CARLO MEGHINI, MIRKO TAVONI,
and MICHELANGELO ZACCARELLO

Mapping the Knowledge of Dante Commentaries in the Digital Context: A Web Ontology Approach

Abstract With digital repositories and databases available since the 1990s, Dante scholarship has always been at the forefront of the digital humanities and the digitization of medieval texts and manuscripts. However, the amount of information available about such aspects is imposing, and its location subject to the extreme dispersion of traditional scholarly publications: commentaries first but also academic journals, miscellanies, and so forth. Rather than being based on traditional word searches, a true advancement of knowledge needs to overcome the rigidity of text-based queries (and in-line markup embedded in text). Such paramount evolution is now made possible by the Semantic Web, an extension of the current web by description standards that help machines to understand and connect the information already available on the web. To achieve this, the latter is mapped using formal description and classification patterns, called ontologies. Ontologies are a key factor in managing meaningful search/data extraction, publishing relevant results on the web, search existing web resources, and offering answers to more sophisticated queries. Due to its vastness and complexity, Dante scholarship has calls for an ontology-based mapping, and specific tools have been designed to express the most difficult and articulate aspects of Dante's literary production, such as its use of biblical, classical, and medieval sources. This paper aims to introduce the aims and scope of a new digital library of Dante commentaries, built according to the aforementioned standards and aiming to refine and extend the ontologies developed for Dante's minor works to the more complex world of the *Commedia*.

Keywords Dante, digital humanities, ontology-based mapping, Semantic Web

Introduction: The Semantic Web, an Opportunity for the Digital Humanities

Since the early 2000s, humanistic researchers have become aware of the severe limitations of text-based queries, subject to the natural

Romanic Review 112:1 MAY 2021
DOI: 10.1215/00358118-8901843

ambiguity of language and returning a large number of nonrelevant hits. Filtering such "background noise" is a time-consuming task, which could be spared by more significant content-based search protocols, such as OntoQuery:

> Traditional search engines depend more or less exclusively on recognition of keywords or patterns of keywords in the text material. By contrast, *Ontoquery* addresses retrieval of pertinent text segments based on the conceptual content of the text. Queries take the form of natural language expressions and the system is primarily intended to retrieve text segments whose semantic content matches the content of noun phrases in the query phrase. This requires for the system to be able to recognise not only lexical synonyms and morphological variants, but also paraphrases—including those expressing conceptual generalisations and specialisations. This, in turn, calls for a partial syntactic and semantic analysis of the natural language queries and of the queried texts. *The semantic analysis is based on a domain-specific ontology for the target domain of the text set up prior to the text analysis*. (Andreasen et al. 2004; our italics)

In the vision of its proponents (Berners-Lee, and Lassila), the Semantic Web would be a global information network similar to the web, but different in one important aspect. While web pages are human-to-human messages that convey information using natural languages (text, images, graphics, and the like), Semantic Web pages are machine-readable messages, technically called *linked data*, that convey information using an artificial language for the formal representation of knowledge, the Resource Description Framework (abbreviated as RDF). As such, the Semantic Web is not expected to replace the web but rather to extend it by complementing the informal knowledge carried by web pages with the formal knowledge carried by linked data.

The rationale behind the pursuing of the Semantic Web is the same as that of the web: to improve the quality of life of people. But the means are different. The web tries to achieve such goal by increasing the amount of information accessible to human beings. To this end, it makes information available to any person at the lowest possible cost. In contrast, the Semantic Web tries to achieve the same goal by *increasing the amount of automation*. The expectation is that by making a significantly large quantity of formally expressed information available to artificial agents, it will be possible to multiply the number of such agents in *carrying out trivial, time-consuming, and error-prone tasks*, freeing humans from such tasks and letting them use their time for the more intellectual activities.

Since its inception, the Semantic Web vision has been pursued by the World Wide Web Committee (abbreviated as W3C), the "international community that develops open standards to ensure the long-term growth of the Web" (W3C, www.w3.org/). The language RDF, mentioned above, sits at the center

of this development, providing a syntax and a semantics for expressing knowledge on the web. The pragmatics have been instead provided by the founder of the Web himself, Tim Berners-Lee, in the form of four simple rules that should be followed in producing linked data (the memo that gives these rules can be retrieved at www.w3.org/DesignIssues/LinkedData.html).

Another fundamental ingredient for the realization of the Semantic Web are common vocabularies fixing the terms to be used in linked data and their meanings. Any such vocabulary is called an *ontology*, borrowing the term from philosophy, but using it in a more engineering sense as the formal specification of a system of categories via logical axioms. Three different kinds of ontologies are typically recognized: *Top-level* ontologies describe very general concepts like space, time, matter, object, event, and action, which are independent of a particular problem or domain. *Domain* ontologies and *task* ontologies describe, respectively, the vocabulary related to a generic domain or a generic task or activity, by specializing the terms introduced in a top-level ontology. *Application* ontologies describe concepts depending both on a particular domain and task, which are often specializations of both the related ontologies.

The growth of the Semantic Web can be appreciated by assessing the growth of linked data datasets accessible on the web: "In October 2007, datasets consisted of over two billion RDF triples, which were interlinked by over two million RDF links. By September 2011 this had grown to 31 billion RDF triples, interlinked by around 504 million RDF links."[1] What about the digital humanities (DHs)? Can they benefit from the Semantic Web family of technologies? This study argues that this is, in fact, the case, since the DHs can indeed take advantage of an increased level of automation, like any other area of science. To this end, they need to make data, and above all knowledge, available to machines, so that machines can perform the trivial, time-consuming, and error-prone tasks that are required to advance the state of the art in DHs.

A case in point is the DanteSources project, whose ultimate goal is to reconstruct the evolution of Dante's cultural background by analyzing the references to primary sources that Dante cites in his works. To achieve this goal, the project has built a linked data dataset of Dante's works and of references to primary sources of these works, extracted from commentaries (synopsis in Tavoni et al., "L'esplorazione delle fonti"). A web application allows users to explore the dataset in various ways and to visualize statistical information, such as the total amount of references to a certain work, or to an author, or a topic, in each of the parts of Dante's works. Such information is gathered from several heterogeneous sources and offered to the user via a single access point as a coherent whole aggregated in different ways. This gathering and offering is a task that a machine can perform in the best way, leaving to the scholar the interpretation of

the gathered information to the end of reconstructing the evolution of Dante's cultural background, a task clearly beyond the capabilities of any artificial agent.

