An RDF/OWL2 model to represent complex, annotated and temporally limited relations between concepts in a cultural heritage context

Marielle St-Germain [a,*,c], Ludovic Font [b,c], Alexandre Naud [c] and Michel Gagnon [b,c]
[a] *EBSI, Université de Montréal, QC, Canada*
E-mail: marielle.st-germain@umontreal.ca
[b] *Département de Génie Informatique, École Polytechnique de Montréal, QC, Canada*
E-mails: ludovic.font@polymtl.ca, michel.gagnon@polymtl.ca
[c] *Ministère de la Culture et des Communications du Québec, QC, Canada*
E-mails: marielle.st-germain@umontreal.ca, ludovic.font@polymtl.ca, alexandre.naud.1@ulaval.ca, michel.gagnon@polymtl.ca

Abstract. When representing cultural knowledge, the limitations of usual RDF models quickly arise. Indeed, metadata must often be added to triples, in order to convey provenance and temporal information, among others. Furthermore, real-world knowledge is often complex and requires additional information, such as the representation of a choice between different possibilities. Several individual solutions exist, such as the CIDOC Conceptual Reference Model for cultural heritage in general, the Wikidata model for the inclusion of metadata on triples, and the OWL-Time ontology for a fine representation of temporal relations. By combining them, we created a model that allows the representation of some of the most complex aspects of cultural knowledge, i.e. that makes it possible to add provenance, bibliographic and temporal information not only to any concept, but also to the relations between objects. Even though the actual scope of the project is limited to objects and materials that composed them, we are convinced that the resulting model could be useful to represent any aspect of cultural knowledge in general.

Keywords: ontology, metadata, temporal information, cultural heritage, linked data

1. Introduction

1.1. Context

The one thing one is never short of in the cultural world is data. By combining the knowledge bases of of the cultural sectors, the resulting amount of available data is staggering. Therefore, many projects have arisen in the last years to organize, classify and provide an easy access to such a vast knowledge. Since 2014, The Plan culturel numérique du Québec[1], piloted by the Ministère de Culture et des Communications, helps cultural actors to position themselves in the digital world, through more than 120 different innovative measures. One of these aims to enhance Quebec's digital platform for cultural heritage, the Répertoire du patrimoine culturel du Québec[2], through data aggregation and integration from multiple providers[3]. This project acted as a starting point of another endeavour,

*Corresponding author. E-mail: marielle.st-germain@umontreal.ca.

[1] http://culturenumerique.mcc.gouv.qc.ca/
[2] http://www.patrimoine-culturel.gouv.qc.ca
[3] http://culturenumerique.mcc.gouv.qc.ca/faire-du-repertoire-du-patrimoine-culturel-du-quebec-rpcq-la-plateforme-commune-de-

the development of an ontology to represent Quebec's cultural heritage[4]. Part of a series of actions aiming at combining different perspectives from the cultural heritage domain, it is realized in cooperation with different cultural actors, as one of the goal is to break down data silos and achieve interoperability. An analysis of vocabularies and thesauri used in various cultural heritage disciplines and specialties was conducted, more specifically on materials, their characteristics and how they are represented and used in relational databases as well as in the literature on the domain. The idea was to really make explicit the meaning behind each element of the vocabularies used by the different actors. The purpose was not to prioritize one classification over another. By describing their characteristics, the goal was rather to see all of them coexist. Noting the value and the potential behind these data, the Ministère de la Culture et des Communications decided to work on a proof of concept, founded on a collaboration with Polytechnique Montréal[5], a consultant in information science and a consultant in the cultural heritage domain. The main objective was to develop a model to represent this subset, based on linked data technologies.

During its course, several major conception decisions were made, because a large part of the available data was too complex to be easily represented in RDF/OWL. Among them, two considerations in particular led to the creation of a non-trivial RDF/OWL model: 1) the representation of complex relations, such as relations between more than two elements (where the relation can be annotated), or that come with a source information; 2) the universal and imprecise temporality characterized by fuzziness, meaning that most instances and relations exist only during a determined timeframe, that is rarely known precisely.

In this paper, our goal is to present the solution we chose to address these difficulties, hoping that the lessons we learned during this process will be useful for others facing similar difficulties in the particular context of dealing with cultural heritage information and its uniqueness. The motivation behind its development and examples are presented in Section 2. The core of our resulting model is explained in Section 3, whereas the specific temporal aspect is detailed in Section 4. In Section 5, we present an example of the representation of complex information available on a material in our model. The various existing possibilities that were envisioned during the process are presented in Section 6. Finally, a conclusion and short discussion is offered in Section 7.

diffusion-et-de-mise-en-valeur-du-patrimoine-numerise-quebecois-2/

[4]http://culturenumerique.mcc.gouv.qc.ca/elaborer-une-ontologie-du-patrimoine-culturel-quebecois/

[5]https://www.polymtl.ca/

1.2. The dataset

In this project, the goal was to create a model to represent how materials are defined. By "materials", we mean the constituents of which something is made or composed. In other words, it must be able to answer questions such as: what are the characteristics of sedimentary rocks? What differentiates soft steel and stainless steel? If a material has a red-brown oxidation color, what could it be besides iron?

To build this model, we rely on a dataset containing 1 439 materials and 71 material characteristics such as the chemical state, the transparency, or the color. The value of these characteristics can be quite simple, but also more complicated, for instance when the color comes from a specific chemical element, or when the material is composed of several other materials in specific, but not always precise, proportions. Furthermore, these characteristics can be a defining feature of the material (steel *must contain* iron by definition), an alternative (ceramic can be *either translucent or opaque*, but never transparent), or even a possibility (aluminum *sometimes* has a polished coating).

