OXFORD UNIVERSITY PRESS | Digital Scholarship in the Humanities

A Formal Representation of the Divine Comedy's Primary Sources: The Hypermedia Dante Network Ontology

Journal:	*Digital Scholarship in the Humanities*
Manuscript ID	DSH-2021-0035.R1
Manuscript Type:	Short Paper
Date Submitted by the Author:	01-Jul-2021
Complete List of Authors:	Bartalesi, Valentina; CNR, ISTI Pratelli, Nicolò; CNR, ISTI Metilli, Daniele; CNR, ISTI Tomazzoli, Gaia; University of Pisa, Dipartimento di Filologia, Letteratura e Linguistica Livraghi, Leyla; University of Pisa, Dipartimento di Filologia, Letteratura e Linguistica Zaccarello, Michelangelo; University of Pisa, Dipartimento di Filologia, Letteratura e Linguistica Meghini, Carlo; CNR, ISTI
Keywords:	Ontology, Semantic Web, Divine Comedy, Dante Alighieri, Digital Humanities

SCHOLARONE™
Manuscripts

A Formal Representation of the Divine Comedy's Primary Sources: The Hypermedia Dante Network Ontology

Valentina Bartalesi

ISTI-CNR, via Moruzzi 1, 56124 Pisa, Italy. E-mail: valentina.bartalesi@isti.cnr.it

Nicolò Pratelli

ISTI-CNR, via Moruzzi 1, 56124 Pisa, Italy. E-mail: nicolo.pratelli@isti.cnr.it

Daniele Metilli

ISTI-CNR, via Moruzzi 1, 56124 Pisa, Italy & Università di Pisa, Largo Bruno Pontecorvo 3, 56127

Pisa, Italy. E-mail: daniele.metilli@isti.cnr.it

Gaia Tomazzoli

Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Piazza Torricelli 2, 56126,

Pisa, Italy. E-mail: gaia.tomazzoli@fileli.unipi.it

Leyla M. G. Livraghi

Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Piazza Torricelli 2, 56126,

Pisa, Italy. E-mail: leyla.livraghi@fileli.unipi.it

Michelangelo Zaccarello

Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Piazza Torricelli 2, 56126,

Pisa, Italy. E-mail: michelangelo.zaccarello@unipi.it

Carlo Meghini

ISTI-CNR, via Moruzzi 1, 56124 Pisa, Italy. E-mail: carlo.meghini@isti.cnr.it

Abstract

Hypermedia Dante Network (HDN) is a three-year Italian National Research Project (PRIN), started in 2020, that aims to enrich the functionalities of the DanteSources Digital Library in order to efficiently represent knowledge about the primary sources of Dante's *Comedy*. DanteSources allows users to retrieve and visualize the list and the distribution of Dante's primary sources that have been identified by recent commentaries of five of Dante's minor works (i.e. *Vita nova*, *De vulgari*

eloquentia, *Convivio*, De *Monarchia*, and *Rime*). The digital library is based on a formal ontology expressed in RDFS language. Based on the DanteSources experience, the HDN project aims to formally represent the primary sources of the *Divine Comedy* whose identification is based on several commentaries included in the Dartmouth Dante Project (DDP) corpus. To reach this goal, we restructured and extended the DanteSources ontology in order to provide a wider and more complete representation of the knowledge concerning the primary sources of the *Comedy*. In this paper we present the result of this effort, i.e. the HDN ontology. The ontology is expressed in OWL and has as reference ontologies the CIDOC CRM and its extension FRBRoo, including its in-progress reformulation LRMoo. We also briefly describe the semi-automatic tool that will be used by the scholars to populate the ontology.

1 Introduction

Since 2013, ISTI-CNR and the Department of Philology, Literature and Linguistics of the University of Pisa have collaborated on the application of Semantic Web technologies to represent the primary sources of Dante Alighieri's oeuvre. As a first result of this collaboration, in 2015 we released the DanteSources1 Digital Library (Bartalesi et al., 2018). Using DanteSources, users can retrieve and visualize the list and distribution of Dante's primary sources that have been identified by recent commentaries of five of the poet's so-called minor works: Vita nova, De vulgari eloquentia, Convivio, Monarchia, and Rime. The digital library is based on a formal ontology expressed in RDFS (Brickley and Guha, 2014) and aimed at representing Dante's works and the knowledge about his primary sources as reported in some authoritative commentaries (Bartalesi and Meghini, 2017).

Taking into account the DanteSources experience, the Hypermedia Dante Network (HDN) project2 (Meghini et al., 2020) will formally represent the primary sources of the Divine Comedy as they are identified by several commentaries digitized by the Dartmouth Dante Project (DDP). HDN is a three-years Italian National Research Project, started in 2020, and aimed at enriching the functionalities of the DanteSources Digital Library in order to efficiently recover knowledge about the primary sources of Dante's Comedy. To reach this goal, the DanteSources ontology was restructured and extended in order to provide a broader representation of the knowledge concerning Dante's primary sources.

With the HDN project, ISTI CNR and the Department of Philology, Literature and Linguistics of the University of Pisa are pursuing the goal of digitizing Dante's works and related knowledge. The ultimate goal is to build a digital library that responds to two main needs: firstly, to make available to the humanistic scientific community a set of data that respect the FAIR principles, and on which scientific hypotheses can be formulated and tested, possibly also through digital agents; secondly, to encourage the creation of innovative applications based on information technologies that allow the use of works of Italian literature. The HDN project moves from different needs and motivations, first of all the need to overcome the limitations imposed by the paper support, in which traditionally, and to a large extent also currently, knowledge of Italian literature is expressed. The comments to Dante's works represent a real treasure, often of similar importance to that of the same works to which they refer, and of which they constitute a fundamental complement. The massive digitization of texts and related exegetical materials undertaken over the last few decades has not solved the problem, but has only partially alleviated it. This is because the tool used for digitization, the eXtensible Markup Language (XML), is not suitable for the double integration of the works with comments and comments between them, since it is based on a single hierarchical structure. As will be seen in the course of this article, the integration of exegesis relating to the intertextual relations with the original

72 structure of the Commedia requires the coexistence of multiple structures within a single knowledge
73 base: this configuration is impossible if the coding tool requires you to choose a single structure.

