LE STUDIUM
Loire Valley
Institute for Advanced Studies

LE STUDIUM Multidisciplinary Journal

www.lestudium-ias.com

FELLOWSHIP FINAL REPORT

Ontology-based data modeling for cultural heritage

Emilio M. Sanfilippo[1, 2, 3]

[1] LE STUDIUM Institute for Advanced Studies, 45000 Orléans, France
[2] ISTC-CNR Laboratory for Applied Ontology, via alla cascata 56/c, 38123, Povo, Trento, Italy
[3] Intelligence des Patrimoines, Centre d'Études Supérieures de la Renaissance, University of Tours, 37000 Tours, France

REPORT INFO

Fellow: Dr. **Emilio M. Sanfilippo**
From: Laboratory of Digital Sciences of Nantes, École Centrale of Nantes, CNRS UMR 6004, France
Host laboratory in region Centre-Val de Loire: Intelligence des Patrimoines, Centre d'Études Supérieures de la Renaissance, University of Tours
Host scientist: **Prof. Benoist Pierre**
Period of residence in region Centre-Val de Loire: April 2019 - August 2020

Keywords:
Ontology, Semantic Web, Digital Humanities, Cultural Heritage, CIDOC-CRM

ABSTRACT

Institutions working in cultural heritage contexts have at their disposal large quantities of data, which need to be properly handled and classified to be used in applications, shared, and possibly even integrated. For these purposes to be achieved, shared conceptual models - in the form of ontologies - are needed. CIDOC-CRM is the reference standard for managing cultural heritage data. However, some of its modeling choices do not meet the modeling criteria of formal ontology. The purpose of this report is to document the research work done during my fellowship at Le Studium in 2019-2020 concerning the use ontologies for cultural heritage. This covers the ontological analysis of (some portions of) CIDOC and the development of an extension of the latter model for research purposes in the scope of scientific projects at the IPAT-CESR.

1- Introduction

Institutions[1] working in the context of cultural heritage have at their disposal large quantities of data, which need to be handled and classified to be used in applications, shared, and integrated [4, 5, 15]. These tasks are gaining increasing relevance nowadays, especially in the light of initiatives related to the FAIR principles for data management, the acronym standing for Findable, Accessible, Interoperable, and Reusable [22]. Research institutions and stakeholders wish indeed to make their data available to others - typically through the use of Web platforms - in such a way to allow the further processing of the data, e.g., to find inter-relations between multiple datasets or to get a broader and integrated view on the investigated phenomena [14].

From a computer science perspective, this scenario opens various research challenges, among which the need of a reference conceptual model. Ideally, such a model should work as a sort of *lingua franca* allowing machines to similarly classify the data, therefore facilitating data publishing, sharing, and integration.

The development of reference conceptual models for data management is commonly pursued with ontologies. Informally, an ontology is a model which approximates the intended meaning of the vocabulary of terms used to describe data and experts' knowledge [7][2]. When multiple communities adopt the same ontology, their datasets have higher chances of interacting in an efficient manner [20].

[1] This report is partially based on the research work presented in [17].

[2] From a more general perspective, an ontology is a model of human knowledge K about a domain of interest D representing what exists in D according to K [7].

For cultural heritage data management, the standard ontology is the so-called CIDOC Conceptual Reference Model (CIDOC-CRM, ISO 21127, hereafter CIDOC) [2]. CIDOC has been adopted in several research projects worldwide and it constitutes the conceptual architecture for many institutions to handle data in information systems.

Despite its large exploitation, CIDOC is weakly axiomatized and some of its modeling choices remain opaque. Existing works like [13] have improved its formal representation but they have only partially contributed to its theoretical foundations. For instance, as we will see in the next sections, the ontology adopts a representational approach at the intersection between (philosophical) three- (3D) and four-dimensionalism (4D), which - apart from being controversial from a theoretical standpoint [21] - does not seem to bring any advantage from a modeling perspective.

The purpose of this report is to document the research work done during my fellowship at Le Studium in the academic year 2019-2020 about the use ontologies for cultural heritage. This covers a first ontological analysis of (some portions of) CIDOC based on well-known approaches in applied ontology. In particular, I rely on both the OntoClean methodology [6] and theories of formal ontology (e.g., 3D, 4D, etc.) to analyze the ontology and its structure. Since many of the latter theories have been already adopted in foundational ontologies like UFO [8] and DOLCE [11], among others, we will rely on these ontologies, too, for the analysis.

The report is structured as follows. We present and analyze in Section 2 - Section 4 some of the core modeling elements of CIDOC. Section 5 presents the ontology IPATO developed in the scope of the fellowship by extending CIDOC to meet specific modeling scenarios. Section 6 presents an academic case study. Finally, Section 7 concludes the paper by addressing some topics of interest in collaboration with the host institution to strengthen the proposal hereby presented.

2- CIDOC-CRM: General overview

CIDOC (version 6.2.1)[3] [2] consists of 94 taxonomically organized classes and 168 horizontal relations (called *properties*).[4] It is mainly conceived and maintained in a semi-formal and application-independent notation, although the ontology is also largely exploited in Semantic Web environments through languages like RDF(S) and OWL (see, e.g., [15]). For each class, the specification provides:

- Its parent and child classes (if the latter are present), where only *direct* taxonomic relations are specified in first-order logic (FOL);
- A natural language definition, which is associated to comments and examples to facilitate the understanding of the class;
- In some cases, the horizontal relations by which the class can be linked to other classes.

Similarly, for each relation the specification provides:

- Domain and range information (in bothnatural language and FOL);
- Taxonomic relations (with respect toother relations);
- Natural language comments andexamples;
- Cardinality restrictions (called *quantification*).