To the best of our knowledge, there exists no ontology allowing the representation of this kind of information. Therefore, since the ultimate goal of our model is to allow for a precise description of cultural heritage resources, we chose to base it on the CIDOC Conceptual Reference Model (CRM) (ISO 21127). The CIDOC CRM provides a formal structure as well as definitions useful to describe concepts used in the cultural heritage documentation and the relationships that unite them [1]. Even though its representation of materials is barebones, it provides a framework that interlinks most general concepts that are used in our model, such as *E21 Person*, *E5 Event* or *E53 Place*.

2. Modeling issues

Even though it is perfectly suited for cultural heritage data, a crucial aspect that is not well handled by the CIDOC CRM is the fact that our model represents a set of definitions, and not statements on real things. For instance, the triple *[Steel, contains, Iron]* represents a definition of what is the concept of *Steel*. Using our model, a knowledge base could define that *[Eiffel Tower, madeOf, Steel]*, where the object *Steel* does not represent the actual, tangible steel that the Eiffel Tower is made of, but instead the general concept of steel, defined in our model by a set of triples. Among these, there is the information that steel can contain chromium, nickel, lead, or tin, even though the steel of the Eiffel Tower may or may not contain any of these elements. Representing the information about the specific composition of the steel of the Eiffel Tower would require a different model that is out of the scope of this project.

This particularity creates its share of challenges. Among these, the aforementioned distinction between a mandatory and optional feature of a material is an information we definitely need to represent in the model. For instance, the triple *[Steel, contains, Iron]* is mandatory for the definition of steel, whereas *[Steel, contains, Lead]* is optional, as some kinds of steel contain lead, but others do not. The first triple is obviously needed in the model in order to provide an accurate depiction of what are the characteristics of steel, and the second one, even though optional, also provides an important information: if an unknown material contains lead, then it could be steel. The exploration of the various materials that can contain lead, whether the inclusion is mandatory or optional, allows the user attempting to identify a material to narrow down the list of possible candidates. We refer to this distinction as the "qualification" of the relation.

As we mentioned in Section1.2, a third possibility also occurs, when several options are available, but are mutually exclusive. For example, any type of ceramic can be translucent or opaque, but not both (by definition of translucence and opacity), and not transparent. Here, the two triples *[Ceramic, levelOfTransparency, Translucent]* and *[Ceramic, levelOfTransparency, Opaque]* are the two possibilities. Any material that enters in the ceramic family must be either one or the other. Because we model the definition of ceramic in general, both possibilities must be considered. Our model does not represent the situation "This ceramic is translucent, and that other ceramic is opaque", but the situation "Ceramic in general, by definition, can be either translucent or opaque". Therefore, the need arises to link these two triples, to represent that they are both part of the definition of ceramic as two possible, mutually exclusive, alternatives. It can be tempting to use this model to indicate that *[Ceramic, levelOfTransparency, Transparent]* is forbidden. However, even though in this example there are only three possible states (opaque, translucent and transparent), in most cases, there are many possible states, such as when indicating the cristalline structure of a material. In such cases, it would be cumbersome to represent each invalid state, along with each valid one, by a triple. Therefore, in our model, the fact that ceramic cannot be transparent is inferred because the only possible states that appear in the data are *translucent* and *opaque*.

Another major point we encountered is the notion of source. Indeed, the definition of a material can vary. Even though *Steel* is quite universally defined, for more specific materials, such as *North Italian style earthenware*, the definition is often variable from one archaeologist to another. Therefore, it becomes important to be able to indicate the source of any information. For instance, in our data, *North Italian style earthenware* is characterized by the presence of a brown glaze. However, it is possible that, in another dataset, this earthenware is characterized by the presence of a glaze that can be either brown or red. We must therefore be able to indicate the source of these definitions.

Finally, it is also often necessary to represent temporal information. Indeed, the definition of a material can change, and it is important for the archaeologist to know what was the definition of a material at a specific point in time. For instance, the first reported creation of lead-crystal glass was around 1670, by Ravenscroft [2]. However, it took until the middle of the eighteenth century for glass-makers to discover the optimal recipe, that corresponds to the current definition of lead-crystal glass. Therefore, the definition of what is the material "lead-crystal glass" has changed at a point in time, and our model must be able to represent that. The CIDOC CRM offers a way to represent temporality based on the notion of event. However, this model is insufficient to represent the temporality of a triple, for example the fact that the triple *[Lead-crystal glass, containsAgent, Lead oxide]* is true only from around 1750. Before that, a glass could be considered as lead-crystal even if it did not contain lead oxide. Furthermore, when dealing with historical knowledge, temporal information is far from perfectly accurate. In the previous example, "around 1750" is definitely not an accurate date. It then becomes necessary to represent imprecise temporal data, such as the notions of "around", "before", "after" and "between". Therefore, it is necessary to expand the notion of temporality from

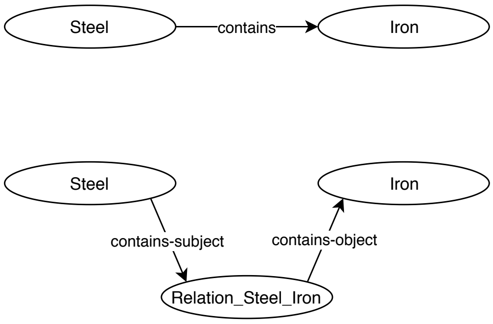

Fig. 1. Representation of a standard RDF triple in the Wikidata model

what is allowed by the CIDOC CRM to represent our data.

