verspille, I. (2017). The Current Europeana Dataset. http://research.europeana.eu/about-our-data/the-current-europeana-dataset.

Guarino, N., Oberle, D. & Staab, S. (2009). What Is an Ontology? In *Handbook on Ontologies* (pp. 1–17). Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-92673-3_0.

Hollink, L., Schreiber, G., Wielinga, B. & Worring, M. (2004). Classification of user image descriptions. *International Journal of Human-Computer Studies*, *61*(5), 601–626. doi:10.1016/j.ijhcs.2004.03.002.

Hooland, S.v. & Verborgh, R. (2014). *Linked data for libraries, archives and museums*. Facet Publishing.

Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M. & Kettula, S. (2005). MuseumFinland – Finnish museums on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *3*(2-3), 224–241. doi:http://dx.doi.org/10.1016/j.websem.2005.05.008.

Isaac, A. (2013). *Europeana Data Model Primer*. Europeana.

Isaac, A. (2014). *Definition of the Europeana Data Model* (5.2.6 ed.). Europeana.

Janowicz, K., Hitzler, P., Adams, B., Kolas, D. & Vardeman II, C. (2014). Five stars of Linked Data vocabulary use. *Semantic Web Journal*, *5*(3), 173–176. doi:10.3233/SW-140135.

Knoblock, C.A., Szekely, P., Fink, E., Degler, D., Newbury, D., Sanderson, R., Blanch, K., Snyder, S., Chheda, N., Jain, N., Raju Krishna, R., Begur Sreekanth, N. & Yao, Y. (2017). Lessons Learned in Building Linked Data for the American Art Collaborative. In C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange and J. Heflin (Eds.), *Proceedings of the 16th International Semantic Web Conference*. ISWC '17 (pp. 263–279). Cham: Springer. doi:10.1007/978-3-319-68204-4_26.

Lagoze, C. & Hunter, J. (2006). The ABC Ontology and Model. *Journal of Digital Information*, *2*(2).

Matsumura, F., Kobayashi, I., Kato, F., Kamura, T., Ohmukai, I. & Takeda, H. (2012). Producing and Consuming Linked Open Data on Art with a Local Community. In J. Sequeda, A. Harth and O. Hartig (Eds.), *Proceedings of the 3rd International Workshop on Consuming Linked Data*. CEUR Workshop Proceedings. CEUR-WS.org.

Mouromtsev, D., Haase, P., Cherny, E., Pavlov, D., Andreev, A. & Spiridonova, A. (2015). Towards the Russian Linked Culture Cloud: Data Enrichment and Publishing. In F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux and A. Zimmermann (Eds.), *Proceedings of the 12th Extended Semantic Web Conference*. ESWC '15 (pp. 637–651). Cham: Springer International Publishing. doi:10.1007/978-3-319-18818-8_39.

Obrst, L., Gruninger, M., Baclawski, K., Bennett, M., Brickley, D., Berg-Cross, G., Hitzler, P., Janowicz, K., Kapp, C., Kutz, O., et al. (2014). Semantic Web and Big Data meets Applied Ontology. *Applied Ontology*, *9*(2), 155–170. doi:10.3233/AO-140135.

Studer, R., Benjamins, R. & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1), 161–197. doi:10.1016/S0169-023X(97)00056-6.

Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R. & Goodlander, G. (2013). Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink and S. Rudolph (Eds.), *Proceedings of the 10th Extended Semantic Web Conference*. ESWC '13 (Vol. 7882, pp. 593–607). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-38288-8_40.

Valentine, C. & Isaac, A. (2015). Enhancing the Europeana Data Model (EDM). *White Paper*.

Wielinga, B.J., Schreiber, A.T., Wielemaker, J. & Sandberg, J. (2001). From thesaurus to ontology. In *Proceedings of the 1st international Conference on Knowledge Capture*. K-CAP '01 (pp. 194–201). New York, NY, USA. ACM. doi:10.1145/500737.500767.