State of the Art: Up-to-Date Digital Resources for Dante Scholarship

With digital repositories and databases available since the 1990s, Dante scholarship has always been at the forefront of the digital humanities, and of the digitization of medieval texts and manuscripts. However, the amount of information available about such aspects is imposing and its location subject to the extreme dispersion of traditional scholarly publications: commentaries first, but also academic journals, collected volumes, encyclopedias, and other general repertoires. The first significant digital research project on Dante was developed in the 1980s, and more than thirty years after the publication of its first prototype, the Dartmouth Dante Project (dante.dartmouth.edu/about.php, hereafter DDP) still constitutes an indispensable resource for anyone studying the *Divine Comedy*. Founded by Robert Hollander and today codirected with Simone Marchesi of Princeton University, DDP provides the full text of more than seventy-five commentaries to the *Comedy*, from Jacopo Alighieri (1322) to Nicola Fosca (2015), into a searchable database accessible online.[2]

The Società Dantesca Italiana offers on its website (www.dantesca.it) encyclopedic information on the life, chronology, and works of Dante, and a rich collection of integral reproductions of manuscripts of the *Comedy* (www.danteonline.it/italiano/codici_indice.htm). A fundamental step forward to allow users to orient themselves in the uncontrollable forest of Dante's bibliography was made through the collaboration agreement signed between the Società Dantesca Italiana, which created and since 1999 has maintained the previously known *Bibliografia Dantesca Internazionale* in Italian, and the Dante Society of America, which created and has since 1952 maintained its own *Annual Dante Bibliography*. Thanks to this partnership agreement, signed in 2017 by the then presidents of the two societies, Marcello Ciccuto and Albert Russell Ascoli, and the cooperation of their respective bibliography committees, users around the world now have free and open access to a bibliographical resource without equal in the realm of Dante studies (bibliografia.dantesca.it/media/biblio/info_eng.html): the International Dante Bibliography (Bibliografia Dantesca Internazionale), with a completely updated interface and search engine and daily bibliographic database updates with new entries added regularly.

Digital Dante (digitaldante.columbia.edu/), an editorial web project carried out by Teodolinda Barolini and her team at Columbia University, offers original research and ideas on Dante in three different contexts: (1) the *Commento Baroliniano* to the *Divine Comedy*, written expressly for Digital Dante; (2) *Intertex-*

tual Dante, a vehicle for intertextual study of the *Divine Comedy* developed by Julie Van Peteghem and featuring her original scholarship on Dante and Ovid; and (3) image, sound, history, and text, the categories through which original pieces contributed by artists, philosophers, and scholars from around the world are presented. Digital Dante does not seek to be characterized as an ordinary scholarly resource for research on Dante, but rather as a virtual place where scholarly research opens up and confronts the reactions of contemporary culture stimulated by Dante, "aiming Dante's missiles in the direction of the present day," in line with Osip Mandelstam's mandate: "It is unthinkable to read the cantos of Dante without aiming them in the direction of the present day. They were made for that. They are missiles for capturing the future" (Fenton 2005).

In the decades between the pioneering study by Brieger, Meiss, and Singleton on the *Illuminated Manuscripts of the Divine Comedy* (1969) and the recent volumes on *Dante visualizzato* edited by Arqués Corominas and Ciccuto (2017) and Ciccuto and Livraghi (2019), the field of studies in the visualization and iconography of Dante's poem in relation to the figurative culture of the fourteenth and fifteenth centuries has experienced a flourishing development. This line of studies will find a worthy representation online on the occasion of the worldwide celebrations for the seventh centenary of Dante's death (1321). The Illuminated Dante Project (www.dante.unina.it/public/frontend), promoted by the University of Naples "Federico II" and the General Direction of the State Libraries of Italy with the collaboration of the Centro Pio Rajna and the Casa di Dante in Rome (principal investigator Gennaro Ferrante), intends to provide a systematic survey and an accurate description of the early illustrations of Dante's *Divine Comedy*, accompanied by the largest archive of high-resolution images of the poem, in which both linguistic and figurative codes of the *Divine Comedy* will interact. So far, the Illuminated Dante Project has created a finding list of about 280 fourteenth- and fifteenth-century manuscripts held in libraries, museums, and archives worldwide.

DanteSearch (dantesearch.dantenetwork.it/) is, as has already been noted above, a research tool through which it is possible to query the complete corpus of Dante's vernacular and Latin works lemmatized and endowed with morphological annotation and, in the case of the *Comedy*, the *Convivio*, and the *Rime*, also with syntactic annotation. The first prototype of this resource was created in the early 2000s at the University of Pisa, under the direction of Mirko Tavoni, as part of the national research project that led to the establishment of the Biblioteca Italiana (www.bibliotecaitaliana.it/), a digital library of more than 1,600 texts representing the Italian cultural and literary tradition from the Middle Ages to the twentieth century, in integral editions based on the most authoritative reference editions, coded in XML-TEI, and freely searchable and downloadable.

The morphological and syntactic annotation system of DanteSearch was implemented, in accordance with the XML-TEI standard, by Elena Pierazzo. As for the syntax, it consists of a classification system covering all the phrases, in line with the categories of the *Grande grammatica italiana di consultazione* by Renzi, Salvi, and Cardinaletti (1988–1995) and then of the *Grammatica dell'italiano antico* by Salvi and Renzi (2010). A classification system was created by Sara Gigli, who in her PhD dissertation (Gigli) applied it to the entire text of the *Divine Comedy*. In a second step, the syntactic coding was extended to the *Convivio* and the *Rime*. Marta D'Amico, in her doctoral dissertation (2014), enriched the syntactic coding of the *Comedy* by distinguishing between diegetic and mimetic parts of the text, so as to make DanteSearch suitable for targeted research on the representation of the spoken language and on the language of dialogue, in line with the book by Paolo De Ventura entitled *Dramma e dialogo nella "Commedia" di Dante* (see also Tavoni, "Lingua parlata"). DanteSearch offers a unique opportunity to query Dante's texts with maximum flexibility, combining lexical queries as well as extremely detailed morphological and syntactic queries, as illustrated in Tavoni's 2015 paper "*DanteSearch*: Istruzioni per l'uso." The syntactic markup of Dante's Latin works according to the standards of the Universal Dependencies project (universaldependencies.org/) is being studied in collaboration with the LiLa: Linking Latin project (lila-erc.eu/#page-top) directed by Marco Passarotti at the Catholic University of the Sacred Heart of Milan, as well as the linking of Dante's Latin works lemmatized in DanteSearch with the LiLa knowledge base of linguistic resources for Latin.

