Formal Ontology in Information Systems
B. Brodaric and F. Neuhaus (Eds.)

doi:10.3233/FAIA200664

Ontological Analysis and Modularization of CIDOC-CRM

Emilio M. SANFILIPPO [a,b,1], Béatrice MARKHOFF [c], and Perrine PITTET [b]
[a] *Le Studium Loire Valley Institute for Advanced Studies, Orléans and Tours, France*
[b] *CESR – UMR 7323, University of Tours, France*
[c] *LIFAT, University of Tours, France*

Abstract. The CIDOC-CRM ontology is a standard for cultural heritage data modeling. Despite its large exploitation, the ontology is primarily maintained in a semi-formal notation, which makes it difficult to homogeneously exploit it in digital environments. In addition, the ontology consists of several classes and relations, whereas one sometimes wishes to reuse it but only partially. The purpose of the paper is to contribute to the use of CIDOC by strengthening its foundations. On the basis of formal ontology theories, we propose a first analysis of the ontology to enhance its conceptual structure. We also present a preliminary modularization of CIDOC aimed at enhancing both its formalization and usage.

Keywords. CIDOC-CRM, cultural heritage data modeling, modularity

1. Introduction

The CIDOC Conceptual Reference Model (hereafter CIDOC) is a standard ontology (ISO 21127) for cultural heritage data modeling [1]. CIDOC has been adopted in several research projects and it constitutes the conceptual architecture for archives, libraries, and museums, among other institutions, to organize data in information systems [2].

Despite its large exploitation, CIDOC is only weakly axiomatized and some of its modeling choices remain opaque. Existing works like [3] have improved its formal treatment but they have only partially contributed to improve its conceptual framework. For instance, as we will see in the next sections, the ontology adopts a representational approach at the intersection between three- (3D) and four-dimensionalism (4D), which – apart from being controversial from a theoretical standpoint [4] – does not seem to bring any advantage from a modeling perspective. In addition, by working with end-users in the exploitation of the ontology, we have observed that the intended meaning of some of its elements is open to alternative interpretations (e.g., the class *E5 Event*),[2] which is a fact running the risk of compromising its uniform usage across applications.

The purpose of the paper is to contribute to the exploitation of CIDOC by strengthening its ontological and formal foundations. We attempt in this way at making the on-

[1] Corresponding Author: CESR - Université de Tours, 59, rue Néricault-Destouches, 37020 Tours, France. Email:emiliosanfilippo@gmail.com (permanent address).

[2] Each class in CIDOC is prefixed by a unique ID starting with 'E', whereas relations' IDs start with 'P'.

tology more robust and transparent to its users. In order to achieve this goal, we present a first ontological analysis of (some parts of) CIDOC based on well-known approaches in applied ontology. In particular, we rely on both the OntoClean methodology [5] to analyze the taxonomic relations of CIDOC and theories of formal ontology (e.g., 3D, 4D, etc.) to improve its overall conceptual framework. Since many of the latter theories have been already adopted in foundational ontologies like UFO [6] and DOLCE [7], among others, we will rely on these ontologies, too, to analyze CIDOC.

The paper is structured as follows. We present and analyze in Sect. 2–Sect.5 some of the core modeling elements of CIDOC. On the basis of the analysis, we propose in Sect. 6 a modularization of the ontology which revises an existing formalization. By splitting CIDOC in various (inter-connected) modules, we attempt to allow for its *selective* reuse depending on specific application scenarios. Sect. 7 concludes the paper by addressing future work needed to strengthen our proposal.

2. Overview of CIDOC-CRM

The CIDOC ontology (version 6.2.1)[3] [1] consists of 94 taxonomically organized classes and 168 horizontal relations (called *properties*). It is mainly conceived and maintained in a semi-formal and application-independent notation, although the ontology is nowadays largely exploited in Semantic Web environments through languages like RDF and OWL (see, e.g., [8,9]). For each class, the original specification provides 1) its parent and child classes (if the latter are present), where only *direct* taxonomic relations are specified in first-order logic (FOL); 2) a natural language definition, which is associated to comments and examples to facilitate the understanding of the class; 3) in some cases, the horizontal relations by which the class can be linked to other classes. Similarly, for each relation the specification provides 1) domain and range information (in both natural language and FOL); 2) taxonomic relations (with respect to other relations); 3) natural language comments and examples; 4) 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" [1, p.XIII]. Hence, given a relation associated with a cardinality, it is not mandatory to comply with the latter when the ontology is represented in a specific formal notation.[4]

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 liberal 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 O_1, implements cardinalities as they are given in [1]; the second one, O_2, 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 O_2's models are O_1'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.

[3] CIDOC version 6.2.1 is the most recent stable version of the ontology; see http://www.cidoc-crm.org/versions-of-the-cidoc-crm, last accessed March 2020.

[4] In the work presented in [3], cardinalities are interpreted as suggested in [1].

Figure 1 shows the most general classes of CIDOC.[5] We discuss the representation of persistent items and spacetime volumes in Sect.3, temporal entities and time spans in Sect. 4, dimensions in Sect. 5. The analysis of places is left to future work.