According to CIDOC, the latter "are provided for the purpose of semantic clarification only, and should not be treated as implementation recommendations" [2, p.XIII]. Hence, given a relation associated with a cardinality, it is not

[3] At the time of the fellowship, the version 6.2.1 of CIDOC was the most stable version publicly available; see http://www.cidoc-crm.org/versions-of-the-cidoc-crm. Last accessed August 2020.

[4] Each class in CIDOC is prefixed by a unique alphanumeric ID identified with the letter E, whereas relations' IDs are identified with P.

mandatory to comply with the latter when the ontology is represented in a specific formal notation.[5]

For the sake of clarity, consider the following example. The class *E5 Event* is subsumed by *E2 Temporal Entity*. Among others, the relation *P11 had participant* is used to relate *E5 Event to E39 Actor*. The cardinality of P11 is set to (0,n) on both sides. CIDOC is however open to alternative interpretations. This choice is unfortunate since divergent formalizations may lead to scarcely interoperable data models.

For instance, consider two alternative formalizations; the first one, call it O1, implements cardinalities as they are given in [2]; the second one, O2, where the cardinality of P11 is restricted to (1,n) on the side of *E39 Actor* so that an instance of E5 must have at least one actor as participant. While O2's models are O1's models, too, the vice-versa does not hold. In this sense, by leaving open to users the choice of how to interpret cardinalities, the CIDOC's approach runs the risk of making it hard for applications to interoperate.

Figure 1 (see Appendix) shows the most general classes of CIDOC.[6]

The distinction between *E77 Persistent Item* and *E2 Temporal Entity* is the core dichotomy of CIDOC. Instances of the former are *endurants* keeping their identity through time [2, p.35], whereas instances of the latter are *perdurants* unfolding in time [2, p.2]. These classes are therefore disjoint[7]. Also, CIDOC adopts a so-called *event-oriented approach* (in the terminology of [4]), according to which the representation of events is fundamental in the scope of the ontology. For example, representing a person's birth date means, first, to represent the person's birth event and,second, to label the time span of this event by a date.

We discuss in the next sections the analysis of persistent items -conceptual objects included- for their relevance in the scope of our research. The interested reader can refer to [17] for a broader analysis of the ontology, codification, and modularization in OWL.

3- Analysis of Persistent Items

We analyze in this section the taxonomy of persistent items, see **Figure 2** at the end of the report. We first provide a general overview of the taxonomy by introducing some of its classes and we then analyze the taxonomy while introducing the remaining classes.

Looking at **Figure 2**, CIDOC models a high-level distinction between *E39 Actor* and *E70 Thing*. Instances of *E39 Actor* are either individual persons (*E21 Person*) or groups (*E74 Group*) "who have the potential to perform intentional actions" [2, p.20].

The class *E40 Legal Body* extends *E74 Group* to model "institutions or groups of people that have obtained a legal recognition [...] and can act collectively as agents" [2, p.21].

E70 Thing is a generic class subsuming different types of entities. A first distinction is between man-made (*E71 Man-Made Thing*) and non-man-made things (*E19 Physical Object*, *E26 Physical Feature*); as the terminology suggests, only the former are intentionally produced by actors. A second distinction is between *E18 Physical Thing* and *E28 Conceptual Object*. Instances of the former class exist in space, whereas instances of the latter are "non-material products of our minds" [2, p.16] such as natural languages (*E56 Language*), the "contents" of physical books (*E89 Propositional Object*), or types (*E55 Type*, e.g., material types), among others. According to CIDOC, conceptual objects

[5] In the work presented in [13], cardinalities are interpreted as suggested in [2].

[5] CIDOC also includes E59 Primitive Value at the same level of E1 CRM Entity to represent data types.

[7] Apart from the disjointness between E77 and E2, there is only another disjointness declaration in CIDOC between *E18 Physical Thing* and *E28 Conceptual Object*, see Section 3.

"exist as long as they can be found on at least one [physical] carrier or in at least one human memory" (ibid.). Since *E28 Conceptual Object* is not subsumed by *E18 Physical Thing*, its instances do not reside in space.[8]

To comment on the taxonomy, first, the distinction between *E39 Actor* and *E70 Thing* is not sharp. Looking at **Figure 2**, *E21 Person* is subsumed by *E20 Biological Object*, which is subsumed by E70. In addition, the scope of E70 is broad enough to cover E39 and all its subclasses.

Second, *E72 Legal Object* subsumes all physical things, among other classes. Its instances are material or immaterial items to which legal rights, such as property rights, apply. In our understanding, from a formal ontology perspective, *E72 Legal Object* models *anti-rigid properties* in the sense of OntoClean [6], i.e., properties that entities only *possibly* satisfy and whose acquisition or loss does not alter their identity. For instance, a cultural artefact like a statue is subject to legal rights only in the scope of a specific socio-legal context; hence, it could stop being a legal object, e.g., if brought to a different context, while remaining always the same statue. On the other hand, *E18 Physical Thing* seems to model *rigid properties*, namely, properties that entities *necessarily* satisfy and whose loss *does* affect identity. Assuming these considerations along with the formal treatment of anti-/rigidity in OntoClean, physical things can not be subsumed by legal objects.

Finally, the class *E92 Spacetime Volume* deserves some discussion. CIDOC has inherited this class from the CRMgeo [9], which extends CIDOC for geo-spatial applications. According to [2], E92 "comprises 4 dimensional point sets (volumes) in physical spacetime [...]. An instance of *E92 Spacetime Volume* is either contiguous or composed of a finite number of contiguous subsets" [2, p.41].