State of the Art: Up-to-Date Lexical Resources for Dante and Medieval Scholarship

A lexical resource not specifically focused on Dante, but essential for any research on ancient Italian starting with Dante, is the *Tesoro della Lingua Italiana delle Origini* (TLIO), the historical vocabulary of ancient Italian created by the Opera del Vocabolario Italiano (OVI), a Consiglio Nazionale delle Ricerche (CNR) institute directed by Paolo Squillacioti (www.ovi.cnr.it/index.php/it/). Director of this institute in previous decades was Pietro Beltrami, to whom we owe the fundamental contribution in his creation of this resource (see Leonardi and Maggiore). The resource consists first of all in the *Corpus OVI dell'italiano antico*, made up today of 2,916 texts, practically all the published texts written in an Italian vernacular from the Origins until around 1374, the year of Petrarch's death conventionally assumed as the final date of the historical phase of the Italian language called "ancient Italian." All texts can be queried—but not downloaded—at gattoweb.ovi.cnr.it/(S(yiwtiusqvwjdbswbyfczzvfj))/CatFormo1.aspx. The *Corpus OVI* contains several sub-corpora within it: the cor-

pus of early lyric poetry, of *volgarizzamenti*, of early Venetian, Sicilian, Sardinian texts, and so on (www.ovi.cnr.it/index.php/it/risorse/interroga-il-corpus). Based on this corpus, the OVI editorial staff draws up the entries of the TLIO, intended for online publication and available for consultation at tlio.ovi.cnr.it/TLIO/. The TLIO, which is updated with new entries every four months, today has about 40,000 entries published online, out of an estimated total of 57,000: it has therefore reached about 70 percent of the total.

In 2015 the Accademia della Crusca and the Istituto Opera del Vocabolario Italiano launched the Vocabolario Dantesco project (www.vocabolariodantesco.it/), which intends to be an innovative and updated tool to allow a fuller understanding of Dante's lexicon in relation to the language of his time and of previous and subsequent generations, and to the Latin and Romance literary traditions. Paola Manni on behalf of the Accademia della Crusca and Lino Leonardi on behalf of OVI are responsible for the Vocabolario Dantesco, which is intended as a computer resource that is freely accessible online.

The structure of the Vocabolario Dantesco entries is modeled on that of TLIO, so as to guarantee the user the integrated use of the two tools together and in comparison to one another. The Vocabolario Dantesco is constantly updated, and the entries produced by the editorial staff and validated by the scientific commission of the project are published on the website www.vocabolariodantesco.it/lemmario.php. From the Vocabolario Dantesco project the parallel Vocabolario Dantesco Latino project was created in order to complete the scientific treatment of Dante's lexicon in both his languages, in close relationship to and with full sharing of the same standards with the Vocabolario Dantesco. The project, coordinated by Gabriella Albanese of the University of Pisa, has as its founding bodies the Accademia della Crusca and the Opera del Vocabolario Italiano as well as the Società Dantesca Italiana (www.dantesca.org/); the Fondazione Ezio Franceschini; the Istituto di ricerca sulla cultura testuale dell'Europa medievale (www.fefonlus.it/index.php/it/); the Società Internazionale per lo Studio del Medioevo Latino (SISMEL: www.sismelfirenze.it/); the Department of Philology, Literature and Linguistics of the University of Pisa (www.fileli.unipi.it/); and the Institute of Information Science and Technologies "A. Faedo" of the CNR (ISTI-CNR: www.isti.cnr.it/). ISTI-CNR will set up and maintain the website of the Vocabolario Dantesco Latino (www.vocabolariodantescolatino.it), in which the entries produced by the editorial staff will be published as soon as they are validated by the scientific committee of the project. The research group coordinated by Carlo Meghini at ISTI-CNR will oversee the project's development according to the philosophy of the Semantic Web by relating the contents of the Vocabolario Dantesco Latino, lemma by lemma, with all the resources related to medieval Latin on the Web.

At the University of Pisa, in parallel development with that of DanteSearch, two other resources have been created in recent years to focus on the texts that make up Dante's library: DaMA and DanteSources. The database DaMA, Dante Medieval Archive (dama.dantenetwork.it/), created by Gabriella Albanese and Paolo Pontari, collects the main classical, late ancient, and medieval sources, both in Latin and the vernacular, of Dante's works. As we have seen above, DanteSources (dantesources.dantenetwork.it/), created in collaboration with ISTI-CNR, displays in the form of graphs and spreadsheets the list and distribution of the texts, authors, and sets of texts cited by Dante in his works (currently available: *Vita Nova*, *Monarchia*, *Convivio*, and *De vulgari eloquentia*). DanteSources is the project in which the development of a knowledge base in the direction of the Semantic Web, which constitutes the cornerstone of the entire HDN (Hypermedia Dante Network, discussed below) project, has been pushed further so far, in the terms that are fully illustrated in the other sections of this essay.

Sources, Places, Structures: Specific Web Ontologies for Dante Studies

As it has been argued in section 1, RDF is the simply structured language recommended by the W3C for representing knowledge on the Web. RDF uses a simple format: its basic unit of representation is a triple, consisting of a subject, a predicate, and an object. A triple represents a natural language statement that expresses that a binary relation, represented by the triple's predicate, holds between two individuals, represented by the triple's subject and object. For instance, a triple may express the statement that Dante is the author of the *Convivio* by using Internationalized Resource Identifiers (IRIs), a generalized version of Universal Resource Identifiers (URIs) that may include non-ASCII characters. Our statement will use an IRI for Dante as subject, an IRI for *Convivio* as object, and an IRI for the authorship relation as predicate.