To summarize, the model must provide: 1) the qualification of relations, i.e. the distinction between "optional", "mandatory" and "one choice mandatory"; 2) the indication of the source of a triple; and 3) the ability to place in the dimension of time any element of the dataset, including triples.

3. The instantiated relation model

3.1. The Wikidata approach

All our needs require the addition of metadata to statements (triples). A good approach for this is the one used on Wikidata [3], where each statement can also be annotated with various information, such as temporal data or source. In the RDF representation of the Wikidata model, each statement is represented by two triples. For example, the fact that steel contains iron is represented by the two triples : *[Steel, contains-subject, Relation Steel Iron]* and *[Relation Steel Iron contains-object, Iron]*, as shown in Figure 1. The central node, *Relation Steel Iron*, can then be annotated with generic metadata (source, date), or more specific metadata (such as the minimal proportion of iron that must be present in steel). In this model, we call the central node *Relation Steel Iron* an instantiated relation.

Because each instantiated relation is represented by a node in the graph, it becomes quite straightforward to add all the needed information. Indicating the source of a triple only requires the addition of a property *documentedIn* (see Figure 2). The same approach is also used to provide temporal information on an object (this object was created at this date) and on a statement (this statement became true at this date). The exact way we handle temporality is detailed in Section 4. Finally, in-

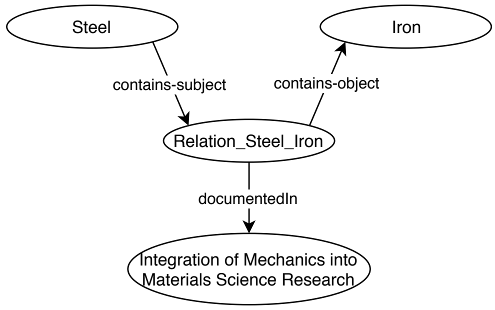

Fig. 2. Indication of the source of a relation between two objects

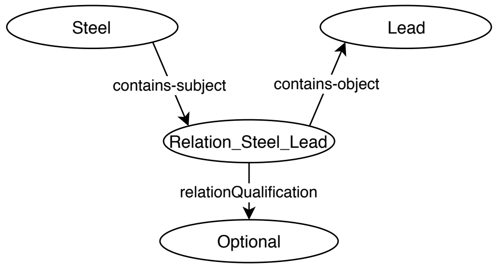

Fig. 3. Representation of an optional relation

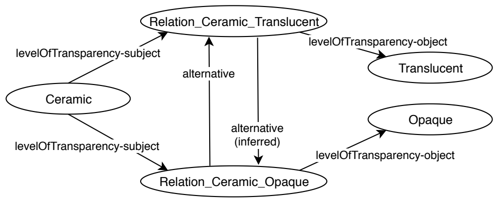

Fig. 4. Representation of two alternative relation

dicating if a relation is mandatory or optional can be done with a simple property *relationQualification* that can take one of the two values *Mandatory* or *Optional* (see Figure 3). The representation of a choice between relations is a bit more tricky, since it requires the creation of a transitive and symmetric property *alternative* that will link all the possible choices for a given relation. For instance, see Figure 4. In this example, since the relation *alternative* is symmetric, only one has to be provided, and the second can be inferred.

3.2. Property chains

Even though the Wikidata approach has several major advantages, its direct application, by instantiating relations, has a major drawback. In the previous example, we may want to represent that the property *contains* is a subproperty of *P46 is composed of* (actually, P46 has a slightly different meaning, but we will

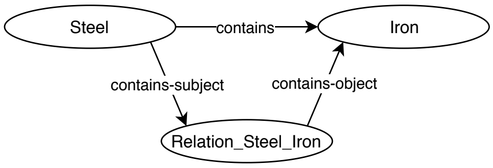

Fig. 5. Representation of a triple in our model, including the instantiated relation

```
select ?material where {
<Steel> <contains> ?material.
}
```

Fig. 6. A SPARQL query to obtain all materials that can be present in steel

consider that it is not the case for the sake of the example). However, in the Wikidata model, *contains* exists, but has no semantic connection to the two properties *contains-subject* and *contains-object*, and neither of those is a subproperty of *P46*. We, however, want to represent semantic relation between the **chaining** of *contains-subject* and *contains-object*, and the direct property *contains*, and therefore its subproperty in the CIDOC CRM, *P46*.

Therefore, it seems natural to use the notion of property chains that appeared in OWL2. This allows to indicate that one property (*contains*) is a subproperty of the chaining of two or more other properties (*contains-subject o contains-object*). Then, in the situation described in the first paragraph of this section, the triple *[Steel, contains, Iron]* can be automatically inferred. Then, we can indicate that *contains* is a subproperty of *P46 is composed of*, and still have the metadata on the statement. The resulting graph is displayed in Figure 5.

