Applied Ontology 16 (2021) 111–135
DOI 10.3233/AO-210246
IOS Press

Ontologies for information entities: State of the art and open challenges

Emilio M. Sanfilippo
Laboratory for Applied Ontology (ISTC-CNR), via alla cascata 56/C, 38123, Povo, Trento, Italy
E-mail: emilio.sanfilippo@cnr.it

Abstract. Information entities are used in ontologies to represent engineering technical specifications, health records, pictures or librarian data about, e.g., narrative fictions, among others. The literature in applied ontology lacks a comparison of the state of the art, and foundational questions on the nature of information entities remain open for research. The purpose of the paper is twofold. First, to compare existing ontologies with both each other and theories proposed in philosophy, semiotics, librarianship, and literary studies in order to understand how the ontologies conceive and model information entities. Second, to discuss some open research challenges that can lead to principled approaches for the treatment of information entities, possibly by getting into account the variety of information entity types found in the literature.

Keywords: Foundational ontology, information entity, YAMATO, DLP, IAO, CIDOC-CRM, FRBR

Accepted by: Veda Storey

1. Introduction

Applied ontologies often need to represent things like narrative fictions, musical scores, anagraphic records, images or engineering specifications like design models or process plans. One of the peculiarities of these entities is that they can exist in different supports without reducing to them. For instance, from the fact that John's copy of Dante's *Divine Comedy* (*Comedy* from now on) is 1 kg heavy, we do not conclude that Dante's poem is 1 kg heavy. It is also commonly claimed that people read the *same* poem when they read the *Comedy* but in alternative translations. The underlying idea is that a poem can be present in multiple supports (e.g., paper-made or digital books) while being conveyed by alternative texts in different languages. A similar consideration can be done, e.g., for a design model that exists as both a computer file stored on a usb stick and a blueprint on paper. Considering these examples from an ontological stance, one wonders about what kind of entities narrative fictions, engineering specifications, etc. are and how one can make sense of ordinary or experts' talks about them. For instance, is it possible to distinguish a poem from its text or are they the same thing?

These concerns have been widely addressed in applied ontology (Bateman, 2019; Smith and Ceusters, 2015; Mizoguchi, 2004), as well as in linguistics (Arapinis and Vieu, 2015), philosophy (Goodman and Elgin, 1986; Thomasson, 2015), semiotics (Eco, 2016), literary (Eggert, 2019; Pierazzo, 2016; Shillingsburg, 2010) and librarianship studies (Smiraglia, 2001). The terminology is heterogeneous and for the sake of clarity we will talk of *information entities* to refer to, e.g., what two copies of the *Comedy* share, i.e., the entity which is intuitively called the *content* of the copies. As we will see, information entities have been differently conceived in ontologies and, as a consequence, there is no guarantee that ontologies

refer to the same things, not even when they use similar terminologies. Nor there is a study comparing existing modeling alternatives to analyze their dis-/similarities, advantages, and shortcomings.

We present in the paper a first step within a larger research effort aimed at digging into the ontological foundations of information entities. In particular, we present a review of the state of the art that compares existing ontologies with both each other and theories proposed in other domains. These latter studies prove useful to both understand how information entities have been conceived out of the scope of applied ontology and provide conceptual tools for analyzing the ontologies. To carry out the analysis, we adopt a descriptive approach to understand the commitments and modeling choices of the ontologies. When comparing the ontologies, we discuss their differences and similarities by stressing the consequences of their modeling patterns. Finally, on the basis of the review, we present some open challenges for research. Note that we primarily focus on information entities like narrative fictions although, by the end of the paper, we will make considerations which are hopefully applicable to information entities at large.

The paper is structured as follows. Section 2 presents some theories in philosophy that are relevant in the scope of our investigation, whereas we report in Section 3 about librarianship and literary studies.[1] In Section 4 we consider the way in which some ontologies conceptualize and represent information entities. We focus on ontologies that either contextualize information entities within foundational frameworks or specifically target their representation. We compare the ontologies in Section 5 and discuss some open challenges for research in Section 6. Section 7 concludes the paper.

2. *Works* in philosophy

The ontological nature of literary and musical *works* (we will simply say works when the context is clear) is a central topic in the philosophical speculation about art (Thomasson, 2004; Livingston, 2016). Differently from items like statues or potteries, works exhibit indeed properties that make them quite peculiar entities. For instance, if the specific spatio-temporal object that is Michelangelo's *David* is destroyed, humankind has lost one of its greatest masterworks. Differently, a work can have multiple copies or performances; hence, Dante's *Comedy* does not stop existing when one of its copies is destroyed, and it exists as long as some of its copies exist, or – someone says – as long as it exists in someone's memory (Thomasson, 1999, 2004). The difference in the persistence conditions of copies and works is a common argument to claim for their ontological difference. A similar consideration can be done for music. According to some philosophers, a musical work can be realized in performances happening at specific times and places, although, first, this is not necessarily the case; second, a musical work outlasts any of its performances. Thomasson (2004), for instance, argues that "[n]or can the work of music (or literature, drama, or dance) be identified with the totality of all such performances (or copies), since that would entail, for example, that the work is not complete until long after the composer's death, when the last performance is finished, and that the work itself would have been different had last night's performance been cancelled" (p. 82).

There exist various proposals about the ontological status of works (Davies, 2007; Livingston, 2016; Thomasson, 2004, 2015). Contemporary Platonists claim that works are universal entities lacking spatio-temporal dimensions and existing everlastingly. In this account a musical work exists even if no performance or score-copy are ever produced, and an artist discovers rather than creates a work.[2] An objection to this view is that works – as they are commonly understood – are human-made entities which would

[1] From a terminological perspective, experts of these domains talk of *works* to refer to information entities.

[2] The assumption is that, differently from creation, something can be discovered only if it already exists.

not exist if they were not intentionally created. An alternative position, called quasi-Platonism by Goehr (1992), argues that works retain some aspects of their Platonic status (e.g., they can be instantiated) while being created and being therefore temporally bounded (Goehr, 1992, p. 14 and pp. 44–68). In this view, works are *types* whose instances can be performances in the case of musical works or physical copies in the case of literary works. On a similar line, Thomasson (2004) considers works as *abstract particulars* (also called *abstract artifacts*). The difference between quasi-Platonic types and abstract particulars is not relevant for our purposes. It is sufficient to note that in Thomasson's view, works are particulars rather than types because they are created; and they are abstract because they lack physical dimensions.[3] In addition, both Thomasson (2004) and authors committed to quasi-Platonism like Levinson (1980) and Margolis (1974) stress the social dimensions of works arguing that they can not be disentangled from intentional, cultural, and historical properties. Finally, idealist philosophers claim that works are ideas in the mind of their creators that "[. . .] once formed, find objectified expression [in the case of music] through score-copies or performances and are, thereby, made publicly accessible" (Goehr, 1992, p. 18). In this view, however, "[w]orks are not identified with the objectified expressions, as one might expect them to be, but with the ideas themselves" (Goehr, 1992, p. 18). A common criticism is that by identifying works with private ideas, they lose their public dimension, which is commonly experienced in ordinary life. For example, a person listening – in principle – to a concert of Beethoven's *Symphony No.9* does not in fact listen to one of Beethoven's masterpieces, because the latter existed only in Beethoven's mind. Because of this and other criticisms, according to Goehr (1992) the idealist position has not been well received in the landscape of analytic philosophy. It has nevertheless received attention in other domains, as we will see.[4]

Before moving to the next section, we spend some words on the ontology of literary works. A fundamental problem addressed in philosophy is the relation between a work and its text, namely, whether the two entities coincide or need to be distinguished. Goodman and Elgin (1986) claim that a work is a text, the latter understood as a linguistic structure consisting of a configuration of spaces, punctuation marks, and letters in a specific language. The authors also argue that a work does not correspond to the interpretation of a text. This because a single text can be differently interpreted, therefore should a work correspond to interpretations, we would not have a single work but a plethora of works, a consequence that Goodman and Elgin (1986) do not accept. Currie (1991) calls *textualism* this position. Against it, Wilsmore (1987) proposes the example of Virgina Woolf's novel with title *To the Lighthouse*, which was published "in both England and America at the same time with small but important differences in text" (p. 312). Because of these differences, we would have two different works according to Goodman and Elgin (1986). Wilsmore (1987) argues that this is counterintuitive and "[r]eaders on both sides of the Atlantic believe they have read the very same work by Virginia Woolf" (p. 312). What is then a literary work if not its text? Various answers have been given to this question (Davies, 2007; Davies and Matheson, 2008; Thomasson, 2015). Thomasson (1999), for example, distinguishes between *texts*, *compositions*, and *literary works*: the first are – *à la* Goodman and Elgin (1986) – sequences of symbols in a language; the second refer to texts "as created by a certain author in certain historical circumstances" (p. 64); the third are things like novels or poems, among others, which "are not mere strings of symbols but rather require a certain community of individuals with the right language capabilities and background assumptions to read and understand the literary work" (p. 65). Also, Thomasson (1999) argues that "[o]ne and the same composition can serve as the foundation for two different literary works in the

[3]Foundational ontologies like DOLCE (Borgo and Masolo, 2009) and GFO (Herre, 2010) cover *concepts*, which can be understood as both types existing in time and abstract particulars in the sense of Thomasson (2004).

[4]For discussions on further developments of idealism in analytic philosophy, see the paper of Cray and Matheson (2017).

context of different readerships" (p. 65). Therefore, contrary to Goodman and Elgin (1986), Thomasson (1999) considers literary works as interpretations. In the same lines, Shillingsburg (2010) argues that "[w]e should be suspicious of locutions like 'the work itself', for the work exists only in *our construct* of it. While the text and the document [i.e., the texts' support] are clearly material, the work is a *mental construct*" (p. 179; emphasis is ours). This perspective, sometimes called *interpretationism* (Davies, 2007), needs to clarify what interpretations are, e.g., whether it is possible for different agents to share the same interpretation or what are the criteria to judge interpretations (Carroll, 2015).

We will come back to the distinction between works and texts in Section 5 and Section 6.1.

3. *Works* in librarianship and literary studies

Scholars of librarianship studies have looked at the notion of work as a useful concept for cataloging. For instance, when developing bibliographic systems, it is common to classify multiple physical items under a common entity, which happens to be called work indeed. This has brought the community to question the notion of work and to seek for foundational theories supporting cataloging practices.