74 In this paper, we describe the HDN ontology, that was integrated with the Narrative Ontology
75 (Meghini et al., 2021) and has as reference ontologies the CIDOC CRM (Doerr, 2003) and its
76 extension FRBRoo (Doerr et al., 2008), including its in-progress reformulation, LRMoo (Riva and
77 Žumer, 2017). Furthermore, we briefly introduce the semi-automatic tool we developed for scholars
78 to use in populating the ontology and building HDN knowledge base.
79 To make this article easier to read, we have drawn our examples only from commentaries
80 written in English (and namely those published by Rev. Tozer in 1901, by John Ruskin in 1903, by
81 John S. Carroll in 1904, and by C.H. Grandgent in 1909-13). The English translation of Dante's
82 Divine Comedy quoted in this article is Hollander 2000-2007.

83
84 The paper is organized as follows: in Section 2 we report the state of the art of digital projects
85 about Dante. Section 3 describes our conceptualization of the features of the Comedy's primary
86 sources that we are interested in representing. In Section 4 we describe the HDN ontology that has
87 been developed to express said conceptualization in a formal way. In Section 5 we briefly introduce
88 the semi-automatic tool we have built to populate the ontology. Finally, Section 6 reports our
89 conclusions and future work.
90

2 Related Works

91

92 The first two major projects for the study of Dante Alighieri and his works through software tools
93 were developed in the 1980s and 1990s: the Dartmouth Dante Project[3] (DDP) and the Princeton Dante
94 Project[4]. In the 1990s, several Websites about Dante appeared on-line, such as Digital Dante[5] and The
95 World of Dante[6]. These first projects consisted in the publication of Dante's texts with some
96 comments. Along with these contents, multimedia documents such as images and videos were often
97 provided in order to create richer Web portals. This is the case of Danteworlds[7] and Dante Online[8]. In
98 the 2000s, the advent of standard formats for the representation of information, and especially those
99 developed by the Text Encoding Initiative (Cummings, 2013), was the scientific base for the Dante
100 Medieval Archive[9] and, more broadly, for large collections of texts in XML-TEI format – such as the
101 Perseus Digital Library[10] or the Biblioteca Italiana[11]. The DanteSearch[12] project, on the other hand,
102 provided a complete lemmatization of Dante's works, both in Latin and in the vernacular. At present,
103 two ongoing projects on Dante's lexicon – the Vocabolario Dantesco[13] and the Vocabolario Dantesco
104 Latino[14] – use DanteSearch as their point of departure. Another interesting project which is currently
105 in the making is the Illuminated Dante Project[15], that aims to provide a catalogue of early
106 illuminations of Dante's *Comedy,* which may disclose references to the poem or the poem's
107 commentary. Manuscript illuminations will be described and provided to users through a high-
108 definition archive, whose images are digitized according to the IFLA[16] and FADGI[17] standards, thus
109 adopting copyright licenses that allow re-use, and thanks to web-interoperability protocols such as
110 IIIF (Snydman et al., 2015).

111 The advent of the Semantic Web and the Linked Open Data paradigm has fostered the development of
112 Digital Humanities (DH) projects that use these technologies to give a formal representation of the
113 collected knowledge, to make the knowledge more interoperable and produce FAIR data (Wilkinson
114 et al., 2016). In recent years, several projects focused on different aspects of DH have successfully
115 used Semantic Web technologies. We report here some recent projects we studied and consider

successful examples. The first example is the WarSampo knowledge graph (KG), a shared semantic infrastructure and a Linked Open Data (LOD) service (Koho et al., 2021b) for publishing data about the Second World War, focusing on Finnish military history. To create aggregated global views of the war, a shared ontology and data infrastructure was developed to harmonize information from different sources. This made it possible to share data between publishers and application developers, support data analysis in DH research, and develop data-driven intelligent applications. The metadata schema was an extension of CIDOC CRM, supplemented by various military history domain ontologies. Semantic Web technologies have recently been used for exploring a corpus of data in the History of Science field, that is the Henri Poincaré correspondence (Bruneau et al., 2021). This is a corpus of letters sent and received by the mathematician. The edition of this correspondence is a long-term project that begun during the 1990s. Since 1999, a website is devoted to publishing online this correspondence with digitized letters. In 2017, an RDFS knowledge base was created, which is accessible through SPARQL queries, to make the corpus more interoperable. Another application of the LOD paradigm in DH is the Mapping Manuscript Migrations (MMM) project (Koho et al., 2021a). MMM harmonized and published heterogeneous premodern manuscript metadata as Linked Open Data. A semantic portal and a LOD service were created to integrate distinct manuscript datasets into a shared platform for search and discovery. Another interesting example of the use of the LOD in DH is the Digitised Manuscripts to Europeana (DM2E) project. Within DM2E, a Linked Data source (Baierer et al., 2017) of digitised manuscripts for the digital humanities was developed. The Linked Data source provides metadata and links for direct access to digitized content from various cultural heritage institutions across Europe. The data model used in the project is a specialization of the Europeana Data Model (Doerr et al., 2010) and satisfies specific requirements from the domain of manuscripts and old prints, as well as from developers who want to create applications on top of the data. The LOD approach is used also for exploring a corpus of data in the ancient geography field. The Pelagios - Enable Linked Ancient Geodata In Open Systems (Simon et al., 2016) is a collective of projects connected by a shared vision of a world in which the geography of the past is as interconnected, interactive and interesting as the present. The Pelagios project created a Linked Open Data source to interlink online resources that refer to places in the historical past. Finally, Semantic Web technologies are also used to formally represent and exploring Dante's primary sources. DanteSources[1] (Bartalesi et al., 2018) (2013–2016) was the first project to provide a semantic representation of the knowledge related to the primary sources of Dante's so-called minor works. The Hypermedia Dante Network project[2] plans to extend the experience of DanteSources, focusing on the semantic representation of the primary sources of Dante's Comedy. To do so, HDN employs the standard technologies of the Semantic Web, such as RDF (Schreiber and Raimond, 2014), OWL (W3C OWL Working Group, 2012), SPARQL (W3C SPARQL Working Group, 2013). Following the Linked Data paradigm, HDN aims at reusing contents and knowledge, making them standardized and freely accessible. This is why the commentaries used by HDN to identify Dante's primary sources are taken from the corpus collected by the DDP project, which includes commentaries from the 14th to the 20th century written in Latin, Italian and English.

3 Conceptualization

To conceptualize Dante's use of primary sources and their identification and description in ancient and modern commentaries, a key notion is that of *exegesis*. Exegesis is the critical explanation or interpretation of a text, including the study of the historical and cultural background of the author, the work, and its intended audience (Britannica, 2020). The primary sources of a text, therefore, provide

160 important knowledge about said background, allowing the scholar to have a more complete
161 understanding of the text itself.

