The Linking Dante project

Carlo Meghini, Cesare Concordia and Gaia Tomazzoli
ISTI CNR, La Sapienza

MITE final workshop: From critical analysis to formal representation: literary characters, interpretations, and ontologies

Outline

- Objective
- Approach
- Ontologies
- Results

Objective

To create a digital library (DL) providing access to the intertextual relationships in Dante's Commedia, selected by humanist scholars in a large number of commentaries, and fed into the DL via an apposite tool.

The tool: the PRIN Hypermedia Dante Network (2017).

The system extends DanteSources, a system built by a previous PRIN project, providing access to the intertextual relationships in Dante's works other than the Commedia.

COVID-due revised objectives

To include in the target DL also knowledge on the morphological and syntactical structure of the Commedia, separately offered by DanteSearch.

And to offer access to all these kinds of knowledge *simultaneously* through a unique point.

This bigger project is called LiDa for **Li**nking **Da**nte.

The DL developed by LiDa is based on a LOD datasets and aims at being part of the Semantic Web of literary resources, a concept that still struggles for acceptance in the Digital Humanities, in spite of the vast consensus on LOD.

Approach

On the other hand, there seems to be a not-so-vast consensus on what lies in between the user and the LOD. This aspect is very important. It is not a technological problem: it involves the understanding of computers and of computer-based information systems.

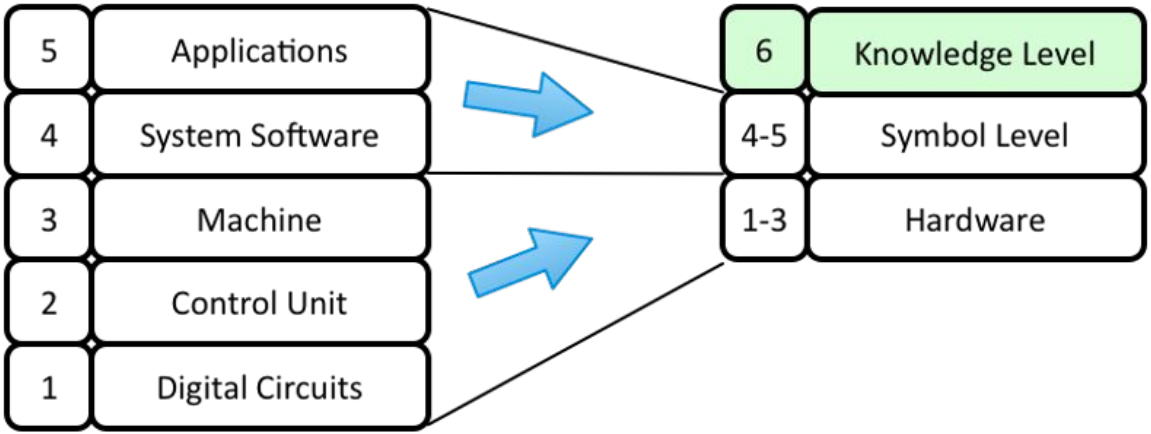
I think the correct understanding is the one proposed by Allen Newell in a seminal talk, given August 19, 1980, at Stanford University, as the first presidential address of the American Association for Artificial Intelligence.

Purpose: to give guidance to the AI community in its path on to the design and the implementation of the synthetic mind, addressing questions like:

- What is the nature of knowledge?
- How is it related to representation?
- What is it that a system has, when it has knowledge?

The Knowledge Level Hypothesis

The Knowledge Level Hypothesis. There exists a distinct computer system level, lying immediately above the symbol level, which is characterized by knowledge as the medium and the principle of rationality as the law of behavior.



In Newell’s words

“The system at the knowledge level is the agent. An agent is composed of a set of actions, a set of goals and a body. The agent processes its knowledge to determine the actions to take. The behavior law is the principle of rationality: Actions are selected to attain the agent's goals”.

To treat a system at the knowledge level is to treat it as having knowledge and goals, and believing it will do whatever is within its power to attain its goals, in so far as its knowledge indicates.