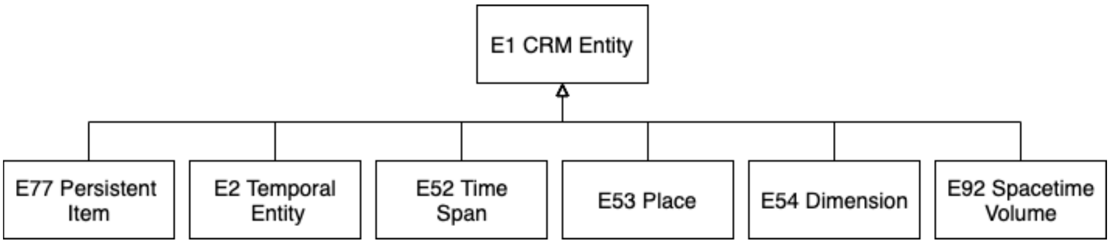

Figure 1. Upper-level taxonomy of CIDOC (v.6.2.1)

Before moving to the next sections, note that 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 [1, p.35], whereas instances of the latter are *perdurants* unfolding in time [1, p.2]. These classes are therefore disjoint.[6] Also, CIDOC adopts a so-called *event-oriented* approach (in the terminology of [2]), 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.

3. Analysis of Persistent Items

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

Looking at Fig. 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" [1, 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" [1, 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" [1, 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 "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*, it follows that its instances do not reside in space.[7]

[5] CIDOC includes also *E59 Primitive Value* at the same level of *E1 CRM Entity* to represent data types. We comment on E59 in Sect. 6.

[6] 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 Sect.3.

[7] The analysis of conceptual objects is left to future work.

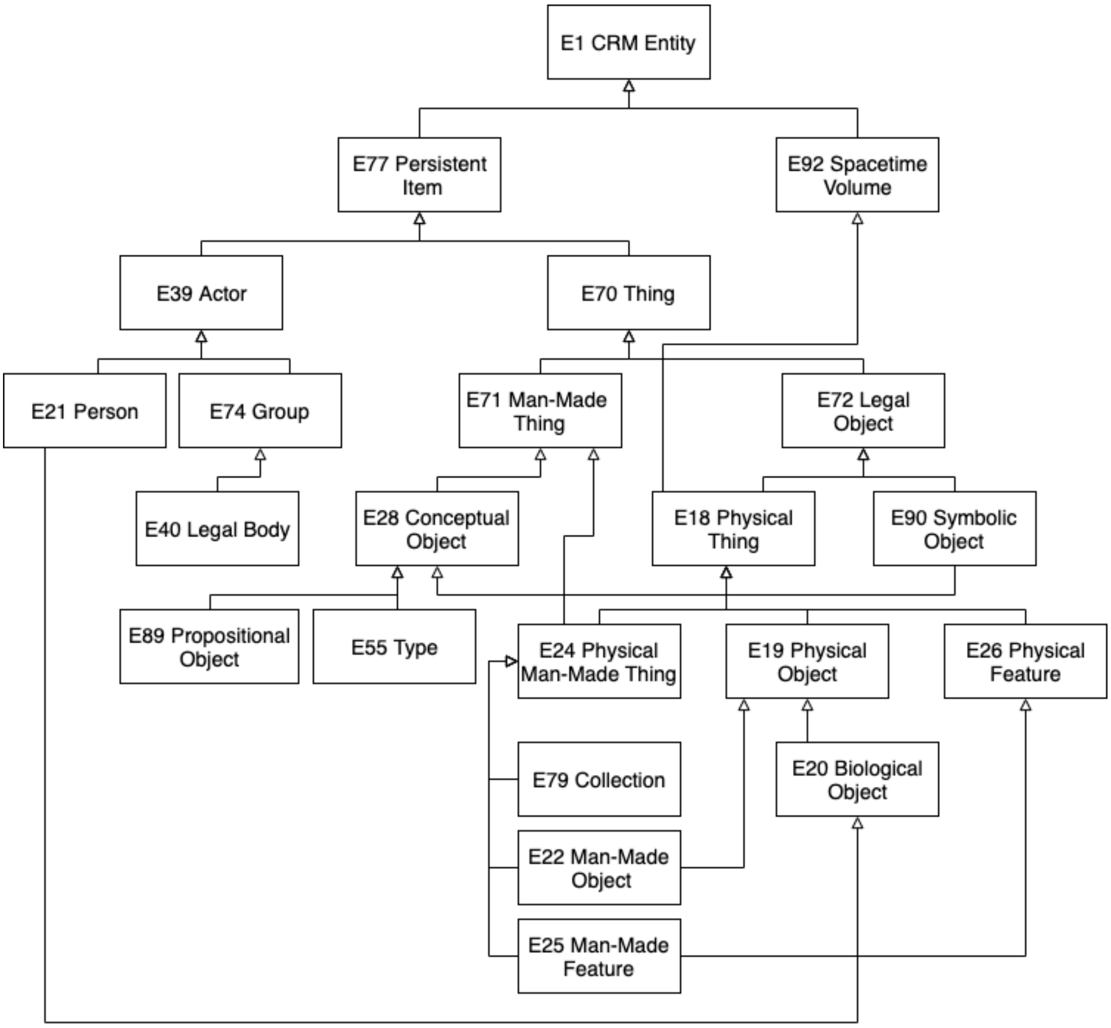

Figure 2. Partial taxonomy of persistent items in CIDOC (v.6.2.1)

To comment on the taxonomy, first, the distinction between *E39 Actor* and *E70 Thing* is not so sharp. Looking at Fig. 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, amongst 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 [5], i.e., properties that entities only possibly satisfy and whose acquisition or loss does not alter their identities. For instance, a human being is subject to legal rights and duties in the scope of a specific socio-legal system, independently from which she always remains a human being for the entire duration of her life. On the other hand, it is reasonable to assume that *E18 Physical Thing* models *rigid* properties, i.e., 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 [10], which extends CIDOC for geo-spatial applications. According to [1], 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 " [1, p.41]. Apart from *E4 Period* (see Sect. 4) 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. " [10].