It has also been argued that ontologies play a fundamental role in the realization of the Semantic Web vision, as they offer the terms to be used as subjects, predicates, and objects in triples. Without such vocabularies, any linked data dataset would remain confined within the community that has created it (and is able to dereference the IRIs used in the dataset), defeating the vision of a common, global data space. The terms of an ontology can be conveniently divided into IRIs for representing particulars, such as individuals, things, time periods, space regions and the like, and IRIs for representing universals, that is, the general categories of discourse; these are usually divided in classes (e.g., people, object, time, space, and the like) and properties (to be a friend of, or the father of, or the author of, and so on). IRIs for particulars are provided by specialized repertories/indexes, such as author lists (e.g., the Virtual International Author-

ity File or the Getty Union List of Artist Names), thesauri (e.g., the *Getty Art and Architecture Thesaurus* or the *Library of Congress Subject Headings*), and gazetteers (e.g., *PeriodO*, a gazetteer of time periods or the GeoNames geographical database: perio.do/en/). In contrast, IRIs for universals are provided by ontologies, such as the CIDOC Conceptual Reference Model (www.cidoc-crm.org/) or the Dublin Core Metadata Initiative. (dublincore.org). Sometimes the term *ontology* is used for both kinds of vocabularies, those for particulars and those for universals, but a greater accuracy would be desirable.

Ontologies are an essential tool in communication; they can be seen as places where meanings can be agreed on between speakers of different languages and cultures. They help to achieve the goal set by Wittgenstein in the preface to his *Tractatus* (4.11), on specific domains: "Everything that can be said can be said clearly" (Wittgenstein 53). Due to their independence from any technology, ontologies are ideal places where the humanist and the IT technologist can meet and collaborate to realize DH tools and apps. Such convergence is ideal because the language used in an ontology is logic, that is, discourse in its purest form: as such, logic is the natural candidate to play the role of lingua franca for the communication between the humanist scholar and the IT expert. Logic may also be seen as the medium through which an agreement on the terms of discourse is reached; conveniently encoded, such agreement may be "read" and used by machines in the proper way. In this sense, ontologies help divide the territory where the DH endeavor takes place: the definition of the meaning of the terms and of the tasks required by the system pertains to the humanist scholar; the selection of the best suited technologies and the usage of these technologies to realize the task pertains to the IT expert. Much damage is done when these two roles are confused.

One case in point is the Text Encoding Initiative, aimed at giving guidelines for "representing the structural, renditional, and conceptual features of texts."[3] While the initiative has the merit of having analyzed in detail the many aspects of text and having provided a representation for these aspects, the choice of markup, and in particular of XML (eXtensible Markup Language), to express this representation weakens the result, making it dependent on a particular technology. Paul Eggert believed that traditional markup places annotation in the middle of the text, interfering with the text itself and, consequently, with any other way of annotating the same text. XML adds its own limitations by imposing a single structure, thus preventing the expression of any other structure.

As could be foreseen, XML is nowadays much less fashionable than it was a couple of decades ago, with JavaScript Object Notation (JSON www.json.org/json-en.html) quickly coming into fashion, only to be replaced probably by

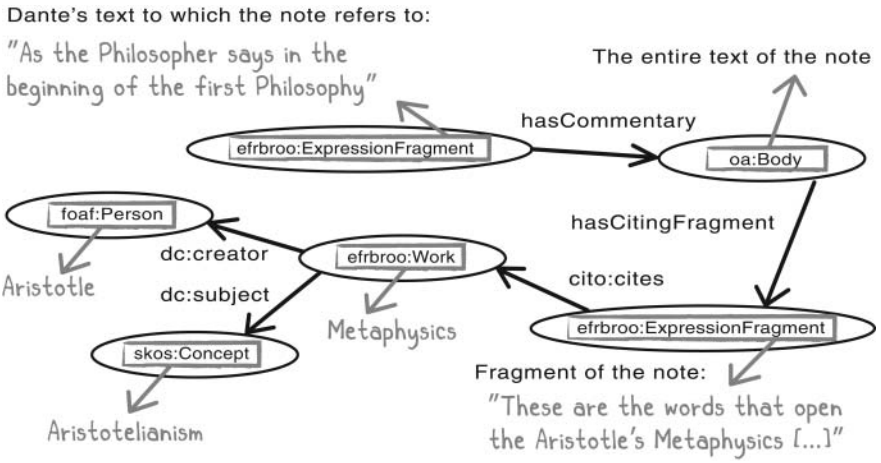

FIGURE 1. The Knowledge Base on which DanteSources relies is structured as an RDF graph that is a set of RDF triples. For the exchange of this graph, or fragments of it, with the other component of the system, an XML encoding of the RDF triples is used. Fragments of Dante's minor works or of their commentaries are encoded as RDF Literals containing just the plain text.

something else in the near future. Indeed, not having a semantics, XML cannot be used as a representation language, but only as a notation for encoding a data structure. In fact, XML has been recommended by the W3C as the official encoding for RDF and OWL, but these languages are endowed with a semantics independent of this encoding.

In DanteSources, the division of territory mentioned above has led to the collaborative development of the ontology used by the system to encode Dante's minor works and their references to primary sources. Humanist scholars have given their definitions, and IT experts, in the role of "knowledge engineers," have expressed these definitions in the language OWL 2 DL, reusing terms from other ontologies to maximize interoperability. In particular, the following main ontologies have been used (see fig. 1):

- the FRBRoo "object-oriented" model (www.cidoc-crm.org/frbroo/home-0) for the representation of the structure of Dante's works;
- the Dublin Core set, for their bibliographic metadata;
- Simple Knowledge Organization System (SKOS: see www.w3.org/2004/02/skos/), for concepts and their lexical expression;
- the Web Annotation Ontology (www.w3.org/ns/oa) and Open Annotation Core Data Model (www.openannotation.org/spec/core/core.html), standards recommended by the W3C consortium, for citations;
- the CIDOC Conceptual Reference Mode (www.cidoc-crm.org/), as a backbone ontology for the integration of all of the above.