Apart from *E4 Period* and *E18 Physical Thing*, this class subsumes *E93 Presence*, i.e., "*snapshots* of a Spacetime volume, i.e. intersections of a Spacetime volume with all space restricted to a particular time-span, such as the extent of the Roman Empire during 33 B.C. " [9].

If we interpret it properly, instances of E92 correspond to *four-dimensional worms* in the sense of ontological four-dimensionalism (4D) [18]. This seems clear from its definition as something that has both temporal and spatial extents but also from the examples in [2, 9]; e.g., the fact that an individual spacetime volume can be "cut" in different parts, each one standing for a spatio-temporal snapshot of the entity at stake. An example provided in the documentation is the temporal extent of the Roman Empire during 33 B.C.

If this consideration is correct, CIDOC mixes 4D with a standard three-dimensionalism (3D) view. From a foundational perspective, this approach is controversial. Despite the hot debate on 4D and 3D in formal ontology, these remain alternative and incompatible positions (see [21] for somediscussion).

The situation is not better from a modeling perspective, since the benefits of introducing spacetime volumes is unclear. According to [2], a reason for having these entities is to simplify data models; e.g., to represent "an [instance of] *E18 Physical Thing* without representing each instance of it together with an instance of its associated spacetime volume" [2, p.12]. What the specification seems to suggest is that one can represent physical (or temporal) entities without necessarily modeling their spatial or temporal locations. This is because they inherit their spatio-temporal dimension by being instances of E92. In our view, this consideration is not fully correct. First, it can be relevant for application purposes to

[8] See Section 4 for the analysis of conceptual objects.

explicitly model, e.g., the space region occupied by an individual object at a certain time. Second, even by assuming the distinction between space regions, temporal regions, perdurants, and endurants, it is not necessary - at the instance level - to represent all (spatial, temporal) regions which an object occupies during its entire life or all perdurants where it participates.

Based on this analysis, **Figure 3** showsthe restructuring of the taxonomy of persistent items. Classes with dashed lines are new[9]; also, the taxonomy does not include *E70 Thing*, *E72 Legal Object*, and *E92 Spacetime Volume*. Some comments are due.

First, *E18 Physical Thing* is now directly subsumed by *E77 Persistent Item* and it is disjoint with *Non-Physical Thing*. This latter class is introduced to sharply distinguish between physical and non-physical items.

Non-Physical Man-Made Thing extends *Non-Physical Thing* to explicitly classify non-physical items resulting from human actions.[10]

E70 Thing has been removed because it was only a generic umbrella without a specific intended meaning. The class *E71 Man-Made Thing* is directly subsumed by *E77 Persistent Item*. It is neither disjoint nor subsumed by E18 or *Non-Physical Thing*, because it subsumes both physical and non-physical man-made entities.

Second, looking at physical things, we introduce *Aggregation* to distinguish between general collections of physical things (e.g., all objects on my desk) and instances of *E78 Collection*, among others. Aggregations should not be confused with physical objects having multiple and physically connected parts suchas potteries or statues (both instances of *E22 Man-Made Object*). Aggregations bear indeed unity conditions other than topological ones. For instance, according to [2], museum collections, which are represented as specific types of aggregations in **Figure 3**, are "assembled and maintained by one or more instances of E39 Actor over time for a specific purpose and audience" [2, p.36]. An example is the collection of the British Museum, which qualifies as a collection because it consists of objects collected and owned by the museum, and possibly used during its exhibitions. Its unity could be therefore defined in legal terms.

E74 Group and *E40 Legal Body* are both subsumed by *Aggregation*, following CIDOC's understanding of groups as collections of individual persons satisfying (non-topological) unity conditions.[11] In addition, both *E74 Group* and *E21 Person* are subsumed by *E39 Actor*, which is a direct subclass of *E18 Physical Thing*. The revision of CIDOC concerning agents is based on and simplifies the ontology of groups and institutions presented in [3,16]. In these works, the authors distinguish between arbitrary collections of individuals and social groups. In addition, differently from CIDOC, the approach in [3,16] allows to explicitly represent the membership conditions that individuals must satisfy to form groups. This approach could be adopted to enhance the ontology of actors in CIDOC, which remains only weakly characterized at the current state.

Third, *E92 Spacetime Volume* has been removed from the taxonomy because of its ambiguity. However, since CIDOC covers both places, temporal regions, and temporal entities, even by removing E92, one still has the possibility of linking persistent items to space, time, and temporal entities.

[9] Following CIDOC minimality principle (see [2,p.XVI]) each new inserted class is used either as domain or range for a relation.

[10] The disjointness between *Non-Physical Man-Made Thing* and *E24 Physical Man-Made Thing* can be logically derived. It is included in the diagram to facilitate understanding.

[11] Since CIDOC understands legal bodies as groups with legal status, legal bodies constituted by single persons are not covered by the ontology. An extension in this direction could be needed.

Finally, by conceiving legal objects as social roles, instances of *E72 Legal Object* can be represented in different ways. A proposal, based on [12], consists in introducing a new class, *Social Role*, for properties like *being a student* or *being a professor* that entities satisfy within specific contexts. From this perspective, legal objects can be (roughly) understood as roles that entities acquire in socio-legal systems or events. Following [12], the property of *being a legal object* is reified in the domain of discourse as an instance of *Social Role*, whereas the CIDOC's relation *P2 has type* can be used to link an entity to it (e.g., a statue *has type* legal object); alternatively, a new relation can be easily introduced.

4- Analysis of conceptual objects

E28 Conceptual Object is an overarching modeling element standing for "non-material products of our minds that [...] are created, invented or thought by someone, and then may be documented or communicated between persons" [2, p.16]. Its instances can exist in multiple carriers at the same time, including paper, canvases, and human memory. Conceptual objects exist if at least one of their carriers exists and can not be directly destroyed, i.e., destroying a conceptual object means destroying all its carriers.