The usage of property chains presents another advantage. Indeed, two layers of complexity emerge, where the "simple" data is still present, but with additional metadata if needed. This allows to represent complex data, but at the same time to cater to non-expert users. If an archaeologist not formed in SPARQL just wants to know what materials are contained in steel, without any metadata, then the query will be very simple, as in Figure 6. If, however, someone requires additional information about these statements, then he can query the instantiated relation and all relevant metadata, as in Figure 7. It should be noted that the correct reasoning on property chains requires the usage of an OWL2-compliant reasoner.

```
select ?material where {
<Steel> <contains-subject> ?relation.
?relation <contains-object> ?material.
?relation <relationQualification> <Mandatory>.
}
```

Fig. 7. A SPARQL query to obtain all materials that must be present in steel

The main disadvantage in the usage of property chains is the limitations it puts on OWL restrictions. Indeed, in the OWL2 specification, a property defined as a property chain is not **simple**, and, therefore, cannot have a restricted cardinality (min, max, exact), cannot be functional, inverse functional, irreflexive, asymmetric or disjoint. This is done to guarantee decidability of the basic reasoning problems[6]. Therefore, the expression of cardinality restrictions is slightly more complicated when using property chains. For instance, to represent that an alloy must contain at least one metal, we must indicate that each alloy must be connected by the property *contains-subject* to at least one *Relation*, and that *Relation* must itself be connected by the property *contains-object* to exactly one *Metal*.

4. Temporality

4.1. Context

In the cultural heritage discourse, the concept of time and temporality (e.g., historical events, periods) is often central and challenges surrounding its representation and its characteristics are numerous. One of the specificities of temporal cultural data is the presence of fuzzy boundaries and exact boundaries of time intervals (e.g., somewhere between 1920 and 1932, during 10 to 50 years) [4]. Furthermore, as pointed out by [5], past events or facts are represented with different wordings close to, or deriving from, natural language (e.g., before the second half of the 15th century, at the beginning of the year 1805). This reality is problematic when using management systems requiring the use of specific date formats like ISO 8901, for instance. Typically, different types of metadata (e.g., descriptive, administrative) are provided to describe cultural and historical information in collection management systems

[6]https://www.w3.org/TR/owl2-syntax/, section 11.2

and we can find precise temporal information as well as vague and unknown information.

Increasingly, during the last years, we noted a certain shift in the cultural heritage domain, more particularly in the museum context, from item-centric cataloguing to event-centric cataloguing of artifacts [6][7]. The CIDOC CRM is a good example of that principle because it revolves around the primitive concept of event, as a semantically meaningful way of representing the links between *things* and *actions* and their impact. This reality gives more importance to the choices we make when modelling temporality.

4.2. OWL-Time vs CIDOC CRM

To implement this reality, we explored two approaches. The first one is to use the temporality as represented in the CIDOC CRM, and the second is to use the OWL-Time ontology, recommended by the W3C [7]. Both are based on Allen's interval algebra to represent relations between temporal intervals [8]. This algebra allows a powerful and versatile representation of intervals. It becomes possible to represent fuzzy relations. For instance, in 1805, a new type of coating for ceramic, the Albany slipware, started being used. "In 1805" is already fuzzy, at we don't know the exact day or month, only the year. This is quite easy to represent using Allen's algebra, by indicating that the first use of the Albany slipware is an interval that is entirely contained (*time:intervalDuring*) in the interval representing the year 1805. Furthermore, this coating gradually stopped being used, and it is therefore difficult to pinpoint exactly the date at which it completely stopped being produced. We only know that it happened at some point between 1925 and 1975. This is a bit harder, but still possible to represent using Allen's relations. The last use of the Albany slipware could be represented by an interval that is contained in another interval, representing the period 1925 to 1975. This second interval itself begins in 1925, and ends in 1975. This situation is represented in Figure 8. In this figure, the representation of a year must be done by two instances: a temporal interval, such as *1925 as interval*, and an instance of *time:DateTimeDescription*, such as *1925 description*, that contains all the information on the date 1925 itself, i.e. that it corresponds to 1925 AD in the Gregorian calendar.

With the usage of these properties on time intervals, it is also possible to represent even less precise information. For example, "before the second half of the 15th century" can be represented, by indicating that the interval started at some point during another interval, and that second interval has no beginning, but only an end in 1450. Finally, it is possible to represent "around 1500" with no further information on the accuracy, by indicating that the interval contains the year 1500. Furthermore, we created three properties, two that allow the direct connection of an interval to its begining and ending description, as in Figure 8 (*intervalStartsInDateTime* and *intervalFinishesInDateTime*), and one that indicates that an interval approximates a certain date, *intervalApproximatesDateTime*. Without entering in too much detail, these three properties allow to obtain all temporal information on a topic with only one SPARQL query, without knowing in advance if the information is fuzzy or uses intervals. Overall, the usage of Allen's algebra allows to represent in a mathematically accurate way the fuzziness born from incomplete information.

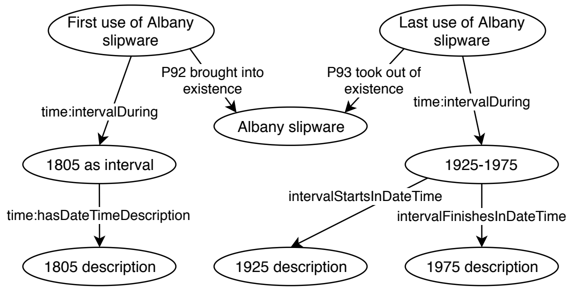

Fig. 8. Representation of fuzzy temporal data

As mentioned, both the CIDOC CRM and the OWL-Time ontology allow a representation of the Allen's relations. The main difference, for our usage, is that OWL-Time provides a detailed representation of actual dates. It is possible to have a date only accurate to a year (for instance, "in 1990"), to a month, or to a day, and to indicate a time zone. In this domain, the CIDOC CRM only provides *E61 Time Primitive* [8], that is described as "This class comprises instances of E59 Primitive Value for time that should be implemented with appropriate validation, precision and references to temporal coordinate systems to express time in some context relevant to cultural and scientific documentation." In other words, it requires the user to implement their own representation of what is a "date". Therefore, instead of re-implementing what is already

[7]https://www.w3.org/TR/owl-time/

[8]http://www.cidoc-crm.org/Entity/e61-time-primitive/version-6.2.2

in the OWL-Time ontology, we chose to represent our temporal data by importing it into our model. Then, it seemed simpler to use the whole OWL-Time ontology, and as a consequence to bypass the CIDOC representation of temporal intervals. Both models, being based on Allen's relations, are equivalent for our needs.