Smiraglia (2001) provides an overview of multiple theories by tracking at the same time the history of the work concept when used for documentary goals. The author argues that scholars have slowly converged to a similar understanding about what works are. This is summarized in the following definition: "Work is the *intellectual content* of a bibliographic entity; any work has two properties: a) the propositions expressed, which form *ideational content*; and b) the expression of those propositions (usually a particular set of *linguistic* (musical, etc.) *strings*), which form *semantic content*" (Smiraglia, 2001, p. 42; emphasis is ours). To clarify the definition, first, a bibliographic entity is an item (e.g., a book, article, etc.) that a librarian needs to classify. Second, Smiraglia (2001) distinguishes between the *material* and the *intellectual* dimensions of a bibliographic entity. The former is about, let us say, its physicality, e.g., when and where it was printed, on which paper type and format, etc. The latter concerns the notion of work itself, i.e., an intellectual content consisting of ideational and semantic contents. An ideational content is a set of ideas that an author (or a group of authors) wishes to communicate, whereas the semantic content concerns the linguistic expressions (e.g., sentences, strings) by which the ideational content is conveyed. It remains however unclear how ideational contents and expressions' meanings are related, e.g., whether the two entities coincide or whether there remains an ontological gap between them in a sort of idealistic approach (see Section 2).

In the case of literary studies, including those on scholarly editing, Pierazzo (2016) argues that verbal documents like books consist of multiple dimensions, among which the *linguistic dimension* concerns the language in which a document is written, and the *semantic dimension* what its words mean. Pierazzo (2016) claims that "people can read from the same document and understand slightly or radical different things, depending on their culture, their understanding, their disposition, their cirmustances" (p. 42). The notion of work used by Pierazzo (2016) is inherited from the theory presented by Eggert (2009), who conceives it as an *organizing category* that is helpful to classify similar documents. In the words of Eggert (2009), "[a] work retains its function as a *pragmatic agreement* for organising our remembered experiences of reading documents that are closely related bibliographically" (p. 235; emphasis is ours). Following these lines, according to Pierazzo (2016), "it is [. . .] necessary to postulate the existence of such an entity [i.e., a work] in order to account for the fact that we are able to use the label, for example, *Pride and Prejudice*, for many objects that present more or less the same sequence of words even when inscribed onto different documents, using different fonts, over different materials laid out differently

with respect to the first edition which in turn may be represented by many different objects (or items) that instantiate it" (p. 47).

To conclude, the notion of work emerging from this section is that of a category useful to classify resembling bibliographic items. When one goes deeper in the analysis, reference to authors' ideas is sometimes done, as in the case of Smiraglia (2001). Finally, it is interesting to note that the criteria to grasp the similarities between bibliographic items are highly heterogeneous and there is only little agreements among various perspectives (Smiraglia, 2001, ch. 3). For example, when translating a work in a different language, does one obtain a new work which stands in some similarity relations with the translated one, or does one obtain a different text for the same work? Or, when a work is re-published with the addition of a new chapter, which was not included in previous publications, does one create a new work or is this case better understood in terms of *change* in the same entity? In the words of Pierazzo (2016): "How much variations among the different texts and documents can be tolerated before it will be possible to define two different works? When can we speak of two versions of the same work or of two distinct works?" (p. 46). Pierazzo's (2016) reply is that "[. . .] there is not straight answer to this question, as this will depend on the interpretation of the special user that is the editor" (p. 46). Eggert (2009) holds a similar position when claiming that "[. . .] the decision of the scholarly editor that a particular work can tolerate a certain amount of variation before its variant texts and presentations constitute a different one is an interpretative act" (p. 145). Perhaps more radically, "[a] theory of the work that might ground what editors do still does not exist" (Eggert, 2009, p. 228).

4. Ontologies for information entities: State of the art

We review in this section the state of the art in applied ontology about information entities by also mapping the ontologies to the materials presented in the previous two sections. We express critical remarks about each ontology, therefore the reader is invited to go thorough the entire section. Otherwise, one may move to Section 5 where the results of our analysis are summarized and compared.

Section 4.1 considers the representation of information entities in the scope of the Yet Another More Advanced Top-level Ontology (YAMATO). The Information Artifact Ontology (IAO) is considered in Section 4.2, whereas Section 4.3 looks at the treatment of information entities in the scope of DOLCE Lite Plus (DLP). Section 4.4 considers the CIDOC Conceptual Reference Model (CRM) and Section 4.5 the Functional Requirements for Bibliographic Records (FRBR). By reviewing the ontologies, we consider only how they treat information entities; the reader should refer to the ontologies' documentation for broader views on their conceptual structures. Also, to facilitate the understanding of the state of the art, we render some portions of the ontologies in the Unified Modeling Language (UML) Class Diagram notation.

4.1. Yet another more advanced top-level ontology (YAMATO)

Information entities are represented in YAMATO (Mizoguchi, 2004, 2010) through four main classes: *Representation Form*, *Content*, *Representation*, and *Representing Thing*. The first three classes are subsumed by *Semi-Abstract*, i.e., their instances exist in time but not in space. Differently, instances of *Representing Thing* are physical objects. Figure 1 shows a partial view on the taxonomy of YAMATO.

According to Mizoguchi (2010), a representation is a "content-bearing thing. That is, anything which has content as its essentials [. . .]" (p. 9). Each representation consists of a form and at least one content. Contents are *propositions* that someone wishes to covey through a representation, whereas forms

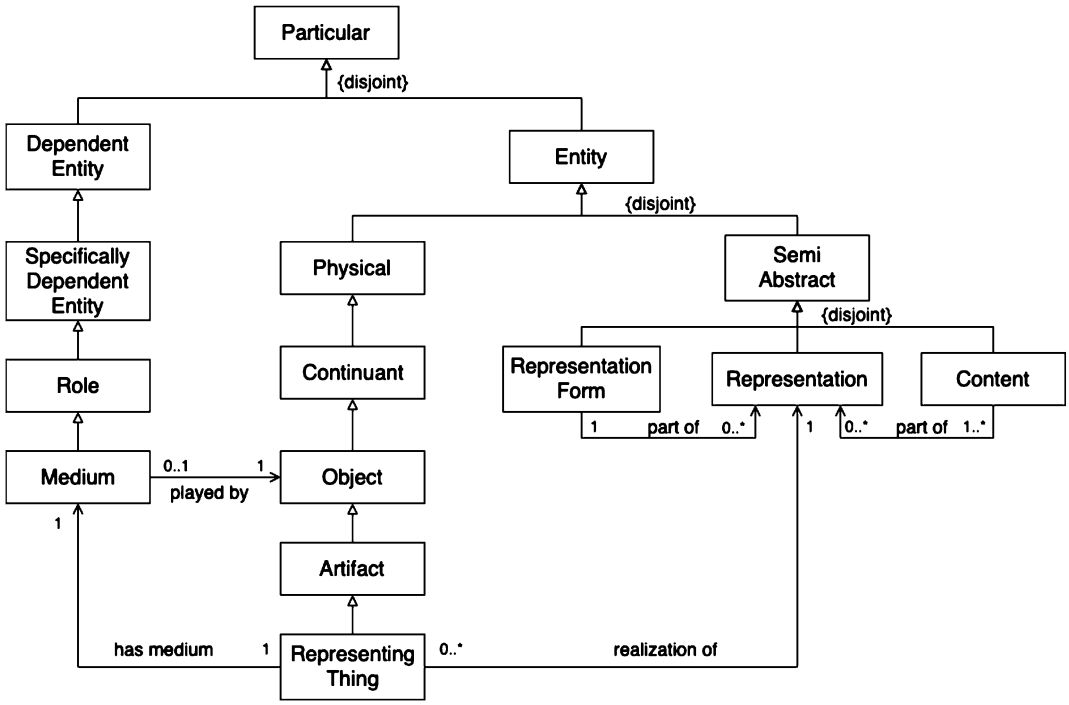

Fig. 1. Partial view on the taxonomy of YAMATO.

are the symbol systems in which contents are conveyed. Examples of contents are plans, symphonies, algorithms, and novels; examples of representation forms are natural or programming languages. Intuitively, a representation 'codes' a content into a form, e.g., it corresponds to the text of (a certain edition of) a novel. Representations are not specific expressions written on individual papers; that is, they are sign-*types* rather than sign-*tokens* (Wetzel, 2018), e.g., the text-type that certain physical copies of the *Comedy* share. A representing thing is composed by a representation and a *representation medium*, the latter being the *role* played by an object which realizes a representation. For example, a physical copy of the *Comedy*, call it *DC copy id 21*, is a representing thing consisting of (i) an individual physical object made of paper and ink (among other materials) with the role of representation medium, and (ii) a representation comprising a narrative (content) expressed in (roughly speaking) Italian (form), see Fig. 2.[5] If the form changes, e.g., when translating the *Comedy* from Italian to English, YAMATO allows users to model different representations sharing the same content while having different forms (see Fig. 3).

Mizoguchi and Toyoshima (2017) present a first-order axiomatization of some of the notions in YAMATO. They introduce an ontological dependence link between forms and contents. In this view, when form f and content c are both part of the same representation r, then c depends on f. Contents, however, are not dependent entities *tout court*, since they are neither necessarily linked to representations, nor to representing things. In the words of Mizoguchi (2010), "[. . .] something exists as content independently of its representation" (p. 10). This raises the question of what is the ontological nature of contents. Mizoguchi (2004, 2010) interchangeably talks of *contents*, *propositions*, and *meanings*. Considering that contents are semi-abstracta, YAMATO seems to hold a position similar to both quasi-Platonism and

[5]To facilitate readability, we use in Fig. 2 the inverse of the *part of* relation shown in Fig. 1.

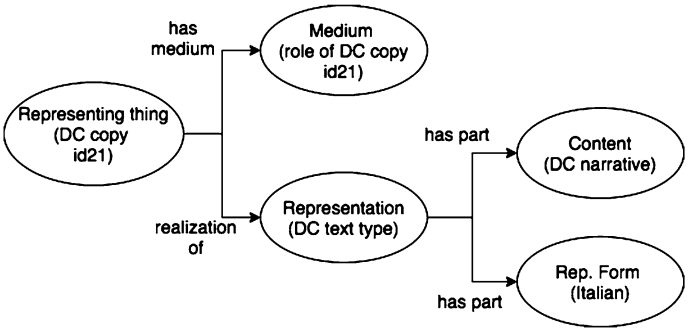

Fig. 2. Modeling the *Comedy* (DC) according to YAMATO.

