Texts and Works: Ontology-based Modeling Patterns*

Emilio M. Sanfilippo[1,3,*], Laura Antonietti[2] and Elena Pierazzo[3]

[1]CNR ISTC Laboratory for Applied Ontology, Trento & Catania, Italy
[2]University of Siena, Italy
[3]CESR University of Tours, France

Abstract

The classification of texts in digital libraries or catalog systems is a longstanding subject in knowledge representation. In these contexts, the concept of *work* is pivotal, serving to group multiple texts for easier retrieval. However, the principles guiding this classification are often contingent on differing interpretations of what constitutes a *work*. This variability can present challenges in information system design, since different groups may organize the same texts differently, potentially causing issues when exchanging or integrating data. The purpose of this paper is to show that the intended meaning of *work* cannot be given for granted because scholars disagree on what a *work* amounts to. Instead of a monolithic ontology, we propose alternative *ontology-based modeling patterns*, which can be used in different application contexts by considering their match with users' requirements, as well as their advantages and shortcomings.

Keywords

texts, documentary works, ontology patterns, digital libraries, catalog systems

1. Introduction

In contexts relative to the creation of digital libraries or catalog systems for research and application in the digital humanities, one is often confronted with the formal modeling of data about texts along with their versions and editions [1, 2]. When designing a model like a computational ontology to handle these sorts of data, one needs to understand how to organize the texts in a way that is compliant with scholars' knowledge and is systematic from a computational perspective. For example, one needs to find ways to make sense of the relationships between texts, possibly by capturing their similarities and distinctions. Assume, for instance, that we need to organize two texts by the same author, published in different contexts but very similar to each other, as the second text was derived from the first one by the author with slight revisions. Should we classify the texts as two different entities, possibly sharing some sort of derivation relation? Or should we conceive and classify them as "representatives" of a single, more abstract entity, such as the *work* that they both exemplify? This is a standard scenario in philological and literary studies, where scholars commonly analyze how authors have reworked their texts over the years, resulting in multiple (and not necessarily equivalent) versions.

To propose a more realistic example that we will recall throughout the paper, in the context of Italian XX century literature, the literary critic and writer Franco Fortini[1] continuously re-published during his life several of his texts with different publishers, sometimes with some variations in the texts. For instance, the poem *Al di là della speranza* appeared in 1959 as part of the collection (by the same author) *Poesia ed errore* published by *Feltrinelli*, as well as in 1987 as part of *Versi primi e distanti* by the publisher *All'insegna del pesce d'oro*, among other publication contexts. Fortini eventually published a shortened version of the poem; for example, in 1974 within *Poesie scelte. 1939-1973* by Mondadori, and in 1978 within *Una volta per sempre* by Einaudi (see Table 1 for an excerpt of data). In addition, even single

Proceedings of the Joint Ontology Workshops (JOWO) - Episode X: The Tukker Zomer of Ontology, and satellite events co-located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), July 15-19, 2024, Enschede, The Netherlands

*Corresponding author.
†These authors contributed equally.
emilio.sanfilippo@cnr.it (E. M. Sanfilippo); laura.antonietti2@unisi.it (L. Antonietti); elena.pierazzo@univ-tours.fr (E. Pierazzo)

[1]Wikipedia page for Franco Fortini: https://it.wikipedia.org/wiki/Franco_Fortini.

collections were published different times by different publishers. The collection *Poesia ed errore*, for instance, was published again in 1969 by Mondadori with a modification in the the title – *Poesia e errore* (the conjunction *ed* is changed into *e*); the addition of an introductory comment by the author; the removal of some poems and therefore a novel organization of the collected poems. Similar editorial cases are present worldwide across authors and epochs.

The management of data about texts is a well-known task in knowledge representation. The FRBR model – Functional Requirements for Bibliographic Records – , nowadays called LRM – Library Reference Model [3] –, stands out as a fundamental reference for these sorts of modeling objectives, being, according to Holden [4], "one of the most influential cataloging publications of the past thirty years." There exists an extended literature on FRBR/LRM,[2] its advantages, and shortcomings from both a practical and theoretical perspective [4, 5, 6, 7, 8]. In particular, LRM's notion of *work*, represented through the modeling element *F1 Work* has been the subject of multiple analyses, since it is not plain clear what a *work* is, or the manner in which one should distinguish between different *works* instead of modeling multiple linguistic realizations (i.e., *F2 Expressions*) for the same *work*, just to mention some discussions [4, 8, 9]. What the documentation says is that a *work* stands for "intellectual ideas conveyed in artistic and intellectual creations" [3, p. 29] having the goal of "bringing together intellectually equivalent Expressions in order to display to a user all available alternatives of the same intellectual or artistic content" (ibid.). As intuitive as these comments may appear, from a scholarly perspective, reference to "intellectual ideas" or the "content" of linguistic expressions remain too generic and can be understood in different ways. Additionally, even from the pragmatic perspective of data classification, comments on LRM's *work* like the ones reported above leave it open to interpretation whether, for example, two texts by the same author but with some variations should be classified as the same or different *works*. Alternative models have been proposed, spanning from BibFrame[3] by the US Library of Congress, to BIBO – the Bibliographic Ontology[4] – by the Dublin Core Metadata Initiative, as well as project-specific ontologies (e.g., [10, 2]), which often reuse some of the previous ontologies while adapting them to specific purposes or formal languages.