162 In the HDN project, we are interested in the study of primary sources of the *Divine Comedy*,
163 and we rely on the authoritative commentaries collected by the Dartmouth Dante Project (DDP).
164 Dante's commentators have analyzed and explained Dante's text, and in doing so they have
165 consistently identified *references* to previous authors or works that Dante may have read, as well as
166 concepts, theories or other relevant features that might have inspired him (e.g. a physical location, a
167 common place, etc.).

168 *Reference* is the main concept that we aim to describe and represent in our ontology. We
169 define a Reference as a statement following this structure:

170 A says that the knowledge about B can be enriched through C

171

172 Where:
- 173 A is a fragment of text that asserts the reference, and is therefore called the *source* of the
174 reference. The source typically belongs to an authoritative commentary: it can be, for
175 example, a fragment of text from Rev. Tozer's commentary on *Inf.* 1, 73-5: "*Aeneas is
176 described by this epithet in Virg. Aen. I. 544, 545*, 'quo iustior alter Nec pietate fuit'";
- 177 B is a fragment of text that is clarified by the reference and is therefore called the *subject* of
178 the reference. In our project, the subject belongs to Dante's *Comedy*, such as, with respect to
179 Tozer's commentary quoted above, verses 73-4 from *Inf.* 1: "*Poeta fui, e cantai di quel giusto
180 / figliuol d'Anchise che venne di Troia, / poi che 'l superbo Ilión fu combusto*" ('I was a poet
181 and I sang / the just son of Anchises come from Troy / after proud Ilium was put to flame');
- 182 C is the *object* of the reference, i.e. the entity to which the reference points, according to the
183 source. In other words, the object of the reference is the textual or conceptual entity that the
184 source considers useful for explaining the subject: in this case, a fragment of text from
185 Virgil's *Aeneid*, book I, ll. 544-5: "*quo iustior alter / Nec pietate fuit*".

186 In most of the cases that interest us, the object is a fragment of text, as reported in the example
187 above. However, there are some instances where this is not the case, and namely:

- 188 when the object is not a single work, but a set of works that is well characterized either (i)
189 from an extensional point of view, such as "*i maestri parigini*" (the Parisian masters): in such
190 occurrences, the object of the reference is said set of works; or (ii) from an intensional point
191 of view, such as "Neoplatonism": in this case the object of the reference is a concept;
- 192 when the object is an entity – such as a character, a work, or a theory – not necessarily
193 referable to a specific text fragment.

194 On the ground of this definition, we identified three subtypes of reference:

1. *External support*, a reference where the source identifies an object that supports its
196 interpretation of the subject, but that is not to be considered as a primary source of the subject
197 that is being annotated. A commentary cites an external support for demonstrative purposes,
198 for example to argue that the author of the subject may or may not know a certain text; see,
199 for example, Carroll's comment on *Purg.* 30, 16-8 ("*cotali in su la divina basterna / si levar
200 cento, ad vocem tanti senis, / ministri e messaggier di vita etterna*", 'there, on the sacred
201 chariot, rose up / *ad vocem tanti senis*, one hundred / ministers and messengers of life

202 eternal') puts forward three biblical passages to show that "they are the Angel-guardians of
203 the Church, and in the present connection are probably regarded as the medium of Divine
204 Revelation. The law of Moses was 'ordained through angels' [*Acts*. VII. 53; *Gal*. III. 19; *Heb*.
205 II. 2]".

206 2. *Loci paralleli*, a reference where the source identifies an object – be it an idea, a character, or
207 a stylistic feature – that displays close similarity with the subject, without necessarily
208 implying a direct derivation (i.e. the development of such features in the two texts might be
209 polygenetic). For example, in commenting the expression "selva oscura" ('dark wood', *Inf*. 1,
210 2), Rev. Tozer notes that "similarly in *Conv*. IV. 24, ll. 123-6, Dante uses the term 'the wood
211 of error of this life' for the world and its temptations";

212 3. *Citation*, a reference where the source suggests that the subject directly descends from the
213 object. These are clearly the most specific references, and they are at the centre of our
214 ontology; when Dante meets Beatrice in the garden of Eden, he acknowledges to Virgil
215 "conosco i segni de l'antica fiamma" ('I felt / the overwhelming power of that ancient love',
216 *Purg*. 30, 48), citing Virgil's own verses from the *Aeneid*, as noted by Tozer: "a translation of
217 Virgil's words in *Aen*. IV. 23, 'Agnosco veteris vestigia flammae'".

218 In addition, we identified three features that references might possess: its type, its content, and
219 the relationship between subject and object. These features are always relevant to citations, but not
220 necessarily to external supports and *loci paralleli*.

221 As to the type, the ontology describes three types of citations, which were already formalized by the
222 DanteSources ontology (Bartalesi et al., 2018):

223 1. *explicit citations*, that is references explicitly made by Dante; an example can be found in
224 *Purg*. 24, 49-51, where Dante explicitly cites his *canzone* "Donne ch'avete intelletto
225 d'amore" (*Vita Nova* XIX, 4): "*Ma dì s'i' veggio qui colui che fore / trasse le nove rime,
226 cominciando / 'Donne ch'avete intelletto d'amore'*" ('But tell me if I see before me / the one
227 who brought forth those new rhymes / begun with Ladies that have intelligence of love');
228 Carroll, in his commentary, pinpoints the reference by saying: "This is the first line of the first
229 Canzone of the *Vita Nuova*".

230 2. *strict citations*, that is references to a specific work and fragment as identified by a scholar;
231 e.g. Rev. Tozer's commentary claims that in writing *Par*. 33, 31-3 ("*perché tu ogne nube li
232 disleghi / di sua mortalità co' prieghi tuoi, / sì che 'l sommo piacer li si dispieghi*", 'so that
233 your prayers disperse on his behalf / all clouds of his mortality and let / the highest beauty be
234 displayed to him') "Dante was thinking of Virg. *Aen*. II. 604-6, 'Aspice, namque omnem,
235 quae nunc obducta tuenti Mortales hebetat visus tibi et umida circum Cali at nubem eri iam'".

236 3. *generic citations*, that is references to a concept or set of works put forward by a scholar, e.g.
237 when Carroll, in explaining Dante's tercets about the possible intellect ("*quest'è tal punto, /
238 che più savio di te fé già errante, / sì che per sua dottrina fé disgiunto / da l'anima il possibile
239 intelletto, / perché da lui non vide organo assunto*", 'This is the point / at which a wiser man
240 than you has stumbled / in that his teaching rendered separate / the possible intellect from the
241 soul, / because he could not find the organ it could live in', *Purg*. 25, 62-6,), notes that
242 "scholastic philosophy drew a distinction between the 'possible intellect' and the 'active
243 intellect'".