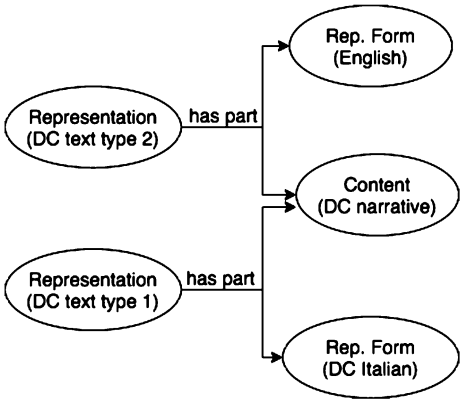

Fig. 3. Modeling translations in YAMATO (the two representations have part the *same* content).

Thomasson (2004), see Section 2. As we have seen, however, Thomasson (2004) argues that an object of art like a work of literature exists as far as at least one of its carriers exists, whereas this condition does not hold for contents in YAMATO. We further comment on YAMATO in Section 5 and Section 6.1.

4.2. Information Artifact Ontology (IAO)

The Information Artifact Ontology (IAO) is developed on the basis of the Basic Formal Ontology (BFO; Arp et al., 2015) and it has the class *Information Content Entity* (ICE) at its grounds. For our purposes, we rely on Arp et al. (2015) and Smith and Ceusters (2015), and the meta-data in the IAO OWL file (IAOwl; Goldstein et al., 2020).[6] Fig. 4 shows the core classes of the IAOwl subsumed by BFO.[7]

According to Smith and Ceusters (2015), an ICE is "an entity that is (1) generically dependent on (2) some material entity and which (3) stands in a relation of aboutness to some entity" (p. 1); examples are novels, journal articles, and graphs, among others. The general idea is to conceive ICEs as non-material

[6]Last accessed July 2020.

[7]Looking at Fig. 4, we slightly revise the UML notation in order to make a clear distinction between IAO's (dotted classes) and BFO's classes. We use classes' labels from the IAOwl rather than their alpha-numerical identifiers. Also, looking at the figure, the *bearer of* and *concretizes* relations simplify the axiom: $MaterialInformationBearer \sqsubseteq MaterialEntity \sqcap \exists bearerOf.\exists concretizes.ICE$.

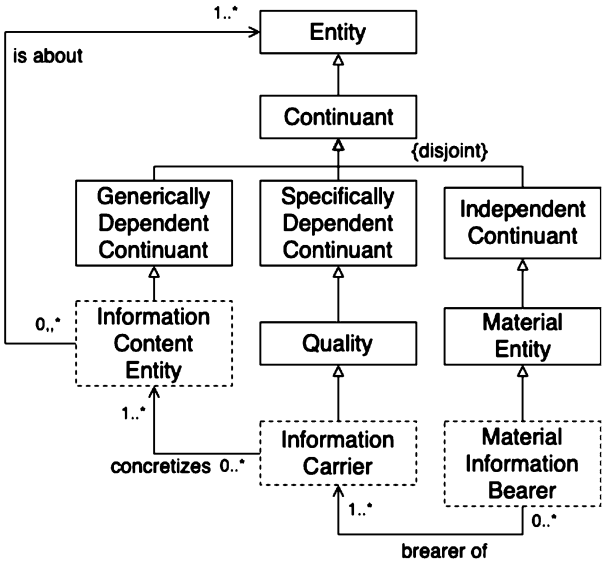

Fig. 4. Partial view on the taxonomy of the IAO (dotted classes) subsumed by BFO.

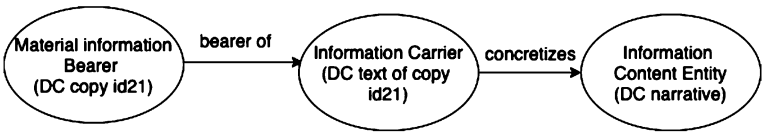

Fig. 5. Modeling the *Comedy* (DC) according to the IAO.

entities that in order to exist have to be related to some material entities, these latter called *material information bearers* in the IAOwl (Goldstein et al., 2020).[8] More precisely, an ICE is *concretized* in a quality[9] inhering in the material information bearer upon which the ICE generically depends. The ICE class specializes in various subclasses, among which *Textual Entity* to explicitly capture, e.g., novels, and *Figure*.

Figure 5 shows the example of the *Comedy* according to the IAO. The *Comedy*'s narrative is an ICE (a textual entity indeed) concretized in the marks of ink inhering in a physical book, i.e., the *DC copy id21*. For the case of translations, it is not clear how it can be represented in the scope of the IAO; if text-types are language-dependent (similarly to the proposal of Goodman and Elgin, 1986), the *Comedy* in Italian and the *Comedy* in English are two different *information content entities*. It remains unclear whether the IAO can capture their similarities, intuitively, that the two texts tell the same story.

The relationship of *aboutness* between ICEs and the entities they refer to lays at the heart of the IAO. According to Smith and Ceusters (2015), "[a]boutness corresponds to what is otherwise referred to by means of the expressions 'reference' or 'denotation' but generalized to include not merely linguistic reference but also the relations of cognitive or intentional directedness that are involved, for instance, when a nurse is *measuring a patient's pulse rate* or a doctor is *observing a rash on a patient's thigh*.

[8]Smith and Ceusters (2015) use the term *information artefact* for material information bearers.

[9]The quality is called *information quality entity* by Smith and Ceusters (2015) and *information carrier* in the IAOwl (Goldstein et al., 2020).

These processes are about, respectively, a *pulse* and a *rash*" (p. 2). Additionally, Smith and Ceusters (2015) claim that when an ICE is about some entities, the latter exist. This principle is called *veridicality*. In the words of Smith and Ceusters (2015): "[A]n ICE must in every case be *about* some portion of reality, where the aboutness in question must always be veridical, so that 'being about' is a success verb" (p. 3).

Some comments are due. First, in the introduction to BFO (Arp et al., 2015), the authors claim that "[t]he pattern of letters of the alphabet and associated spacing which is the novel *Robinson Crusoe* is concretized in the patterns of ink in this (and that) particular *copy* of the novel. [. . .] The novel *Robinson Crusoe* is a generically dependent continuant instance, an *abstract pattern*, made concrete through the acts involved in printing successive copies" (p. 106). The authors distinguish therefore between *individual patterns* and *abstract patterns*: the former are the patterns of ink inhering in a particular book and the latter are the ICEs themselves. Smith and Ceusters (2015) make a similar distinction when claiming that "pattern can [. . .] be understood in two senses – as referring either (i) to what is shared or communicated (between original and copy, between sender and receiver) or (ii) to the specific pattern before you when you are reading from your copy of Tolstoy's novel" (p. 1). This view is adopted by the IAOwl (Goldstein et al., 2020), too; e.g., according to the meta-data of *Textual Entity*, the concretizations of this class are individual patterns carried by some supports. The documentation about the IAOwl (Goldstein et al., 2020), however, does not specify what patterns are, nor BFO covers abstracta. Smith and Ceusters (2015) claim that ICEs are not abstract entities like propositions "in the logical parlance" (p. 2), since ICEs are created and therefore they exist in time. It remains therefore unclear in which sense ICEs are abstract patterns. By subsuming *Textual Entity* directly under ICE, ICEs seem to correspond to sign-types, a position recalling Goodman and Elgin (1986). In fact, differently from both YAMATO and philosophical positions *à la* Thomasson (1999), there is no distinction in the IAO between the content of, e.g., a novel and its text-type; a novel *is* a text-type for the IAO.

Second, the theory of *aboutness* in the ontology deserves clarifications. Considering the sentence "Barack Obama is President of Russia", Smith and Ceusters (2015) claim that there is a corresponding ICE "which is about Barack Obama, his being president, and Russia" (p. 3) However, the ICE "*is not about* any corresponding configuration [in reality], simply because there is no corresponding configuration" (Smith and Ceusters, 2015, p. 3; emphasis is ours). Since the whole ICE fails to be about any configuration, it is not clear why it is an ICE given that ICEs' *aboutness* is required to be veridical.

A further issue with *aboutness* is related to applications domains where ICEs are not necessarily about some things. In engineering, for instance, design models or process plans (considering them as ICEs) can fail to be about (actual) individuals, especially if they are created before the creation of the physical products or the events that they describe (Masolo and Sanfilippo, 2020).[10] We will further comment on *aboutness* in Section 6.2.

4.3. DOLCE Lite Plus (DLP)

We consider in this section the work done by Behrendt et al. (2005), Gangemi and Mika (2003), and Presutti and Gangemi (2016) and formally represented in the ontologies DOLCE+DnS Ultralite (DUL) and DOLCE Lite Plus (DLP). Both these ontologies depart from the DOLCE ontology (Masolo et al.,

[10] Schulz et al. (2014) have raised some concerns on the veridicity of IAO's *aboutness* in the biomedical domain.

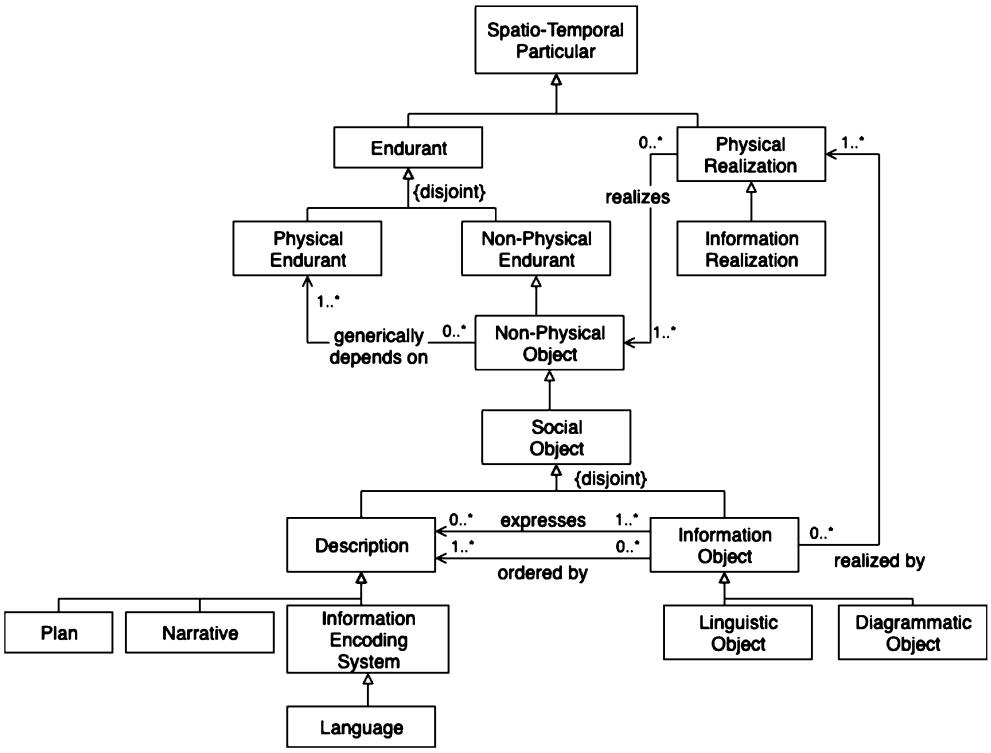

Fig. 6. Partial view on the taxonomy of DLP.

