Saggio per Romanic Review https://french.columbia.edu/content/romanic-review

Meghini, Carlo*; Tavoni, Mirko; Zaccarello, Michelangelo****

*) CNR-ISTI (Istituto di Scienza e Tecnologie dell'Informazione) "A. Faedo"

**) Università degli Studi di Pisa

Mapping the Knowledge of Dante Commentaries in the Digital Context: a Web Ontology Approach[1]

Abstract

With the advent of the digital medium, access to the diverse range of Dante scholarly sources has been increasingly supported by indexing and text searches. Rather than being based on traditional word search, a true advancement of knowledge needs to overcome the rigidity of text-based queries (and in-line markup embedded in text) and connect knowledge to a broader range of classes and categories, in connection to larger portions and/or abstract features of the primary texts. 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 descriptions and classification patterns drawn from standardized and formalized vocabularies called ontologies. Applied to the specific knowledge expressed by various domains, available from subject repositories or specifically designed, ontologies are a key factor to manage a meaningful search / data extraction and publish relevant results on the web in the profitable format of Linked Data. When ontology-based searches can be run on existing web resources, the latter may become more 'understandable' by the machine, offering more accurate answers to more sophisticated queries. Due to its extraordinary vastity and complexity, Dante scholarship has soon called for an ontology-based mapping, based on description standards such as the family of RDF (Resource Description Framework) languages, and specific tools have been designed to express the most

[1] **NB.** Authors have closely collaborated throughout this paper: however, paragraphs 1, 4 and 6 are mainly authored by Carlo Meghini; paragraphs 2 and 3 by Mirko Tavoni and paragraphs 5 and 7 by Michelangelo Zaccarello. The structure of this paper is then as follows:

1. Introduction: the Semantic Web, an opportunity for the Digital Humanities (CM);
2. State of the Art: Up-to-date Digital Resources for Dante Scholarship (MT);
3. State of the Art: Up-to-date Lexical Resources for Dante and Medieval Scholarship (MT)
4. Sources, Places, Structures: Specific Web Ontologies for Dante Studies (CM);
5. A Digital Library for Dante Commentaries: the Hypermedia Dante Network (MZ);
6. Hypermedia Dante Network (HDN): Designing and Building the Digital Infrastructure (CM);
7. The Editorial Issue of Dante Commentaries: towards the HDN Digital Library (MZ).

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 based on the award-winning tool *Dante Sources*, released in 2015, the Hypermedia Dante Network (HDN), aiming to refine and extend the ontologies developed for Dante’s minor works to the more complex world of the *Comedy*. At this early stage, the research team is designing and developing a custom-made search engine (ontology reasoner), using the existing annotation of Dante’s minor works to parse the vast knowledge available in a high number of commentaries to Dante’s *Comedy*, old and new. Funded by the Italian Minister of University and Research (MUR), the HDN team can count on some of the most important scholarly institutions for Dante studies, including the Società Dantesca Italiana of Florence (whose Director M. Ciccuto is participating in the project).

1. 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 ambiguity of language and returning a large amount of non-relevant hits; filtering such “background noise” is a time-consuming task, which could be spare 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, p. 200, my italics)

In the vision of its proponents (Berners-Lee et al. 2001), 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”. 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 https://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, action, etc., 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”[2]. What about the DH? Can they benefit from the Semantic Web family of technologies? The present paper argues that this is in fact the case, since the Digital Humanities (DHs) can indeed take advantage from an increased level of automation, like any other area of science. To this end, they need to make available data and, above all, knowledge to machines, so that machines can perform the

[2] https://en.wikipedia.org/wiki/Linked_data. 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.