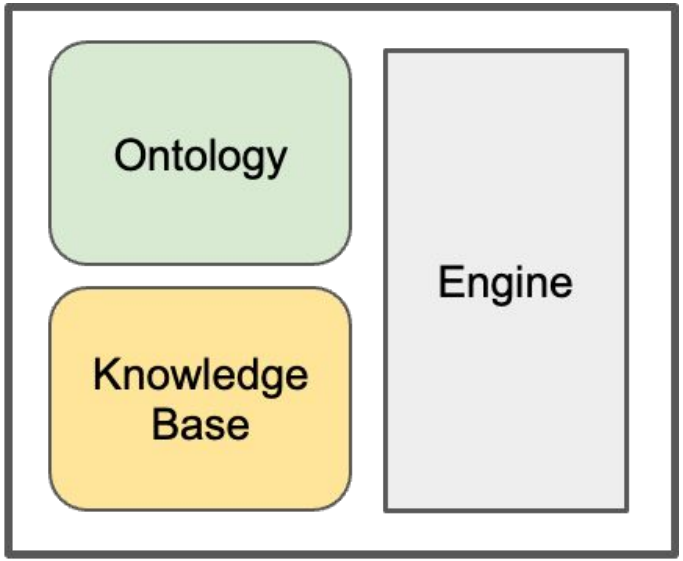
Why it is useful? The knowledge level permits predicting and understanding behavior **without having an operational model of the processing that is actually being done by the agent**.

The symbol level: the declarative view

At the *heart* of the symbol level there are two information structures:

- an ontology
- a knowledge base

These structures are used by a standardized piece of code, the Engine, to carry out the main tasks of the system.



The symbol level

Ontologies use the language of science, *i.e.* logic (*and, or, not, every, some, implies,* etc.) and mathematics (*at least, no more than, transitive, symmetric, reflexive*, etc.), to provide rigorous, unambiguous definitions of the terms of a universe of discourse.

They are an essential element in communication, as they help to say clearly what can be said *without involving the symbol level*.

A knowledge base is a set of assertions giving facts related to a *particular state of affairs* or *a particular epistemic state*.

These terms are sometimes used in different ways, for instance by OWL.

Back to Dante

LiDa is mainly an integration project.

Ontologies play a fundamental role in information integration as they provide the common interlingua that allows seeing the systems to be integrated as specialized agents able to cooperate by sharing that common interlingua.

The project has followed a two-stage approach:

- developing a common conceptualisation about the facts represented in the systems to be integrated
- expressing the “gluing” facts and the models of the systems to be integrated in a single application ontology, as standard as possible

Back to Dante

To develop the common conceptualisation of the systems to be integrated, we have worked bottom-up by de-abstraction: re-introducing the facts that have been suppressed upon implementing those systems but are relevant to connect the knowledge embodied by them.

To develop the glue, we have worked top-down. We started from the observation that a text is a structured place where several kinds of phenomena happen: linguistic (morphological, syntactical), inter-textual (citations, external supports, *loci paralleli*), rhetorical (metaphores) and many other.

The *where* is the fundamental notion that allows to connect all these phenomena, precisely what mark-up languages suppress.

Back to Dante

The other observation is that phenomena can be classified at three levels:

- the linguistic level
- the occurrence level
- the conceptual level

Structural entities and dialogues are mainly occurrence entities, while the other entities explicate at all levels.

Aspect	Entities
Structure	cantica, canto, verso
Morphology	form, lemma, morphologic category
Syntax	dialogue, period, sentence, syntagm, complement
Rhetorics	metaphor

Back to Dante

The three levels are linked by the *instantiation* and the *occurrence* relationships: each linguistic entity is an instance of the corresponding conceptual entity and manifests itself in an occurrence.

- For example, the fourth element of the fragment "Nel mezzo del cammin di nostra vita," is a form occurrence and an instance of the form (whose written representation is) "cammino", which in turn refers to the concept of "cammino."

Entities are linked by a relation of composition (*e.g.*, each sentence is composed by a sequence of form occurrences) and localized in text fragments (each entity occurs in one or more text fragments) determined by their start and end points.