If we interpret it properly, instances of E92 correspond to *four-dimensional worms* in the sense of ontological four-dimensionalism (4D) [11]. This seems clear from its definition as something that has both temporal and spatial extents but also from the examples in [1,10]; e.g., the fact that an individual spacetime volume can be cut in different parts, each one standing for a spatio-temporal 'snap-shot' of the entity at stake like the Roman Empire during 33 B.C. If this consideration is correct, CIDOC mixes 4D with a standard three-dimensionalism (3D) view.[8] From a foundational perspective, this approach is controversial. Despite the hot debate on 4D and 3D in formal ontology, these remain indeed alternative and perhaps even incompatible positions (see [4] for some discussion). The situation is not better from a modeling perspective, since the benefits of introducing spacetime volumes is unclear. According to [1], 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" [1, 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 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 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.

On the basis of this analysis, Fig. 3 shows the 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 any 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 such as potteries or statues (both

[8] Recall that *E2 Temporal Entity* and *E77 Persistent Item* are disjoint classes.

[9] Following CIDOC's minimality principle (see [1, 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.

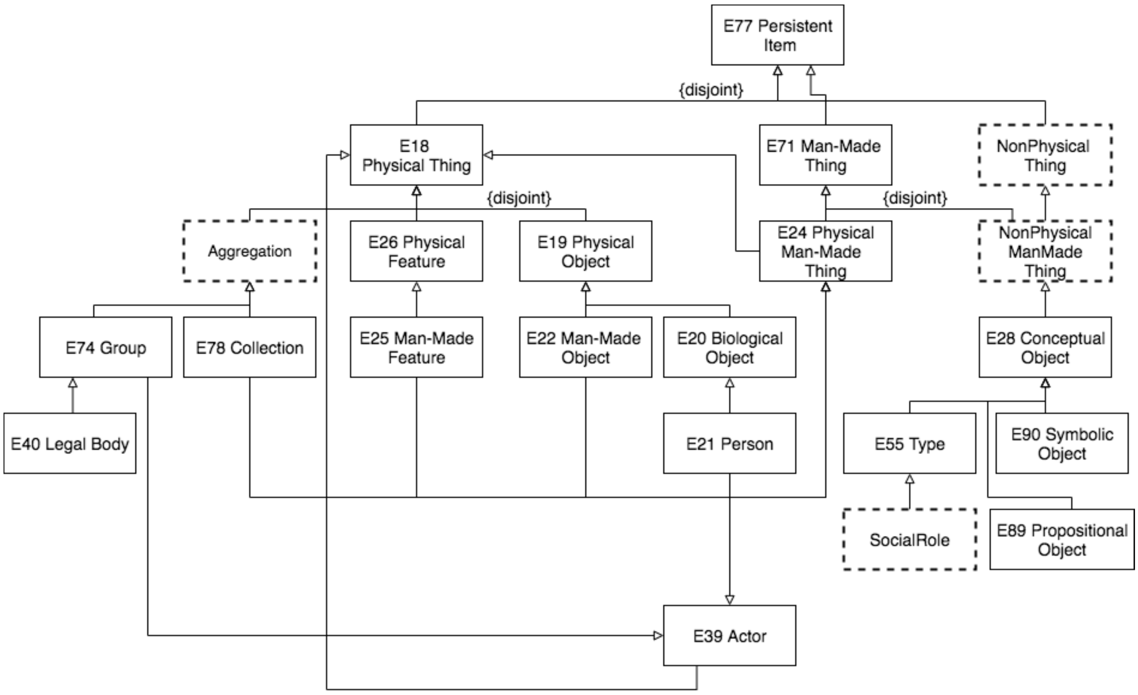

Figure 3. Revised taxonomy of persistent items

instances of *E22 Man-Made Object*). Aggregations bear indeed unity conditions other than topological ones. For instance, according to [1], museum collections, which are represented as specific types of aggregations in Fig. 3, are "assembled and maintained by one or more instances of E39 Actor over time for a specific purpose and audience" [1, 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 collection 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 [12,13]. In these works, the authors distinguish between arbitrary collections of individuals and social groups. In addition, differently from CIDOC, the approach in [12,13] 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.

Finally, by conceiving legal objects as social roles, instances of *E72 Legal Object* can be represented in different ways. A proposal, based on [14], consists in introducing a

[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.

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 [14], 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.[12]

4. Analysis of Temporal Entities

Figure 4 shows the highest classes in CIDOC for the representation of temporal entities. For the sake of the analysis, we limit to show the taxonomic relations between these classes while providing a general overview on their subclasses to facilitate the understanding of the modularization of the ontology presented in Sect. 6.