245 The content of a citation is the entity that, according to the source, the subject is deriving from
246 its object. We have identified three types and seven sub-types of content:

1. Textual correspondences: these are references to textual fragments identified by linguistic and stylistic features. These are further divided into:

   (a) *image*, that is to say when the source identifies in the object-text an image on which the subject-text relies for a description or conceptualization, as it is the case with Carroll's reference to Statius' *Thebaid* to explain Dante's comparison about a two-horned flame (in *Inf.* 26, 52-4: "*chi è 'n quel foco che vien sì diviso / di sopra, che par surger de la pira / dov'Eteòcle col fratel fu miso?*", 'Who is in the flame so riven at the tip / it could be rising from the pyre / on which Etèocles was laid out with his brother?'): "Dante compares the two horns into which their flame parted to those which rose from the funeral pyre of Eteocles and Polynices. These two brothers quarrelled over the succession to the throne of their father, OEdipus of Thebes, and slew each other in the war of the Seven against Thebes which followed. Even death could not quench their hatred: the very flames which consumed their bodies divided and refused to mingle [Statius, *Thebaid*, XII. 431]";

   (b) *stylistic feature*, which occurs when the source identifies a linguistic and/or rhetorical precedent for an expression included in the subject-text, e.g. when Rev. Tozer, while commenting Dante's tercet "*Ed essi quinci e quindi avien parete / di non caler – così lo santo riso / a sé traéli con l'antica rete!*" ('[my eyes] walled off from anything around them, enclosed / in their indifference, so did the holy smile / ensnare them in its old, familiar net', *Purg.* 32, 4-6), suggests that "The form of expression was probably suggested by *Ex.* XIV. 22, 'The waters were a wall unto them on their right hand and on their left'".

2. Thematic correspondences: references to textual fragments identified by thematic features. These are further divided into:

   (a) *character*, when the source relates a character mentioned by Dante to a specific object-text, e.g. C. Grandgent does when he traces Dante's knowledge about Semiramis ("*Ell'è Semiramìs, di cui si legge / che succedette a Nino e fu sua sposa: / tenne la terra che 'l Soldan corregge*", *Inf.* 5, 58-60, 'She is Semiramis, of whom we read / that she, once Ninus' wife, succeeded him. / She held sway in the land the Sultan rules') back to the Latin historian Paolo Orosio ("Semiramis, queen of Assyria, of whom Dante had read in the *Historia* of Paulus Orosius, I, iv);

   (b) *episode*, when the source identifies a literary episode as an antecedent for the subject-text, e.g. when Rev. Tozer explains that the episode of the "miserable magpies" mentioned by Dante in his purgatorial invocation ("*e qui Caliopè alquanto surga / seguitando il mio canto con quel suono / di cui le Piche misere sentiro / lo colpo tal, che disperar perdono*", *Purg.* 1, 9-12, 'Here let Calliope arise / to accompany my song with those same chords / whose force so struck the miserable magpies / that, hearing it, they lost all hope of pardon') is to be found in Ovid's *Metamorphoseon libri* ("The story here referred to, which is given by Ovid, *Met.* V. 294 foll., is that of the nine daughters of Pierus, king of Macedonia, who challenged the nine Muses to a contest in singing, on which occasion the latter were represented by Calliope as their champion. The challengers after they were defeated were changed into magpies");

   (c) *topography*, when the source declares that a literary memory lies behind Dante's mention of a place, e.g. when Tozer suggests that Dante might be calling his infernal river 'Flegetonta' ("*fanno Acheronte, Stige e Flegetonta*", 'where they form Acheron, Styx, and Phlegethon', *Inf.* 15, 116) because he "may have got it from the *Culex* (l.

271), which in his time was regarded as a work of Virgil's, or from Servius' Commentary on *Aen*. VI. 265".

3. Conceptual correspondences: these are references to textual fragments identified by conceptual features. These are further divided into:

(a) *motif*, that is to say when the source identifies in the object-text the precedent for the development of one of Dante's motifs, such as Grandgent's reference to *Psalm* 121 in his interpretation of *Par*. 25, 38-9 ("*ond'io levai li occhi a' monti / che li 'ncurvaron pria col troppo pondo*", 'Therefore I raised my eyes up to the hills / whose blazing light had weighed them down before'): "*Ps*. 121 (*Vulg*. 120), 1: 'I will lift up mine eyes unto the hills, from whence cometh my help'";

(b) *theory*, when the source identifies in the object-text a theory upon which Dante's text relies, e.g. when Carroll reminds the reader Dante's choice of placing children dead before their baptism in the limbo ("*Or vo' che sappi, innanzi che più andi, / ch'ei non peccaro; e s'elli hanno mercedi, / non basta, perché non ebber battesmo, / ch'è porta de la fede che tu credi*", 'I want you to know, / before you venture farther, / they did not sin. Though they have merit, / that is not enough, for they were unbaptised, / denied the gateway to the faith that you profess', *Inf*. 4, 33-6) is inspired by Thomas Aquinas: "Without doubt he accepted the merciful view of his master in theology, St. Thomas Aquinas, that such infants are simply excluded from Paradise and suffer no pain even of loss: 'as they are not made capable of possessing the vision of God, they no more grieve for its loss than a bird does that it is not an emperor or a king. Moreover, though not united to God in glory, they are joined to Him by the Share they possess of natural goods, and are able to rejoice in Him by natural knowledge and love' [Appendix to Supplement, q. i. a. 2]".

Finally, the relationship between the subject and the object of the citation is classified according to three categories: correction (or contradiction), extension (or re-elaboration), and confirmation (or homology). In other words, besides just relying on a primary source, the subject-text can use an object-text as its point of departure while contradicting or extending it.

4 The HDN Ontology

The HDN ontology specifies the above conceptualization and is an evolution of the one developed for the DanteSources Digital Library (Bartalesi et al., 2018); it is integrated with the Narrative Ontology (Meghini et al., 2021), and has as reference ontologies the CIDOC CRM (Doerr, 2003) and its extension FRBRoo (Doerr, 2008), including its in-progress reformulation, LRMoo (Riva and Žumer, 2017).

The HDN ontology has been implemented using the OWL language of the Semantic Web (W3C OWL Working Group, 2012).

As a notational convention, the CIDOC CRM adopts the letters "E" and "P" to indicate classes and properties, respectively. On the other hand, FRBRoo (and its recent revision LRMoo) adopts the letters "F" and "R" to indicate classes and properties, respectively.