Ontologies

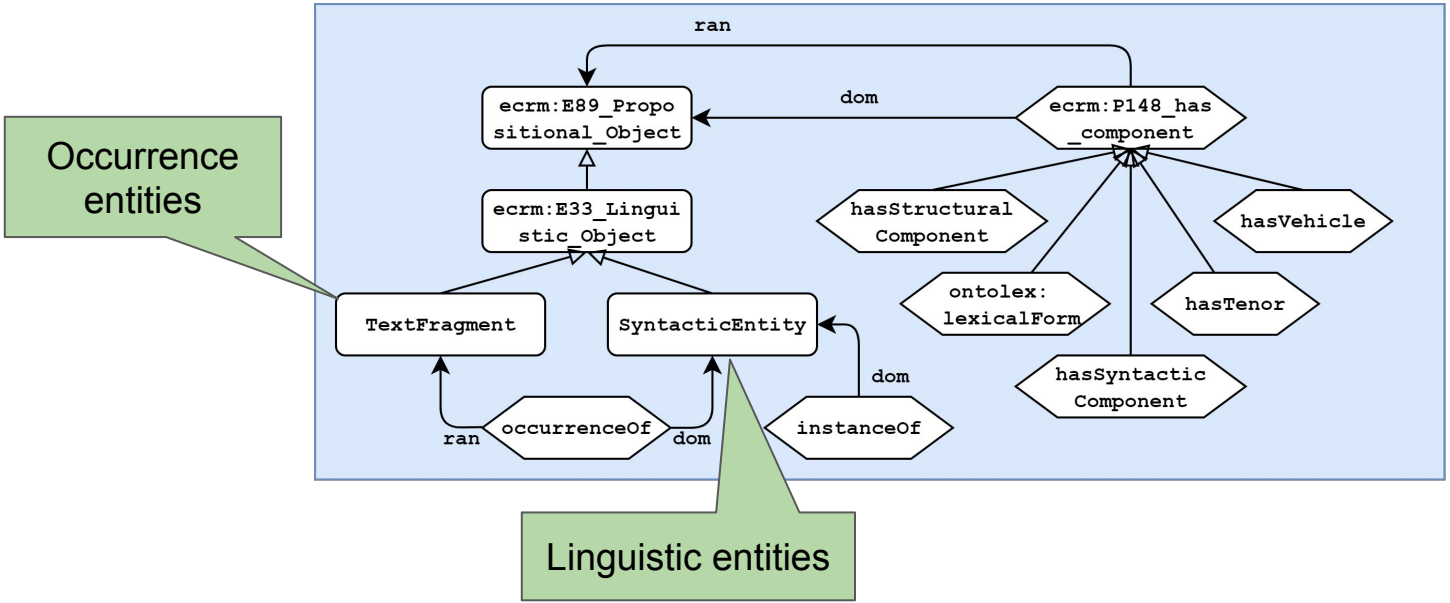
LiDa has developed OLiRes (**O**ntology of **Li**terary **Res**ources), an application ontology expressed in OWL 2 DL and founded on the CIDOC CRM top ontology and on some domain ontologies like Ontolex.

OliRES is a composite, it includes:

- **HDN** (Hypermedia Dante Network): focussed on inter-textual relations
- **MOnt**: Metaphor Ontology
- **ORL:** Ontology of Lexical Resources, for the representation of morphological knowledge
- **SyntIt**: Ontology of ancient Italian

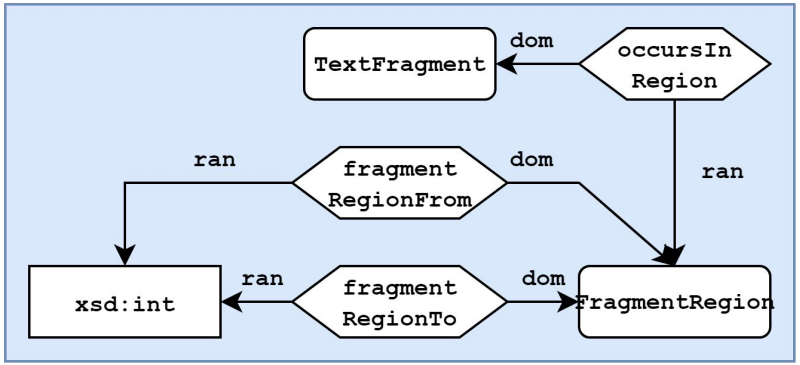
OLiRes

OLiRes provides classes and properties for the three levels (actually, two) and for their composition. All its classes and properties are related to the CIDOC CRM, which is the chosen top ontology.