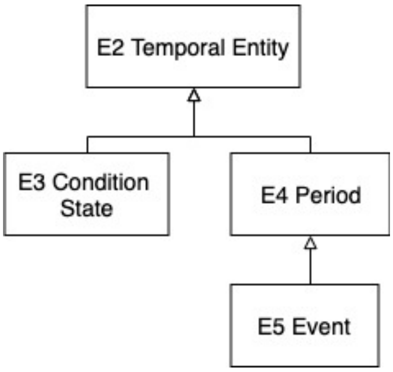

Figure 4. Top-level temporal entities in CIDOC (v.6.2.1)

The class *E3 Condition State* "comprises the states of objects characterized by a certain condition over a time-span" [1, p.3]. An example provided in [1] is the "condition of the SS Great Britain between 22 September 1846 and 27 August 1847 [as being] *wrecked*" (ibid). From a formal ontology perspective, this class matches well with the notion of *state*, e.g., in the DOLCE ontology [7] (e.g., *being sitting*, *being open*, etc.).

E4 Period subsumes all temporal entities other than condition states. It is defined as comprising "sets of coherent phenomena or cultural manifestations occurring in time and space. It is the social or physical coherence of these phenomena that identify a *E4 Period* and not the associated spatiotemporal extent. [...] Often, this class is used to describe prehistoric or historic periods such as the *Neolithic Period*, the *Ming Dynasty* or the *McCarthy Era* [...]" [1, p.3]. E4 subsumes *E5 Event*, whose instances are "changes of states in cultural, social or physical systems, regardless of scale, brought about by a series or group of coherent physical, cultural, technological [...] phenomena" [1, p.5]. E5 directly subsumes *E7 Activity*, i.e., intentional actions performed by actors; *E63 Beginning of Existence*, i.e., events that bring into existence persistent items; and *E64 End*

[12] We model legal object as an individual rather than a class to avoid multiplying roles for specific entities, e.g., the legal-object-role$_1$ of statue$_1$ *vs* the legal-object-role$_2$ of statue$_2$. The reader can refer to [15] for various approaches on the modeling of roles.

of Existence, i.e., events that end the existence of persistent items. These classes are not mutually disjoint (e.g., *E12 Production* is subsumed by both E7 and E63).

Classes like *E66 Formation*, *E66 Dissolution*, *E86 Leaving*, *E85 Joining*, *E67 Birth*, and *E69 Death* are related to actors, in particular, to the formation and dissolution of groups, to persons leaving and joining groups, and to persons' birth and death, respectively. *E11 Modification* and *E65 Creation* are related to the production of physical man-made things and conceptual objects, respectively. *E6 Destruction* models intentional or natural events that destroy physical things. Instances of *E81 Transformation* are events resulting in the destruction of a persistent item and the creation of another item which is different in both nature and identity in comparison to the destroyed one. *E13 Attribute Assignment* concerns the attribution of properties to entities; among its subclasses, it covers measurement events. Finally, *E9 Move*, *E10 Transfer of Custody*, *E8 Acquisition*, and *E87 Curation Activity* are specific to the cultural heritage domain; e.g., they can be useful to describe the transfer of ownership of goods from one museum to others.

Let us now comment, in particular, on the notions of *E4 Period* and *E5 Event*. A first issue is that E4 captures temporal phenomena bearing a cultural nature (e.g., Italian Renaissance, Cubism, etc.). Instances of E5, however, are not necessarily relevant from a cultural standpoint according to CIDOC (see, e.g., the class *E6 Destruction* in [1]). The subsumption of E5 under E4 is therefore misguided. A second issue concerns the mereological structure of periods and events. At first glance, instances of E4 are complex temporal entities consisting of multiple (temporal) parts. At the same time, CIDOC does not take any explicit commitment on the structure of events, which can be either complex or atomic (see [1, p.3]). This is unfortunate because if periods are complex, considering the subsumption of E5 under E4, it cannot be the case for events to be atomic.[13]

On the basis of these considerations, we propose to detach the classes E4 and E5, and to subsume the latter directly under *E2 Temporal Entity*. In this perspective, E5 is a general umbrella for temporal entities that are neither condition states nor periods. A mereological relation of *parthood* between temporal entities can be used to model atomic and complex temporal phenomena (see, e.g., [7]). Finally, E4, E5, and E3 are disjoint.

5. Analysis of Dimensions

The class *E54 Dimension* is directly subsumed by *E1 CRM Entity* (see Fig. 1) to capture "quantifiable properties that can be measured by some calibrated means and can be approximated by values, i.e. points or regions in a mathematical or conceptual space, such as natural or real numbers, RGB values etc." [1, p.26]. The relationship *P43 has dimension* links things to dimensions; *P90 has value* relates dimensions to numeric values, whereas *P91 has unit* models the link between a dimension and its measurement unit, the latter being represented via *E58 Measurement Unit*, a subclass of *E55 Type*.

From a formal ontology perspective, CIDOC's dimensions correspond to a restricted understanding of *qualities* in foundational ontologies like DOLCE or UFO, 'restricted' because limited – at first glance – to classes of qualities for sizes, e.g., lengths or widths.

[13] It should be noted that the distinction between events and periods is partially a question of scale of observation: "Viewed at a coarse level of detail, an *E5 Event* is an instantaneous change of state. At a fine level, the *E5 Event* can be analysed into its component phenomena within a space and time frame, and as such can be seen as a *E4 Period* [1, p.4] (emphasis is ours). CIDOC however lacks a framework to handle granularity.

Also, similarly to these ontologies, CIDOC assumes that a dimension characterizes a single entity. In addition, a dimension can have exactly one value. It is not however clear whether changes in dimensions' values affect changes in their identities.