The purpose of this paper is to contribute to the design of ontologies for text classification in catalogues or digital libraries. In particular, by analyzing the notion of *work*, our goal is to show that there can be multiple approaches to represent texts and their interrelations, with each approach answering certain requirements and coming along with benefits and shortcomings. From this perspective, instead of proposing a single overarching ontology *à la* LRM or any other alternative model, leveraging on research in knowledge representation, our goal is to discuss alternative *ontology patterns* [11] to classify texts. Our result is therefore both methodological and pragmatic, because by discussing some crucial aspects relative to texts and *works*, we deliver alternative modeling strategies to be used in applications.

The remaining of the paper is structured as follows. In Section 2 we provide an analysis of the notion of *work* based on studies relevant for our purposes. On the basis of the analysis, we present in Section 3 three modeling patterns; by the end of the section, we discuss and compare them. Finally, Section 4 concludes the paper summarizing the contribution and addressing future work on our the proposal.

Table 1
Some of the full and shortened editions of Franco Fortini's poem *Al di là della speranza*

Version type	Poem Collection	Publisher	Publication Year
Full	*Poesia ed errore*	Feltrinelli	1959
	Versi primi e distanti	All'insegna del pesce d'oro	1987
Shortened	*Poesie scelte. 1939-1973*	Mondadori	1974
	Una volta per sempre	Einaudi	1978

[2]For simplicity, we will simply speak of LRM in the remaining of the paper, including the literature on FRBR and its object-oriented version, $FRBR_{oo}$.
[3]https://www.loc.gov/bibframe/.
[4]https://www.dublincore.org/specifications/bibo/.

2. *Work*: A debated notion

The notion of *work* for the organization of entities like texts is so pervasive that it might appear odd trying to disentangle what a *work* is (see the review of the literature in [8]). However, upon closer examination, one can discern that things are not as simple as they initially appear, especially if considering that *work* is interpreted differently across various fields. As an overview, we will now discuss some research relevant to our investigation.

In literary studies and philosophy [12], some scholars argue that a *work*, in the specific sense of a *verbal work*, can be defined as a *textual entity* comprising a sequence of words in a language and various sorts of signs. This stance, sometimes referred to as *textualism*, contrasts with the perspective of others who assert that a *work* is a text's "content", i.e., what a text conveys. Advocates of this viewpoint generally agree on considering a text's content as its *meaning* but then differ in what they understand with this. In addition, it is debated whether the meaning of a text has to be uniquely identified with respect to what the author of the text wished to convey or whether it should be rather conceived with respect to readers, including all sorts of intermediate positions between these two extremes. In the first of these last two cases, reference to what we may call the *authorial work* is notoriously problematic, especially if considering that scholars often lack empirical evidence for what authors had in mind to convey through their productions. In addition, the role of authors to interpret texts has been hotly debated in XX century literary criticism [13]. The second option, call it the *readership work* view, ascribes an interpretive nature to *works* in relation to readers' habits, culture, etc [14]. This option leaves open the possibility of having even a single text differently classified according to different interpreters, especially when the latter disagree on the way in which a text is interpreted.

In philology, scholars must decide how to classify texts when editing them, in particular, whether multiple texts by the same author are to be considered as representative of a single *work* or not. In the words of Pierazzo's [15, p. 46] one may ask: "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?." Pierazzo herself replies that "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" [15, p. 47]. On similar lines, according to Eggert (quoted in Creider [5]): "[T]o make the editorial decision that a particular work can tolerate a certain amount of variation before its version texts and presentations constitute a different work is to engage in an interpretive act: the 'work' emerges as a principle allowing the editor and the reader to regulate that variation." *Work* seems to be here understood as a sort of regulative idea presupposing the interpretation of some agents like scholarly editors to group together texts presenting some high degree of resemblance. Hence, not only a *work* is a scholarly construct but it is possible that, because of disagreements, scholars come to different *works* for the same texts, a view that resembles the readership perspective previously mentioned. As claimed by Shillingsburg [16], "[w]e should be suspicious of locutions like 'the work itself,' for the work exists only in our construct of it."

Finally, in the context of librarian studies, the concept of *work* emerged at least in the XIX century when scholars started to organize catalogs not by listing specific documents but by classifying their contents [4, 17]. From this perspective, "[a] book to be cataloged has two aspects. It is a bibliographical, or physical, entity; and, at the same time, it is a vehicle by which intellectual content is presented" (quoted in Holden [4]). Accordingly, borrowing the terminology from [4, 17], one may talk of *documentary* or *bibliographic work* as an information retrieval entity that is functional to catalog texts.[5]

As pragmatic as this last view could be, one still needs to decide what are the "boundaries" of a *documentary work* to take a principled stance on which texts can be classified under a single *work*, and

[5] A similar proposal has been recently presented by De Berardinis et al. [10], where instead of *work* the authors talk of *information entity*.

therefore how to differentiate between different *works*. Also, the notion of *documentary work* used in the cataloging community is itself contentious. Holden [4], for instance, stresses the lack of a stable and universal notion and talks of it as a conceptual and intangible entity in connection to authors' creative ideas, begging back previously mentioned issues about how to empirically access what the author of a text had in mind to convey. Smiraglia [7], on the other hand, seems to abstract from both authors' ideas or readers' interpretations, assuming that a text carries a content as, say, a matter of fact.