Among its subclasses (see Figure 4), *E90 Symbolic Object* stands for "[...] identifiable symbols and any aggregation of symbols, such as characters, identifiers, traffic signs, emblems, texts, data sets, images, musical scores, multimedia objects, computer program code or mathematical formulae that have an objectively recognizable structure and that are documented as single units" [2, p.41]. In addition, symbolic objects "[do] not depend on a specific physical carrier [...], and can exist on one or more carriers simultaneously" (ibid.). An example is "the Italian text of Dante's *Divina Commedia* as found in the authoritative critical edition *La Commedia secondo l'antica vulgata* a cura di Giorgio Petrocchi, Milano: Mondadori, 1966-67" (ibid.). As a symbolic object, this entity is not the specific text printed on an individual book; rather, it corresponds to the text-type shared by all the physical copies of the *Comedy* edited by Petrocchi. Symbolic objects can also be symbols without specific meanings, "for example an arbitrary character string" (ibid.).

Another subclass of *Conceptual Object* is *E89 Propositional Object* whose instances are "*immaterial items*, including but not limited to stories, plots, procedural prescriptions, algorithms, [...] or images that are, or represent in some sense, sets of propositions about real or imaginary things and that are documented as single units or serve as topic of discourse" [2, p.40]. Examples are "the ideational contents of Aristotle's book entitled *Metaphysics*" or "[t]he image content of the photo of the Allied Leaders at Yalta published by UPI, 194"' (ibid).

Looking at **Figure 4**, the class E73 *Information Object* is subsumed by both *Symbolic Object* and *Propositional Object*; the intended meaning is that its instances are propositional objects encoded in some symbolic form. It includes various subclasses, among which *E33 Linguistic Object* and *E36 Visual Item*.

Visual items are "the intellectual or conceptual aspects of recognisable marks and images" [2, p.19]. An example is the Coca-Cola logo, which is not the individual logo printed on a specific Coca-Cola can but the "underlying prototype" (ibid.) appearing in all Coca-Cola cans.

Linguistic objects are "identifiable expressions in natural language or [other] languages" that are independent "from the medium or method by which they are expressed" [2, p.18]. Examples are "the text of the *Jabberwock* by Lewis Carroll" or "the lyrics of the song *Blue Suede Schoes*".

The relation *has language* links the class *Linguistic Object* to *E56 Language*, whereas *has translation* links instances of *Linguistic Object* to each other with the restriction that

"[w]hen a Linguistic Object is translated into a new language it becomes a new Linguistic Object, despite being conceptually similar to the source object" [2, p.70].

Finally, *is carried by* (not shown in Figure 4) links symbolic objects (information objects included) to their physical supports, e.g., physical books.

To comment on these modeling elements, first, *Propositional Object* and its subclasses are understood as ideational contents but the documentation does not clarify what this means. This perspective recalls (certain types of) idealist theories in philosophy according to which the ideational content of, e.g., Aristotle's *Metaphysics* - to recall one of CIDOC's examples - is a mental object (an *idea*) possibly embodied in a text (see [19]).

Second, *Information Object* and its subclasses are subsumed by both *Symbolic Object* and *Propositional Object*. An information object is therefore *both* an intellectual content *and* a symbolic form. It is not a case that the identity of linguistic objects is bound to their languages, so that - as we saw above - the translation of a linguistic object brings about the creation of a new linguistic object. What remains surprising is the choice of relating *Information Object* to *Symbolic Object* and *Propositional Object* via taxonomic relations.

This approach has practical disadvantages, e.g., Bouef et al. [2] claim that linguistic objects' translations share the same content. However, by identifying a linguistic object with both a content and a symbolic form, one lacks a way to identify and represent the content shared by multiple linguistic objects.

To deal with the latter issue, a first proposal - based on the ontology called (FRBR) [1] - is to detach the subsumption of *Information Object from Symbolic Object*. In this sense, an information object corresponds to the content of, e.g., a novel or essay, whereas *Symbolic Object* is used to represent its corresponding text. In this approach, one has therefore the flexibility of representing multiple texts sharing the same content. The former issue, relative to the ontological characterization of contents, is more challenging, especially considering that various theories have been proposed in both philosophy and applied ontology without however reaching a high level of formal robustness and conceptual transparency. Further work towards this direction is therefore needed[12].

5- IPAT Ontology

The analysis of CIDOC discussed in the previous sections (see also [17]), as well as application requirements emerging from projects within the scope of the research group *Intelligence des Patrimoines* (IPAT) at the CESR University of Tours have motivated the development of a new ontology for cultural heritage, called IPAT-O(ntology)[13].

IPAT-O is developed and maintained in OWL, which is the standard W3C language mostly used for the use of ontologies in software applications. The ontology is based on a revised and extended version of CIDOC that includes modeling elements which are either imported from existing Semantic Web resources or are created *ex novo* for our purposes. For instance, in order to facilitate the publishing of data on the Web, their sharing, and possibly integration with other data, we use portions of Dublin Core[14], Friend of a Friend (FOAF)[15], Dbpedia[16], Eurovoc[17], and Geonames[18], among others. An example is the use of modeling elements

[12] The ontological analysis of information objects is discussed in: Sanfilippo, E.M, Ontologies for Information Entities. State of the Art and Open Challenges, vol. 16, no. 2, pp. 111-135, 2021.

[13] The ontology is available at: https://github.com/emiliosanfilippo/IPAT-O.