A drawback in the CIDOC's conceptualization of dimensions is the restriction of their values to numerical terms only, whereas one may wish to represent also *qualitative* values.[14] For instance, representing a man-made object's color, one may wish to say that it is red without specifying its exact shade in quantitative terms. Our proposal is to revise CIDOC on the basis of the work done in [7,16], therefore, by allowing for the representation of dimensions' qualitative values, too. This is done by introducing the class *Qualitative Quality Space*, which provides a way to organize and represent qualities' values in terms of, e.g., mereological or topological structures, among others (see the `cidoc:dimension-module` described in Sect. 6).

6. Towards the Modularization of CIDOC

We discuss in this section a preliminary modularization of CIDOC; we do not cover the entire input ontology and future work in this regard is required. By the end of the section, we present examples about cultural heritage data modeling showing the (potential) advantages of using CIDOC in different inter-connected modules.

Before presenting the modular structure, let us recall some core ideas about ontology modularization. Following [17] "ontology modularization can be interpreted as decomposing potentially large and monolithic ontologies into (a set of) smaller and interlinked components (modules)." An ontology module M corresponds to "[...] a subset of a source ontology O, $M \subset O$, either by abstraction, removal or decomposition, or module M is an ontology existing in a set of modules such that, when combined, make up a larger ontology" [18]. Also, despite the amount of research work, at the current state of the art "*there is no universal way to modularize an ontology*" [19] (emphasis is ours). Hence, according to the same authors, "the choice of a particular technique or approach should be guided by the requirements of the application or scenario relying on modularization" (see [18] for similar considerations in a more recent publication).

For our application and research purposes the modularization of CIDOC is primarily aimed at facilitating its *selective* use. For example, when modeling (social) groups, one may be interested in their members without necessarily describing the events by which the groups are created (or destroyed). Similarly, when working with man-made objects, one may wish to represent only their physical structure without necessarily relating them to temporal information. Because of usability requirements, we rely on Semantic Web (SW) languages, namely, the Web Ontology Language (OWL). Recall that OWL is indeed the leading formalism for the exploitation of ontologies in the Digital Humanities (see, e.g., [20]). In addition, by using OWL, we aim at enhancing the (computational) representation of the ontology. For this purpose, we reuse and (partially) revise the Erlangen release of CIDOC,[15] which formalizes the latter (version 6.2.1) in OWL.

In addition to usability criteria, the modularization of the ontology has been driven by *functional* and *subject similarity* considerations between its various modeling elements. Accordingly, we group classes (and relations) which are aimed at a common goal

[14] This is a further restriction of CIDOC in comparison to DOLCE or UFO.

[15] `https://github.com/erlangen-crm/ecrm`, last accessed in March 2020.

(e.g., facilitating the integration of other modules) or at covering the same portion of reality. For example, considering persistent items (see Fig. 3), one can distinguish between physical things that are not man-made (*Aggregation*, *E19 Physical Object*, and *E26 Physical Feature*) from their man-made counterparts. On the same lines, looking at temporal entities, one can identify and distinguish between, e.g., events concerning the creation or destruction of man-made things (e.g., *E11 Modification* and *E6 Destruction*, among others), and similar events about actors (e.g., *E67 Birth*, *E69 Death*, etc.).

Moving to the technique for the modularization, Kahn and Keet [18] present various automatic approaches based on computational techniques. We have adopted a *manual* approach (an option discussed in [18] as well), because, as a result of the analysis presented in the previous sections, we modularize but also revise CIDOC. We therefore need to look at its conceptual and formal structure and change it wherever necessary.

At the current development stage, the modular architecture comprises 18 modules including the module called `cidoc:whole` which is the union of all modules used to build the whole ontology.[16] For data organization in, e.g., RDF triplestores, this module should be always imported for first to guarantee the integration and interoperability of data instantiating the other modules. For the sake of shortness, we provide here only a general overview of the modules; Tables 1 – 4 give a schematic view on the entire library, including the structure of imports (`owl:imports`).

Besides *E92 Spacetime Volume*, which has been removed, all classes in Fig. 1 constitute the `cidoc:top-module`. This also includes the new class *Qualitative Quality Space* (see below) to represent non-numerical dimensions' values (e.g., the space of weights having values such as *heavy*, *medium*, *light*, etc.). The purpose of the `cidoc:top-module` is to represent the highest classes of the ontology to allow for the consistent integration of all other modules; e.g., to guarantee the disjointness between persistent items and time-spans when these are integrated.

Table 1. General modules

Module name	**Goal**	**Direct imports** (*owl:imports*)
`cidoc:top-module`	To represent the highest classes of the ontology to allow for the consistent integration of all other modules	–

`cidoc:whole module`	The union of all modules in the CIDOC's library	`cidoc:top-module;`
		`cidoc:persistent-item-whole-module;`
		`cidoc:temporal-entity-whole-module`

We spend some words on the `cidoc:dimension-module` to explain its differences with the standard CIDOC. First, the module covers the classes *E54 Dimension*, *Qualitative Quality Space*, and *E77 Persistent Item*; the latter is used to characterize dimensions in relation to E77's instances. For instance, one may characterize a pottery as bearing a color-dimension with value *black*, the latter being a region within a space for colors. Note that the intended meaning of *Qualitative Quality Space* is more restricted than the notion

[16] The library of CIDOC's modules is available at: `https://github.com/emiliosanfilippo/cidoc-modularization`. The repository also contains some diagrams to facilitate the understanding of the modular architecture.