4.1 Classes

In this Section we describe the main classes of our ontology and the relative axioms. Due to the focus of the HDN project, we highlighted Dante's work by introducing the Dante's Work class as a subclass of F2 Expression. The different parts of Dante's works are also instances of F2 Expression, but they are characterized by a more specific type that describes their structural nature as it has been determined by the author. The types of parts defined by the ontology are, in a descending order of granularity: Cantica, Canto, and Verso.

In addition to these three structural divisions of Dante's *Comedy*, we also define two more classes, Sentence and Text Fragment, that allow us to represent other subsections of the text. The class Sentence represents a full sentence, and it is independent from any structural division (e.g. it may be shorter than a verse, or, on the contrary, span over two or more verses). The class Text Fragment is more generic and represents any possible fragment of text (e.g. a single word).

We also define a class Position that is used to indicate the position of a specific Text Fragment (e.g. a word) within the F2 Expression that contains it.

The commentaries to Dante's text, e.g Rev. Tozer (1901) or John Ruskin (1903), are represented as instances of the Commentary class, which we defined as a subclass of F2 Expression. The subparts of the commentaries, including individual notes, are represented as instances of E90 Symbolic Object, e.g. the IRI that identifies the note from John Ruskin (1903), Inf. I, 16: "*For classical and patristic texts that present mankind as upright and up-looking, and thus as different from (and vastly superior to) the beasts, see Mazzoni, Saggio di un nuovo commento alla "Divina Commedia": "Inferno" –Canti I-III (Florence: Sansoni, 1967), pp. 61-68.*".

The ontology defines a class for each category of reference, and namely: Reference, External Support, Loci Paralleli, and Citation. Loci Paralleli and External Support are subclasses of Reference. Citation is a subclass of Loci Paralleli. Reference is defined as a subclass of E73 Information Object. The taxonomy of the classes seen so far is shown in Figure 1 (white boxes represent classes from other ontologies, while arrows connect classes to their superclasses).

Insert Figure 1 here

The citation types and subject-to-object relationships are individuals that are represented as instances of the E55 Type class. To distinguish them, the ontology includes two subclasses of E55 Type:

- Citation Type, having as instances: genericCitation, strictCitation, and explicitCitation;
- Subject-Object Relationship, having as instances: correction/contradiction, extension/re-elaboration, and confirmation/homology.

The subject of the Citation can vary significantly, therefore we consider it as an instance of E1 CRM Entity, the most general class of CRM.

The content of the reference is an instance of one of the following classes:

- E33 Linguistic Object, for textual correspondences – i.e. images and stylistic features;
- E21 Person, for characters;
- E5 Event, for episodes;
- E53 Place, for topographical elements;
- Sentence, for motifs;
- E89 Propositional Object, for theories.

380 Note that the ontology does not provide specific classes for those entities that lie on the first level of
381 our categorization (i.e. textual, thematic and conceptual correspondences), because this level has an
382 illustrative role that will be used only for the entry or visualization of our knowledge.

4.2 Properties

384 In this Section we describe the main properties of our ontology and their relative axioms.

385 *Relationship Between a Text Fragment and the Work It Belongs to*

386 An instance of Text Fragment is linked to the part of the work to which it belongs by the property
387 R15i is Fragment of, which has E90 Symbolic Object as domain and F2 Expression as range, e.g. the
388 IRI corresponding to the text fragment "*Nel mezzo del cammin di nostra vita*", ("Midway upon the
389 journey of our life") R15 is fragment of the IRI corresponding to the I Canto of Inferno. Note that R15
390 is a subproperty of P106 is composed of. The work to which the fragment belongs can be a work of
391 Dante's, a work by another author or a commentary.
392   The part of the work to which a fragment of text belongs is linked to the part that includes it
393 by the property P148 is component of, which has E89 Propositional Object both as domain and as
394 range. Note that F2 Expression is a subclass of E73 Information Object, which is in turn a subclass of
395 E89. The same property will be used to access the entire work through its structural components. For
396 example, a fragment of the *Divine Comedy* composed of two verses, e.g. the IRI that identified the
397 verses "*Nel mezzo del cammin di nostra vita/mi ritrovai per una selva oscura*" ("Midway upon the
398 journey of our life / I found myself within a forest dark"), is linked to its canto, e.g. the IRI of the
399 canto I, by P148. Then, the canto is linked to the relevant cantica, e.g. the IRI of Inferno, and the
400 cantica will be linked to the *Divine Comedy* through the same property P148. This part of the
401 ontology is represented in Figure 2 (labelled arrows connect the domain and the range of the property
402 expressed by the label).
403

404 Insert Figure 2 here
405

406

407 The location of the text fragment within the work to which it belongs is represented by the indication
408 of both the position where the fragment begins and the position where it ends; these two positions are
409 expressed by the two properties:
410

411   ● fromPosition, which indicates the beginning position;
412   ● toPosition, which indicates the end position.
413

414 Both properties have Text Fragment as domain and Position, a class we defined, as range. Each
415 position is represented by two coordinates: (i) the unitary part (the verse for the *Divine Comedy* or the
416 note when it comes to the commentaries) and (ii) the offset, a positive integer that indicates the
417 position of the word within the part: one for the first word, two for the second, and so on.
418 Consequently, each instance of the Position class is characterized by two properties:
419

420   ● positionUnit, which has Position as domain and F2 Expression as range;
421   ● positionOffset, which has Position as domain and xsd:int as range.
422

423 Figure 3 shows the classes and properties introduced to indicate the fragment's position:

424
425
426   Insert Figure 3 here
427

Relationship Between a Work, or a Fragment, and Their Content

429   Every part of a literary work, as well as every fragment of a text, has a textual content. From an
430   ontological point of view, the work or fragment coincides with its textual content. Formally, it is
431   common practice to represent the two entities as different. Our ontology conforms to this practice,
432   therefore it introduces the property "has content", which binds an instance of the class E90 Symbolic
433   Object to an RDF literal, i.e. an instance of rdfs:Literal, which represents its textual content. For
434   example, the IRI https://hdn.dantenetwork.it/ontology/resources/cantica/1/canto/1/verso/1 has content
435   "Nel mezzo del cammin di nostra vita". The property "has content" is a subproperty of P3 has note.