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 *Dante Sources* project whose ultimate goal is to reconstruct the evolution of Dante's cultural background by analyzing the references to primary sources that the Poet does 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 (a synopsis in Tavoni et al. 2017). 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 topic, in each of the parts of the works of Dante. Such information is gathered from several heterogeneous sources and offered to the scholar via a single access point as a coherent whole aggregated in different ways. This gathering and offering is a trivial, time-consuming and error-prone task that a machine can perform in the best way, leaving to the scholar the interpretation of the so obtained information to the end of reconstructing the evolution of the Dante's cultural background, a task clearly beyond the capabilities of any artificial agent.

2. 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 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, 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* (https://dante.dartmouth.edu/about.php: DDP), still constitutes an indispensable resource for anyone studying the *Divine Comedy*. Founded by Robert Hollander and today co-directed by Simone Marchesi of Princeton University, DDP provides the full text of more than 75 comments to the *Comedy*, from Jacopo Alighieri (1322) to Nicola Fosca (2015), into a searchable database accessible online.[3]

The Società Dantesca Italiana offers on its website (www.dantesca.org/) 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 scholars 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 has maintained since 1999 the previous *Bibliografia Dantesca Internazionale* in Italian, and the Dante Society of America, which created and has maintained since 1952 its *Annual Dante Bibliography*. Thanks to this

[3] As such, DDP is an extremely important partner of this project: with 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

partnership agreement, signed in 2017 by the presidents of the two societies, Marcello Ciccuto and Albert Russell Ascoli, and the cooperation of their respective bibliography committees, scholars around the world now have free and open access to a bibliographical resource without equal in the realm of Dante Studies (https://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 (https://digitaldante.columbia.edu/), an editorial web project carried out by Teodolinda Barolini 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) *Intertextual Dante*, a vehicle for intertextual study of the *Divine Comedy* developed by Julie Van Peteghem and featuring her original scholarship on Dante and Ovid; 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 want to be characterized as an ordinary scholarly resource for research on Dante, but 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".

In the decades between the pioneering study by Brieger-Meiss-Singleton on the *Illuminated Manuscripts of the Divine Comedy* (1969) and the recent volumes on *Dante visualizzato* edited by Arqués Corominas, Ciccuto and Livraghi (2017, 2019), the field of studies on 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 the death of Dante in 1321. The *Illuminated Dante Project* (http://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), aims to provide a systematic survey and an accurate description of the early illustrations of Dante's *Divine Comedy*, accompanied by the biggest high-definition image archive of the *Divine Comedy*, 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 14th and 15th centuries manuscripts held in libraries, museums and archives worldwide.

DanteSearch (https://dantesearch.dantenetwork.it/) is 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, limited to 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 1600 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-Cardinaletti (1988-1995) and then of the *Grammatica dell'italiano antico* by Salvi-Renzi (2010). A classification system created by Sara Gigli, who in her PhD thesis (see Gigli 2015) 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 thesis (2010), 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 on *Dramma e dialogo nella* Commedia *di Dante* (2007; and see also Tavoni 2020). All in all, *DanteSearch* offers a unique opportunity to query Dante's texts with maximum flexibility, combining lexical queries and extremely detailed morphological and syntactic queries, as illustrated in Tavoni 2015. The syntactic markup of Dante's Latin works according to the standards of the *Universal Dependencies* project (https://universaldependencies.org/) is being studied in collaboration with the *LiLa: Linking Latin* project (https://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.

3. 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 CNR Institute directed by Paolo Squillacioti (http://www.ovi.cnr.it/index.php/it/). Director of this Institute in previous decades was Pietro Beltrami, to whom we owe a fundamental contribution in the creation of this resource: see Leonardi-Maggiore, Eds. 2016. The resource consists first of all in the *Corpus OVI dell'italiano antico*, made up today of 2916 texts, practically all the published texts written in an Italian vernacular from the Origins until around 1374, Petrarch's date of death conventionally assumed as final date of the historical phase of the Italian language called "ancient Italian". All texts can be queried - but not downloaded - at http://gattoweb.ovi.cnr.it/(S(yiwtiusqvwjdbswbyfczzvfj))/CatForm01.aspx. The *Corpus*

OVI contains several sub-corpora within it: the corpus of early lyric poetry, of *volgarizzamenti*, of early Venetian, Sicilian, Sardinian texts, etc.: (http://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 *Tesoro della Lingua Italiana delle Origini* (TLIO), intended for online publication and available for consultation at http://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% of the total.