However, what the CIDOC CRM offers that OWL-Time has not is the notion of event[9]. Indeed, the CIDOC CRM represents temporal information by considering that, when something has a start or an end at some point in time, these can be represented by events, modeling the beginning and end of existence of that thing. This is interesting, since it allows to provide more information on the circumstances that led to the apparition or disparition of things. For instance, the event representing the apparition of lithyalin glass is an invention event, also indicating that it was invented by Friedrich Egermann.

To link the notion of event as represented by the CIDOC CRM and of time interval as represented by OWL-Time, we considered *E4 Period*, the parent class of *E5 Event*, to be a subclass of *time:Interval*. Since *E4 Period* is already a child of *E2 Temporal Entity*, which is quite similar in many aspects to *time:Interval*, this decision was the most logical way to link the two. Then, we can directly indicate that the event "invention of lithyalin glass" occurred at some point during the year 1816, represented by using the description provided by the OWL-Time ontology.

Another element that is important to mention is that the CIDOC CRM proposes to model reality by observing the traces on material evidence about past events or periods. In the context of this project, the problem with that focus is the fact that, first, we are trying to represent materials as abstract concepts and, second, as [5] underline, physical traces on heritage artifacts are not always present, more particularly in the case of intangible heritage.

4.3. Inclusion to the instantiated relation model

As we explained in the previous section, combining the CIDOC CRM and OWL-Time representations allows a fine-tuned representation of temporal information on any object. Furthermore, when we consider that, in our model, relations are also objects (as explained in Section 3), we can now indicate that a **relation** between two objects is itself only true during

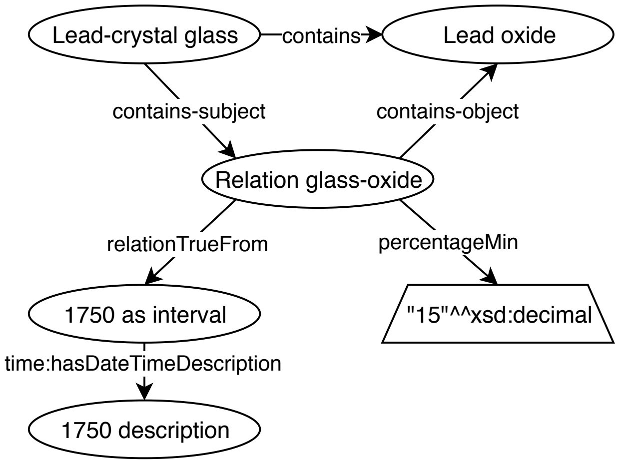

Fig. 9. Representation of the temporality of a relation without an event

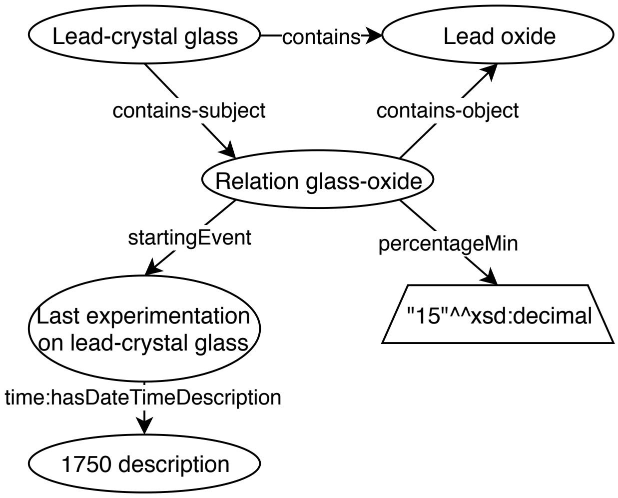

Fig. 10. Representation of the temporality of a relation with an event

a certain period. If we go back to the example given in Section 2, where the craftsmen determined only in 1750 that lead-crystal glass must contain at the very least 15 percent of lead oxide, we can represent this situation as shown in Figure 9. Here, the properties *relationTrueFrom* and *relationTrueUntil* (not used in this example), indicate that a relation started and finished, respectively, being true at a given point in time.

An important point to note here is that, unlike previously, there is no notion of event to represent the temporality of this relation. This is because, in most cases, the only information we have about the beginning of a relation is its date, and, in this context, it seems artificial to create an empty event to provide this temporal information. However, in some cases, we

[9] http://www.cidoc-crm.org/Entity/e5-event/version-6.2

do know more. In the previous example, the new definition of lead-crystal glass (as containing at least 15 percent of lead oxide) occurred as the result of several experimentations. An alternative representation of this sitution can therefore be created, using an event, as shown in Figure 10. The relations *startingEvent* and *finishingEvent* are the counterpart of *relationTrueFrom* and *relationTrueUntil*, when the relation is started or ended by an explicit event instead of a simple date.

One should note that, because *E5 Event* is a subclass of *time:Interval*, and *startingEvent* is a subproperty of *relationTrueFrom*, the first representation is a logical consequence of the second and, therefore, one SPARQL query can cover both cases. This is another example of a situation where it is possible to simply represent a situation, but, if the need arises, the model is flexible enough to allow for more detailed information.