To summarize, we have identified at least three core notions of *work*: (1) as text, (2) as text's content, and (3) as documentary entity. Within the second view, content can be understood with respect to either (2.1) authors (*authorial work*) or (2.2) readers (*readership work*), including scholarly editors, but also in terms of, say, (2.3) a pure abstract meaning that is independent from any interpretation act (*abstract content work*). Also, *work* in the sense of (3) presupposes something to be classified, i.e., it presupposes texts grouped together in virtue, *again*, of their contents, these understood in the lines of (2). These various notions are sometimes confused and mixed together like in the case of LRM where *F1 Work* refers to both "intellectual ideas", which could be understood in the sense of (2.1)-(2.3), and a documentary entity (3) grouping texts. Interpretation of *work* in the sense of (1) is probably the one that is less adopted for information modeling whereas interpretations (2) and (3) are mostly found. In these latter two views, the notion of *work* remains contentious and raises concerns about how to empirically access authors' ideas (2.1), how to distinguish between different sorts of readers' interpretations (2.2), e.g., those that are legitimate from those that are less legitimate, or how to single out the pure abstract meaning of a text independently from any interpretation (2.3). *Work* in the sense of (3) is meant to have a more pragmatic rather than "substantial" flavor since a *work* is a sort of "entry point" to group resembling texts in an information system. As we have seen, however, also in this case there remains space for debate as it presupposes (a version of) view (2). In addition, it bears a strong interpretive nature on the lines of other views since it is a matter of someone's choice to decide to classify certain texts together because of the similarity in their contents. From this perspective, the notion of documentary *work* is not very different from the notion of *work* used in, e.g., philological studies.

At this point one may ask whether the modeling of *works* is so essential, if alternative approaches, possibly getting rid of this thorny notion are available, what are their advantages and shortcomings. For this purpose, we will present in the next section some modeling patterns for classifying texts; the idea is to offer alternatives strategies that may fit different users' requirements instead of a single ontology fitting all possible desiderata.

3. Ontology Patterns for Texts and Works

In the context of ontology engineering, an ontology pattern describes a recurring modeling problem arising in specific ontology development contexts, along with a solution for the problem [11]. By developing and comparing alternative patterns, information system designers can rely on different approaches for their modeling tasks while leveraging their respective advantages and shortcomings.

For the sake of the presentation, we utilize the UML Class Diagram notation to convey the core structure of the patterns while using RDF for the example graphs. The formal representation of the patterns in OWL2 is available on a GitHub repository.[6] Additionally, since our focus lies on conceptual analysis at this stage, we refrain from incorporating existing Semantic Web resources by importing their IRIs into the patterns, leaving this integration to future work. The section is organized as follows: we first introduce the patterns and then compare them in Section 3.1.

Approach 1 – Situated Texts. A first approach consists in the representation of what we might call *situated texts* (s-text for shortness), intuitively, texts that are always situated (*embodied*, one may say) in a context like a publication context.[7] For instance, the full-version of *Al di là della speranza*

[6] https://github.com/emiliosanfilippo/text-work.

[7] Recall that the sorts of texts scholars might be interested in are not necessarily published.

published in 1987 within *Versi primi e distanti* is a s-text in the sense of being published by a publisher at a certain time, place, etc. Considering the data in Table 1, according to this first approach, one would distinguish four different s-texts for the poem, each one being related to a specific collection, publisher, and publication year, among other information.

The diagram in Figure 1 is a simplified representation of this approach. As said, we show only some modeling elements in the diagrams while a full formal specification can be found in the OWL files. A s-text can be included in another s-text, e.g., a poem that is part of a collection;[8] this can be done through the relation *includes*. The cardinality of the relation *hasPublisher* between *SituatedText* and *Publisher* is set to *zero-one*, *zero* being the case of a s-text that is not published, in which case a s-text could be present but in a different context, e.g., in an archive. Limiting the maximum cardinality to *one* stems from the fact that, according to Approach 1, each s-text can be associated with a single publisher (refer to Approach 2 for a deviation from this constraint.)

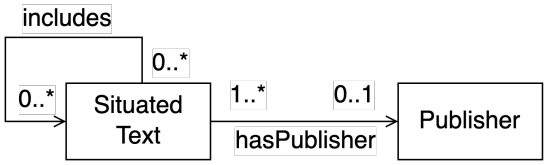

Figure 1: Approach 1 – Partial representation for pattern based on situated texts

The RDF graph in Figure 2 shows only an excerpt of data from Table 1 with *data:ALD1* as a s-text with title *Al di là della speranza* included in the collection *data:PoesiaEdErrore*, which is further related to publication information. A similar approach can be adopted for the representation of the other data in Table 1. For instance, representing the poem in the context of *Versi primi e distanti* published in 1987 by All'insegna del pesce d'oro leads to the creation of another poem, say, *data:ALD2*, with the same title as *data:ALD1* but this time related to *Versi primi e distanti* etc.