2003) under different aspects, hence the considerations hereby expressed do not apply to DOLCE. Since DUL simplifies DLP, we consider only the latter throughout the paper.[11]

DLP represents entities like novels, figures or plans by distinguishing between *Information Object* and *Description*, both subsumed by *Social Object*, which is in turn a subclass of *Non-Physical Object*, see Fig. 6. Instances of the latter class lack spatial locations and generically depend on physical endurants; social objects are entities that are socially created (Masolo et al., 2003, 2004). *Description* covers, among other things, information encoding systems such as natural or formal languages, and classes for novels, regulations, plans, etc.

The *Information Object* class links a description to both a language and a support (*Information Realization*). The relationship *ordered by* is used to relate information objects to "the languages, codes, grammars, etc. that they are ordered by", whereas *expresses* is used to link information objects to the "content (meaning, conceptualization) they represent." *Linguistic Object* specializes *Information Object* to explicitly refer to information objects encoded in natural languages. Among other classes, it subsumes *Text*, i.e., "a complex linguistic object [ordered by] a language and still independent from a particular physical realization." In this view, therefore, information objects are sign-types, whereas classes like *Plan* or *Narrative* stand for their contents. The supports of information objects are represented by *Information Realization*, which is subsumed by *Physical Realization*. This latter class models "any physical

[11] If not otherwise specified, quotes are taken from the OWL file of DLP (named DLP_397.owl). The reader can refer to Gangemi and Peroni (2016) for ontology design patterns for information entities based on DLP/DUL.

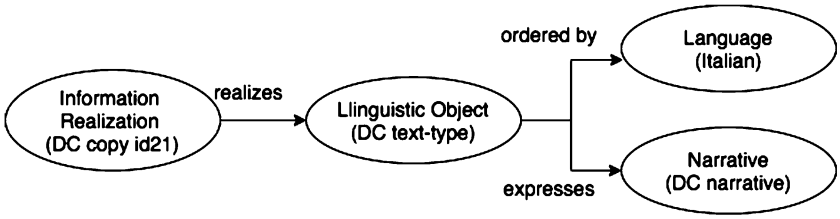

Fig. 7. Modeling the *Comedy* (DC) according to DLP.

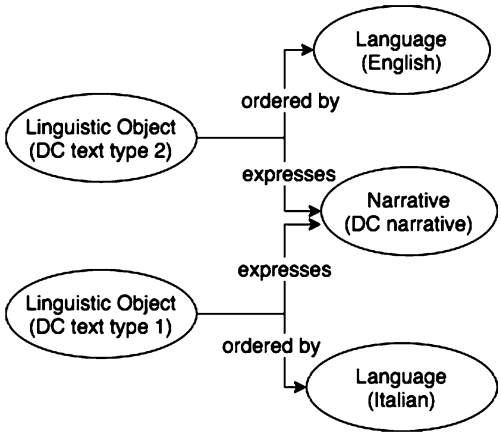

Fig. 8. Modeling translations in DLP (the two linguistic objects express the *same* narrative).

particular that realizes a non-physical endurant" and it covers a broad range of entities such as "physical endurants, physical qualities, physical regions, perdurants [. . .], or situation[s]."

Considering the example of the *Comedy*, in order model one of its copies one distinguishes between (i) an information realization (the specific copy), (ii) a text-type (linguistic object), (iii) the *Comedy*'s narrative (description), and (iv) the language ordering the linguistic object, see Fig. 7. DLP does not take an explicit stance on the representation of translations. However, similarly to YAMATO, the same narrative can be expressed by multiple linguistic objects but in different languages, see Fig. 8.

Finally, the relationship *is about* (not shown in Fig. 6) is used to model the entities to which information objects are about, "e.g. Dante's *Comedy* is about facts like Dante's travel to the hereafter." It is not however mandatory for information objects to be about some entities.

Some comments are due, first, concerning *Description*. In DLP this class is defined as "a social object which represents a conceptualization (e.g., a mental object or state), hence it is generically dependent on some agent and communicable." This approach is borrowed from the work of Masolo et al. (2004) where descriptions are introduced to characterize *concepts* and *roles*. DLP keeps this understanding of descriptions but it also broadens its scope to cover information encoding systems. This move results in ambiguity. For example, the ontology includes the axiom $Description \sqsubseteq \exists expressedBy.InformationObject$, hence *Information Encoding System* – as a subclass of *Description* – inherits it. It is not however clear what *expressed by* means in this case, namely, that a language is expressed by some information object.

Second, by distinguishing between descriptions like narratives and information objects like text-types, DLP – similarly to YAMATO – understands the former as contents. This raises the question of what contents are. Behrendt et al. (2005) talk of *propositional content*, Gangemi and Peroni (2016) of *meaning*

and *interpretation*, among others. However, the authors do not dig into the conceptual foundations of information entities and, despite they refer to semiotic studies, they do not say how to conceive meanings or propositions; e.g., it remains unclear what sorts of entities they are (see the work of Bateman, 2019, for a possible manner of conceiving meanings in the scope of DOLCE). We further discuss about DLP and meanings in Section 5 and Section 6.1.

4.4. CIDOC Conceptual Reference Model (CIDOC-CRM)

The CIDOC-CRM ontology (CIDOC hereafter; Bruseker et al., 2017; Doerr, 2003) is a standard (ISO 21127) to manage cultural heritage data.[12] It covers the modeling of historical documents, photos, newspapers, and books, among others; hence, the representation of information entities lays at its grounds.

From a high-level perspective, CIDOC includes various classes, among which *Persistent Item* (class identifier E77) and *Temporal Entity* (E2). The former are "items that have a persistent identity, sometimes known as *endurants* in philosophy" (Le Boeuf et al., 2015, p. 35). The latter are things that "happen over a limited extent in time [. . .] [and] are also called perdurants" (Le Boeuf et al., 2015, p. 2). Figure 9 provides an overview of some of the CIDOC's classes that we present in this section.

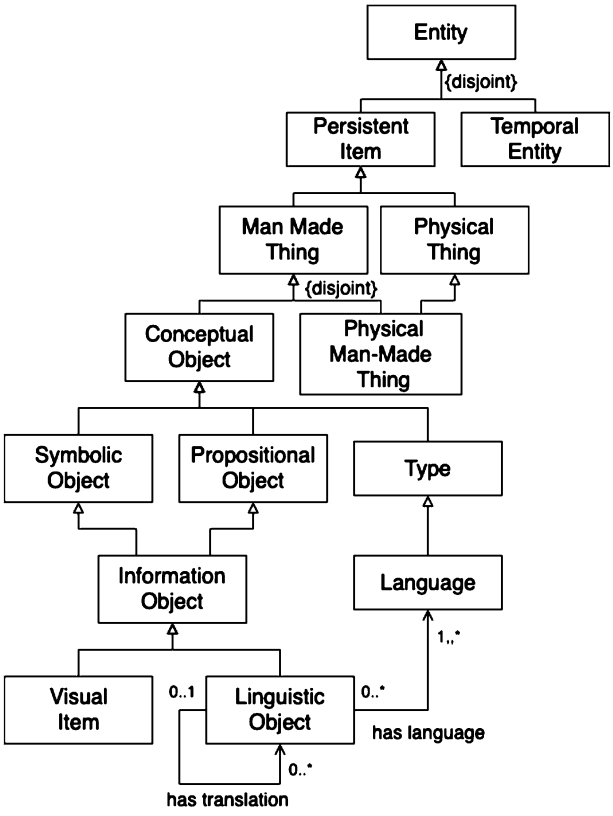

Fig. 9. Partial view on the taxonomy of CIDOC-CRM.

[12]The CIDOC ontology undergoes regular changes. When writing this paper, we considered the version presented by Le Boeuf et al. (2015). The reader can refer to Doerr (2003) and Doerr et al. (2008) for an overview of CIDOC's history.

Conceptual Object (E28) 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" (Le Boeuf et al., 2015, p. 16). Its instances can exist in multiple carriers at the same time, including paper, canvases, and human memory. Conceptual objects exist as long as at least one of their carriers exists and can not be directly destroyed, i.e., destroying a conceptual object means destroying all its carriers.

Amongst its subclasses, *Symbolic Object* (E90) 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" (Le Boeuf et al., 2015, p. 41). In addition, symbolic objects "[do] not depend on a specific physical carrier [. . .], and can exist on one or more carriers simultaneously" (Le Boeuf et al., 2015, p. 41). 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" (Le Boeuf et al., 2015, p. 41). As 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" (Le Boeuf et al., 2015, p. 41).

Another subclass of *Conceptual Object* is *Propositional Object* (E89) 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" (Le Boeuf et al., 2015, p. 40; emphasis is ours). 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, 1945" (Le Boeuf et al., 2015, p. 40).

Looking at Fig. 9, the class *Information Object* (E73) is subsumed by both *Symbolic Object* and *Propositional Object*; the intended meaning is that its instances are propositional objects encoded in a symbolic form. It includes various subclasses, among which are *Linguistic Object* (E33) and *Visual Item* (E36). Visual items are "the intellectual or conceptual aspects of recognisable marks and images" (Le Boeuf et al., 2015, 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" (Le Boeuf et al., 2015, p. 19) appearing on 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" (Le Boeuf et al., 2015, 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 *Language* (E56), 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" (Le Boeuf et al., 2015, p. 70). Finally, *is carried by* (not shown in Fig. 9) links symbolic objects (information objects included) to their physical supports, e.g., physical books.

Recalling the example of the *Comedy*, according to CIDOC one distinguishes between (i) a linguistic object, (ii) the language in which it is encoded, and (iii) a physical man-made thing (a physical book), see Fig. 10. Note that the linguistic object is both a propositional object, i.e., the *Comedy*'s ideational content, and a symbolic object, i.e., the *Comedy*'s text-type.