Properties to Describe a Work or Commentary

437   The works that we represent, including both Dante's works and commentaries about them, are
438   described using classes and properties from the CRM. In particular:

439   ● The relation between a Commentary and the Dante's Work it refers to is captured by the
440   property P129 is about, e.g. the IRI the identifies the commentary C.H. Grandgent (1909-13)
441   P129 is about the IRI that identifies the Divine Commedy;
442   ● The title of a Dante's Work or Commentary is expressed through the property P102 has title,
443   e.g. the IRI the identifies the commentary C.H. Grandgent (1909-13) P102 has title "*La
444   Divina Commedia di Dante Alighieri*, edited and annotated by C. H. Grandgent".
445   ● The author of a Dante's Work or Commentary is connected to it through an event of type F28
446   Expression Creation, which is linked to an instance of E39 Actor by the property P14 carried
447   out by.

Properties for References

449   Each reference has three constituent elements: the source, the subject, and the object. The link among
450   these parts and the reference to which they belong are described by the following properties, and are
451   graphically represented in Figure 4 (blue arrows connect properties to their super properties):
452

453   ● hasRefSource links a reference to its source, that is to the text of the commentary that
454   contains it. This property has Reference as domain, and Sentence as range. This property is
455   functional, but not inverse functional, because the same note can contain two or more
456   references. Semantically, this property links a natural language sentence to the symbolic
457   representation of its meaning.
458   ● hasRefSubject links a reference to its subject, and it has Reference as domain, and Text
459   Fragment as range. This property is functional, but not inverse functional, because the same
460   note can contain two or more references. The property has a mereological nature because it
461   links the whole (the reference, an E73 Information Object) to a part of it (the subject, an
462   instance of E90 Symbolic Object), so it is a subproperty of P106 is composed of, which has
463   E90 as domain and range.
464   ● hasRefObject links a reference to its object, that is the entity reused by the subject-text
465   according to the source. The domain of the property is Reference, and the range is E1 CRM

466   Entity. This property is functional, but not inverse functional, because the same entity can be
467   the object of two or more references. The property is a subproperty of P129 is about.

468

469   Insert Figure 4 here

470

471   *Relationship Between Citation and Type*

472   The Citation Type (i.e. explicit, strict, and generic) applies only to the Citation. To link the Citation to
473   its type we use the property P2 has type, which has E1 CRM Entity as domain and E55 Type as range.

474   *Relationship Between Loci Paralleli and Its Content*

475   The link between *loci paralleli* and its content is represented by the hasRefContent property, which
476   has LociParalleli as domain and E1 CRM Entity as range. Note that the property also applies to
477   citations, since these are a special case of *loci paralleli*. hasRefContent is a sub property of P129 is
478   about, which has E89 Propositional Object as domain and E1 CRM Entity as range.

479

480   To represent the different categories of content (i.e. textual, thematic, and conceptual
481   correspondences), the ontology defines the following subproperties of HasRefContent. All these
482   subproperties have LociParalleli as domain:

483

484   ● HasRefContentImage, having E33 Linguistic Object as range;
485   ● HasRefContentStilema, having E33 Linguistic Object as range;
486   ● HasRefContentCharacter, having E21 Person as range;
487   ● HasRefContentEpisode, having E5 Event as range;
488   ● HasRefContentPlace, having E53 Place as range;
489   ● HasRefContentMotive, having Sentence as range;
490   ● HasRefContentTheory, having E89 Propositional Object as range.

5 The Tool for the Ontology Population

492   In order to facilitate the process of populating the ontology and building our knowledge base,
493   we implemented a semi-automatic tool. The population of the ontology will be carried out by ten
494   experts in the field of Dante studies, who participate in the HDN project. We started from 51
495   commentaries, that are stored in a non-standard format, following the encoding originally applied by
496   the DDP project. Therefore, we had to develop a custom software to pre-process the DDP data and
497   extract the knowledge contained therein.
498   To achieve this goal, we developed a Java software to convert the DDP format into a JSON
499   (JavaScript Object Notation) format, and then we used a Python software to apply regular expressions
500   in order to identify specific textual fragments and structural information. This pre-processing phase on
501   the text of Dante's commentaries allowed us to automatically extract the following information: (i)
502   the fragment of Dante's text to which the note refers (the subject of the Reference); (ii) the entire text
503   of the note; and the localization of the subject within the Divine Comedy, namely (iii) the name of the
504   cantica; (iv) the number of the canto; (v) the number of the verse(s) where the subject is expressed.
505   Once extracted, these pieces of information were stored in a JSON file.
506   After this first phase of pre-processing, we developed a tool with a Web interface in order to
507   facilitate Dante scholars in adding knowledge to our knowledge base. The tool was developed using a

508 Python backend with the Django18 framework, and a frontend built with HTML5, JavaScript, and the
509 Bootstrap19 library. It takes as input the JSON file, where the knowledge extracted from the
510 commentaries is stored, and automatically shows the relevant information in the corresponding fields
511 of the interface, as shown in Figure 5.

512

513 Insert Figure 5 here

514

515 After analyzing the notes on Dante's works reported by the commentaries, scholars use the Web
516 interface to insert the following knowledge: (i) the specific fragment of the note where the reference is
517 expressed, (ii) the kind of the reference (i.e. citation, *loci paralleli* or external support); (iii) the type
518 of citation, if requested; (iv) the content of the reference; (v) the subject-object relationship (i.e.
519 correction, extension, confirmation); (vi) the title of the primary source, (vii) the source's author,
520 (viii) the thematic area of the primary source, (ix) the text of the primary source, (x) the link to the
521 primary source; and (xi) possible notes.

522 With respect to primary sources, authors, thematic areas, and two types of reference content
523 (i.e. character and topography), scholars also add the corresponding Internationalized Resource
524 Identifiers (IRIs), and in doing so they are facilitated by auto-completion menus. For primary sources
525 and authors, and for the location, the IRIs were imported from the Wikidata[20] knowledge base. For
526 characters, on the other hand, the IRIs were imported from the Enciclopedia Dantesca[21], an
527 authoritative encyclopedia on Dante and his works. Thematic areas were imported from the Nuovo
528 Soggettario[22], a standard thesaurus maintained by the Italian National Central Library. When IRIs
529 were not available in Wikidata, Enciclopedia Dantesca or the Nuovo Soggettario, custom IRIs are
530 automatically assigned by the population tool.

531 To implement the auto-completion menus of the interface, we reused, where possible, the
532 JSON lists elaborated during the development of the DanteSources digital library. In particular, the
533 auto-completion menus were implemented in DanteSources for what it concerns: the title of primary
534 sources, their authors and thematic areas, and the corresponding IRIs. In the JSON file, a mapping
535 between primary sources, authors, and IRIs was also defined. Therefore, when a scholar selects a
536 primary source title, the interface automatically completes the related fields (i.e. author, IRIs). We
537 also create two completely new menus for characters and topographies. This functionality allowed
538 experts to reduce (i) the time for populating the ontology and (ii) the possibility to make mistakes
539 while inserting the data manually. The knowledge that is inserted by scholars through the tool
540 interface is later converted into an OWL graph according to our ontology model, by means of a
541 triplifier written in Java.