The Accademia della Crusca and the Istituto Opera del Vocabolario Italiano launched in 2015 the *Vocabolario Dantesco* project (http://www.vocabolariodantesco.it/), which intends to be an innovative and updated tool to allow a full 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 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 and their profitable comparison. 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 http://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 of culture, in close relationship and with full sharing of the same standards with the *Vocabolario Dantesco*, without prejudice of what is specific to the Latin language. The project, coordinated by Gabriella Albanese of the University of Pisa, has as its founding bodies, in addition to the Accademia della Crusca and the Opera del Vocabolario Italiano, in the Società Dantesca Italiana (https://www.dantesca.org/), the Fondazione Ezio Franceschini. Istituto di ricerca sulla cultura testuale dell'Europa medievale (http://www.fefonlus.it/index.php/it/), the Società Internazionale per lo Studio del Medioevo Latino (SISMEL: http://www.sismelfirenze.it/), the Department of Philology, Literature and Linguistics of the University of Pisa (https://www.fileli.unipi.it/) and the Institute of Information Science and Technologies "A. Faedo "of the CNR (ISTI-CNR: https://www.isti.cnr.it/). ISTI-CNR will take care of setting up and maintening 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. But, most importantly, the research group coordinated by Carlo Meghini at ISTI-CNR will take care of evolving the project in the direction and 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 with the development of *DanteSearch*, focused on Dante's language, two other resources, focused on the texts that made up Dante's library, have been created in recent years: *DaMA* and *DanteSources*. The database

DaMA. Dante Medieval Archive (http://dama.dantenetwork.it/), created under the responsibility of Gabriella Albanese and Paolo Pontari, collects the main classical, late ancient and medieval sources, both Latin and vernacular, of Dante's works. *DanteSources* (http://dantesources.dantenetwork.it/), created in collaboration with ISTI-CNR under the responsibility of Mirko Tavoni and Carlo Meghini (and winner of the DH Awards 2015 as *Best DH tool or suite of tools*: http://dhawards.org/dhawards2015/results/), displays in the form of graphs and spreadsheets the list and distribution of the texts, authors and sets of texts cited by Dante in some of his works: currently *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 project, has been pushed further so far, in the terms that are fully illustrated in the surrounding paragraphs of this presentation.

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

As it has been argued in paragraph 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 bersion of *Universal Resource Identifiers* (URIs) that may include non-ASCII chracters. 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 visione, 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 Authority 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: https://perio.do/en/). In contrast, IRIs for universals are provided by ontologies, such as the CIDOC Conceptual Reference Model (http://www.cidoc-crm.org/) or the *Dublin Core Metadata Initiative*

(DCMI). (https://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 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 1999, p. 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 l*ogic*, i.e. 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 endeavour 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"[4]. While the initiative has enormous merits, having analyzed in detail the many aspects of text and having provided a representation for these aspects, the choice of mark-up, and in particular of XML (*eXtensible Markup Language*), to express this representation weakens the result, making it dependent on a particular technology. Mark-up places annotation in the middle of the text, interfering with the text itself and, consequently, with any other way of annotating the same text (Eggert 2005). XML adds its own limitations by imposing a single structure, thus preventing the expression of any other structure. These limitations are very heavy for a humanist scholar, and, above all, unnecessary.