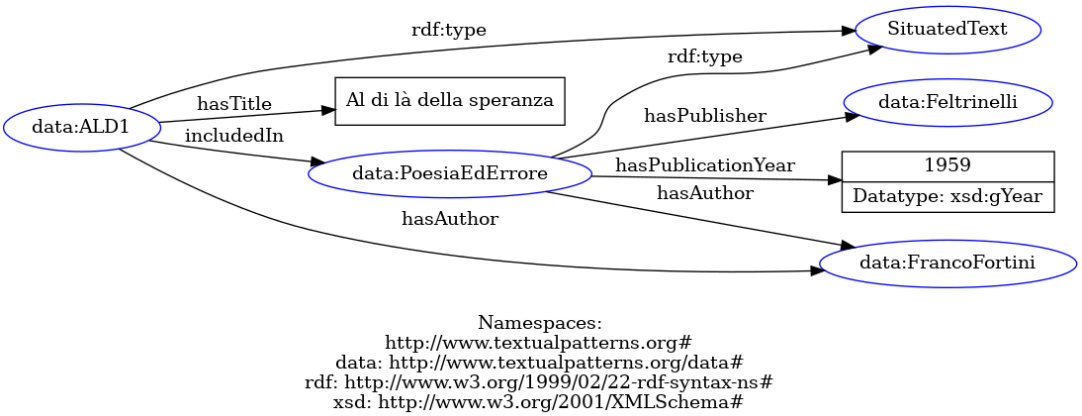

Figure 2: Example of RDF graph based on Approach 1

To comment on the way in which the data is represented in the RDF graph in Figure 2, first, information about authors (*hasAuthor*) is added to both the single poem and the entire collection. This is because there can be collections grouping poems of multiple authors. Second, in the case of texts that are included in other texts like a poem in a collection, publishing information may be directly related to the larger text (as in Figure 2) with the assumption that it propagates to all the texts it includes. For the case of OWL2 object properties, this propagation could be enabled through the use of an object

[8]The patterns do not distinguish between a single text like a poem and a collection of texts but they can be easily extended to cover these cases in an explicit manner; see, e.g., the approach presented in [18].

property chain like:

$$includedIn \; o \; hasPublisher \sqsubseteq hasPublisher \tag{1}$$

This property chain says that: when (text) x is included in (text) y, and y has publisher z, then x has publisher z, too. Otherwise, this sort of information can be explicitly included for both texts, especially when reasoning over property chains cannot be achieved, e.g., for computational restrictions.

Depending on application requirements, Approach 1 can be extended to model relations of *derivation* between situated texts. For instance, in the case of Fortini, we know that the shortened-version of *Al di là della speranza* was derived from the full-version. Derivation implies therefore also a temporal order among texts. This is clear when looking at the data in Table 1, where the first publication of the full-version precedes all publications of the shortened-version of the poem.

SPARQL query 1 (example). Having data represented according to this pattern, to search for alternative versions of a text like *Al di là della speranza*, one needs to go through texts as shown in the following SPARQL query. Table 2 shows the results for query 1.[9]

```
Select ?poem ?versionType ?collectionTitle ?publisherName ?pbYear Where
  { ?poem a :SituatedText;
    :hasTitle "Al di là della speranza";
    :hasVersionType ?versionType;
    :includedIn ?collection.
  ?collection a :SituatedText;
    :hasTitle ?collectionTitle;
    :hasPublisher/rdfs:label ?publisherName;
    :hasPublicationYear ?pbYear.}
```

Table 2
Examples of results for SPARQL query 1

poem	versionType	collectionTitle	publisherName	pbYear
data:ALD1	:Full-version	"Poesia ed errore"	"Feltrinelli"	"1959"
data:ALD2	:Full-version	"Versi primi e distanti"	"All'insegna del pesce d'oro"	"1987"
data:ALD3	:Shortened-version	"Poesie scelte 1939-1973"	"Mondadori"	"1974"
data:ALD4	:Shortened-version	"Una volta per sempre"	"Einaudi"	"1978"

Approach 2 – Texts. The second approach consists in the representation of texts as sequence of words in a language without binding them to the (publication) contexts where they possibly occur. In this sense, as we will see, Approach 2 allows for a more abstract representation than Approach 1.

Considering the diagram in Figure 3, we find it useful to introduce *Situation* as a modeling element originating from previous research in ontology [19]. From a technical perspective, situations were introduced by Gangemi et al. [20] to model relations with arity higher than two to deal with the restricted expressivity of Semantic Web languages. From a conceptual perspective, a situation can be understood as a state of the world including entities that satisfy certain conditions. For instance, in the case of what we call, following [10], a *publication situation*, it represents the context when and where a text was published. The introduction of situations is useful for our purposes, because they allow to represent single texts in the context of multiple publications (more discussion below).

In the diagram in Figure 3, the class *Text* is related to both the general class *Situation* and its subclass *PublicationSituation*. Considering the data in Table 1, for the poem of *Al di là della speranza*, according to this approach, one would distinguish between two texts, one for the full-version, call it *ALD12*, and one for the shortened-version *ALD34*, each of them included in certain poem collections, these being texts themselves. To stress it once again, the presence of only two texts for the poem, instead of four as in Approach 1, is now due to the independence of a text from its publication situation. Hence, the *same* poem can be now published by different publishers at different times.