Table 2. Modules about places and dimensions

Module name	**Goal**	**Direct imports** (*owl:imports*)
`cidoc:place-module` module	To represent places (*E53 Place*)	–
`cidoc:dimension-module module`	To represent dimensions (e.g., *E54 Dimension*, *Qualitative Quality Space*)	–

of *quality space* in [7], where the authors use such spaces for both qualitative and quantitative values. In our case, the latter are simply represented through OWL *data properties* and their *value spaces* (e.g., integers) to express numerical values. This approach weakens the expressivity of the ontology in comparison to [7] (e.g., we can not say that 8kg is a value within a space for weights), but it takes the benefits of a Description Logic based formalism to model quantitative dimensions' values. In addition, end-users can introduce data properties like *hasWeightInKg* to characterize the intended meaning of numerical values attached to dimensions (see [16]). With this approach, differently from the original spirit of CIDOC, dimensions can be now characterized in terms of either quantitative or qualitative values.

The taxonomy of persistent items (see Fig. 3) is split into 6 modules, see Table 3. Since the taxonomy covers both physical and non-physical entities, man-made and non-made-made entities, the `cidoc:persistent-item-top-module` is created to provide the most general classes and, therefore, to facilitate the consistent integration of more specific modules. Also, this module is (indirectly) imported by all modules about persistent items besides the `cidoc:concept-module`.[17]

Table 3. Modules about persistent items

Module name	**Goal**	**Direct imports** (*owl:imports*)
`cidoc:persistent-item-top module`	To integrate modules about persistent items	–

`cidoc:physical-thing-module`	To represent non-man-made physical things (e.g., *E19 Physical Object*)	`cidoc:persistent-item-top-module;`
		`cidoc:place-module`

`cidoc:artifact-module`	To represent physical man-made entities (e.g., *E22 Man-Made Object*)	`cidoc:physical-thing-module`
`cidoc:actor-module`	To represent actors (e.g., *E21 Person*, *E74 Group*)	`cidoc:physical-thing-module`
`cidoc:concept-module`	To represent non-physical conceptual entities (e.g., *E28 Conceptual Object*)	–
`cidoc:persistent-item-whole-module`	The union of all persistent items modules	All modules about persistent items

The modular architecture of temporal entities is organized in 8 modules, see Table 4. The `cidoc:temporal-entity-top-module` covers the most general classes for tem-

[17] The design of the `cidoc:concept-module` is incomplete because further work on the analysis of conceptual entities is required.

poral entities plus the direct subclasses of *E5 Event*, i.e., *E7 Activity*, *E63 Beginning of Existence*, and *E64 End of Existence*, as well as *E52 Time Span*. This module is imported by all modules about temporal entities to guarantee their consistent integration. Looking at the table, note that modules about temporal entities import modules about persistent items. Following [21], an alternative approach would consist in splitting between persistent items and temporal entities, and creating *bridging modules* for their integration. We avoid this approach, first, to keep a simple modular architecture and to avoid the proliferation of modules, second because the representation of temporal entities in cultural heritage scenarios often requires the representation of their participants (see, e.g., [10]).

Table 4. Modules about temporal entities

Module name	**Goal**	**Direct imports** (*owl:imports*)
`cidoc:temporal-entity-top module`	To integrate modules about temporal entities	`cidoc:persistent-item-top-module`; `cidoc:place-module`
`cidoc:actor-activity-module`	To represent activities related to the life of individual actors or groups (e.g., *E67 Birth*, *E68 Dissolution*)	`cidoc:temporal-entity-top-module`; `cidoc:actor-module`
`cidoc:attribute-assignment-activity-module`	To represent activities for attributes assignment (e.g., *E16 Measurement*)	`cidoc:temporal-entity-top-module`
`cidoc:creation-activity-module`	To represent the creation of conceptual objects (e.g., *E65 Creation*)	`cidoc:temporal-entity-top-module`; `cidoc:concept-module`
`cidoc:cultural-heritage-activity-module`	To represent temporal entities relative to cultural heritage (e.g., *E87 Curation Activity*	`cidoc:temporal-entity-top-module`; `cidoc:actor-module`
`cidoc:modification-activity-module`	To represent the production, modification or destruction of physical entities (e.g., *E79 Part Addition*, *E6 Destruction*)	`cidoc:temporal-entity-top-module`
`cidoc:move-activity- module`	To represent movements of physical objects (*E9 Move*)	`cidoc:temporal-entity-top-module`
`cidoc:temporal-entity-whole module`	The union of all temporal entities modules	All modules about temporal entities

Let us now add some comments. First, CIDOC employs relations which contain disjunctive terms. An example is *P53 has former or current location* between *E18 Physical Thing* and *P53 Place*. This subsumes the relation *P55 has current location* whereas no counterpart for *has former location* is available. From a semantic perspective, the meaning of *having former location* is not the same as *having current location*. It is therefore unclear why a unique modeling element is used, since a relation like P53 can easily lead to misunderstandings. In the ontology modules, we have not reused CIDOC's relations employing disjunctions; rather, we have split each of these relations in further relations while maximizing the reuse of existing elements (e.g., we reuse P55 but not P53).