Let us comment on CIDOC. First, *Propositional Object* and its subclasses are understood as ideational contents but the documentation does not clarify what this means. This perspective recalls idealist views

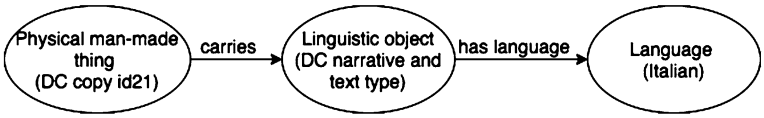

Fig. 10. Modeling the *Comedy* (DC) according to CIDOC-CRM.

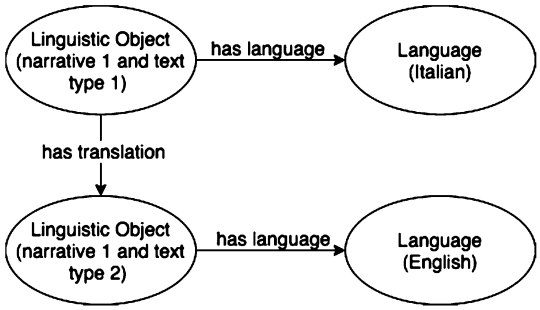

Fig. 11. Modeling translations in CIDOC-CRM (it is *implicitly* assumed that the two linguistic objects share the same narrative).

(see Section 2), e.g., the ideational content of Aristotle's *Metaphysics* – to recall one of CIDOC's examples – as ideas expressed in a text. However, similarly to the case of Smiraglia (2001) discussed in Section 3, it remains unclear how an ideational content relates to the meaning of the text expressing it.

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 involves 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., Le Boeuf et al. (2015) 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 explicitly identify and represent the content shared by multiple linguistic objects. For instance, looking at Fig. 11, it is only implicitly assumed that the two linguistic objects share the same narrative, since the narrative itself can not be represented as an entity on its own that is distinguished from the linguistic objects. As we will see in the next section, this is a key point of departure between FRBR and CIDOC.

4.5. Functional Requirements for Bibliographic Records (FRBR)

The Functional Requirements for Bibliographic Records (FRBR) models concepts and relations to manage bibliographic information. It was initially developed independently from CIDOC but the two ontologies have been eventually integrated in the FRBR object-oriented version (FRBR_{OO}; Bekiari et al., 2015). We first introduce the four core concepts of FRBR, namely, *Work*, *Expression*, *Manifestation*, and *Item*, and we then show how they have been restructured in FRBR_{OO} according to CIDOC.

Following the documentation (IFLA, 1998), a *Work* is "a distinct intellectual or artistic creation" (p. 17). The specification adds that "[w]hen we speak of Homer's *Iliad* as a work, our point of reference is not a particular recitation or text of the work, but the *intellectual creation* that lies behind all the various expressions of the work" (IFLA, 1998, p. 17, emphasis is ours). The documentation on FRBR_{OO} (Bekiari et al., 2015) claims that "[a]n instance of [. . .] Work begins to exist from the very moment an

individual has the initial idea that triggers a creative process in his or her mind. [...] Unless a creator leaves one physical sketch for his or her Work, the very existence of that instance of [...] Work goes unnoticed, and there is nothing to be catalogued" (p. 27).

Expressions stand for "the intellectual or artistic realization[s] of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, object, movement, etc., or any combination of such forms. An expression is the specific intellectual or artistic form that a work takes each time it is realized" (IFLA, 1998, p. 19). Bekiari et al. (2015) add that "[e]xpressions cannot exist without a physical carrier, but do not depend on a specific physical carrier and can exist on one or more carriers simultaneously. Carriers may include human memory" (p. 55). Expressions stand therefore for sign-types rather than sign-tokens and they must be carried on at least one carrier to exist. Carriers correspond to the concept of *Item*, e.g., printed books or CDs (Bekiari et al., 2015, p. 58).

Finally, the notion of *Manifestation* captures the "shared characteristics of copies of a particular publication, edition, release, etc." (IFLA, 1998, p. 23). For example, consider a novel published in hardcover and paperback. Instances of the hardcover manifestation and instances of the paperback manifestation share the same work and expression, but they differ in a number of physical characteristics, e.g., the numbers of pages and the way in which the (tokens of) expressions are arranged throughout the corresponding items. In this sense, FRBR's manifestations can be understood as items' layouts.

Figure 12 shows the subsumption of FRBR_{OO}'s classes under CIDOC. *Work* (class identifier F1) is subsumed by *Propositional Object* (E89), *Expression* (F2) by *Information Object* (E73), *Item* (F5) by *Information Carrier* (E54); *Manifestation* is renamed *Manifestation Product Type* (F3) and it is subsumed by *Type* (E55). In addition, as can be seen from the figure, FRBR_{OO} has been extended with

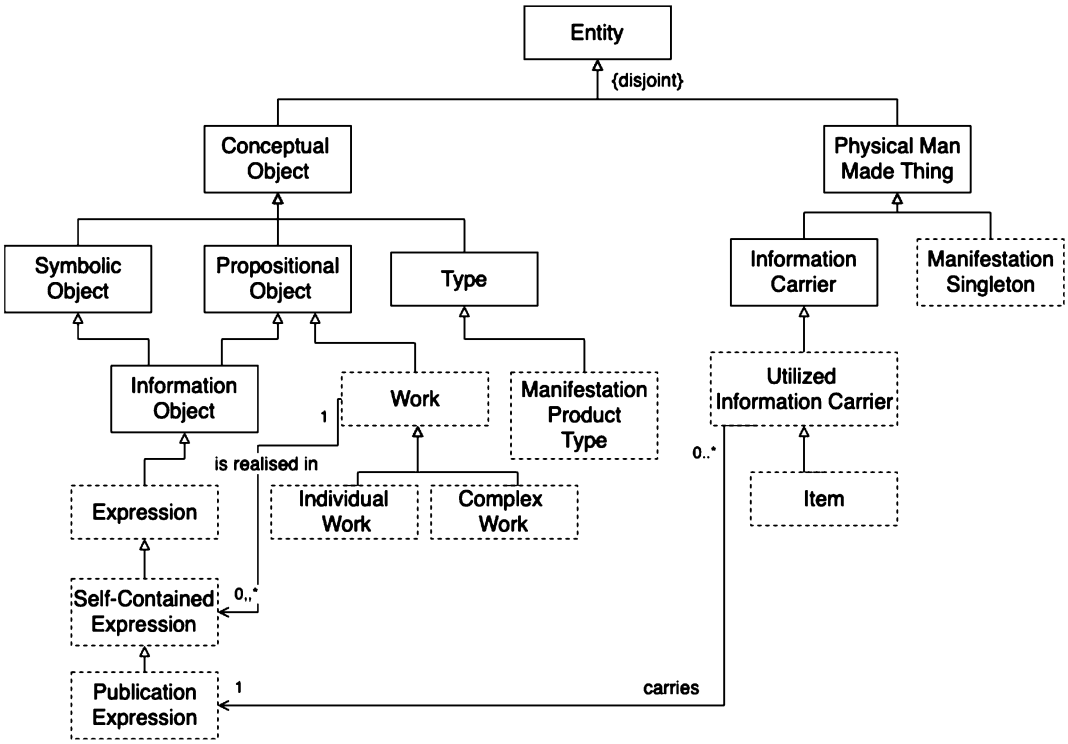

Fig. 12. Partial view on the taxonomy of FRBR_{OO} (dotted classes) subsumed by CIDOC-CRM.

classes which are not included in FRBR. *Work* now covers *Individual Work* (F14) and *Complex Work* (F15), among others, while *Expression* is extended in *Self-Contained Expression* (F22), just to mention the most relevant classes for our investigation.

A self-contained expression is an expression that – in the intention of its creator – conveys a whole work; an example is the Italian text of Dante's *Inferno* in Petrocchi's edition (Bekiari et al., 2015, p. 68). Figure 12 also shows *Publication Expression* (E24) to make explicit the link with carriers. In particular, a publication expression "comprises complete sets of signs present in publications, reflecting publishers' final decisions as to both selection of content and layout of the publication" (Bekiari et al., 2015, p. 69). The relation *carries* holds between *Utilized Information Carrier* (F54) and *Publication Expression*, where the former is a general class for entities carrying expressions. Items distinguish from *manifestation singletons* (F4), which are objects conceived without siblings, e.g., a manuscript. Looking at Fig. 12, differently from *Item*, *Manifestation Singleton* is not subsumed by *Utilized Information Carrier* (F54).

The class *Individual Work* stands for "works that are realised by one and only one self-contained expression, i.e., works representing the concept as expressed by precisely this expression" (Bekiari et al., 2015, p. 63). *Complex Work* is understood as a work that comprises other works as members. As recognized by Bekiari et al. (2015), this leads to a double interpretation of the intended meaning of this class. On the one hand, it stands for works "composed of several structural parts" (Bekiari et al., 2015, p. 26). An example is "Dante's textual work entitled *Divina Commedia* [which] has member Dante's textual work entitled *Inferno*" (Bekiari et al., 2015, p. 90). On the other hand, it captures "[t]he conceptual unity observed across a number of complete signs, which makes it possible to organise publications into bibliographic families" (Bekiari et al., 2015, p. 26), see example below.

Figure 13 shows the example of the *Comedy* according to FRBR_{OO}. The model distinguishes between (i) a physical copy of the *Comedy* (item) carrying a (ii) publication expression (e.g., the text-type of Petrocchi's edition) such that the latter realizes (iii) an *individual work* and is encoded in a (iv) language.[13] The representation of translations is depicted in Fig. 14, where the conceptual unity – to recall FRBR_{OO}'s words – shared between the two individual works, one expressed in Italian and one in English, is captured by the fact that they are members of the same complex work.

To comment on both FRBR and FRBR_{OO}, differently from CIDOC, works are explicitly represented as ideational contents, i.e., propositional objects that are not necessarily realized into expressions. For the case of translations, differently from CIDOC, one now has the possibility of representing multiple expressions sharing the same work. The notion of work remains however highly ambiguous. As we have seen, FRBR explicitly takes an idealist interpretation by which a work is an idea in the mind of its

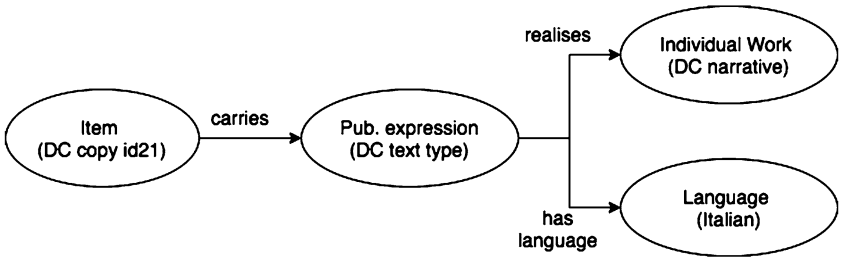

Fig. 13. Modeling the *Comedy* according to FRBR_{OO}.