[9] SPARQL queries have been tested on a local machine with an instantiation of GraphDB by Ontotext (free release). From the query results, note that we use modeling elements that are not shown in the diagrams; see the OWL files for more insights.

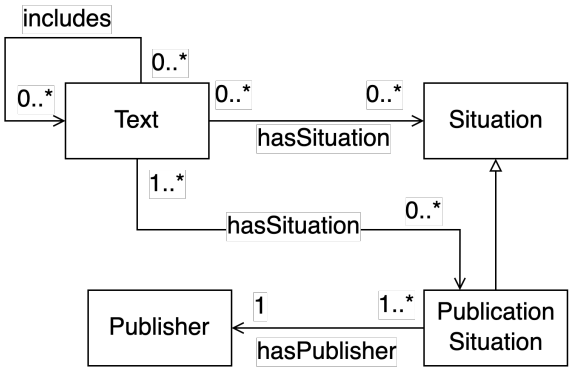

Figure 3: Approach 2 – Partial representation of pattern based on texts and situations

The RDF graph in Figure 4 shows the the full-version *data:ALD12* of the text related to the collection *Poesia ed errore*. The modeling element for the latter, *data:PoesiaEdErrore*, is related to the publication situation *data:pubSit1* making explicit the information relative to the publisher and the publishing year. A similar approach can be adopted to represent, e.g., *data:ALD12* in the context of *Versi primi e distanti* through a second publishing situation related to the publisher *All'insegna del pesce d'oro* and the publishing year 1987. The relation between *data:ALD12* and the situations where it occurs can be either explicitly represented or automatically inferred through an OWL 2 object property chain like:

$$includedIn \; o \; hasSituation \sqsubseteq hasSituation \tag{2}$$

The property chain says that when (text) x is included in (text) y, and y is situated in (situation) z, then x is situated in z, too.

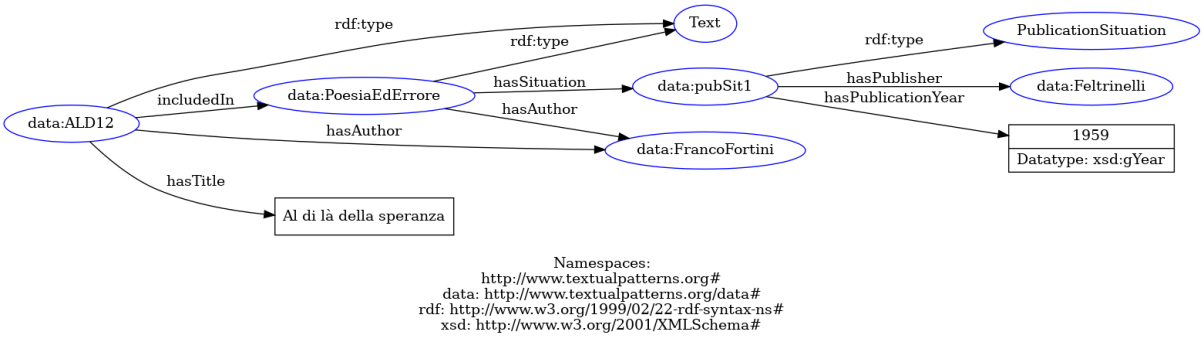

Figure 4: Example of RDF graph based on Approach 2

To comment on situations, assume that *Poesia ed errore* is published by another publisher. If we bind the text directly to publishers, this will lead to a confusing modeling approach since the text would be linked to multiple data without clear relations. Also, creating two entities for *Poesia ed errore* with respect to the publishers would align with Approach 1. It should be clear that, in full compliance with Approach 2, situations allow for modeling single texts but in different contexts.

SPARQL query 2 (example). To retrieve alternative data about a text, one can query the data as shown in the SPARQL query below. Differently from query 1, we now go through publication situations to retrieve publishing information relative to texts. Table 3 shows the results for query 2 where we have only two texts for the poem in two versions. That is, *data:ALD12* (full-version) is included in both *Poesia ed errore* and *Versi primi e distanti*, and *data:ALD34* is included in both *Poesie scelte 1939-1973* and *Una volta per sempre*.

```
Select ?poem ?versionType ?collectionTitle ?publisherName ?pbYear where
  {?poem a :Text;
    :hasTitle "Al di là della speranza";
    :hasVersionType ?versionType;
    :includedIn ?collection.
  ?collection a :Text;
    :hasTitle ?collectionTitle;
    :hasSituation ?s.
  ?s a :PublicationSituation;
    :hasPublisher/rdfs:label ?publisherName;
    :hasPublicationYear ?pbYear }
```

Table 3
Examples of results for SPARQL query 2

poem	versionType	collectionTitle	publisherName	pbYear
data:ALD12	:Full-version	"Poesia ed errore"	"Feltrinelli"	"1959"
		"Versi primi e distanti"	"All'insegna del pesce d'oro"	"1987"
data:ALD34	:Shortened-version	"Poesie scelte 1939-1973"	"Mondadori"	"1974"
		"Una volta per sempre"	"Einaudi"	"1978"

Approach 3 – Texts and Works. In a third approach, one may add a further level of abstraction to collect multiple texts under a common "modeling umbrella." For instance, as we have seen in Approach 2, ALD12 and ALD34 stand, respectively, for the full- and shortened-versions of *Al di là della speranza*. One may therefore wish to say that they are both texts for a single entity, i.e., the poem of *Al di là della speranza*. The notion of *work* could be functional for this goal but as we have seen throughout Section 2, the problem is that its introduction leads to some sort of ambiguity in the models.