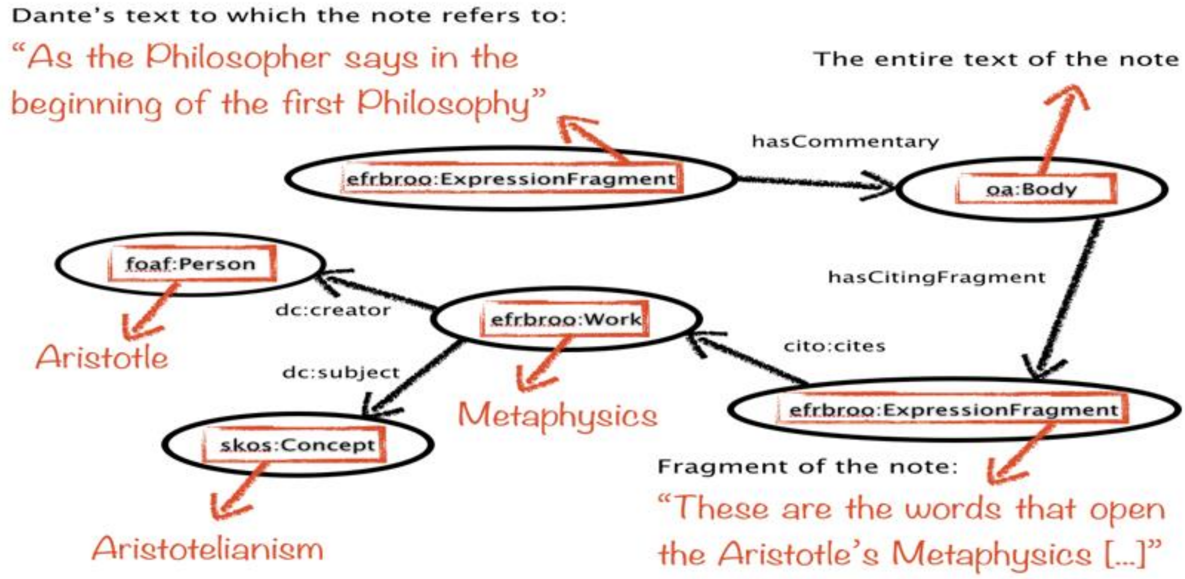
As could be foreseen, XML is nowadays much less fashionable than it was a couple of decades ago, with *JavaScript Object Notation* (JSONm https://www.json.org/json-en.html) quickly taking over, only to be replaced by something else in the near future. Indeed, not having a semantics, XML cannot be used 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. Other encodings are defined as well, for instance Turtle.

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. The humanist scholars have

[4] Up-to-date guidelines for text annotation are available at https://tei-c.org/guidelines/.

given their definitions and the IT experts, in the role of knowledge engineers, have expressed these definitions in the OWL 2 DL language, reusing terms from other ontologies to maximize interoperability. In particular, the following main ontologies have been used (see picture below):

- the FRBRoo “object-oriented” model (http://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 https://www.w3.org/2004/02/skos/), for concepts and their lexical expression;
- the Web Annotation Ontology (https://www.w3.org/ns/oa) and Open Annotation Core Data Model (http://www.openannotation.org/spec/core/core.html), standards recommended by the W3C consortium, for citations;
- the CIDOC Conceptual Reference Mode (http://www.cidoc-crm.org/) as a backbone ontology for the integration of all of the above.


Dante’s text to which the note refers to:
“As the Philosopher says in the beginning of the first Philosophy”
The entire text of the note
hasCommentary
efrbroo:ExpressionFragment
oa:Body
foaf:Person
hasCitingFragment
dc:creator
efrbroo:Work
cito:cites
Aristotle
dc:subject
Metaphysics
efrbroo:ExpressionFragment
skos:Concept
Fragment of the note:
“These are the words that open the Aristotle’s Metaphysics [...]”
Aristotelianism

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 other components 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.