[14] www.dublincore.org/specifications/dublin-core/dcmi-terms/

[15] http://xmlns.com/foaf/spec/

[16] https://wiki.dbpedia.org/services-resources/ontology

[17] https://op.europa.eu/en/web/eu-vocabularies

[18] https://www.geonames.org/

imported from FOAF and Dcat[19] to represent (meta-)data about research projects and datasets, and published through the HeritageS platform developed at IPAT; see **Figure 5** and **Figure 6**.

6- An example about bibliographic data

The case study presented in this section is about the management of data extracted from historical documents of the Renaissance like deeds. The data have been made available in the context of the project *Ressources Numériques pour l'édition des archives de la Renaissance* (RENUMAR)[20], supported by the *Région Centre - Val de Loire*. The purpose of our study is the transformation of the RENUMAR's dataset in a Semantic Web data structure in order to facilitate its publication on the Web platform HeritageS.

The data collected by the project are encoded in (a specific dialect of) TEI-XML. Therefore, the first step of our work has been the analysis of the data model to understand its intended meaning.

For this goal to be achieved, interaction with the domain experts involved in the project has played a major role. We have also agreed with the experts to extract only some meta-data about RENUMAR's documents, e.g., their titles or production dates, while relying on the project's database to access the document's text or editorial comments. Once this goal has been achieved, correspondences in the form of mapping rules between the dataset and the ontology have been created to convert the original data in RDF, therefore to map the data to the ontology.

Consider the following example[21]:

- **Autorité(s) émettrice(s)** : Anne deBretagne; Marchant;
- **Date** : [1513], 25 juin;
- **Lieu d'établissement de l'acte** : Vincennes;
- **Type d'acte** : Lettre close;
- **Forme de l'acte** : Correspondance politique;
- **Suppor t**: Papier;
- **Collection(s)** : Correspondancesurbaines - Bourges;
- **Institution de conservation** : AD 18,Bourges, France;
- **Cote** : 8 G 434.

In the notation of RDF, the data can be represented as:

- ipato:TIPO631967[22]
  - ipato:created_by ipato:Anne_de_Bretagne,ipato:Marchant;
  - dc:created "1513-06-25"^^rdfs:Literal;
  - ipato:created_in ipato:Vincennes;
  - rdf:type ipato:Lettre_Close;
  - ipato:has_act_form ipato:political_correspondance;
  - ipato:archieved_in ipato:archive_AD_18;
  - ipato:editedBy ipato:David_Rivaud;
  - bf:ShelfMark "8_G_434"^^rdfs:Literal.

To comment on the example, first, ipato namespaces are used for modeling elements which are created ex novo for our purposes; all other namespaces refer to existing Semantic Web vocabularies[23]. Second, ipato:Lettre_Close is directly subsumed by CIDOC's *Information Object*, representing therefore a specific type of document's content. On the other hand, a shelf-mark refers to an

[19] https://www.w3.org/TR/vocab-dcat-2/
[20] http://renumar.univ-tours.fr
[21] The original data is available at: http://renumar.univ-tours.fr/xtf/view?docId=tei/TIPO631967.xml;chunk.id=n1;toc.depth=1;toc.id=n1;brand=default

[22] This modeling element identifies the document as a whole entity. In the RDF syntax, ipato:TIPO631967 is the subject of all RDF triples below.
[23] **rdf** and **rdfs** stand for the RDF and RDFS languages (W3C standards), **dc** for Dublin Core, and **bf** for Bibframe.

identifier attached to a document's support, in this case, a paper sheet in the archive of Bourges. The RDF triple ipato:TIPO631967 bf:ShelfMark "8_G_434"^^rdfs:Literal has to be therefore understood as a shortcut used to simplify the data while meaning that the shelf-mark is actually identifying the document's support. Finally, recall that the original data is in French because of the RENUMAR's research context. For the prototyping phase, we have prioritized the use of English to achieve a large community of scholars. The development of a Semantic Web knowledge base in multiple languages remains a desiderata.

A modeling approach like what just presented has been adopted for all other documents available in the RENUMAR's dataset.

7- Perspectives of future collaborations with the host laboratory

We presented in the report the ontological analysis of CIDOC and the development of IPATO for the data management needs in the scope of various research projects. Cultural heritage knowledge representation and data management are complex tasks and it is not surprising that further work to strengthen our proposal is needed.

Future work in collaboration with the host laboratory could be classified along (at least) four research axes:

1. Conceptual analysis and formal representation of CIDOC: the work presented in the report (and [17]) has only partially addressed the analysis of CIDOC. Further work in this direction is required. The analysis needs to be strengthened by a formal representation of the ontology to enhance its conceptual transparency and formal robustness. For these purposes to be achieved, a formalization in first-order logic is needed.

2. Modularization: as a result of the analysis, CIDOC has been re-engineered and restructured in a modular architecture (see [17]). This requires further work. First, the overall methodology for the modularization needs to be better investigated. Despite the several research efforts towards this direction, the ontology engineering community has not produced indeed standard procedures for modularization, which makes it hard to select a specific approach. Second, once the modularization method has been defined, one would expect the development of algorithmic procedures to automatize the development (extraction) of the modules as much as possible.

3. Test cases and implementation: robust and possibly data-driven test cases are needed to validate the ontology against real-world modeling scenarios. These should include not only the *representation* of data but also their *publishing* on the Web platform developed at the IPAT, and possibly *integration tasks* across datasets. Researchers at the CESR University of Tours have been collecting over the years a plethora of data about disparate cultural heritage areas like history, history of art, archeology, book studies, music, and musicology, etc. Collaboration with them would be therefore a desiderata to further strengthen the results of our study.