For the sake of the discussion, let us consider the view for which *work* captures a modeling element for documentary purposes (interpretation 3 in Section 2). As said, this view is not less controversial than others, as it relies on the way in which texts' contents are interpreted, possibly even in connection to authors' intentions, and then grouped together. Considering its controversial status, it should not surprise if the pattern in Figure 5 can result controversial.[10]

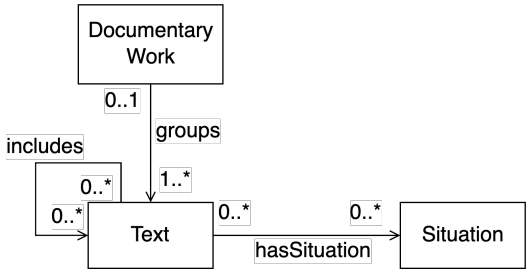

Figure 5: Approach 3 – Partial representation of pattern based on documentary *works*

The RDF graph in Figure 6 shows a partial representation of *Al di là della speranza*. In this case, *data:W1* is the single documentary *work* for the poem, grouping both the text of the full-version *data:ALD12*, and the text of the shortened-version *data:ALD34*. Each text is then included in a specific poem collection and related to a specific publication context through a situation as in Approach 2.

To comment on how data is represented, first, note that Figure 6 shows only the *work* for *Al di là della speranza* but the overall representation covers *works* for the poem collections, too (see the available OWL files). Also, according to the model in Figure 5, a text can be classified by *zero* to *many* documentary *works*, *zero* being the case of a text that is not yet organized in an information system.

[10]The diagram in Fig. 5 provides an integration with Approach 2; something similar can be done with Approach 1.

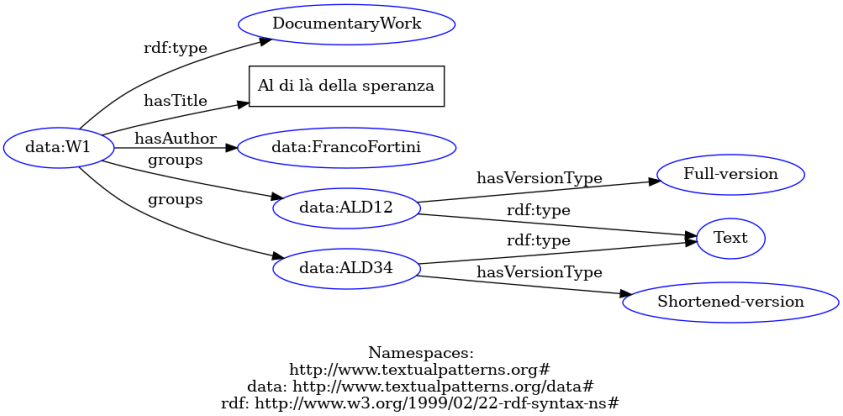

Figure 6: Example of RDF graph based on Approach 3

Second, to avoid the proliferation of elements, one might keep relations of *inclusions* only between texts but, along the lines of LRM [3], something similar could be done at the level of documentary *works*, too, if applications require it. Third, the relationship of *grouping* between *DocumentaryWork* and *Text* does not propagate over *inclusion* between texts. That is, if a documentary *work* x groups a text t_1, and t_1 includes text t_2, x does not group t_2. This choice is due to the idea that texts are grouped by a documentary *work* when they share a common content. For instance, the *work* of *Poesia ed errore*, call it x, groups the text of *Poesia ed errore* (t_1), the latter including the text of *Al di là della speranza* (t_2). Should x grouping both t_1 and t_2, this would mean that both texts convey the same content, which is not the case considering that we have a collection of poems on the one hand, and a single poem on the other hand. Hence, the following OWL 2 object property chain does *not* hold in Approach 3:

$$groups \; o \; includes \sqsubseteq groups \tag{3}$$

Fourth, when a text is grouped by a documentary *work*, one might model, as we did in the graph in Figure 6, information relative to authorship at the *work* level in order to access data by specific authors and avoid repeating the same information for texts. However, as for the previous cases, authorship could be added at the level of the texts, too, if needed by application requirements. Note that from an ontological stance, it would be more appropriate to attach authorship primarily to texts rather than *works*, given that documentary *works* are built only *a posteriori* to organize texts. The choice of relating authorship to *works* can be therefore seen as a practical convention to avoid the proliferation of data.