A Digital Library for Dante Commentaries: The Hypermedia Dante Network

Since the 1980s, the study of Dante commentaries could count on a pioneering resource: the aforementioned Dartmouth Dante Project. The advent of the World Wide Web in the 1990s has boosted such a prospect to a truly global readership. Since then, the DDP has become an indispensable resource for all Dante scholarship, including more and more commentaries and going through a redesign in 2005. More recently, a very useful digital workspace has been brought alongside the DDP, allowing word-to-word collation and comparison of the text, translations, and commentaries: DanteLab is an "online application that allows students and scholars of the *Divine Comedy* to read and compare up to four texts from the site's database simultaneously" (dantelab.dartmouth.edu/about). Shortly thereafter, an articulate multimedia website for teaching Dante was developed by the University of Virginia (www.worldofdante.org/). More recently, as we noted above, Teodolinda Barolini of Columbia University and her collaborators have created a tool that addresses various issues in Dante's *Comedy*, such as interpretation and topography, by means of multimedia resources, images, and sound, with a specific focus on intertextuality (Digital Dante: digitaldante.columbia.edu/). Italy, however, both in terms of the reliable digitization of primary Dante sources and the construction of searchable databases, has not immediately followed suit. Rather, such new tools have more often been developed in the framework of more general archives of medieval Latin and/or Italian texts.[4]

Traditional databases for Dante scholarship are essential and exciting tools to study Dante's *Comedy* and the work's many commentaries. However, it is notable that most search and comparison tasks are to be intended as strictly "text-based," that is, all hits and related information are generated via the input of keyword(s) in various combinations. Given the impressive amount of knowledge and information contained in Dante's commentaries (history, interpretation, intertextuality, etc.), our project intends to map an extensive amount of data via semantic categories, in order to allow artificial intelligence to access and select more sophisticated knowledge, thus supporting various forms of scholarly endeavors. Examples of such benefits for research are (a) reduce the ambiguity of hits returned by "text-only" searches (language itself has considerable margins of ambiguity, i.e., homographs or words with multiple meanings) and (b) allow more meaningful searches targeting specific interpretive issues in the *Comedy*.

In various fields of the digital humanities, web ontologies have proven very useful to achieve these goals, once the underlying logical structure has been aptly designed: a good case in point may be the above-cited DanteSources,

where specific ontologies were designed and developed to express the complex interaction of literary and encyclopedic *auctoritates* in Dante's minor works, both Latin and Italian, and the various uses that were made of them (citation, intertextuality, interdiscursivity, etc.). The source information used was drawn from existing commentaries recently published on works in both Latin and the vernacular (e.g., De Robertis's *Rime di Dante* or Albanese's 2011 edition of Dante's *Ecloge*), making it an ambitious challenge to adapt and expand such a complex and diverse semantic network to the much larger *corpus* of commentaries, old and new, of the *Divine Comedy*. However, such an enormous task could be made easier by (a) a specific agreement with the DDP that would provide the XML-encoded full text of all commentaries currently hosted on its database; (b) a certain flexibility of the digital infrastructure (and of the semantic ontologies) developed by DanteSources, whose staff participates in this project; and/or (c) new government-funded support, with the hiring of specialized staff to design and build a larger digital infrastructure where semantic data may be more easily stored and recovered.

Though relying on a vast, high-quality digital library, it is now apparent that Dante scholarship needs to address a broader range of conceptual issues through a more articulate range of meaningful queries. Such an effort in conceptualization is attainable via appropriate web ontologies (as argued in our introduction) and the establishment of *narratives*, "in the sense of networks of events related to one another and to the Digital Library resources through semantic links" (Bartalesi Lenzi, Meghini, and Metilli, "Conceptualisation" 36). This kind of effort is the only way to overcome the rigidity and ambiguity of traditional text-based queries consisting in a list of keywords. Only in this decade has Italian Dante scholarship produced specific resources for the study of Dante's works in the context of medieval literature and cultures, specifically addressing the intertextuality of Dante's works. The project presented herewith intends to build on the resources listed in section 3 (DanteSources, Dante Search, DaMA) in a continuity of aims, actors, and methods with lasting results.

During the project's initial stages, collaboration with CNR will ensure smooth and uniform progress in the creation of the digital library; such stages will entail important issues of conceptualization, linking heterogeneous data that could not be managed by standard digital libraries, and elaboration of a data model, that is, formal specification of the abstract properties of the objects represented. An important example of this is the creation of narratives, consisting of two main components: networks of events related to one another and to the textual resources (digital library) through semantic links, and textual narrations of those events (Bartalesi Lenzi, Meghini, and Metilli, "Conceptualisation"), where this methodology is applied to Dante's biography.

In each unit, the presence of specifically recruited fellows will establish an ideal context for the critical assessment of existing sources and the development of original research on the various issues arising from Dante's poem and its rich bibliography. Close cooperation with a diverse range of established specialists will provide them with highly specialized training in a methodologically ideal research context, in the form of a permanent seminar, whose output will be promptly published on the HDN tool. Parallel to the content/concept search, many sophisticated kinds of linguistic research may be run on the digital library via up-to-date digital tools such as DanteSearch (Tavoni), using a lexical and morphological markup specifically applied to Dante's works, in Latin and the vernacular.

The construction of our semantic network will implement semiautomatic extraction of web-based knowledge from a number of resources (whose quality is previously ascertained by our researchers): however, in certain cases existing classes—such as those of *WikiData* (www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Classes)—may be useful comparisons to develop appropriate categories. Although these operations are efficiently supported by automatic tools, it is always advisable to maintain a human moderation in the process, because there is a trade-off between the level of automation and accuracy of the information, in the sense that "automatic techniques are prone to introducing errors in narratives" (Bartalesi Lenzi, Meghini, and Metilli, "Conceptualisation" 44). The adoption of SPARQL query syntax guarantees rapid and efficient searches, relying on the many suitable databases that have been created (a list of SPARQL endpoints is provided by www.w3.org/wiki/SparqlEndpoints).