4. Research about the identity of cultural heritage objects. Cultural heritage objects exist in socio-cultural contexts within the scope of specific temporal frames. A statue, for instance, is not only a piece of marble with aesthetic features; it is also the embodiment of the values that a community ascribes to it in the scope of a reference - temporally bounded - society and culture. Changes in the latter can therefore imply even radical changes in the identity of the statue. This has emerged in the protests sparkled worldwide after the murder of George

Floyd in May 2020. Many statues and monuments have been attacked and destroyed because they embody cultures and values (e.g., colonialism, slavery) which contemporary societies do not accept anymore. To make sense of this and therefore to properly characterize cultural heritage objects in the scope of information systems, we need to make explicit their socio-cultural dimension, that is, what it means for, e.g., a statue to be the expression of a culture at a certain time. At the current state of the art, ontologies for cultural heritage (CIDOC included) limit to the handling of meta-data (e.g., production date or data about artists) without characterizing the identity of cultural heritage objects in relation to socio-cultural contexts.

8- Articles published in the framework of the fellowship

2021

Sanfilippo, E.M., Ontologies for information entities: State of the art and open challenges, Applied ontology, vol. 16, no. 2, pp. 111-135, 2021.

2020

Sanfilippo EM, Markhoff B, Pittet P. Ontological Analysis and Modularization of CIDOC-CRM. To appear in Proceedings of the 11th International Conference Formal Ontologies in Information Systems (FOIS); 2020.

Masolo C, Sanfilippo EM. Technical Artefact Theories: A Comparative Study and a New Empirical Approach. Review of Philosophy and Psychology. 2020 Apr 27:1-28.
DOI: https://doi.org/10.1007/s13164-020-00475-9

2019

Masolo C, Sanfilippo EM, Lamé M, Pittet P. Modeling concept drift for historical research in the digital humanities. In 1st International Workshop on Ontologies for Digital Humanities and their Social Analysis (WODHSA). CEUR workshop proceedings vol. 2518, 2019

Lamé M, Pittet P, Ponchio F, Markhoff B, Sanfilippo EM. Heterotoki: non-structured and heterogeneous terminology alignment for Digital Humanities data producers, CEUR workshop proceedings, vol. 2375, 2019

Sanfilippo EM, Kitamura Y, Young RI. Formal ontologies in manufacturing. Applied Ontology. 2019 Apr 25;14(2):119-25.
DOI: https://doi.org/10.3233/AO-190209

Sanfilippo EM, Terkaj W, Borgo S. Resources in Manufacturing. In 10th International Workshop on Formal Ontologies meet Industry (FOMI). CEUR workshop proceedings vol. 2518, 2019

Guarino N, Sanfilippo EM. Characterizing IOF terms with the DOLCE and UFO ontologies, In 10th International Workshop on Formal Ontologies meet Industry (FOMI) 2019. CEUR workshop proceedings vol. 2518, 2019

Romero F, Sanfilippo EM, Rosado P, Borgo S, BenaventS. Feature in product engineering with single and variant design approaches. A comparative review. In Procedia Manufacturing 2019, vol. 41.
DOI:
https://doi.org/10.1016/j.promfg.2019.09.016

9- Acknowledgements

We gratefully acknowledge the financial support provided by the *Intelligence des Patrimoines* programme and the Région Centre-Val de Loire (ARD 2020 programme).

I wish to thank both Le Studium and the host institution for the great research opportunity I had during the fellowship. I'm particularly grateful to all colleagues, Le Studium fellows included, with whom I had the pleasure of

collaborating. This includes colleagues at the CESR, laboratories at the CNRS, the University of Tours, and the MSH Val de Loire.

10- References

1- Bekiari C, Doerr M, La Boeuf P, Riva P. Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism; 2015.

2- Boeuf PL, Doerr M, Ore CE, Stead S, et al. Definition of the CIDOC Conceptual Reference Model. Version 6.2.1. ICOM/CIDOC Documentation Standards Group CIDOC CRMSIG. 2015

3- Bottazzi E, Ferrario R. Preliminaries to a DOLCE ontology of organisations. International Journal of Business Process Integration and Management. 2009 Jan 1;4(4):225-38.

4- Bruseker G, Carboni N, Guillem A. Cultural heritage data management: the role of formal ontology and CIDOC CRM. InHeritage and Archaeology in the Digital Age 2017 (pp. 93-131). Springer, Cham.

5- Carriero VA, Gangemi A, Mancinelli ML, Marinucci L, Nuzzolese AG, Presutti V, Veninata C. ArCo: The Italian cultural heritage knowledge graph. InInternational Semantic Web Conference 2019 Oct 26 (pp. 36-52). Springer, Cham.

6- Guarino N, Welty CA. An overview of OntoClean. In Handbook on ontologies 2004 (pp. 151-171). Springer, Berlin, Heidelberg.

7- Guarino N, Oberle D, Staab S. What is an ontology?. In Handbook on ontologies 2009(pp. 1-17). Springer, Berlin, Heidelberg.

8- Guizzardi G. Ontological foundations for structural conceptual models; 2005.

9- Hiebel G, Doerr M, Eide Ø. CRMgeo: A spatiotemporal extension of CIDOC-CRM. International Journal on Digital Libraries. 2017Nov 1;18(4):271-9.

10- Marlet O, Rodier X. A way to express the reliability of archaeological data: data traceability at the Laboratoire Archéologie et Territoires (Tours, France). International Journal on Digital Libraries. 2019 Aug 16:1-0.

11- Bekiari C, Doerr M, La Boeuf P, Riva P. Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism; 2015.