5. 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 afore-mentioned *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 associated to 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” (http://dantelab.dartmouth.edu/about). Shortly after, an articulate multimedia website for teaching Dante was developed by the University of Virginia (http://www.worldofdante.org/). More recently, Teodolinda Barolini of Columbia University has created a tool that addresses various issues in Dante’s *Comedy*, such as interpretation and topography by means of multimedia resources such as images and sound, with a specific focus on intertextuality (*Digital Dante*: https://digitaldante.columbia.edu/). In Italy, however, both the reliable digitization of primary Dante sources and the construction of searchable databases has not immediately followed suit, and has more often happened in the framework of more general archives of Medieval Latin and/or Italian texts.[5]

Traditional databases for Dante scholarship are essential and exciting tools to study Dante’s *Comedy* and its many commentaries: however, it is notable that most tasks of search and comparison are to be intended as strictly “text-based”, i.e. 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 attempts to map such extensive amount via semantic categories, in order to allow artificial intelligence to access and select more sophisticated knowledge, thus supporting various forms of scholarly endeavour. 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. omographs 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 *Dante Sources*, 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 2005 for Dante’s *Rime* or Albanese 2011 for the *Ecloge*), making it an ambitious challenge to adapt and expand such complex and diverse semantic network to the much larger *corpus* of commentaries, old and new, to the *Divine Comedy*. However, such enormous task could me 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

[5] Cases in point may be the *Archivio della Latinità Italiana del Medioevo*, ALIM: http://alim.unisi.it/, and the *Biblioteca Italiana*, BIBIT: http://www.bibliotecaitaliana.it/ respectively.

infrastructure (and of the semantic ontologies) developed by *Dante Sources*, whose staff participates in this project; (c) new government-funded financial support, with the hiring of specialized staff, to design and build a larger digital infrastructure where semantic data may be 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 conceptualization effort is attainable via appropriate web ontologies (as argued in the *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-Meghini-Metilli 2017, p. 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 par. 3 (*Dante Sources*, *Dante Search*, *DaMA*) in a continuity of aims, actors, methods and with durable results.

During the project’s initial stages, collaboration with CNR in the coordinating unit will ensure a 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, i. e. 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 (digital library) resources through semantic links, and textual narrations of those events (Bartalesi-Meghini-Metilli 2017, 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 grant them 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 the *Dante Search* (Tavoni 2011), using a lexical and morphological mark-up specifically applied to Dante’s works, in Latin and the vernacular.

The construction of our semantic network will implement semi-automatic 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* (https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Classes) - may be useful terms of comparison 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 level of automation and accuracy of the information, in the sense that automatic techniques are prone to introducing errors in narratives” (Bartalesi, Meghini, Metilli 2017, p. 44). Anyhow, 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 https://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 to 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 following paragraph will tackle this topic in further detail.

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

The HDN project aims at expanding the work carried out by its predecessor *DanteSources*, extending the works considered to include also the *Divina Commedia* (DC), and possibly also extending the knowledge gathered in the underlying digital library beyond the references to its primary sources. These extensions require three important aspects to be re-considered in the design and the implementation of the system, and this Section will discuss them. Firstly, it must be noted that the project will use knowledge extracted by over 30 commentaries to the DC, while the commentaries used for Dante’s minor works were only the most recent (1-2 for each work). Such increased scope requires a shift in methodology, if we want to complete the work in the project lifetime: the purely manual approach followed in *DanteSources* must be abandoned in favour of a semi-automatic one. In *DanteSources*, a team of scholars went through a commentary to detect the fragments of the text that asserted a citation to a primary source in the corresponding work of Dante. Once detected, each fragment was used to fill in the fields of a record reporting the citation in detail:

- textual fragment of Dante’s Work containing the citation,
- textual fragment of the commentary asserting the citation,
- cited textual fragment (whenever possible),
- kind of citation (concordanza stringente, citazione esplicita, concordanza generica),
- position of the cited fragment within the cited work, and
- bibliographic record of the cited work.