[13]The relation *has language* (P72) is inherited from CIDOC.

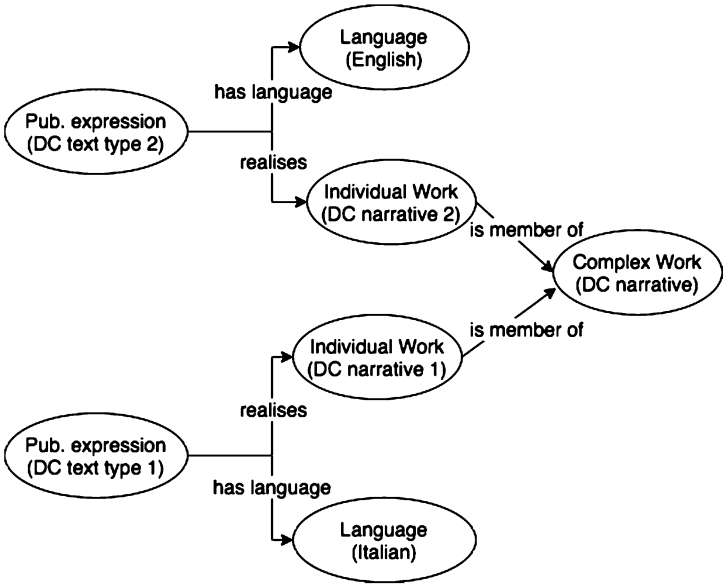

Fig. 14. Modeling translations in FRBR_{OO}.

creator and, as such, it may not be encoded in any publicly accessible expression.[14] When considering the distinction between *Individual Work* and *Complex Work*, we saw that the latter is sometimes interpreted as a documentary entity which is created to group multiple resembling works. This recalls the notion of work discussed by Eggert (2009), Pierazzo (2016), and Smiraglia (2001) with respect to classification theories in literary and librarianship studies (see Section 3). From this perspective, the overall notion of work in the ontology is – in the terminology of Weber (1997) – *semantically overloaded* because it is used to classify more than one ontological thing, i.e., ideas and documentary entities.

5. Summary and comparison

We now compare the ontologies introduced in Section 4 and map them to the materials presented in Section 2 and Section 3. We further discuss the various positions in Section 6.1 to both analyze their differences and similarities, and dig into the theoretical foundations for the modeling of information entities.

Table 1 provides a summarizing view on some of the properties used in the ontologies to characterize the classes in the *Class* column.[15] The ontologies rely on different and non-aligned vocabularies, hence we keep using the term *information entity* to refer to the classes in the table.[16] We consider whether information entities exist in time or space, generically depend (GDP) on some other entities,[17] are about

[14]The idealist view seems the standard interpretation for FRBR's works. For instance, according to Peroni and Shotton (2012), an example of FRBR's work is "the ideas in Lewis Carroll's head concerning Alice's *Adventures in Wonderland*"; see also the papers of Choffé and Leresche (2016), and Damiano et al. (2019).

[15]We talk of FRBR as including FRBR_{OO} when it is not necessary to distinguish between the two ontologies.

[16]In the case of DLP, we consider descriptions like novels or plans but not information encoding systems (see Section 4.3).

[17]Generic dependence between two properties ϕ and ψ means that it is necessarily the case that for each instance x of ϕ there is at least one instance y of ψ (upon which x depends); see the papers of Galton (2014) and Masolo et al. (2003).

Table 1
Comparing how information entities are characterized in the ontologies

Ontology	Class	In time	In space	GDP	Aboutness	Unity
YAMATO	Content	Yes	No	No	No commitment	No commitment
IAO	Information Content Entity	Yes	No	Yes	Necessary condition + veridicity	No commitment
DLP	Description (Novel, etc.)	Yes	No	Yes	Not necessary condition	No commitment
CIDOC	Propositional Object	Yes	No	Yes	Not necessary condition	Weak commitment
FRBR	Work	Yes	Yes (?)	Yes	Not necessary condition	Weak commitment

some thing, and carry some sort of unity. By reviewing the state of the art, these are the most general properties by which information entities are characterized.

Looking at the table, all ontologies agree on the temporal existence of information entities. DLP explicitly ascribes them a social nature. Similarly, propositional objects and works in CIDOC and FRBR, respectively, are human-made and, therefore, temporally bounded. YAMATO and the IAO make similar assumptions although their formal characterizations are not committed towards the artifactual or social nature of contents (in YAMATO) and information content entities (in the IAO).

Concerning existence in space, information entities in YAMATO, DLP, and the IAO lack spatial locations. In the case of FRBR, if works are ideas existing in someone's head, it might be legitimate to assume that works bear a spatial nature. In the case of CIDOC, we have seen that the notion of propositional object is vaguely conceived in terms of ideational contents. The class of conceptual objects (including propositional objects) is disjoint with the class of (man-made) physical entities, and only the latter are located in space. Propositional objects do not therefore exist in space.

In the case of dependence, it is only for YAMATO that information entities enjoy an independent nature. Better said, they depend on their creators for their existence; once created, the ontology assumes their independence from both supports and expressions (see Section 4.1). All other ontologies are clear in that information entities generically depend on their carriers.

Considering aboutness, it is only for the IAO that being about some thing is a necessary condition for information entities. YAMATO does not cover aboutness; DLP, CIDOC, and FRBR_{OO} include about-like relations but being about is not necessary for descriptions, propositional objects or works. In addition, differently from the IAO, these latter ontologies do not commit to the veridicity of aboutness (see Section 4.2). As a consequence, DLP, CIDOC, FRBR, and YAMATO are more flexible than the IAO for application domains where aboutness may fail (we further comment on aboutness in Section 6.2).

Concerning unity conditions, neither YAMATO, nor the IAO, nor DLP take an explicit commitment to the wholeness of information entities. This may depend on the fact that their notions of information entity need to be further characterized to specific application domains. Differently, CIDOC and FRBR refer to information entities' unity but only in generic terms. For instance, the CIDOC's class *Man-Made Thing*, subsuming *Proposition Object* (see Fig. 9), is defined as something that can be documented as a single unity (Le Boeuf et al., 2015, p. 33), but nothing is said about the kind of unity that is meant to be satisfied. In the case of FRBR, the documentation (IFLA, 1998) argues that the demarcation line between multiple works can be differently understood in different cultures. In addition, recall the ambiguity concerning the class of *Complex Work* (see Section 4.5). Representing both structured works having parts (e.g., a whole narrative and its chapters) and multiple resembling works, the notion of *Complex Work* assumes a sort of conceptual unity among the parts and the whole (Bekiari et al., 2015, pp. 63–64). Nothing more is however said. Finally, all ontologies cover sorts of parthood relations between information entities.

Apart from these general considerations, we have seen that information entities are differently understood in the ontologies. YAMATO and DLP are the only ones talking of propositions and meanings.

Neither of the two, however, takes an explicit stance on what these latter entities are. This is not surprising considering the variety of theories about meanings or propositions (Carroll, 2015; McGrath and Frank, 2018; Speaks, 2018). At a more general level, both YAMATO and DLP understand contents and descriptions, respectively, as non-physical entities (YAMATO uses the notion of semi-abstract).[18] Differently from these approaches, as said, there is no reference to meanings in the IAO. According to Arp et al. (2015) and Smith and Ceusters (2015), information content entities are abstract patterns, the latter being only vaguely characterized. CIDOC collapses multiple entities in the same modeling element, since linguistic objects are both texts and (ideational) contents. FRBR_{OO} distinguishes between these two entities but a work is both an idea and a documentary entity, i.e., an organizing category in the terms of Eggert (2009). This approach is not coherent; documentary entities can not be ideas indeed, because they are the sorts of things used in digital systems to organize similar bibliographic entities.

A further interesting aspect to compare the ontologies concerns the relation between information entities and the signs representing them. YAMATO, DLP, CIDOC, and FRBR explicitly distinguish between (i) information entities, (ii) signs like texts, and (iii) the signs' encoding forms (e.g., natural languages). In particular, YAMATO's representations, FRBR's expressions, DLP's and CIDOC's information objects are similarly understood as, e.g., the text of (a certain edition of) the *Comedy*. We use the term *expression* to refer to them. Importantly, expressions are sign-types rather than sign-tokens. Differently from these approaches, the IAO does not distinguish between expressions and their contents. As said, information content entities are expressions, recalling in this way the understanding of literary works of Goodman and Elgin (1986), see Section 2. We have seen that these modeling alternatives have certain consequences when representing translations. YAMATO and DLP share a similar pattern (cf. Fig. 3 and Fig. 8). CIDOC is not able to explicitly capture the similarity in the content shared by multiple texts (cf. Fig. 11). This problem is overcome in FRBR_{OO} by introducing complex works (cf. Fig. 14), although the class *Complex Work* is semantically overloaded. Finally, it is not clear how the IAO would handle the representation of translations.

To conclude, the state of the art review suggests that both CIDOC and FRBR need to be revised, and multiple interpretations for the same modeling element have to be avoided. The approaches adopted by YAMATO and DLP, on the one hand, and the IAO, on the other hand, deserve clarifications about meanings and abstract patterns, respectively.

6. Discussion

Section 6.1 discusses the interpretations of information entities found in the literature, whereas Section 6.2 and Section 6.3 make some (preliminary) considerations on the notions of aboutness and expression-token, respectively.

6.1. Information entities: A pluralistic approach

We have identified four interpretations about information entities, depending on whether they correspond to (1) semi-abstracta, (2) meanings, (3) ideas or (4) documentary entities. YAMATO and DLP lay between (1) and (2); the IAO is committed to (1); FRBR adopts (3) with FRBR_{OO} committed to (4), too; CIDOC seems to lay between (1) and (3).

[18] Recall from Section 2 that Thomasson (1999, 2004) talks of abstract particulars in the sense of particulars existing in time but not in space, whereas other philosophers introduce classes for types that are created and are therefore present in time.

Although idealistic positions like (3) meet various criticisms (see Section 2), one can hardly dismiss the relevance of authors' ideas to make sense of ordinary talks about works. For example, when claiming that an author came with the idea of writing a novel (or composing a symphony) after a tragic experience or that the *Comedy* is influenced by Dante's view on society and religion. This does not mean to identify the *Comedy* with Dante's ideas but to link the two when possible.