SPARQL query 3 (example). The query below shows how data about the documentary *work* of *Al di là della speranza* is retrieved. Differently from the previous query, note that we now go through an instance of *DocumentaryWork* and the texts it groups. Results are shown in Table 4, where one can clearly see that the single documentary *work data:W1* groups both texts *data:ALD12* and *data:ALD34*.

```
  Select ?poemWork ?poemText ?versionType ?collectionTitle ?publisherName
?pbYear Where
   { ?poemWork a :DocumentaryWork;
     :hasTitle “Al di là della speranza”;
     :groups ?poemText.
   ?poemText a :Text;
     :hasVersionType ?versionType;
     :includedIn ?collectionText.
   ?collectionText a :Text;
     :hasTitle ?titleCollection;
     :hasSituation ?s.
   ?s a :PublicationSituation;
     :hasPublisher/rdfs:label ?publisherName;
     :hasPublicationYear ?pbYear }
```

Table 4
Examples of results for SPARQL query 3

poemWork	pomeText	versionType	collectionTitle	publisherName	pbYear
data:W1	data:ALD12	:Full-version	"Poesia ed errore"	"Feltrinelli"	"1959"
			"Versi primi e distanti"	"All'insegna..."	"1987"
	data:ALD34	:Shortened-version	"Poesie scelte 1939-1973"	"Mondadori"	"1974"
			"Una volta per sempre"	"Einaudi"	"1978"

3.1. Discussion

Having introduced three alternative approaches for the organization of texts in information systems, we now compare them and discuss some of their advantages and shortcomings.

With respect to the model of LRM, Approaches 1-3 cut across the *work-expression-manifestation* distinctions. In particular, Approach 1 targets texts-in-context and, in a sense, it focuses on the level of *manifestations* only. However, texts cannot be equated to LRM's *manifestations*; first, because they do not bear links with entities like *works* or *expressions*, since the latter are not present in Approach 1; second, because *manifestations* in LRM are primarily devoted to the representation of published texts whereas, as said, a text may not be necessarily published in our case. Approach 2 focuses on *expressions*, whereas *situations* enables reference to, e.g., publication information. Texts cannot be equated with LRM's *expressions* for the lack of *works* and *manifestations* in the scope of the presented approach. Finally, Approach 3 extends Approach 2 with the modeling of documentary *works*. As said, the notion of LRM's *work* is affected by ambiguities and has been interpreted in various ways [8]. Although the notion of documentary *work* in Approach 3 is not without problems, in a minimalist sense, it captures a modeling element that is built *a posteriori* from texts that are meant to convey a common content. From this perspective, its intended meaning is more specific when compared to LRM.

In our perspective, Approach 1 is a simple modeling pattern for representing texts with information, e.g., related to publication, when available. A possible disadvantage is that it multiplies texts even when scholars may consider them as being the same. For instance, as we have seen in Table 2, we have four texts for *Al di là della speranza* whereas scholars may tend to think of *data:ALD1* and *data:ALD2* as a unique text (the same for *data:ALD3* and *data:ALD4*).

Approach 2 leads to a more abstract representation in comparison to Approach 1, since texts are distinguished from the contexts where they occur (to be published). For instance, this approach makes sense that a single text can be published at different times by different publishers, possibly even by being included in different (larger) texts. The use of *situations* results useful, because it makes clear the distinction between a text and the situations where it occurs. In our understanding, Approach 2 can be functional in application domains where one is interested in representing, e.g., information relative to publications but also in scholarly domains to capture different sorts of relations between texts, including, e.g., relations of philological derivation.

In the specific interpretation of *documentary works*, these are introduced in Approach 3, which results therefore more abstract than Approaches 1-2. Considering Table 4, it should be clear that this view is able to convey what is assumed to be the common documentary *work* organizing multiple texts. However, as we have discussed, the introduction of *works* as a pragmatic modeling choice should not be taken for granted. Once it is used, there remains plenty of choices to decide what are its boundaries, e.g, how to classify texts in a homogeneous way across multiple information systems while avoiding ad-hoc solutions and relying instead on robust principles that can facilitate data interchange and systems' interoperability. To recall one of our examples, Fortini's reworked the collection of *Poesia ed errore*, firstly published in 1959, publishing it again in 1969 with variations in the collected poems. Should the editions of 1959 and 1969 be classified as a single documentary *work*? Recalling the contributions of Pierazzo and Shillingsburg (see Section 2), there is no absolute reply to this question, as it depends on how the texts are interpreted. Hence, different scholars may come up with different opinions and criteria, leading to heterogeneous modeling choices.

Overall, Approach 2 remains in our understanding the best solution avoiding the introduction of the vague notion of *work* and the difficulties it leads to. As said, on the basis of, e.g., historical information about texts, one may include different sorts of relations between them without "reifying" their supposed conceptual similarity into a *work*, whose introduction seems to be more deleterious than useful.

4. Conclusion

Texts are commonly classified in information systems based on the notion of *work*. As we have shown, determining what constitutes a *work*, including its boundaries and how to differentiate between different *works*, remains a challenging issue for which no straightforward solution seems to exist. This is because, according to our analysis, the concept of *work* inherently involves interpretation; classifying texts based on their "contents" is a decision made by an agent according to its interpretation criteria. In our view, this would not be a problem if the research community could agree on robust shared principles – a kind of algorithmic procedure – to determine whether two or more texts belong to a common *work*. However, years of research indicate that this is unlikely feasible, as disagreements frequently arise regarding how texts are interpreted and therefore classified.