12- Boeuf PL, Doerr M, Ore CE, Stead S, et al. Definition of the CIDOC Conceptual Reference Model. Version 6.2.1. ICOM/CIDOC Documentation Standards Group CIDOC CRM SIG. 2015

13- Bottazzi E, Ferrario R. Preliminaries to a DOLCE ontology of organisations. International Journal of Business Process Integration and Management. 2009 Jan 1;4(4):225-38.

14- Bruseker G, Carboni N, Guillem A. Cultural heritage data management: the role of formal ontology and CIDOC CRM. InHeritage and Archaeology in the Digital Age 2017 (pp. 93-131). Springer, Cham

15- Carriero VA, Gangemi A, Mancinelli ML, Marinucci L, Nuzzolese AG, Presutti V, Veninata C. ArCo: The Italian cultural heritage knowledge graph. InInternational Semantic Web

Conference 2019 Oct 26 (pp. 36-52). Springer, Cham.

16- Guarino N, Welty CA. An overview of OntoClean. In Handbook on ontologies 2004 (pp. 151-171). Springer, Berlin, Heidelberg.

17- Guarino N, Oberle D, Staab S. What is an ontology?. In Handbook on ontologies 2009 (pp. 1-17). Springer, Berlin, Heidelberg.

18- Guizzardi G. Ontological foundations for structural conceptual models; 2005.

19- Hiebel G, Doerr M, Eide Ø. CRMgeo: A spatiotemporal extension of CIDOC-CRM. International Journal on Digital Libraries. 2017 Nov 1;18(4):271-9.

20- Marlet O, Rodier X. A way to express the reliability of archaeological data: data traceability at the Laboratoire Archéologie et Territoires (Tours, France). International Journal on Digital Libraries. 2019 Aug 16:1-0

21- Wahlberg TH. The endurance/perdurance controversy is no storm in a teacup. Axiomathes. 2014 Dec 1;24(4):463-82.

22- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3(1):1-9

Appendix: Figures

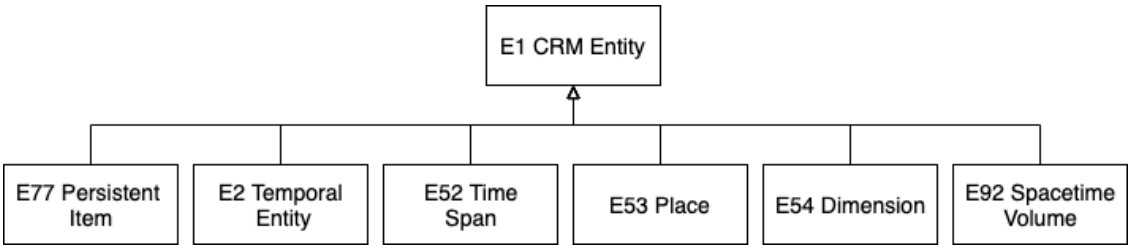

Figure 1: Upper-level taxonomy of CIDOC-CRM (v. 6.2.1)