Position (4) goes in the direction of grouping multiple *resembling* information entities, e.g., translations, editions, arrangements, etc. under a common category. This interpretation presupposes therefore the others; e.g., it assumes the possibility of analyzing the classified entities to grasp their similarities. As said in Section 3, the definition of similarity criteria for literary or musical *works* is not straightforward.

According to position (1), information entities are non-physical entities that – in some interpretations – must exist on at least one support. The IAO extends this view by identifying information entities with abstract patterns and discarding meanings. We have seen that it remains challenging in this approach to grasp the similarities between, e.g., multiple texts without reference to meanings. One may group the texts under a common documentary entity but the similarity between the texts has to be established in terms other than semantic ones.

In position (2), claiming that two or more texts – even in different languages – express the same information entity means that they have the same meaning resulting from interpreting the texts in a certain way (recall the reaction of Wilsmore (1987) to Goodman and Elgin (1986) presented in Section 2). The ontological status of meanings is hotly debated (Speaks, 2018); refer to the papers of Bateman (2019) and Sowa (2015) for some discussions in applied ontology. Also, it is debated whether the identity of literary works depends on the meanings that were intended by their authors or whether one should also consider – *à la* Thomasson (1999) and Shillingsburg (2010) – the meanings that a work acquires in the context of a certain readership (Stecker, 2001). A well-known consequence of this latter position is the proliferation of works for a single text, given that the same text can be interpreted in different (and not overlapping) manners. Methods in linguistics like the *segmented discourse representation theory* (Asher et al., 2003) may be useful to analyze the linguistic meanings of texts on the basis of rigours formal procedures in such a way to obtain interpretations that are intersubjectively shareable. Figurative meanings remain however more challenging to be detected even in these approaches.

Although there are relevant differences between the positions (1)-(4), and some of them seem better than others for specific modeling cases, they seem all relevant to make sense of the variety of information entities found in applications. There are indeed scenarios where one wishes to distinguish between, e.g., multiple translations sharing the same content, in which case the distinction between expressions and meanings could be better suited than an approach relying on expressions only. At the same time, there are scenarios where reference to information entities as meaning may not be needed. For instance, one may claim that two copies of Raphael's fresco *The School of Athens* share a configuration of signs and colors, in which case position (1) seems fine. A similar perspective has been adopted for the analysis of computer programs and softwares, which correspond to patterns according to Wang et al. (2014).[19]

It emerges from these latter considerations the ambiguous way in which information entities have been treated in applied ontology. One and the same category is indeed used sometimes to capture the distinction between meanings and their encoding signs (as it is done in YAMATO and DLP), and some other times to refer to the signs themselves (IAO). This is misleading since the understanding of information entities as meanings needs to be clearly distinguished from the understanding of information entities as signs. In our view, therefore, positions (1)–(4) are not competing and they find their place for the

[19]Musical works are sometimes conceived as abstract patterns (Dodd, 2008).

ontological characterization of different kinds of information entities. Even reference to ideas (position 3) may be in fact useful (see comments above), and the notion of documentary entity (position 4) can be well-suited to capture the similarities between multiple resembling works. It is however challenging to fit these four positions and all kinds of information entities – from literary and musical works to documents, pictures, films, softwares, etc. – within a single, informative category.[20] An alternative approach is to rely on a *pluralistic* framework that recognizes the presence of various kinds of information entities and allows to model them according to different, and possibly related, modeling views. Further work in this direction is needed.

6.2. Aboutness

We have seen that it can be useful to represent the relation between information entities and the things to which they are about.[21] For instance, one may want to say that John's health-record is about John's blood pressure, that Mozart's biography is about Mozart's life, that a design model is about the functionality of the desired but not yet fabricated product.

Considering these and other examples, an information entity (i) is not necessarily about entities that are currently present and it can refer also to past entities (e.g., Mozart), and entities whose ontological nature is fictitious (e.g., Dante's journey through the afterlife); (ii) can be about a broad range of things including (at least) – what foundational ontologies commonly consider as – objects (e.g., Mozart), qualities (e.g., John's blood pressure), events (e.g., Mozart's life), and complex entities like situations (e.g., Mozart in a concert hall playing piano at a certain time); (iii) can refer to some entities but only partially or from a specific perspective (e.g., John's health record is only about his blood pressure taken at a certain time with a specific measurement device); (iv) may lack reference altogether, as in the case of a product to be fabricated. Alternatively, for this latter case, one may commit to possible entities; if *possibilia* are excluded, being about should not be a necessary condition for information entities.

The intended meaning of aboutness is challenging to be characterized (Hawke, 2018; Yablo, 2014). Eco (2016) argues that the aboutness of signs is a semiotic phenomenon grounded in the fact that agents use signs to refer to certain things (see also the paper of Barton et al., 2020, for a similar position in applied ontology). A sign, therefore, is about an entity only indirectly, namely, because there is an agent who refers to that entity and uses the sign to make this reference explicit. This position is shared by Smith and Ceusters (2015), too. Mediating the relation of aboutness between information entities and their referents through agents can be also useful to make sense of the fact that information entities refer to things from specific perspectives; e.g., when a physician creates an health record, she decides what aspects of the patient to capture and what to exclude. Further research on the conceptualization of aboutness is required.

6.3. Expression tokens

We have seen that ontologies conceive the expressions encoding information entities as sign-types (with the exception of the IAO where information entities are sign-types); e.g., the text-type of the *Comedy* edited by Petrocchi as the text that is common to all its physical copies. However, applications in

[20] This modeling and classification effort is even more challenging if one considers the variety of works in the literary or musical domains, including tales in the oral tradition and musical improvisations (Howell, 2002; Talbot, 2000).

[21] We interchangeably use the expressions to *refer to* and *being about* some thing.

domains like cultural heritage require to model also sign-tokens, e.g., the typeface and physical dimensions of specific characters (Cantone et al., 2019; Felicetti and Murano, 2017).[22] A modeling approach for sign-tokens is therefore required.

In the case of the IAO, sign-tokens are qualities, i.e., information carriers inhering in information content entities' supports. Differently, YAMATO treats sign-tokens as two-dimensional drawings, which are ultimately physical objects. Similarly, Felicetti and Murano (2017) model sign-tokens through the CIDOC's class *Man-Made Feature* (E25), which is subsumed by both *Physical Man-Made Thing* and *Physical Feature* (E26). Accordingly, a sign-token is a human-made entity which can exist if and only if it is physically bounded to an (independent) object. Cantone et al. (2019) rely on CIDOC, too, for the representation of epigraphies. Differently from Felicetti and Murano (2017), however, the authors do not commit to the dependent nature of sign-tokens in terms of features. Instead, they use the more general *Physical Man-Made Thing* CIDOC's class. Finally, neither FRBR nor DLP explicitly represent sign-tokens. Being based on CIDOC, FRBR_{OO} can easily adopt modeling patterns similar to Cantone et al. (2019) and Felicetti and Murano (2017).

Both the *quality*-based and the *object*-based understanding of sign-tokens are interesting perspectives. The first one can make sense of sign-tokens as particular patterns of, e.g., ink on paper. Further work in this direction is required; some proposals in the applied ontology community for the modeling of patterns have been presented by Galton (2018), Guarino (2013), and Masolo and Sanfilippo (2017). The second approach (sign-tokens as objects) can straightforwardly make sense of the characteristics commonly attributed to sign-tokens, e.g., colours, physical dimensions, and typeface, just to mention some examples. The attribution of characteristics to tokens may be more problematic for quality-based approaches, at least if qualities can not be further characterized by other qualities.

Finally, both Felicetti and Murano (2017) and the IAO conceive sign-tokens as dependent entities. This is reasonable in our view to characterize the fact that, e.g., a character has to be related to a certain object to exist. In the case of the IAO, an information carrier specifically depends on its support. CIDOC's features, on the other hand, only generically depend on supports. This latter choice results more flexible for cases where the identity of a support changes without affecting the identity of (some of) its sign-tokens; e.g., when the support of an epigraphy like a column is destroyed but the epigraphy is still found in its entirety on some of the column's fragments.

7. Conclusions

We presented a review of the state of the art in applied ontology for the treatment of information entities. The ontologies have been compared with each other and with theories in domains such as the philosophy of art, librarianship and literary studies, where notions similar to the ones used in applied ontology are adopted. As a result, we have identified four different interpretations, namely, information entities as (1) (specific types of) semi-abstract entities, (2) meanings, (3) ideas in someone's head, and (4) documentary entities useful for cataloguing purposes. We have seen that some of these interpretations are further restricted and sometimes even combined. For example, YAMATO combines the first two, whereas the IAO extends the first one by both restricting information entities to sign-types and avoiding reference to meanings. To the best of our knowledge, these four interpretations have not been

[22] We talk sloppily of characters here. The reader can refer to the work of Stokes and Meister (2012) for some insights on paleography.

clearly identified and compared in previous papers. Also, the variety of interpretations has brought some confusion, since one and the same notion has been tacitly used to refer to different things.

On the basis of the analysis, we have discussed some open challenges for the ontological characterization of information entities. We have argued that the four interpretations are not competing and they all find their place for representing different kinds of information entities. It may be indeed misleading to adopt a monolithic approach based on, e.g., the understanding of information entities as meanings for things like figures or softwares, for which a modeling approach based on the representation of (complex) sign-types may be better suited. On the other hand, reference to meanings may be necessary for representing verbal documents and literary works, among others, in such a way to distinguish a semantic content from its text(s). We have therefore suggested to develop a *pluralistic* framework that distinguishes between the different positions found in the state of the art while modeling their inter-relations in order to make sense of the variety of information entity kinds considered in application scenarios.

Future work on our proposal will cover further investigation about the notions needed to characterize information entities, *interpretation acts*, *meanings*, (abstract) *patterns*, and the relation of *aboutness* above all. Future efforts will also cover research and test cases about the representation of specific information entity kinds. In principle, this study will lead to domain-specific formal representations and will help in developing the general framework to handle multiple perspectives.

Acknowledgements

The paper has been mainly written during a postdoc position at the CESR University of Tours (France) thanks to a Le Studium scholarship. I am grateful to all colleagues at the CESR for the stimulating discussions with respect to the Digital Humanities. I am also grateful to colleagues at the ISTC-CNR Laboratory for Applied Ontology for their continuous support. Finally, I wish to thank the reviewers of Applied Ontology for all suggestions and remarks on previous versions of the paper. Nobody but me is responsible for any remaining mistake.