On the basis of the analysis, we presented three alternative ontology-based modeling patterns to organize texts in digital libraries or catalogs. The first two patterns get rid of *works* and rely on the plain modeling of texts. The difference between Approach 1 and Approach 2 depends on the relation between texts and the (publication) contexts where they occur. Approach 3 includes *documentary works* but the advantage of this introduction does not explicitly emerge. In our understanding, Approach 2 is the most promising one: texts are distinguished from their publication contexts resulting in a flexible approach. Also, as said, various sorts of relations can be introduced between texts without however treating their "similarity in content" as an entity on its own.

To strengthen our proposal and in particular to make the patterns functional in Semantic Web applications, their formal elements need to be further characterized and possibly related to existing, available resources. In addition, the patterns need to be enriched with other modeling elements useful to handle texts in applications. Said that, many existing resources rely on LRM's structure (e.g., the suite of SPAR ontologies [2]);[11] therefore reusing them does not come for free as one needs to isolate the elements that do not strictly depend on LRM. In addition, despite we have tried to remain as much as possible close to the way in which texts are treated by experts in different areas, we have relied only on an abstract conceptualization of what they are. However, should one be interested in book or manuscript studies, among others, it would be necessary to introduce elements to model their *physical* level. Data models in this direction exist (e.g., [21]), hence our library of patterns needs to be enlarged to cover them, enabling scholars to include the elements they need.

Acknowledgments

Sanfilippo's work leading to this publication was supported by the project *MITE – Make it explicit: Documenting interpretations of literary fictions with conceptual formal models* funded by the European Union - Next Generation EU. Antonietti's work was supported by the project *BiGraFo – Biblio-grafo. Un catalogo semantico per il Centro di ricerca Franco Fortini*, funded by the University of Siena Research Support Plan 2022 for "Curiosity-driven" projects.

References

[1] M. T. Biagetti, A comparative analysis and evaluation of bibliographic ontologies, in: Challenges and Opportunities for Knowledge Organization in the Digital Age, Ergon-Verlag, 2018, pp. 499–510.

[11] http://www.sparontologies.net/

[2] S. Peroni, D. Shotton, Fabio and cito: Ontologies for describing bibliographic resources and citations, Journal of Web Semantics 17 (2012) 33–43.
[3] C. Bekiari, M. Doerr, P. La Boeuf, P. Riva, LRMoo object-oriented definition and mapping from IFLA LRM (version 0.9.6 september 2023), 2023.
[4] C. Holden, The bibliographic work: History, theory, and practice, in: Cataloging and Classification, Routledge, 2021, pp. 7–26.
[5] L. S. Creider, Cataloging, reception, and the boundaries of a "work", Cataloging & classification quarterly 42 (2006) 3–19.
[6] P. Le Boeuf, A strange model named frbroo, in: The FRBR Family of Conceptual Models, Routledge, 2014, pp. 68–84.
[7] R. P. Smiraglia, Further reflections on the nature of 'a work': An introduction, Cataloging & classification quarterly 33 (2002) 1–11.
[8] E. M. Sanfilippo, Ontologies for information entities: State of the art and open challenges, Applied ontology 16 (2021) 111–135.
[9] M. Pietras, L. Robinson, Three views of the "musical work": bibliographical control in the music domain, Library Review 61 (2012) 551–560.
[10] J. de Berardinis, V. A. Carriero, A. Meroño-Peñuela, A. Poltronieri, V. Presutti, The music meta ontology: a flexible semantic model for the interoperability of music metadata, arXiv preprint arXiv:2311.03942 (2023).
[11] R. d. A. Falbo, G. Guizzardi, A. Gangemi, V. Presutti, Ontology patterns: clarifying concepts and terminology, in: Proceedings of the 4th Workshop on Ontology and Semantic Web Patterns, volume 1188, 2013.
[12] D. Davies, C. Matheson, Introduction, in: D. Davies, C. Matheson (Eds.), Contemporary readings in the philosophy of literature: An analytic approach, Broadview press, 2008.
[13] A. Compagnon, Le démon de la théorie. Littérature et sens commun, Média Diffusion, 2014.
[14] C. Masolo, E. M. Sanfilippo, R. Ferrario, E. Pierazzo, et al., Texts, compositions, and works: A socio-cultural perspective on information entities., in: JOWO, 2021.
[15] E. Pierazzo, Digital scholarly editing: Theories, models and methods, Routledge, 2016.
[16] P. Shillingsburg, How literary works exist: Implied, represented, and interpreted, Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions-Open Book Publishers (2010) 165–82.
[17] R. P. Smiraglia, The history of "the work" in the modern catalog, Cataloging & classification quarterly 35 (2003) 553–567.
[18] E. M. Sanfilippo, R. Freedman, Ontology for analytic claims in music, in: European Conference on Advances in Databases and Information Systems, Springer, 2022, pp. 559–571.
[19] V. A. Carriero, A. Gangemi, M. L. Mancinelli, A. G. Nuzzolese, V. Presutti, C. Veninata, Pattern-based design applied to cultural heritage knowledge graphs, Semantic Web 12 (2021) 313–357.
[20] A. Gangemi, P. Mika, Understanding the semantic web through descriptions and situations, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer, 2003, pp. 689–706.
[21] A. Felicetti, F. Murano, Scripta manent: A CIDOC CRM semiotic reading of ancient texts, International Journal on Digital Libraries 18 (2017) 263–270.