Second, as said in Sect. 2, CIDOC relies on temporal entities to represent information about persistent items such as birth dates. A similar position is adopted in ontologies like DOLCE or UFO. From a data modeling perspective, however, this approach forces users to create entities which may not be required. Our proposal is to introduce *shortcuts* to enhance data modeling tasks, a strategy which is adopted by CIDOC itself [1]. For example, a new binary predicate *createdAt(o,d)* between a physical man-made object and its production date can be defined (in FOL) as in (Def1), where all defining predicates belong to the CIDOC's signature.[18]

Def1 $createdAt(o,d) \equiv PhysicalManMadeThing(o) \wedge Date(d) \wedge \exists e,t(Production(e) \wedge hasProduced(e,o) \wedge hasTimeSpan(e,t) \wedge identifiedBy(t,d))$

Because of expressivity restrictions, definitions similar to (Def1) can not be employed in SW ontologies. One can however use OWL data properties – possibly by importing them from existing SW vocabularies – while characterizing their formal interpretations in external FOL theories.[19] Following this consideration, we have included in the modules some data and object properties to facilitate data representation.

A third observation is about CIDOC's use of *appellations* (e.g., names, dates) and *primitive values* (strings, numbers). As a formalism-independent model, the relevance of these elements can not be dismissed. When choosing a specific formalism, however, they need to be handled with care (see, e.g., [3]). In the case of OWL, it is reasonable to rely on data types and data properties to handle primitive values and appellations, respectively, rather than representing them as domain instances as it is done in existing OWL releases of CIDOC such as the Erlangen release (see above for references). In this way, one can rely on value spaces to characterize values' meanings and can enable the use of algorithmic procedures to manipulate data (e.g., the use of regular expressions on strings or arithmetic operations on numbers). A deeper analysis of appellations is however required to strengthen their representation.

Finally, the use of cardinality restrictions and axioms in CIDOC deserve attention. For example, physical things are characterized by material types in both [1] (therefore in [3]) and the Erlangen formalization; see (Ax1) for a representation in FOL.[20]

Ax1 $PhysicalThing(x) \rightarrow \exists y(Material(y) \wedge consistsOf(x,y))$

Considering that *E18 Physical Thing* subsumes *E26 Physical Feature*, (Ax1) is misguided, at least if CIDOC understands features like holes as *immaterial* entities (as it seems). Hence, we have not included in the modules the entirety of axioms that are present in the CIDOC-Erlangen; further work on their analysis and the analysis of CIDOC's cardinalities is required.

As a first example, let us assume that we need to represent museological data about statues. These can include data about statues' dimensions, creators, creation dates, material types, identifiers, and the museums where they are preserved. To represent these data in our framework it is sufficient to use the `cidoc:artifact-module` and the `cidoc:actor-module`. The former contains the basic modeling elements for statues, whereas the latter is required to represent the statues' creators. Hence, differently from the current release of CIDOC, end-users can now exploit the ontology by reusing only the modules that are relevant for their tasks. In addition, as said, in a data modeling

[18] For simplicity, we omit CIDOC's identifiers. Also, looking at (Def1) some unary predicates can be derived from relations' domain/range restrictions. We include them to facilitate the understanding of the formula.

[19] Recall that the Distributed Modeling Language (DOL) [22] can be used to handle and link alternative formalizations of the same conceptual model.

[20] Looking at (Ax1), *Material* stands for material types and not for amounts of matter in sense of, e.g., [7].

scenario one may not desire the explicit representation of temporal phenomena like the production events leading to the statues and their time-regions. Although ontologically coherent, this approach would lead to verbose data models at the expenses of computational resources. By introducing shortcuts on the line of (Def1) we can link statues to their creation dates and creators while keeping a simple data representation.

As a second example, we consider the design of a domain-specific ontology based on CIDOC. OpenArchaeo is a semantic mediator for archaeological datasets currently hosted by the French infrastructure Huma-Num.[21] It interconnects multiple datasets by using an ontology dedicated to archeology [8,23], which is based on CIDOC plus some of its extensions, e.g., CRMsci[22] and CRMba,[23] among others. One of the most relevant classes is the event of (archeological) site discovery represented by *S19 Encounter Event* from CRMsci, which is a subclass of *S4 Observation*, the latter subsumed by *E13 Attribute Assignment*. The site discovery event (*i*) is carried out by a *E21 Person* who is member of a *E40 Legal Body*; (*ii*) took place on a *E27 Site*, which has a place as location; (*iii*) is linked to a *E52 Time-Span* with dates; and (*iv*) found some artifacts. This ontology was developed by taking into account the whole CIDOC, whereas with our approach one would require the `cidoc:actor-module`, the `cidoc:artifact-module`, and the `cidoc:attribute-assignment-activity-module` including both their imported modules and the `cidoc:top-module`, the latter used to consistently integrate all modules. In principle, from an ontology design perspective, the selective reuse of CIDOC could facilitate the development of the ontology, since one would not need to go through its entire taxonomy. For end-users, this may also facilitate the understanding of the ontology, since many of CIDOC's modeling elements would be left out.

7. Conclusions

In order to foster the use of ontologies for knowledge representation and data management in the area of cultural heritage, we presented in the paper a first ontological analysis and modularization of the CIDOC ontology. We focused on the latter because of its wide use in both research projects and institutions. Our contribution is twofold: first, by analysing CIDOC, the goal is to enhance and make transparent its ontological commitment. As a result, we have proposed to remove some classes from the ontology and to introduce some new modeling elements. Second, by modularizing it, the purpose is to facilitate its selective reuse, maintenance, and extension with domain-specific modules.