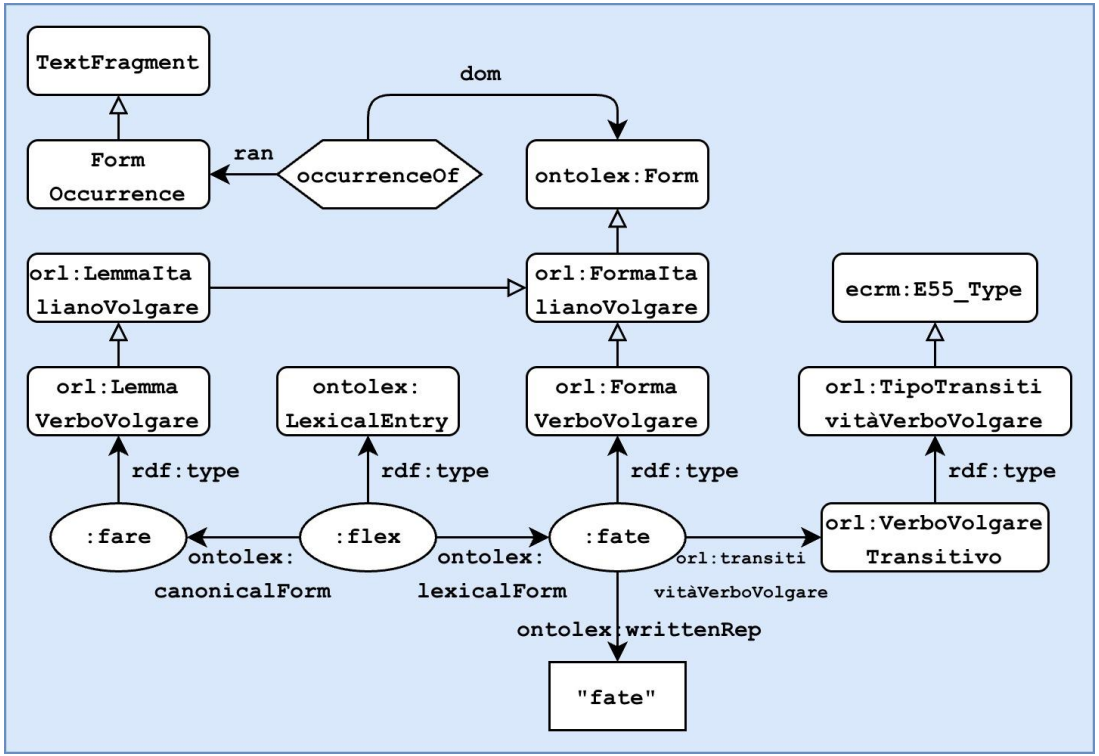


OLiRes

OLiRes also provides the machinery for representing the “places inside the text where phenomena occurs. These places are called text fragments.



ORL


TextFragment
dom
Form Occurrence
ran
occurrenceOf
ontolex:Form
orl:LemmaItalianoVolgare
orl:FormaItalianoVolgare
ecrm:E55_Type
orl:Lemma VerboVolgare
ontolex: LexicalEntry
orl:Forma VerboVolgare
orl:TipoTransitivitàVerboVolgare
rdf:type
rdf:type
rdf:type
rdf:type
:fare
:flex
:fate
orl:VerboVolgare Transitivo
ontolex: canonicalForm
ontolex: lexicalForm
orl:transiti vitàVerboVolgare
ontolex:writtenRep
"fate"