In HDN, we will experiment machine learning techniques to train an automatic classifier to recognize the fragments of the commentaries containing a citation assertion. In this way, the scholars working in th project will not have to go through commentaries to detect such fragments; instead, they will receive from the classifier the commentary with proposed fragments highlighted, and will evaluate whether or not they are valid citation assertions. In order to minimize the potential loss of information, the decision threshold for deciding whether or not a fragment of text contains or a citation will be kept conveniently low. In addition, we will experiment the possibility of recognizing various aspects of the citation within a candidate citation fragment, 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 followed by different commentators; but more systematic experiments are still to begin. At present, we are in the process of reducing the commentaries to an 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 a most widely used ontology in the Cultural Heritage domain. This mapping allows DanteSources to be queried via the classes and properties of the CRM, widening its interoperability in a significant way and in fact placing the digital library underlying DanteSources in the Semantic Web scenario discussed above. Second, the development of the narrative ontology (Narratives 2017) 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 able to serve 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 semi-automatically. It must be noted that the role of the HDN ontology will be (as emphasized above) of meeting point between scholar humanists and IT experts, where the IT experts in the HDN case include also 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 concerns the usage of a Digital Research Infrastructure (DRI for brevity) as the technological backbone of the project. The chosen DRI is D4Science[6], an infrastructure currently serving a community of more than 11000 researchers belonging to 16 subject areas. The Humanities are already present with the PARTHENOS (https://www.parthenos-project.eu/) and the ARIADNEplus (https://ariadne-infrastructure), so HDN project will fit in nicely[7]. The rationale behind this move is that for HDN we need to rely on a

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

[7] A list of the thematic areas served by D4Science can be obtained from: https://services.d4science.org/thematic-gateways

wider set of services than those we used for DanteSources. To mention the most important such services, we need to be able to federate HDN with the most popular identity servers, so that researchers working in the project can re-use their local credentials without creating new ones; we need a protected, secure and capable storage where 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 in the project; similarly, we need virtual research environments where CNR can make available their annotation tool, enriched with the machine learning annotations, to the scholars of the project; after the service developed by the project is launched, we need to make it available through a reliable and efficient Portal where different typologies of users are served and statistics are collected to be periodically analyzed; and we want the results of HDN to be interoperable with the main RI active today in Europe, that is the European Open Science Cloud (EOSC: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud). Finally, we want all of the above without paying any cost or spending any time in 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 will be completed, that infrastructure will be a unique access point to the results of the HDN project and of its predecessors cited above: *DanteSearch*, *DanteSources* and *DaMA*.

7. The Editorial Issue of Dante Commentaries: towards the HDN Digital Library

For a project based on the knowledge expressed by Dante commentaries, old and new, an extraordinary support is given by the existing database 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 gives scholars easier access to the full texts of many important, and, in some cases, difficult to obtain works” (https://dante.dartmouth.edu/about.php). Via appropriate agreements with the Italian *Edizione nazionale dei commenti danteschi* (published by Salerno Editore), such powerful XML-encoded database may be textually updated and extended (in general, source texts are in public domain; however, appropriate agreements will be sought with relevant copyright holders, and a specific entry in the spending plan has been reserved for applicable royalties, so that the research outcome may be fully available in open access).

Broadly speaking, a real progress in the searchability of the database will be made when the texts are made semantically meaningful for the search engine, i.e. semantically mapped via the conceptualization techniques quoted above. The extraction of information from Dante commentaries (or relevant studies, e.g. Corrado) and its encoding in RDF graphs based on appropriately designed or extended ontologies will be made easier thanks to a cooperation agreement (section 4) with

the aforementioned *Dartmouth Dante Project*, which has XML versions of most commentaries to Dante's Comedy from the 14th century to the present day.
The CNR Unit of Pisa will manage the overall IT structure of the Digital Library, making available to the whole project partnership a virtual research environment instrumented with the tool for the extraction of knowledge from commentaries created in *DanteSources*, in continuity with Dante Sources but with a broader range of more specific ontologies; in such context, according to their expertise, members of the unit will assess historical and biographical sources and encode Dante commentaries published 1900-present in the HDN digital library; in addition, it will assess the textual accuracy of primary sources and its variants, e.g. with the markup of variant readings available from existing collations.