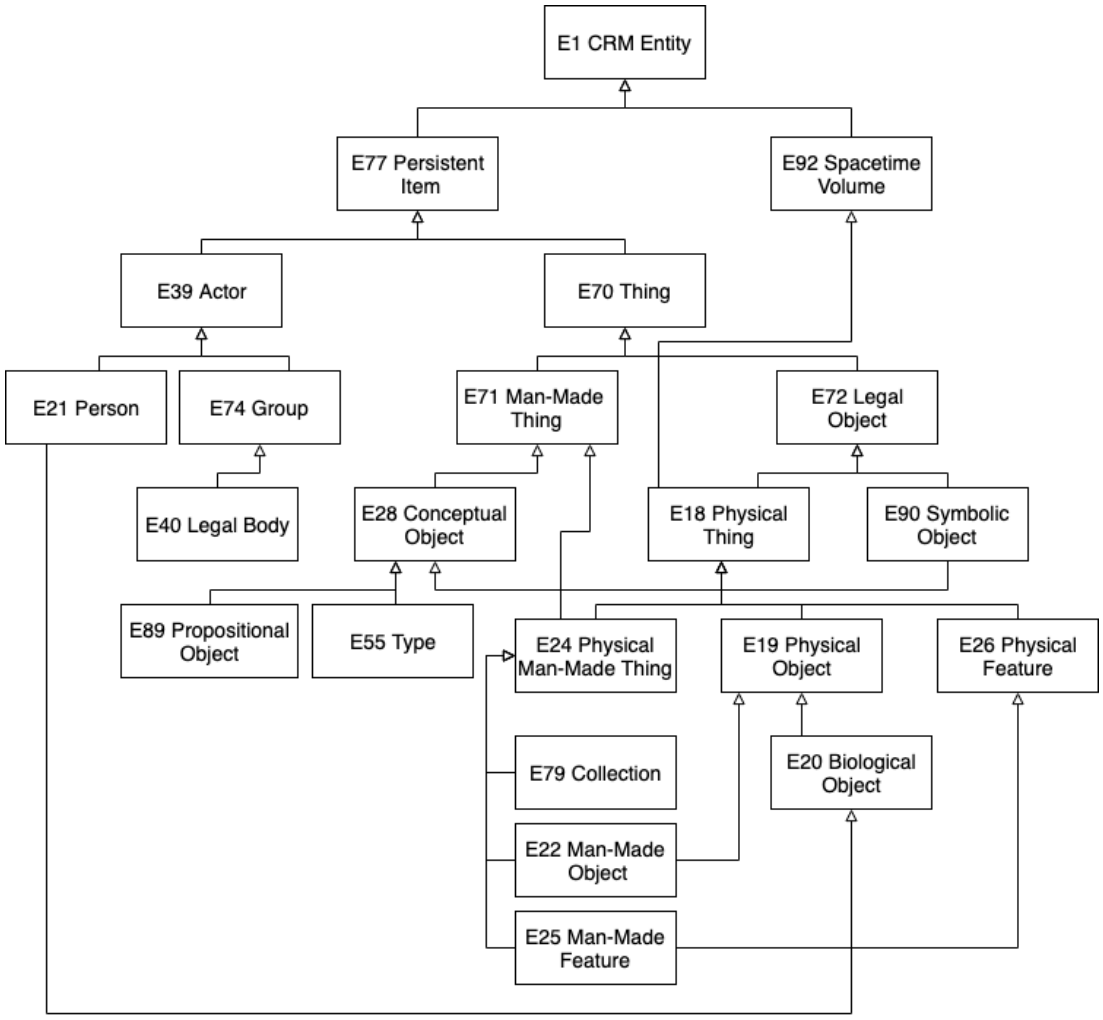

Figure 2: Partial taxonomy of persistent items in CIDOC (v.6.2.1)

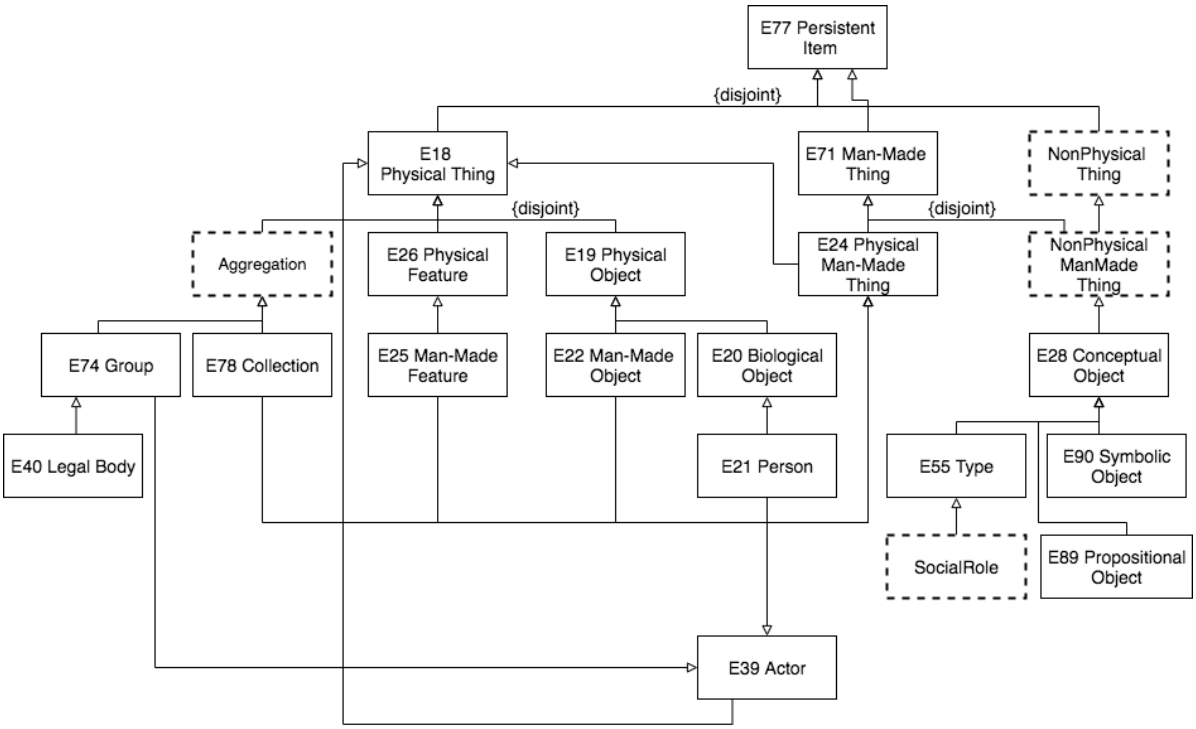

Figure 3: Revised taxonomy of persistent items

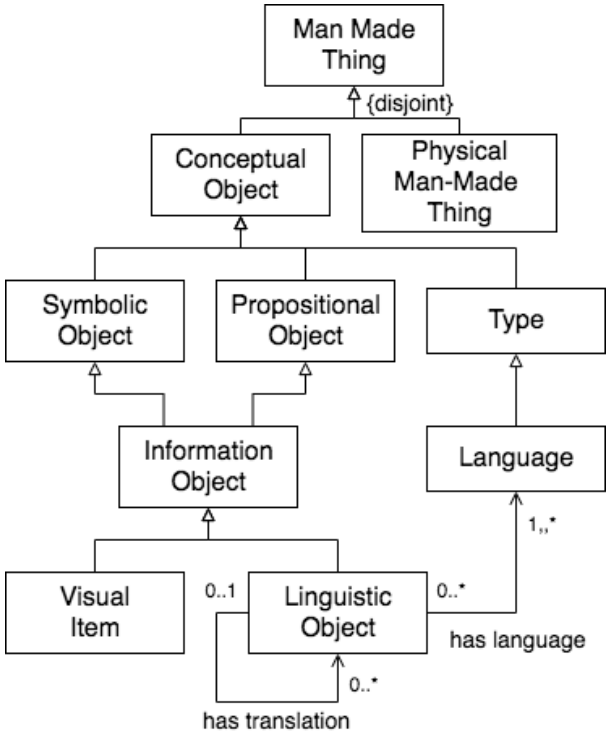

Figure 4: Partial taxonomy of conceptual objects in CIDOC (v. 6.2.1)

Figure 5: Characterization of the class Project (Protégé view)

Figure 6: Characterization of the class Dataset (Protégé view)