5. Exemple of representation of a complex situation

To demonstrate in more details how the model developed in this project can be applied to a real use-case, we provide in this section an example aiming at demonstrating how the model would solve the typical problems related to data in the cultural heritage context.

One of the materials in our data is the North Devon sgraffito ware. Some of its characteristics, as available in our data, are provided in Table 1. As one can see, this data comes from two sources, Brassard and Leclerc (2001) in [9] and Métreau (2016) in [10]. Brassard is coming from a North American background, while Métreau's background is European. Therefore, the decriptions they offer of the same material are sometimes different. However, we still have the need to represent both descriptions despite the contrast.

This small sample of data on the North Devon sgraffito ware provides an example of each of the issues we attempted to solve with our model: the source of an information, and what happens when there are several sources; temporal data; and qualification of relations (with the fact that, according to Brassard, the pattern can be either geometric, or floral).

Figure 11 illustrates the value of the color of the North Devon sgraffito ware according to, on the left, Brassard and, on the right, Métreau. It is possible to represent not only who stated this information, but also in which documentation (in this case, *Identifier la céramique et le verre anciens au Québec* and *Identifier la céramique au Québec*). This is particularly important in the context of this project, but also in the broader context of cultural heritage and digital humanities in general. It is common to encounter different views on the same object and it is essential to be able to represent this reality, not only to be as thorough as possible, but also to achieve a certain dissemination and preservation of information.

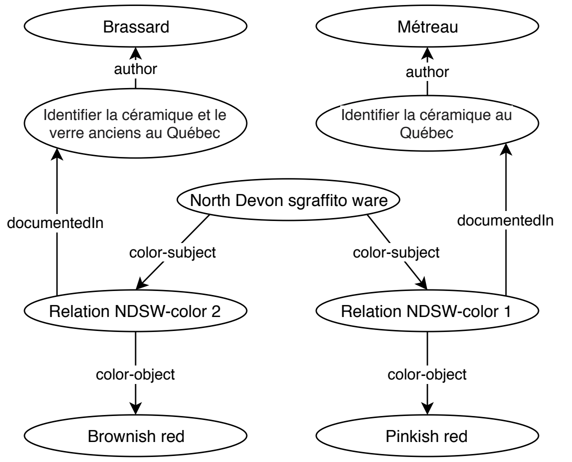

Fig. 11. Representation of the color of North Devon sgraffito ware according to two sources

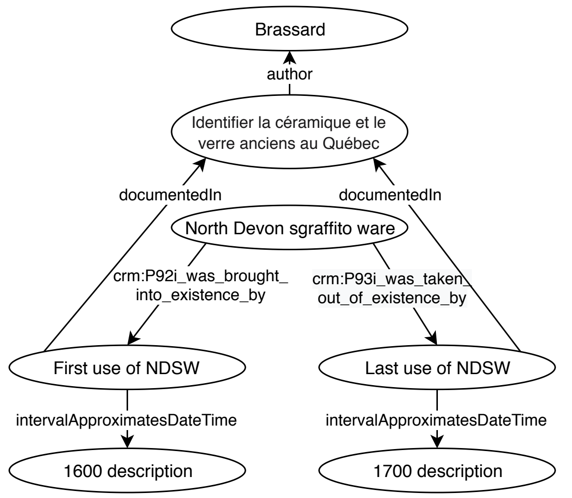

Fig. 12. Representation of the temporality of North Devon sgraffito ware according to Brassard

In Figures 12 and 13, we represented the temporal information available on North Devon sgraffito ware according to the two sources. In this case, the sources provide complementary information, one on the time during which it was produced, and the second on its

Table 1
A sample of the available data on North Devon sgraffito ware

Attributes	Value according to Brassard and Leclerc	Value according to Métreau
Color	Pinkish red	Brownish red
Datation	17th century, stopped being produced in 1700	Golden age at the end of 17th century
Decoration type	Sgraffito	Sgraffito
Decoration pattern	Geometric or floral	N/A

Fig. 13. Representation of the temporality of North Devon sgraffito ware according to Brassardt

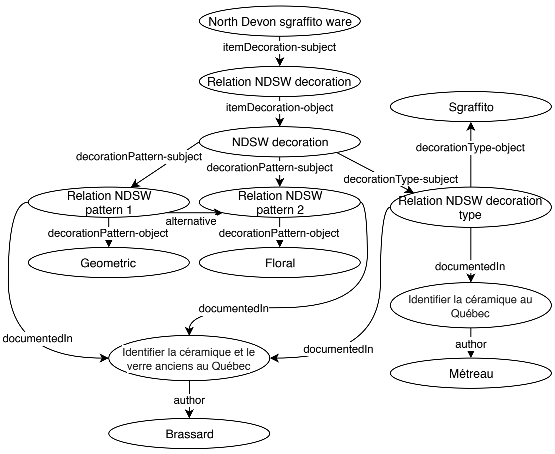

Fig. 14. Representation of the data on the decoration of North Devon sgraffito ware

golden age. In Figure 12, we used the predicate *intervalApproximatesDateTime*, as was introduced in Section 4.2, to indicate that the first use happened around the year 1600, with no more information. If the data was more precise, with information such as "between 1590 and 1610", we would have used an interval, as we did in Figure 13, with the information that the "end of the 17th century" corresponds to the period between 1670 and 1700. The decision we made here that the end of a century is 30 years long is arbitrary, and its only purpose here is to demonstrate the possibilities offered by our model.