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 high number of heterogeneous sources: Manning et al. 2008). 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 (Alnanese 2014). A peculiar feature of the HDN will be 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; advanced / collaborating). Appropriate linking is available with the most authoritative repertories: respectively, the imposing database TLIO, an archive of Italian vernacular texts prior to 1375, managed by the unit CNR/Opera del Vocabolario (currently attending to a *Vocabolario dantesco*, Latin and vernacular), the Società Dantesca Italiana (within its project Bibliografia Dantesca Internazionale, in collaboration with the Dante Society of America, est. 1882). Relevant portions of text will be also associated to multimedia resources, such as maps (2D and 3D) and illustrations (illuminations, engravings), according to period iconography (e.g. Ciccuto 2016, Ferrante 2018, the latter providing an outline of the Poem's early iconography).

The project's conclusion will coincide with the great celebrations of 700 years since Dante's death (1321-2021), an ideal context for the dissemination of its results: not to be restricted to an academic audience, a series of conferences and presentations 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 as president of the Società Dantesca Italiana with its programme for the celebrations, often involving artists and school (www.dantesca.it); G. Ledda with the Dante2021 committee for the Ravenna celebrations (http://www.dante2021.it/); Andrea Mazzucchi through the activities of the Scuola Superiore, Biblioteca dell'Oratorio dei Girolamini, Naples; Alberto Casadei and Michelangelo Zaccarello with the interuniversity ICoN – Italian Culture on the Net platform (www.italicon.edu) aimed at promoting the study of Italian language and culture abroad via the Internet, and so forth.

Once established, the HDN digital library will benefit from a specialized international consortium supporting the Italian units and fostering optimal exchange and cooperation on crucial issues posed by Dante's works. Partners include the University of Notre Dame, boasting a leadership in Dante Studies based on its interdisciplinary vocation (e. g. Barański-Pertile 2015; Cachey 2016), very important to exploit the various possibilities of semantic conceptualization (a synopsis of the various possibilities in Hildebrand et al. 2007). Thanks to a recently-signed framework agreement with Pisa, Notre Dame will help disseminate the project's results in North America. Similar cooperations are being established with other US institutions, such as the University of Princeton, and with various European universities (Trinity College Dublin, Université Savoie-Mont Blanc) in order to broaden the reach of HDN by the paramount deadline of the 2021 7th-centennial celebrations or shortly after.

WORKS CITED

Albanese, Gabriella (ed.). 2014. Dante, *Egloge*, in Opere, vol. II (*Convivio, Monarchia, Epistole, Egloghe*, Milano: Mondadori, 1593-1783.

Arqués Corominas, Rossend - Ciccuto, Marcello, Eds., 2017. *Dante visualizzato. Carte ridenti I: XIV secolo*, Firenze: Cesati.

Andreasen, Troels, Per Anker Jensen, Jørgen Fischer Nilsson, Patrizia Paggio, Bolette Sandford Pedersen, Hanne Erdman Thomsen. 2004, "Content-based text querying with ontological descriptors", *Data & Knowledge Engineering*, XLVIII, 199-219.

Barański, Zygmunt G. and Lino Pertile (eds.). 2015. *Dante in context*, Cambridge, Cambridge University Press.

Bartalesi, Valentina, Carlo Meghini and Daniele Metilli. 2017. *A Conceptualisation of Narratives and Its Expression in the CRM [Conceptual Reference Model]*, «International journal of metadata, semantics and ontologies» (Online) 12, 35-46 (DOI:10.1504/IJMSO.2017.087692).