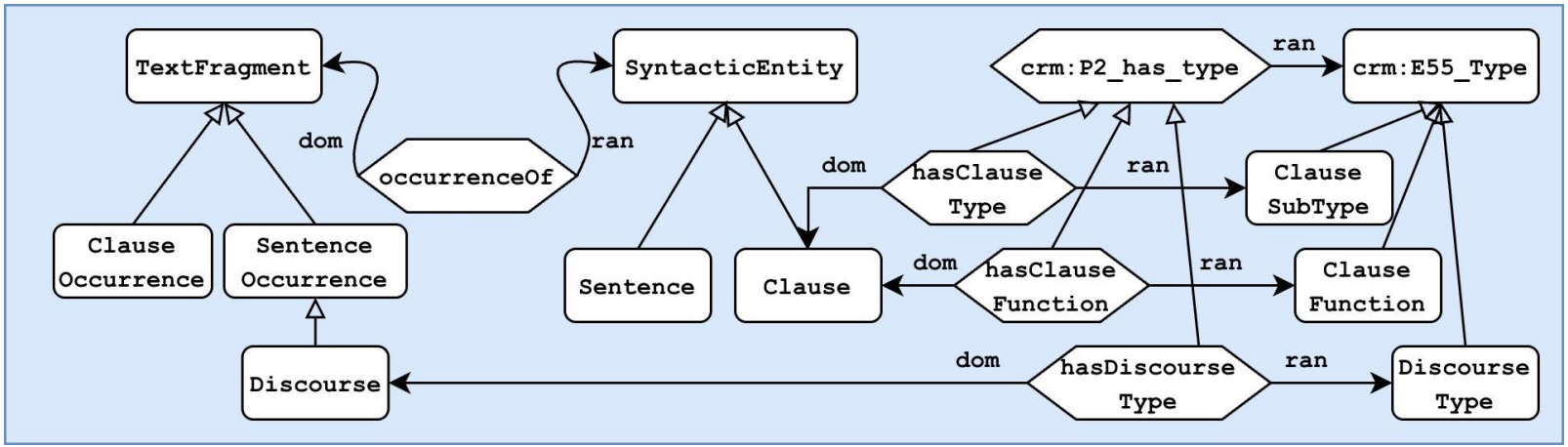
ORL

Also ORL relies on the categorizations of DanteSearch. For instance. the morphological macro-categories of forms are

- Verbo Volgare
- Sostantivo Volgare
- Aggettivo Volgare
- Pronome Volgare
- Articolo Volgare
- Avverbio Volgare
- Preposizione Volgare
- Congiunzione Volgare
- Interiezione Volgare
- Onomastica Volgare
- Citazione Volgare

SyntIt

The basic classes of SyntIt are Syntagm, Clause and Sentence.



SyntIt

Clause types and subtypes, clause functions and discourse types are derived from the categorizations underlying DanteSearch. For instance,

- Clause functions are main, coordinate, subordinate, parenthetical coordinate and pseudo-parenthetical coordinate.
- There are 30 clause macro-types and 200 types.

They are detailed in the documentation of the DanteSearch system, which can be downloaded from the original web site.

The composition of syntagms has been modelled via special properties, as a syntagm can include one or more clauses, breaking the hierarchical structure of these classes. The properties used for this purpose are

- hasHead, linking a syntagm to its head
- hasComponentSyntagm, linking a syntagm to its composing syntagms
- includesClause, linking a syntagm to its composing clauses
- hasModifier, linking a syntagm to its composing modofoers

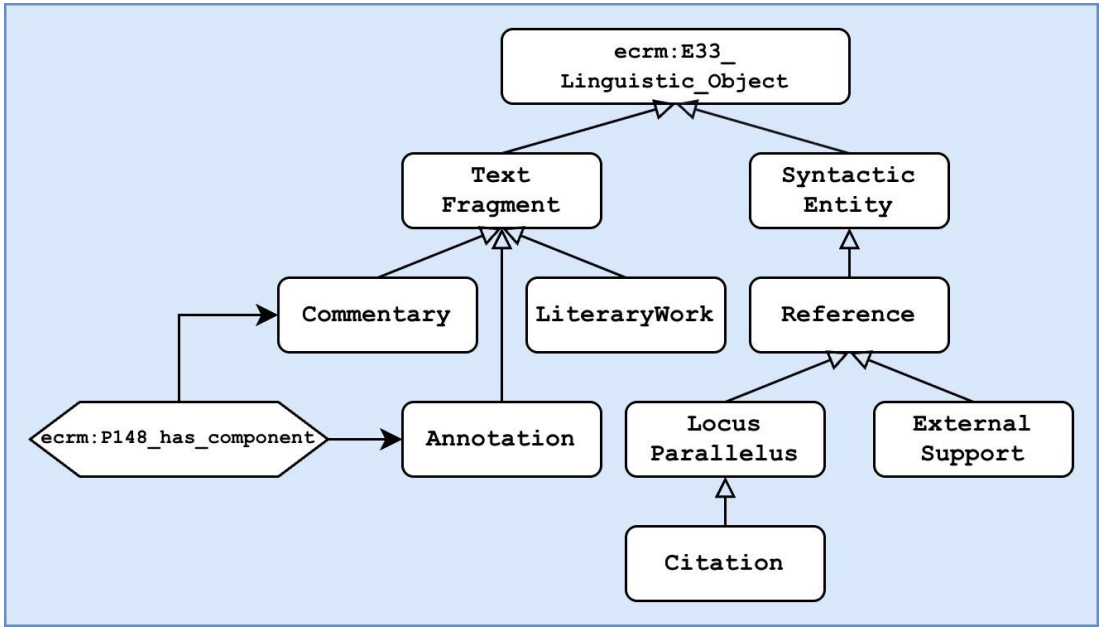
HDN: Ontology of Inter-textual relations