As in all Semantic Web applications, a key factor in the successful design and building of the DL will be the quantity and quality of metadata associated with the resources in order to specify their semantic context in a format suitable for interpretation and for various (automated and human) queries: hence a need for human supervision throughout the process. The next section addresses this topic in further detail.

Hypermedia Dante Network (HDN): Designing and Building the Digital Infrastructure

The Hypermedia Dante Network (HDN) project aims at expanding the work carried out by its predecessor DanteSources, extending the works to include the *Divina Commedia*, and possibly also extending the knowledge gathered in the underlying digital library beyond the references to its primary sources. This section will discuss these extensions, which require three important aspects to be considered in the design and implementation of the system. First,

it must be noted that the project will use information extracted by over thirty commentaries of the *Commedia*, while the commentaries used for Dante's minor works were only the most recent (one or two for each work). Such increased scope requires a shift in methodology if we wish to complete the work in the project's lifetime. The purely manual approach followed in DanteSources must be abandoned in favor of a more semiautomatic process. In DanteSources, a team of scholars went through a commentary to detect those pieces of the text that revealed a citation of a primary source in the corresponding work of Dante. Once detected, each piece was used to fill in the fields of a record reporting the citation in detail:

- textual element of Dante's work containing the citation,
- textual element of the commentary asserting the citation,
- cited textual element (whenever possible),
- kind of citation (three categories: *concordanza stringente*, *citazione esplicita*, *concordanza generica*),
- position of the cited element within the cited work, and
- bibliographic record of the cited work.

In HDN, we will experiment with machine-learning techniques to train an automated classifier to recognize the elements of the commentaries containing a citation assertion. Thus, the scholars working on the project will not have to read through commentaries to detect such textual elements. They will instead receive from the classifier the commentary with proposed elements highlighted, and will evaluate whether they are valid citation assertions. In order to minimize the potential loss of information, the decision threshold for deciding whether a textual element contains a citation will be kept low. In addition, we will experiment with the possibility of recognizing various aspects of the citation within a candidate citation element, such as the cited work and author. We will use the manually annotated commentaries resulting from DanteSources as a training set for the various tasks. Preliminary studies have shown the difficulty of these tasks due to the different styles and techniques followed by different commentators. But more systematic experiments will follow. At present, we are in the process of reducing the commentaries to a homogeneous textual format.

The second aspect on which HDN will expand the work of DanteSources is the ontology underlying the digital library. The ontology used in DanteSources has been developed starting from 2014, and in the meantime several achievements have been accomplished that require reconsidering this ontology. First, the DanteSources ontology has been entirely mapped to the CIDOC CRM, an ISO standard and the most widely used ontology in the Cultural Heritage domain. This mapping allows DanteSources to be queried via the classes and prop-

erties of the CRM, enhancing its interoperability and placing the digital library underlying DanteSources in the Semantic Web scenario discussed above. Second, the development of the narrative ontology (Bartalesi Lenzi, Meghini, and Metilli, "Conceptualisation") has been brought to a significant stage, and the narrative ontology can now be used in HDN to connect the works of Dante and the citations they contain to the life of the poet, thus creating a more extended network of knowledge capable of serving a wider set of requests and the set of pilots that HDN plans to develop. Based on these results, HDN will revisit the DanteSources ontology to connect it to the narrative ontology and to include new categories of knowledge that the project will develop, whether manually or semiautomatically. It should be noted that the role of the HDN ontology will be (as emphasized above) that of a meeting point between scholar humanists and IT experts, where the IT experts in the HDN case also include experts in machine learning, contributing statistically inferred knowledge to the other kinds of knowledge gathered by the project.

Finally, HDN will follow a different approach than *DanteSources*, which concerns the usage of a Digital Research Infrastructure (DRI for brevity) as the technological backbone of the project. The chosen DRI is D4Science,[5] an infrastructure currently serving a community of more than eleven thousand researchers belonging to sixteen subject areas. The humanities are already present with the PARTHENOS (www.parthenos-project.eu/) and the ARIADNEplus (ariadne-infrastructure), so the HDN project should result in an effective interface.[6] The rationale behind this move is that for HDN we need to rely on a wider set of services than those we used for *DanteSources*. To mention the most important of these services: we need to be able to federate HDN with the most popular identity servers so that users working on the project can reuse their local credentials without creating new ones; we need protected, secure, and capacious storage to hold the many information resources needed by the project; we need virtual research environments where researchers can perform their machine-learning experiments in a reliable and efficient way, and share the results with the other researchers working on the project; similarly, we need virtual research environments where CNR can make available their annotation tool, enriched with the machine-learning annotations, to the project's scholars. After the service developed by the project is launched, we need to make it available through a reliable and efficient portal where different kinds of users are served and statistics collected for periodic analysis. Our final goal is to make the results of HDN interoperable with the main RI active today in Europe, that is the European Open Science Cloud (EOSC: ec.europa.eu/research/openscience/index.cfm ?pg=open-science-cloud). And we are keen upon establishing these structures free of charge and free of system administration operations. Shared DRIs, and

D4Science in particular, are designed to respond to these and other needs; therefore it has been a natural choice for the project. The building of the HDN infrastructure (hdn.dantenetwork.it) on top of D4Science has started. At present, the only offered service is the upload of commentaries. When it is completed, that infrastructure will be a unique access point for the results of the HDN project and of its predecessors cited above: *DanteSearch*, *DanteSources*, and *DaMA*.

The Editorial Issue of Dante Commentaries: Toward the HDN Digital Library

For a project based on the knowledge expressed by Dante commentaries, old and new, an important support is given by the existing database we have already mentioned, the Dartmouth Dante Project, which—since the 1980s—has made available the "entire texts of more than 75 commentaries into a searchable database that anyone can access via the World Wide Web. This database gives scholars easier access to the full texts of many important works, and, in some cases, difficult to obtain works" (dante.dartmouth.edu/about.php). Via appropriate agreements with the Italian *Edizione nazionale dei commenti danteschi* (published by Salerno Editore), such a powerful XML-encoded database may be updated and extended.[7]

Broadly speaking, real progress in the searchability of the database will be made when the texts are made semantically meaningful for the search engine, that is, semantically mapped via the conceptualization techniques noted above. The extraction of information from Dante commentaries (or relevant studies) and its encoding in RDF graphs based on appropriately designed or extended ontologies will be made easier thanks to a cooperation agreement with the aforementioned Dartmouth Dante Project (see section 4), which has, for example, XML versions of most commentaries of Dante's *Comedy*.