Lastly, in Figure 14, several interesting structures appear. The relation between the North Devon sgraffito ware and its decoration is a ternary relation, between the material (*North Devon sgraffito ware*), the type of the decoration (*sgraffito*) and the pattern (of which there are two in this case, *Geometric* and *Floral*). To represent it, there is a central node *North Devon sgraffito ware decoration* (abbreviated *NDSW decoration* for the sake of the visual representation), connecting these three elements. Then, each of these relations is also instantiated, which allows a lot of flexibility in the representation. For instance, we can represent that both authors agree on the fact that North Devon sgraffito ware has a decoration of te type sgraffito, but that only Brassard indicates that this decoration can have a geometric or floral pattern. Speaking of which, the structure of alternative possibilities appear here, since, according to Brassard, the pattern used for the decoration of North Devon sgraffito ware can be either geometric or floral.

This example, although it comes from a small sample of data, represents well the flexibility that this model offers. Since we worked with a data-driven approach and our model was built on use cases that took the form of questions, examples of questions we can answer are:

- What do we know about the color of North Devon sgraffito ware?
- In which document Brassard and Leclerc describe the color of North Devon sgraffito ware?
- When was North Devon sgraffito ware produced?
- When was the golden age of North Devon sgraffito ware, and according to who?
- What is the decoration typically found on North Devon sgraffito ware?

– What are the patterns one can found on North Devon sgraffito ware?

6. Related work

6.1. Complex relations

As presented, this project was characterized by several cases where relationships considered complex were encountered. Also, the data-centric approach we advocated demonstrated the need to be able to add information on the relations and on the statements by annotating them (qualitative value of the relation and/or source of the statement). In the past, this reality of representing triples with triples (meta triples) has also been experienced by others, leading to the development of different solutions that we explored and that have influenced our work : the W3C's ontology patterns for representing n-ary relations, RDF reification, the singleton property and, finally, the Wikidata approach presented previously.

In principle, the ontology patterns proposed by the W3C for n-ary relations [11] can help with the issues related to the need to have more than one object to a property. By introducing a new class for a relation, it is thus possible to describe a relation instance with additional information (i.e. qualitative value of the relation). Thereby, the individual becomes an object of a triple and, at the same time, becomes the subject of other triples. The W3C uses blank nodes to represent instances of relations, which is problematic for our project. Indeed, since we are working in a linked open data environment, blank nodes are not considered a valid solution since they are not represented with an Uniform Resource Identifier (URI). Blank nodes are not respecting the principles outlined by Berners-Lee [12] and makes interoperability between datasets from different sources much more difficult.

In comparison with this n-ary relations pattern, RDF reification [13] is designed so that one can add information about statements (object, predicate, subject triples) and not about relation instances. This approach is particularly useful for provenance and context information, but implies the creation of a fourth triple (rdf:statement) acting as a proxy for the statement. This approach, however, has been discussed and is often considered unproductive and lacking formal semantics [3] [14]. Furthermore, because of its limitations, the RDF reification approach has been removed from the normative section of the W3C recommandation [15].

[14] proposed the singleton properties approach for annotating RDF statements. In this case, the idea is to have a property instance representing a unique relationship between a subject and an object in a particular context. However, this model implies a certain ambiguity between classes, properties and instances that does not completely stay true to the definitions and uses recommended by the W3C. Also, while we acknowledge it would possibly work in our context, it would notably complicate the process of writing queries.

Contrary to the different approaches presented previously, the model developed as part of this project allowed us to solve the problems related to the issues mentioned : the need to be able to represent mandatory and optional features of a material as well as mutually exclusive features and, finally, to add a source to a statement.

6.2. Temporality

Different approaches have been proposed to depict this information in a structured and sharable way, most of them built on Allen's interval algebra [8]. [4] examined the similarity or closeness to match query and annotation intervals in cultural heritage applications. They applied the fuzzy set theory to model imprecise time intervals, but they did not take into account certain temporal statements as well as persistent and unknown boundaries [16]. [6] focused on anchored time intervals and proposed a model to represent uncertain temporal information in museum information systems (relational databases). [5], for their part, based their work on the temporal evidence as it is worded in the cultural heritage discourse, putting aside Allen's formal relations. They proposed an OWL temporal ontology that aims at representing in a formal way the temporal hints in historical narratives.

Contrary to these different solutions, our goal was not to build a new ontology to represent imprecise temporal knowledge, but to be able to use already existing ontologies and models to achieve a formal representation of the way temporality was described in our specific dataset. By combining different approaches (the Wikidata approach, property chains in OWL2, OWL-Time, etc.) and working in a data-driven context, we are building almost exclusively from previous work. This is in agreement with the idea of creating ontologies by reusing other ontologies' design patterns instead of proposing a whole new model for an already well-known issue.

Following this train of thought, we also took a look at CRMgeo, a CIDOC CRM extensions. This model was developped to represent spatiotemporal properties of temporal entities and persistent items while ensuring compatibility with the CIDOC CRM [17]. However, this model is based on encoding standards and topological relations defined by the Open Geospatial Consortium in GeoSPARQL and does not focus primarly on temporality and its possible fuzziness.

7. Conclusion and future work

In this paper, we presented a model aiming at representing statements about archaeological knowledge, most specifically materials as concepts (in contradiction with specific physical materials) and their characteristics. Several challenges have arisen during this process, such as the need to represent the qualification of the relation (for instance, the distinction between a optional and a mandatory property of a material). We also had to be able to indicate the source of any information as well as temporal information and all its specificities. By building on solutions developed in the past by others (CIDOC CRM, OWL-Time, the Wikidata approach, etc.), we were able to meet the different needs emerging from the dataset we were working with.