HDN provides classes and properties for sentences of the form:

According to A the knowledge on B can be enriched by C

- A is the **source** of the reference;
- B is the **subject** of the reference, that is a fragment of a work of Dante that is enriched by the reference;
- C is the **object** of the reference, that is the entity referred to by the reference.

HDN: Ontology of Inter-textual relations



Kinds of references

- **Quotations:** There is a relationship of direct dependence between Dante and a text identified as his source; we distinguish between explicit quotations, close quotations, and generic concordances**.**
- ***Loci Paralleli*****:** This includes all cases in which the commentator mentions a text in which a concept, stylistic device, or character displays similarities to Dante's text, without implying a derivation.
- **External supports:** All cases in which the commentator cites a text simply to support his own interpretation of Dante's work.

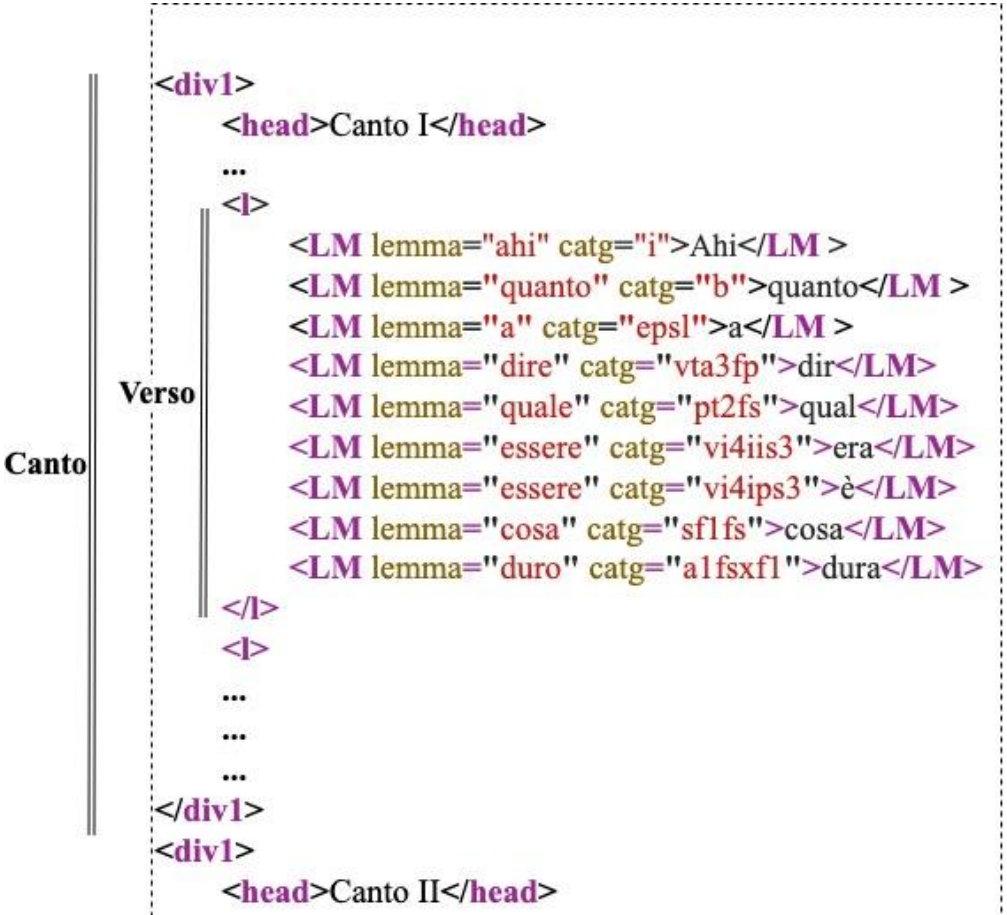
The Knowledge Graph

The LiDa Knowledge Graph has been generated by parsing several XML files:

- those at the basis of DanteSearch
- those collecting the inputs of the scholars on references

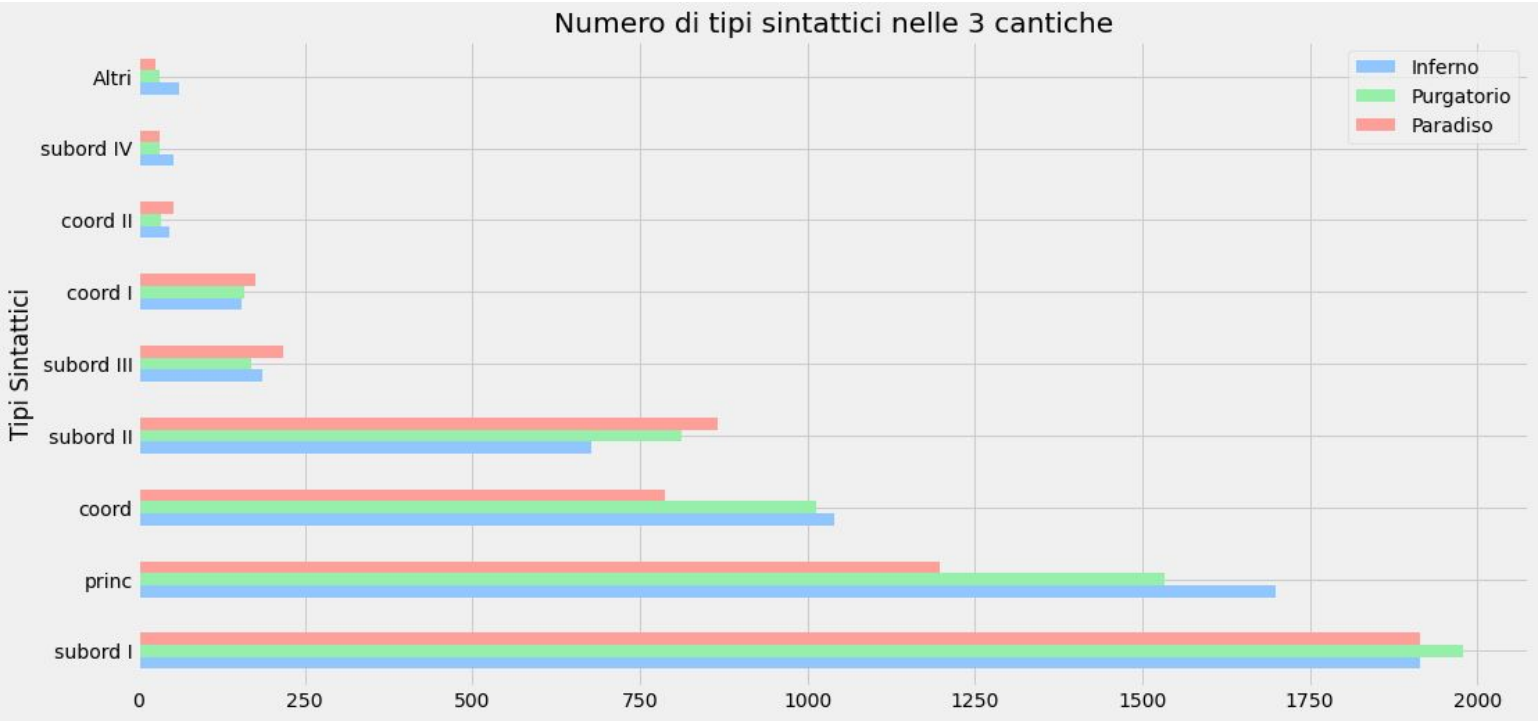
and collecting the LOD at the basis of DanteSources.