542 The tool was released as a test version in December 2020. Two scholars tested the tool for a
543 month, and then sent us comments, suggestions, and reported on the problems they found. We
544 integrated and implemented their suggestions and corrected some mistakes and bugs. In February
545 2021, we released the tool to the ten scholars who are involved in the annotation of the *Comedy*'s
546 primary sources. The tool is accessible through the HDN-Lab[23], that is the Virtual Research
547 Environment (VRE) of the project, hosted on the D4Science[24] infrastructure. When a user logs into
548 the VRE, she/he accesses the commentaries that are assigned to her/him. For each user, the tool
549 reports the date of the last change and makes available a text field in which each scholar can insert

notes, questions or doubts to share with the other scholars. Furthermore, some users can have reviewer role to access all commentaries. If a reviewer makes changes in the annotations made by other scholars, the tool records all changes to a log file. However, during the requirement collection phase, the scholars expressed the preference to discuss and collaborate also out of the tool. The VRE indeed also embeds a social networking environment where the scholars can create posts and receive answers to keep track of the discussed topics. Furthermore, the scholars can also use the VRE embedded collaborative Wiki to write scientific documentation.

6 Conclusions and Future Work

In this article we have presented the ontology we have developed within the Hypermedia Dante Network (HDN) project (2020-2023) to formally represent knowledge about the primary sources of Dante's *Divine Comedy*. The ontology has been created as an extension of the vocabulary previously developed for representing knowledge about the primary sources of Dante's minor works, that we collected in the DanteSources digital library. The HDN ontology uses CIDOC CRM and its extension FRBRoo (including its in-progress reformulation, LRMoo) as reference vocabularies. We have also presented the semi-automatic tool that we have developed, during the first year of the project, for scholars to populate the ontology. The tool has just been released to the experts who are currently annotating the primary sources identified by numerous authoritative commentaries digitized by the Dartmouth Dante Project. The development of the ontology and the population tool are the first two fundamental steps in the creation of a knowledge base about the *Divine Comedy*'s primary sources. The knowledge base will be made available through a web application, where different queries will be performed; their results will be visualized in simple and user-friendly formats, such as tables, graphs and CSV files.

Notes

[1] https://dantesources.dantenetwork.it
[2] https://hdn.dantenetwork.it
[3] https://dante.dartmouth.edu
[4] https://dante.princeton.edu
[5] https://digitaldante.columbia.edu
[6] http://www.worldofdante.org
[7] http://danteworlds.laits.utexas.edu
[8] https://www.danteonline.it/index.html
[9] https://dama.dantenetwork.it
[10] http://www.perseus.tufts.edu/hopper/
[11] http://www.bibliotecaitaliana.it
[12] https://dantesearch.dantenetwork.it
[13] http://www.vocabolariodantesco.it
[14] http://www.vocabolariodantescolatino.it
[15] http://www.dante.unina.it/public/frontend/index
[16] http://www.ifla.org
[17] http://www.digitizationguidelines.gov
[18] https://www.djangoproject.com
[19] https://getbootstrap.com
[20] https://wikidata.org
[21] https://www.treccani.it/enciclopedia/elenco-opere/Enciclopedia_Dantesca

595 [22] http://thes.bncf.firenze.sbn.it
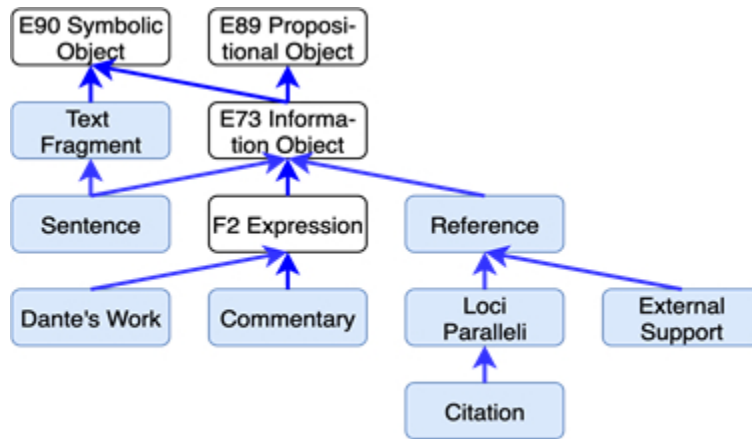596 [23] https://dante.d4science.org
597 [24] https://www.d4science.org

Captions

599 Fig. 1 The ontology classes representing the references and the main structural components of Divine Comedy
600 and Commentaries. The classes we defined are highlighted in light blue and the subclass relations in blue.
601

602 Fig. 2 A graphical view of the properties used to represent the structure of Divine Comedy. The blue arrows are
603 subclass relationships, while the black arrows link the domain to the range of the property with which they are
604 labeled.
605

606 Fig. 3 Classes and properties introduced to indicate the position of a text fragment.
607

608 Fig. 4 The properties that link the three constituent parts of the reference and the reference itself.
609

610 Fig. 5 The initial interface of the tool shows the data automatically extracted from the commentaries.

References

612 **Baierer, K., Dröge, E., Eckert, K., Goldfarb, D., Iwanowa, J., Morbidoni, C., Ritze, D.** (2017). DM2E: A
613 Linked Data Source of Digitised Manuscripts for the Digital Humanities. Semantic Web 733 – 745.

614 **Bartalesi, V. and Meghini, C.** (2017). Using an ontology for representing the knowledge on literary texts: The
615 Dante Alighieri case study. Semantic Web, 8(3), 385-394. IOS Press.

616 **Bartalesi, V., Meghini, C., Metilli, D., Tavoni, M. and Andriani, P.** (2018). A web application for exploring
617 primary sources: The DanteSources case study. Digital Scholarship in the Humanities, 33(4), 705-723.

618 **Brickley, D. and Guha, R.V.** (2014). RDF Schema 1.1. W3C Recommendation. https://www.w3.org/TR/rdf-
619 schema/ (accessed 19 February 2021).

620 **Britannica, The Editors of Encyclopaedia**. "Exegesis". Encyclopedia Britannica, 10 Mar. 2020,
621 https://www.britannica.com/topic/exegesis. Accessed 22 February 2021.

622 **Bruneau, O., Lasolle, N., Lieber J., Nauer, E., Pavlova, S., Rollet, L.** (2021). Applying and Developing
623 Semantic Web Technologies for Exploiting a Corpus in History of Science: The Case Study of the Henri
624 Poincaré Correspondence. Semantic Web, 12(2), 359-378. IOS Press.