We acknowledge that one of the consequence of the model, however, is that it heavily increases the number of triples, making the processing more complex. That said, we feel that the benefits justify the costs, since it allowed us to represent in a formal way the information and to answer use cases. In fact, with this model, we were able to answer more than 100 use cases that took the form of scenario-related informal competency questions [18]. Another element to consider are the cardinality issues for “non-simple properties". Since we use property chain axioms in our model, we can’t combine them with cardinality constraints because it can lead to undecidability. OWL2 specifies that it is only possible to use *simple* properties, that is properties that can’t be inferred from property chains, in cardinality expressions. This can be solved with some tricks but this won’t be detailed here since it is out of scope.

We believe the model that resulted from our work could be extended in another context and on a larger scale. Indeed, in the course of the development of an ontology aiming at representing Quebec’s cultural heritage, the solutions developed during this proof of concept on materials will inevitably prove valuable. The next step planned for this project is to populate the ontology with the significant amount of data stored in the Ministry’s database, which will help us confirm the practicality and performance of this model.

8. Acknowledgements

This research was supported by the Ministère de la Culture et des Communications du Québec and is a result of a collaboration with Polytechnique Montréal. Alexandre Naud and Marielle St-Germain were both hired as consultants by the Ministry. We thank Isabelle Jacques for constant assistance during this work and for insightful comments that helped improve the manuscript.

References

[1] P. Le Boeuf, M. Doerr, C.-E. Ore and S. Stead (eds), Definition of the CIDOC Conceptual Reference Model, ICOM/CIDOC Documentation Standards Group and CIDOC CRM Special Interest Group. http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v6.2.6_Definition_esIP.pdf.

[2] C. MacLeod, Accident or Design? George Ravenscroft’s Patent and the Invention of Lead-Crystal Glass **28**(4), 776–803. doi:10.2307/3105182. https://www.jstor.org/stable/3105182.

[3] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandečić, Introducing Wikidata to the Linked Data Web, in: *The Semantic Web – ISWC 2014*, Lecture Notes in Computer Science, Springer International Publishing, pp. 50–65. ISBN 978-3-319-11964-9.

[4] T. Kauppinen, G. Mantegari, P. Paakkarinen, H. Kuittinen, E. Hyvönen and S. Bandini, Determining relevance of imprecise temporal intervals for cultural heritage information retrieval **68**(9), 549–560. doi:10.1016/j.ijhcs.2010.03.002. https://linkinghub.elsevier.com/retrieve/pii/S107158191000039X.

[5] G. Saygi, J.-Y. Blaise and I. Dudek, Temporal hints in the cultural heritage discourse: what can an ontology of time as it is worded reveal?, in: *Revue des Nouvelles Technologies de l’Information (RNTI)*, p. 14.

[6] M. Nurminen and A. Heimbürger, Representation and Retrieval of Uncertain Temporal Information in Museum Databases, 90–109. doi:10.3233/978-1-60750-992-9-90. http://www.medra.org/servlet/aliasResolver?alias=iospressISSNISBN&issn=0922-6389&volume=237&spage=90.

[7] A. Häyrinen, Towards semantic modelling of cultural historical data, in: *Information Modelling and Knowlege Bases XXII*, IOS Press, pp. 290–308.

[8] J.F. Allen, Maintaining knowledge about temporal intervals **26**(11), 832–843. doi:10.1145/182.358434. http://portal.acm.org/citation.cfm?doid=182.358434.

[9] M. Brassard and M. Leclerc, *Identifier la céramique et le verre anciens au Québec : guide à l'usage des amateurs et des professionnels*, *Cahiers d'archéologie du CELAT* Vol. 12, CELAT, Université Laval. ISBN 2-920576-64-X.

[10] L. Métreau (ed.), Identifier la céramique au Québec, *Cahiers d'archéologie du CELAT*, CELAT. ISBN 978-2-923663-26-5.

[11] W.W.W. Consortium, Defining N-ary Relations on the Semantic Web. https://www.w3.org/TR/swbp-n-aryRelations/.

[12] T. Berners-Lee, Linked Data - Design Issues. https://www.w3.org/DesignIssues/LinkedData.html.

[13] W.W.W. Consortium, RDF Primer. https://www.w3.org/TR/rdf-primer/#reification.

[14] V. Nguyen, O. Bodenreider and A. Sheth, Don't Like RDF Reification? Making Statements about Statements Using Singleton Property **2014**, 759–770. doi:10.1145/2566486.2567973. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350149/.

[15] W.W.W. Consortium, RDF 1.1 Primer. https://www.w3.org/TR/rdf11-primer/.

[16] U. Visser, *Intelligent Information Integration for the Semantic Web*, Springer, Google-Books-ID: VmP2BwAAQBAJ. ISBN 978-3-540-28636-3.

[17] G. Hiebel, M. Doerr and Eide, CRMgeo: A spatiotemporal extension of CIDOC-CRM **18**(4), 271–279. doi:10.1007/s00799-016-0192-4.

[18] S. Peroni, A Simplified Agile Methodology for Ontology Development., in: *Proceedings of the 13th OWL: Experiences and Directions Workshop and 5th OWL reasoner evaluation workshop (OWLED-ORE 2016)*. https://essepuntato.it/papers/samod-owled2016.html.