References

Arapinis, A. & Vieu, L. (2015). A plea for complex categories in ontologies. *Applied Ontology*, *10*(3–4), 285–296. doi:10.3233/AO-150156.

Arp, R., Smith, B. & Spear, A.D. (2015). *Building Ontologies with Basic Formal Ontology*. Mit Press.

Asher, N., Asher, N.M. & Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Barton, A., Toyoshima, F., Vieu, L., Fabry, P. & Ethier, J.-F. (2020). The mereological structure of informational entities. In *Formal Ontology in Information Systems – Proceedings of the 11th International Conference (FOIS)* (pp. 201–215). IOS Press.

Bateman, J.A. (2019). Ontology, language, meaning: Semiotic steps beyond the information artifact. In *Ontology Makes Sense* (pp. 119–135). IOS Press.

Behrendt, W., Gangemi, A., Maass, W. & Westenthaler, R. (2005). Towards an ontology-based distributed architecture for paid content. In *European Semantic Web Conference* (pp. 257–271). Springer.

Bekiari, C., Doerr, M., Boeuf, P.L. & Riva, P. (2015). *Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism*. IFLA: Den Haag.

Borgo, S. & Masolo, C. (2009). Foundational choices in DOLCE. In *Handbook on Ontologies* (pp. 361–381). Springer. doi:10.1007/978-3-540-92673-3_16.

Bruseker, G., Carboni, N. & Guillem, A. (2017). Cultural heritage data management: The role of formal ontology and CIDOC CRM. In *Heritage and Archaeology in the Digital Age* (pp. 93–131). Springer. doi:10.1007/978-3-319-65370-9_6.

Cantone, D., Cristofaro, S., Nicolosi-Asmundo, M., Prado, F., Santamaria, D.F. & Spampinato, D. (2019). An EpiDoc ontological perspective: The epigraphs of the castello ursino civic museum of catania via CIDOC CRM. *Archeologia e Calcolatori*, *30*, 139–157.

Carroll, N. (2015). Interpretation. In *The Routledge Companion to Philosophy of Literature* (pp. 302–312). Routledge. doi:10.4324/9781315708935.
Choffé, P. & Leresche, F. (2016). DOREMUS: Connecting sources, enriching catalogues and user experience. In *24th IFLA World Library and Information Congress* (pp. 1–20).
Cray, W.D. & Matheson, C. (2017). A return to musical idealism. *Australasian Journal of Philosophy*, *95*(4), 702–715. doi:10.1080/00048402.2017.1281323.
Currie, G. (1991). Work and text. *Mind*, *100*(3), 325–340. doi:10.1093/mind/C.399.325.
Damiano, R., Lombardo, V. & Pizzo, A. (2019). The ontology of drama. *Applied Ontology*, *14*(1), 79–118. doi:10.3233/AO-190204.
Davies, D. (2007). *Aesthetics and Literature*. A&C Black.
Davies, D. & Matheson, C. (2008). *Contemporary Readings in the Philosophy of Literature: An Analytic Approach*. Broadview Press.
Dodd, J. (2008). Musical works: Ontology and meta-ontology. *Philosophy Compass*, *3*(6), 1113–1134. doi:10.1111/j.1747-9991.2008.00173.x.
Doerr, M. (2003). The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI magazine*, *24*(3), 75–92.
Doerr, M., Bekiari, C. & LeBoeuf, P. (2008). FRBRoo, a conceptual model for performing arts. In *2008 Annual Conference of CIDOC, Athens* (pp. 15–18).
Eco, U. (2016). Trattato di semiotica generale. *La Nave di Teseo Editore spa.*
Eggert, P. (2009). *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge University Press.
Eggert, P. (2019). *The Work and the Reader in Literary Studies*. Cambridge University Press.
Felicetti, A. & Murano, F. (2017). Scripta manent: A CIDOC CRM semiotic reading of ancient texts. *International Journal on Digital Libraries*, *18*(4), 263–270. doi:10.1007/s00799-016-0189-z.
Galton, A. (2014). On generically dependent entities. *Applied Ontology*, *9*(2), 129–153. doi:10.3233/AO-140133.
Galton, A. (2018). Processes as patterns of occurrence. In *Process, Action, and Experience* (pp. 41–57). Oxford University Press.
Gangemi, A. & Mika, P. (2003). Understanding the semantic web through descriptions and situations. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (pp. 689–706). Springer.
Gangemi, A. & Peroni, S. (2016). The information realization pattern. In *Ontology Engineering with Ontology Design Patterns: Foundations and Applications* (pp. 299–312). IOS Press.
Goehr, L. (1992). *The Imaginary Museum of Musical Works: An Essay in the Philosophy of Music: An Essay in the Philosophy of Music*. Clarendon Press.
Goldstein, A., Ruttenberg, A., Goldfain, A., et al. (2020). Information Artefact Ontology (IAO), https://github.com/information-artifact-ontology/IAO.
Goodman, N. & Elgin, C.Z. (1986). Interpretation and identity: Can the work survive the world? *Critical Inquiry*, *12*, 564–575.
Guarino, N. (2013). Local qualities, quality fields, and quality patterns: A preliminary investigation. *SHAPES*, *2013*, 75–81.
Hawke, P. (2018). Theories of aboutness. *Australasian Journal of Philosophy*, *96*(4), 697–723. doi:10.1080/00048402.2017.1388826.
Herre, H. (2010). General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In *Theory and Applications of Ontology: Computer Applications* (pp. 297–345). Springer. doi:10.1007/978-90-481-8847-5_14.
Howell, R. (2002). Ontology and the nature of the literary work. *The Journal of Aesthetics and Art Criticism*, *60*(1), 67–79. doi:10.1111/1540-6245.00053.
IFLA (1998). Functional requirements for bibliographic records. In *Technical Report, Study Group on the Functional Requirements for Bibliographic Records*.
Le Boeuf, P., Doerr, M., Emil Ore, C. & Stead, S. (Eds.) (2015). Definition of the CIDOC Conceptual Reference Model version 6.2.1. http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.1.pdf.
Levinson, J. (1980). What a musical work is. *The Journal of Philosophy*, *77*(1), 5–28. doi:10.2307/2025596.
Livingston, P. (2016). History of the ontology of art. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.
Margolis, J. (1974). Works of art as physically embodied and culturally emergent entities. *The British Journal of Aesthetics*, *14*(3), 187–196. doi:10.1093/bjaesthetics/14.3.187.
Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003). *WonderWeb Deliverable D18*. Technical report, Laboratory for Applied Ontology ISTC-CNR.
Masolo, C. & Sanfilippo, E.M. (2017). Representing types through image schemas and patterns. In *Workshop on Cognition and Ontologies (CAOS) at the AISB Annual Convention*.
Masolo, C. & Sanfilippo, E.M. (2020). Technical artefact theories: A comparative study and a new empirical approach. *Review of Philosophy and Psychology*, *11*, 831–858. doi:10.1007/s13164-020-00475-9.

Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., Guarino, N., et al. (2004). Social roles and their descriptions. In *KR*, (pp. 267–277).
McGrath, M. & Frank, D. (2018). Propositions. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.
Mizoguchi, R. (2004). Part 3: Advanced course of ontological engineering. *New Generation Computing*, *22*(2), 193–220. doi:10.1007/BF03040960.
Mizoguchi, R. (2010). YAMATO: Yet Another More Advanced Top-level Ontology. In *Proceedings of the Sixth Australasian Ontology Workshop* (pp. 1–16).
Mizoguchi, R. & Toyoshima, F. (2017). YAMATO: Yet-Another More Advanced Top-level Ontology. Applied Ontology (under review). http://www.hozo.jp/onto_library/AO-YAMATO_final.pdf.
Peroni, S. & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, *17*, 33–43. doi:10.1016/j.websem.2012.08.001.
Pierazzo, E. (2016). *Digital Scholarly Editing: Theories, Models and Methods*. Routledge.
Presutti, V. & Gangemi, A. (2016). Dolce+D&S Ultralite and its main ontology design patterns. In *Ontology Engineering with Ontology Design Patterns: Foundations and Applications* (pp. 81–103). IOS Press.
Schulz, S., Martínez-Costa, C., Karlsson, D., Cornet, R., Brochhausen, M. & Rector, A.L. (2014). An ontological analysis of reference in health record statements. In *Formal Ontology in Information Systems – Proceedings of the Eighth International Conference (FOIS)* (pp. 289–302).
Shillingsburg, P. (2010). How literary works exist: Implied, represented, and interpreted. In *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions* (pp. 165–182).
Smiraglia, R.P. (2001). *The Nature of 'a Work': Implications for the Organization of Knowledge*. Scarecrow Press.
Smith, B. & Ceusters, W. (2015). Aboutness: Towards foundations for the information artifact ontology. In *Proceedings of the International Conference on Biomedical Ontology* (Vol. 1515, pp. 1–5). CEUR Workshop Proceedings.
Sowa, J.F. (2015). Signs and reality. *Applied Ontology*, *10*(3–4), 273–284. doi:10.3233/AO-150159.
Speaks, J. (2018). Theories of meaning. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.
Stecker, R. (2001). Interpretation. In *The Routledge Companion to Aesthetics* (pp. 309–319). London: Routledge.
Stokes, P.A. & Meister, J.C. (2012). Modeling medieval handwriting: A new approach to digital palaeography. In *DH*, (pp. 382–384).
Talbot, M. (2000). *Introduction. in the Musical Work: Reality or Invention?* (pp. 1–13). Oxford University Press. doi:10.5949/UPO9781846313615.
Thomasson, A. (1999). *Fiction and Metaphysics*. Cambridge University Press.
Thomasson, A. (2004). *The Ontology of Art*. Malden, MA: Blackwell.
Thomasson, A.L. (2015). The ontology of literary works. In *The Routledge Companion to Philosophy of Literature* (pp. 349–358). Routledge.
Wang, X., Guarino, N., Guizzardi, G. & Mylopoulos, J. (2014). Towards an ontology of software: A requirements engineering perspective. In *Formal Ontology in Information Systems – Proceedings of the 8th International Conference (FOIS)* (pp. 317–329).
Weber, R. (1997). *Ontological Foundations of Information Systems. Coopers & Lybrand and the Accounting Association of*. New Zealand: Australia and.
Wetzel, L. (2018). Types and tokens. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
Wilsmore, S. (1987). The literary work is not its text. *Philosophy and Literature*, *11*(2), 307–316. doi:10.1353/phl.1987.0060.
Yablo, S. (2014). *Aboutness*. Princeton University Press.