625 **Cummings, J.** (2013). The Text Encoding Initiative and the study of literature. In: Siemens, Ray and
626 Schreibman, Susan (eds.), A Companion to Digital Literary Studies, 451-476. Blackwell Publishing.

627 **Doerr, M.** (2003). The CIDOC conceptual reference module: An ontological approach to semantic
628 interoperability of metadata. AI magazine, 24(3), 75.

629 **Doerr, M., Bekiari, C. and LeBoeuf, P.** (2008). FRBRoo, a conceptual model for performing arts. In 2008
630 Annual Conference of CIDOC, Athens (pp. 15-18). International Council of Museums.

631 **Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., and Van de Sompel, H.** (2010). The
632 europeana data model (edm). In World Library and Information Congress: 76th IFLA general conference and
633 assembly. Gothenburg, 10–15 August 2010, pp. 10–15.

634   **Hollander, R.** (2000-2007). The Divine Comedy. Inferno, Purgatorio, Paradiso. Translated by R. Hollander and
635   J. Hollander. Doubleday & Co.

636   **Koho M., Burrows T., Hyvönen, E., Ikkala, E., Page, K., Ransom, L., ... and Wijsman, H.** (2021a).
637   Harmonizing and publishing heterogeneous premodern manuscript metadata as Linked Open Data. Journal of
638   the Association for Information Science and Technology (JASIST).

639   **Koho, M., Ikkala E., Leskinen P., Tamper M., Tuominen J., Hyvönen E.** (2021b). 'WarSampo Knowledge
640   Graph: Finland in the Second World War as Linked Open Data. Semantic Web, 12(2), 359-378. IOS Press.

641   **Meghini, C., Bartalesi, V. and Metilli, D.** (2021). Representing Narratives in Digital Libraries: The Narrative
642   Ontology. Semantic Web Journal. 1 Jan. 2021 : 241 – 264.

643   **Meghini, C., Tavoni M. and Zaccarello, M.** (2020). Mapping the Knowledge of Dante Commentaries in the
644   Digital Context: A Web Ontology Approach. Forthcoming Romanic Review.

645   **Riva, P. and Žumer, M.** (2017). FRBRoo, the IFLA Library Reference Model, and now LRMoo: a circle of
646   development.

647   **Schreiber, G. and Raimond, Y.** (2014). RDF 1.1 primer. W3C Recommendation.
648   https://www.w3.org/TR/rdf11-primer/ (accessed 19 February 2021).

649   **Simon, R., Isaksen, L., Barker, E. T., and de Soto Canamares, P.** (2016). The Pleiades gazetteer and the
650   Pelagios project. Bloomington: Indiana University Press.

651   **Snydman, S., Sanderson, R. and Cramer, T.** (2015). The International Image Interoperability Framework
652   (IIIF): A community & technology approach for web-based images. In Archiving conference (Vol. 2015, No. 1,
653   pp. 16-21). Society for Imaging Science and Technology.

654   **Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... and Mons, B.**
655   (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

656   **W3C OWL Working Group** (2012). OWL 2 Web Ontology Language Document Overview (Second Edition),
657   W3C Recommendation. http://www.w3.org/TR/owl2-overview/ (accessed 19 February 2021).

658   **W3C SPARQL Working Group.** (2013). SPARQL 1.1 overview. W3C Recommendation.
659   https://www.w3.org/TR/sparql11-overview/ (accessed 19 February 2021).

The ontology classes representing the references and the main structural components of Divine Comedy and Commentaries. The classes we defined are highlighted in light blue and the subclass relations in blue

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
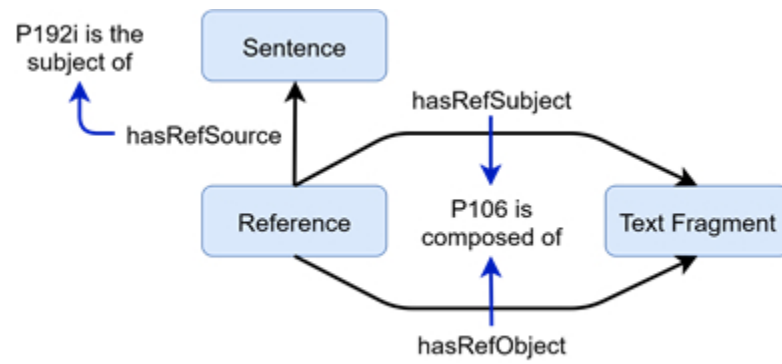50
51
52
53
54
55
56
57
58
59
60
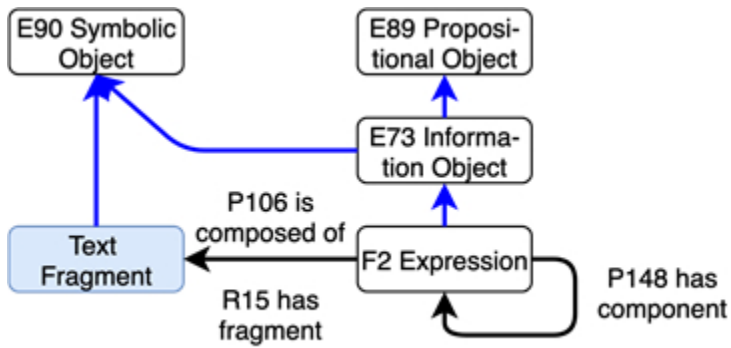
A graphical view of the properties used to represent the structure of Divine Comedy. The blue arrows are subclass relationships, while the black arrows link the domain to the range of the property with which they are labeled
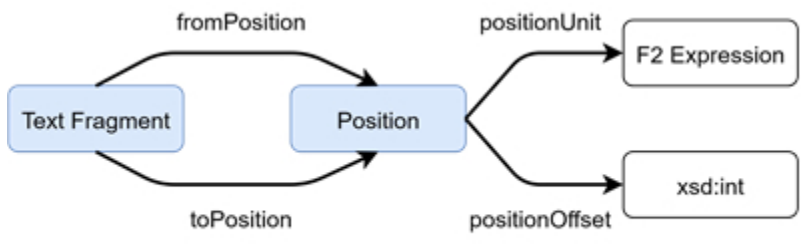
Classes and properties introduced to indicate the position of a text fragment

The properties that link the three constituent parts of the reference and the reference itself

Manuscript submitted to Digital Scholarship in the Humanities for peer review

The initial interface of the tool shows the data automatically extracted from the commentaries