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

Bartalesi Lenzi Valentina, Carlo Meghini and Daniele Metilli. 2017. "A conceptualisation of narratives and its expression in the CRM", in *International Journal of Metadata, Semantics and Ontologies*, vol. 12 (1) pp. 35 - 46.

Bartalesi, Valentina, Carlo Meghini, Daniele Metilli. Mirko Tavoni and Paola Andriani. 2018. "A web application for exploring primary sources: the DanteSources case study.", in *Digital Scholarship in the Humanities*, 33/4, pp. 705-723.

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

Cachey, Theodore J. 2016. «La "Commedia" come 'mappamundi'», *Le forme e la storia*, II, 49-74.

Ciccuto, Marcello. 2016. *Una leggenda dantesca: Matelda nell'Eden*, "Dante und die bildenden Künste", DOI: 10.1515/9783110486117-005

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

De Robertis, Domenico (ed.). 2005. Dante Alighieri, *Rime*, Firenze: Edizioni del Galluzzo.

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

Eggert, Paul. 2005. "Text-encoding, "Theories of the Text, and the *Work-Site*", *Literary and Linguistic Computing*, 20/4, 425-435.

Ferrante, Gennaro.2018. "Il censimento e l'analisi delle immagini della Commedia di Dante (sec. XIV-XV)", *Digitalia*, I, pp. 35-48, link http://digitalia.sbn.it/article/view/2034/1407

Gigli, Sara. 2015. *La codifica sintattica della* Commedia *di Dante*, in Marta D'Amico (ed.), Sintassi dell'italiano antico e sintassi di Dante (Atti del seminario di studi, Pisa 15-16 ottobre 2011), Pisa, Felici, pp. 81-96.

Hildebrand, Michiel, Jacco van Ossenbruggen and Lynda Hardman, *An analysis of search-based user interaction on the semantic web*. 2007.Technical Report. CWI (Centrum Wiskunde & Informatica), link https://www.narcis.nl/publication/RecordID/oai:cwi.nl:12302

Leonardi, Lino and Marco, Maggiore (eds.). 2016. *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: Edizioni dell'Orso.

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

Narratives in Digital Libraries. 2017. Meghini, Carlo et al*. Conceptual model for Dante's biography*, link https://dlnarratives.eu/project.html

Renzi, Lorenzo - Salvi, Giampaolo - Cardinaletti, Anna, Eds. 1988-1995. *Grande grammatica italiana di consultazione*, 3 vols., Bologna: il Mulino.

Salvi, Giampaolo - Renzi, Lorenzo, Eds. 2010. *Grammatica dell'italiano antico*, 2 vols., Bologna: il Mulino.

Tavoni, Mirko. 2011. *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*, Napoli: Il Torcoliere, pp. 583-608.

Tavoni, Mirko. 2015. "*DanteSearch*: istruzioni per l'uso. Interrogazione morfologica e sintattica delle opere volgari e latine di Dante", in Marta D'Amico (ed.), *Sintassi dell'italiano antico e sintassi di Dante*, Pisa: Felici, pp. 59-79.

Tavoni, Mirko. 2020. “Lingua parlata e lingua scritta in Dante: appunti metalinguistici e linguistici”, in *L’antinomia scrito / parlato*, a cura di Franca Orletti e Federico Albano Leoni, Città di Castello: I libri di Emil, pp. 89-115.

Tavoni, Mirko, Paola Andriani, Carlo Meghini, Valentina Bartalesi, Daniele Metilli. 2017. *L’esplorazione delle fonti dantesche attraverso la biblioteca digitale DanteSources*, in Thomas Persico and Riccardo Viel (eds.), *Sulle tracce del Dante minore. Prospettive di ricerca per lo studio delle fonti dantesche*, Bergamo, Società Dante Alighieri / Sestante edizioni, pp. 29-52.

Wittgenstein, Ludwig. 1999. *Tractatus Logico-Philosophicus*, translated by C. K. Ogden, with an Introduction by B. Russell, Mineola (NY), Dover.

.