The CNR Unit of Pisa will manage the overall IT structure of the digital library, making available to the whole project partnership a digital research environment with tools for the extraction of knowledge from commentaries created in DanteSources with a broader range of more specific ontologies. In this context, members of the unit will assess historical and biographical sources and, according to their expertise, encode Dante commentaries published from 1900 to the present in the HDN digital library. Additionally, the team members will assess the textual accuracy of primary sources and their variants through the markup of variant readings available from existing collations.

The project's conclusion will coincide with the commemoration of the seventh centenary since Dante's death (1321–2021), an ideal context for the dissemination of its results: not to be restricted to an academic audience. Through a

series of conferences and presentations, the project will introduce scholars, students, and the general public to the various search options available via the HDN tool, with special attention paid to its potential in supporting teaching activities for schools and universities. At this stage, a paramount role will be played by members of the consortium who are already committed to a diverse range of dissemination of medieval literature and Dante scholarship: Alberto Casadei with ADI (Associazione degli Italianisti, www.italianisti.it), an association which is particularly sensitive to the requirements and needs of schools and teachers (ADI-SD, with Giancarlo Alfano on its board); Marcello Ciccuto, the president of the Società Dantesca Italiana with its programs for the 2021 celebrations often involving artists and schools (www.dantesca.it); Giuseppe Ledda with the Dante2021 committee for the Ravenna celebrations (www.dante2021.it/); Andrea Mazzucchi, in his activities for the Scuola Superiore, Biblioteca dell'Oratorio dei Girolamini in Naples; and Michelangelo Zaccarello, with the interuniversity ICoN (Italian Culture on the Net platform, www.italicon.edu) dedicated to the promotion and study of Italian language and culture abroad via the Internet.

A publication standard recommended by the European Commission, Linked Open Data, offers a number of advantages for data integration and semantic interoperability of resources, fostering innovation and simplifying research by means of integrated queries (as opposed to filtering information coming from a large number of heterogeneous sources; see Manning, Raghavan, and Schütze). The standards developed in the project may be easily adapted and exported to cover a broad range of investigations on large textual *corpora* of Italian literature, with particular reference to phenomena of literary intertextuality; in particular, ontologies elaborated for the HDN library may become classification standards for much of the Italian literature of the Middle Ages and Renaissance, especially for "mixed" Latin-vernacular genres of problematic codification such as the eclogue (Albanese). A peculiar feature of the HDN is its diverse accessibility, suitable for multiple purposes from secondary school teaching to advanced scholarship and research. Such flexibility will be attained by means of the codification of three levels of fruition: general, scholarly/specialized, and advanced/collaborative. Appropriate linking is available with the most authoritative repertories: respectively, the imposing database TLIO, an archive of Italian vernacular texts, managed by the unit CNR/Opera del Vocabolario (currently attending to a *vocabolario dantesco*, Latin and vernacular), and the Società Dantesca Italiana (within its project Bibliografia Dantesca Internazionale, in collaboration with the Dante Society of America, est. 1882). Relevant portions of text also will be associated to multimedia resources, such as maps (2D and 3D) and illustrations (illuminations, engravings), according to period

iconography (Ciccuto; Ferrante; the latter provides an outline of the *Comedy*'s early iconography).

The HDN digital library will benefit from a specialized international consortium supporting the Italian units and fostering optimal exchange and cooperation on crucial issues and questions posed by Dante's works. One such partner is the University of Notre Dame, which boasts a current interdisciplinary focus on Dante studies (Barański and Pertile; Cachey) important in the development of a range of possibilities in the area of semantic conceptualization (for a synopsis, see Hildebrand, van Ossenbruggen, and Hardman). Thanks to a recently signed framework agreement with the University of Pisa, Notre Dame will help disseminate the project's results in North America. Similar arrangements to host the DDP are being established with Simone Marchesi at Princeton University and with Robert Hollander, now retired from Princeton, as well as with various European institutions (Trinity College Dublin, Université Savoie-Mont Blanc) in order to broaden the outreach of HDN by the deadline of the seventh centenary of Dante's death.

CARLO MEGHINI, CNR-ISTI (ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE) "A. FAEDO"
MIRKO TAVONI AND MICHELANGELO ZACCARELLO,
UNIVERSITÀ DEGLI STUDI DI PISA

NOTES

The authors have closely collaborated throughout this study. However, sections 1, 4, and 6 are primarily authored by Carlo Meghini; sections 2 and 3 by Mirko Tavoni; and sections 5 and 7 by Michelangelo Zaccarello.

1. en.wikipedia.org/wiki/Linked_data (accessed January 28, 2021). A triple (consisting of a subject, an object, and a predicate expressing the relationship between the two) is the basic unit of representation of RDF.

2. As such, DDP is an extremely important partner of this project: access to the XML-encoded texts of all commentaries stored will make HDN progress significantly faster and more efficient, allowing to complete the digital library and grant its full accessibility.

3. Up-to-date guidelines for text annotation are available at tei-c.org/guidelines/.

4. Cases in point may be the Archivio della Latinità Italiana del Medioevo (ALIM, alim.unisi.it/) and the Biblioteca Italiana (BIBIT, www.bibliotecaitaliana.it/), respectively.

5. This infrastructure for science is intended "to serve the biological, ecological, environmental, social mining, culture heritage, and statistical communities world-wide" (d4science.org).

6. A list of the thematic areas served by D4Science can be obtained from services.d4science.org/thematic-gateways.

7. It is worth noting that, in general, source texts are in the public domain. Additionally, appropriate agreements will be in place with relevant copyright holders, with applicable royalties in place as well, so that users may consult the works with free and open access.