Future work to strengthen our proposal is required. First, both the analysis and modularization have to be extended to the whole ontology, conceptual objects and relations included. The analysis of relations requires a careful evaluation of their cardinalities to check whether these are consistent with the intended meaning of the related classes. Second, a testing benchmark is necessary to evaluate both the ontology resulting from the analysis and its modular architecture. From a usability perspective, we plan to exploit the ontology modules in research projects and to test their impact on data management practices. Finally, a stable formalization of CIDOC in a language like FOL is a desiderata to unambiguously characterize its elements. This could be based on the work presented

[21] `http://openarchaeo.huma-num.fr/explorateur/sourcesSelect`, last accessed in March 2020.
[22] `http://www.cidoc-crm.org/crmsci/`, last accessed in March 2020.
[23] `http://www.cidoc-crm.org/crmba/`, last accessed in March 2020.

in [3] possibly revised and extended by the work we presented. This formalization could be then used as a foundational basis for the computational treatment of the ontology.

References

[1] 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;.

[2] Bruseker G, Carboni N, Guillem A. Cultural heritage data management: the role of formal ontology and CIDOC CRM. In: Heritage and Archaeology in the Digital Age. Springer; 2017. p. 93–131.

[3] Meghini C, Doerr M. A first-order logic expression of the CIDOC Conceptual Reference Model. International Journal of Metadata, Semantics and Ontologies. 2018;13(2):131–149.

[4] Wahlberg TH. The endurance/perdurance controversy is no storm in a teacup. Axiomathes. 2014;24(4):463–482.

[5] Guarino N, Welty CA. An overview of OntoClean. In: Staab S, Studer R, editors. Handbook on ontologies. Springer; 2009. p. 201–220.

[6] Guizzardi G. Ontological foundations for structural conceptual models; 2005.

[7] Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A. WonderWeb Deliverable D18. Laboratory for Applied Ontology ISTC-CNR; 2003.

[8] Marlet O, Francart T, Markhoff B, Rodier X. OpenArchaeo for usable semantic interoperability. In: Proceedings of the 1st International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH19). vol. 2375. CEUR; 2019. p. 5–14. Available from: http://ceur-ws.org/Vol-2375/paper1.pdf.

[9] Moraitou E, Aliprantis J, Christodoulou Y, Teneketzis A, Caridakis G. Semantic Bridging of Cultural Heritage Disciplines and Tasks. Heritage. 2019;2(1):611–630.

[10] Hiebel G, Doerr M, Eide Ø. CRMgeo: A spatiotemporal extension of CIDOC-CRM. International Journal on Digital Libraries. 2017;18(4):271–279.

[11] Sider T. Four Dimensionalism. An Ontology of Persistence and Time. Oxford University Press; 2001.

[12] Bottazzi E, Ferrario R. Preliminaries to a DOLCE ontology of organisations. International Journal of Business Process Integration and Management. 2009;4(4):225–238.

[13] Porello D, Bottazzi E, Ferrario R. The Ontology of Group Agency. In: Proc. of the 8th Int. Conf. on Formal Ontology in Information Systems. IOS Press; 2014. p. 183–196.

[14] Masolo C, Vieu L, Bottazzi E, Catenacci C, Ferrario R, Gangemi A, et al. Social roles and their descriptions. In: Principles of Knowledge Representation and Reasoning. AAAI Press; 2004. p. 267–277.

[15] Masolo C, Guizzardi G, Vieu L, Bottazzi E, Ferrario R, et al. Relational roles and qua-individuals. In: AAAI Fall Symposium on Roles, an interdisciplinary perspective. AAAI PRESS-MIT PRESS; 2005. p. 103–112.

[16] Sanfilippo EM. Feature-based product modelling: an ontological approach. International Journal of Computer Integrated Manufacturing. 2018;31(11):1097–1110.

[17] Ben Abbès S, Scheuermann A, Meilender T, D'Aquin M. Characterizing Modular Ontologies. In: Proc. of the 7th Int. Conf. on Formal Ontologies in Information Systems. IOS Press; 2012. p. 13–25.

[18] Khan ZC, Keet CM. An empirically-based framework for ontology modularisation. Applied Ontology. 2015;10(3-4):171–195.

[19] dAquin M, Schlicht A, Stuckenschmidt H, Sabou M. Criteria and evaluation for ontology modularization techniques. In: Modular ontologies. Springer; 2009. p. 67–89.

[20] Carriero VA, Gangemi A, Mancinelli ML, Marinucci L, Nuzzolese AG, Presutti V, et al. Arco: the italian cultural heritage knowledge graph. In: International Semantic Web Conference. Springer; 2019. p. 36–52.

[21] Rector A, Brandt S, Drummond N, Horridge M, Pulestin C, Stevens R. Engineering use cases for modular development of ontologies in OWL. Applied Ontology. 2012;7(2):113–132.

[22] Mossakowski T, Codescu M, Neuhaus F, Kutz O. The distributed ontology, modeling and specification language–DOL. In: The road to universal logic. Springer; 2015. p. 489–520.

[23] Marlet O, Rodier X. A way to express the reliability of archaeological data: data traceability at the Laboratoire Archologie et Territoires (Tours, France). International Journal on Digital Libraries. 2019;.