WORKS CITED

Albanese, Gabriella, ed. Dante, *Egloge*. In *Convivio, Monarchia, Epistole, Egloghe*. Vol. 2 of *Opere*, pp. 1593–783. Milan: Mondadori, 2014.

Andreasen, Troels, Per Anker Jensen, Jørgen Fischer Nilsson, Patrizia Paggio, Bolette Sandford Pedersen, and Hanne Erdman Thomsen. "Content-Based Text Querying with Ontological Descriptors." *Data and Knowledge Engineering* 48 (2004): 199–219.

Arqués Corominas, Rossend, and Marcello Ciccuto, eds. *Dante visualizzato. Carte ridenti I: XIV secolo*. Florence: Cesati, 2017.

Barański, Zygmunt G., and Lino Pertile, eds. *Dante in Context*. Cambridge: Cambridge University Press, 2015.

Bartalesi Lenzi, Valentina, Carlo Meghini, and Daniele Metilli. "A Conceptualisation of Narratives and Its Expression in the CRM (Conceptual Reference Model)." *International Journal of Metadata, Semantics and Ontologies* 12, no. 1 (2017): 35–46.

Berners-Lee, Tim, James Hendler, and Ora Lassila. "The Semantic Web." *Scientific American Magazine*, May 1, 2001. www.scientificamerican.com/article/the-semantic-web/.

Brieger, Peter, Millard Meiss, and Charles S. Singleton, eds. *Illuminated Manuscripts of the Divine Comedy*. 2 vols. Princeton, NJ: Princeton University Press, 1969.

Cachey, Theodore J. 2016. "La *Commedia* come 'mappamundi.'" *Le forme e la storia* 2 (2016): 49–74.

Ciccuto, Marcello. 2016. "Origini poetiche e figurative di una leggenda dantesca: Matelda nell'Eden." In *Dante und die bildenden Künste. Dialoge—Spieglungen—Transformationem*, edited by Maria Antonietta Terzoli and Sebastian Schuetze, 49–80. Boston: De Gruyter, 2016.

Ciccuto, Marcello, and Leyla M. G. Livraghi, eds. *Dante visualizzato. Carte ridenti II: XV secolo. Prima parte*. Florence: Cesati, 2019.

D'Amico, Marta. "La sintassi della frase complessa nella Commedia di Dante." PhD diss., University of Pisa, 2014.

De Robertis, Domenico, ed. *Rime*, by Dante Alighieri. Florence: Edizioni del Galluzzo, 2005.

De Ventura, Paolo. *Dramma e dialogo nella 'Commedia' di Dante. Il linguaggio della mimesi per un resoconto dall'aldilà*. Naples: Liguori, 2007.

Eggert, Paul. "Text-Encoding, Theories of the Text, and the Work-Site." *Literary and Linguistic Computing* 20, no. 4 (2005): 425–35.

Fenton, James. "Hell Set to Music." *Guardian*, July 16, 2005. www.theguardian.com/books/2005/jul/16/classics.dantealighieri.

Ferrante, Gennaro. "Il censimento e l'analisi delle immagini della *Commedia* di Dante (sec. XIV–XV)." *DigItalia* 1 (2018): 35–48. digitalia.sbn.it/article/view/2034/1407.

Gigli, Sara. "La codifica sintattica della Commedia di Dante." Sintassi dell'italiano antico e sintassi di Dante (Atti del seminario di studi, Pisa, 15–16 ottobre 2011), edited by Marta D'Amico, 81–96. Pisa: Felici, 2015.

Hildebrand, Michiel, Jacco van Ossenbruggen, and Lynda Hardman. *An Analysis of Search-Based User Interaction on the Semantic Web*. Technical Report. CWI (Centrum Wiskunde and Informatica), 2007. www.narcis.nl/publication/RecordID/oai:cwi.nl:12302.

Leonardi, Lino, and Marco Maggiore, eds. *Attorno a Dante, Petrarca, Boccaccio: La lingua italiana. I primi trent'anni dell'Istituto CNR Opera del Vocabolario Italiano 1985–2015. Convegno internazionale sotto l'Alto Patronato del Presidente della Repubblica, Firenze, 16–17 dicembre 2015*. Alessandria, Italy: Edizioni dell'Orso, 2016.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*, Cambridge: Cambridge University Press, 2004.

Renzi, Lorenzo, Giampaolo Salvi, and Anna Cardinaletti, eds. *Grande grammatica italiana di consultazione*. 3 vols. Bologna: Il Mulino, 1988–1995.

Salvi, Giampaolo, and Lorenzo Renzi, eds. *Grammatica dell'italiano antico*. 2 vols. Bologna: Il Mulino, 2010.

Tavoni, Mirko. "*DanteSearch*: Il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica." In *Lectura Dantis 2002–2009. Omaggio a Vincenzo Placella per i suoi settanta anni*, 583–608. Naples: Il Torcoliere, 2011.

Tavoni, Mirko. "*DanteSearch*: Istruzioni per l'uso. Interrogazione morfologica e sintattica delle opere volgari e latine di Dante." *Sintassi dell'italiano antico e sintassi di Dante*, edited by Marta D'Amico, 59–79. Pisa: Felici, 2015.

Tavoni, Mirko. "Lingua parlata e lingua scritta in Dante: Appunti metalinguistici e linguistici." *L'antinomia scritto / parlato*, edited by Franca Orletti and Federico Albano Leoni, 89–115. Castello, Italy: I libri di Emil, 2020.

Tavoni, Mirko, Paola Andriani, Carlo Meghini, Valentina Bartalesi, and Daniele Metilli. "L'esplorazione delle fonti dantesche attraverso la biblioteca digitale DanteSources." In *Sulle tracce del Dante minore. Prospettive di ricerca per lo studio delle fonti dantesche*, edited by Thomas Persico and Riccardo Viel, 29–52. Bergamo, Italy: Società Dante Alighieri / Sestante edizioni, 2017.

Wittgenstein, Ludwig. *Tractatus Logico-Philosophicus*, translated by Charles Kay Ogden, with an introduction by Bertrand Russell. Mineola, NY: Dover, 1999.