DanteSearch morphological annotation



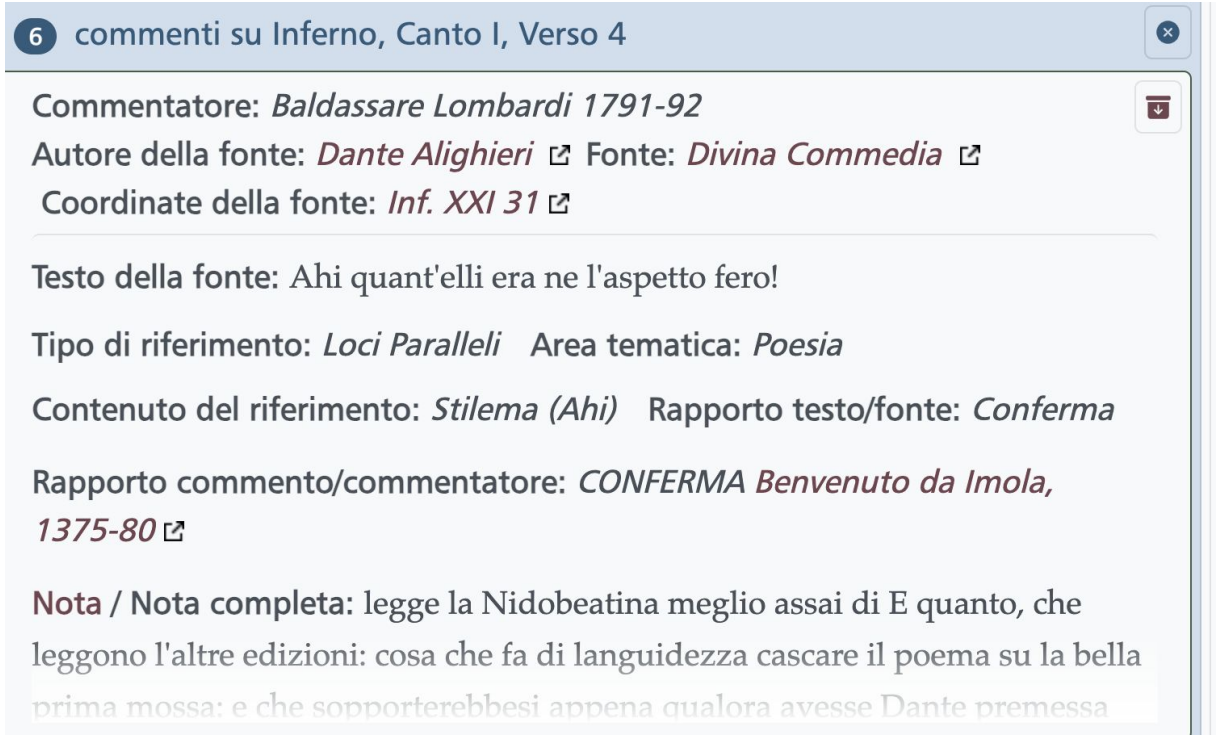
DanteSearch syntactical annotation

Distribution of clause types in the Comedy



References

The KG includes 14687 references from 18 commentaries, most of them on *Inferno*.



Accessing the LiDa KG

The LiDa KB is accessible to humans via a GUI, allowing navigating the text by selecting different views:

- Forms: hovering the cursor displays morphological and lexical information.
- Sentences: Colors delimit and characterize the sentences and their component phrases; hovering the cursor displays the syntactic function of each sentence.
- Dialogues: Dialogues are highlighted; hovering the cursor displays the speaker and type of dialogue (direct, reported, thought).

In addition to browsing, there are two search modes:

- Simple: allows searching form occurrences across the entire work.
- Advanced: allows defining complex search filters, combined by Boolean operators, spanning the structural, morphological, syntactical and metaphorical dimensions.

Accessing the KG

More importantly, the LiDa KB is accessible to digital agents via a SPARQL endpoint GUI, allowing interacting with the KG through the standard SPARQL API which supports queries and updates.

The GUI uses this endpoint to access the graph.

Any other agent can use the endpoint to extract any subset of the graph and use it for whatever purpose, for instance for creating a more specialized GUI.

Resources

- Ontologies and datasets:
  - https://github.com/LinkingDante/data

- Web app for querying and browsing the *Commedia*:
  - https://lida.dantenetwork.it/

- SPARQL endpoint:
  - https://lida.dantenetwork.it/sparql

People

- Valentina Bartalesi, Nicolò Pratelli and Daniele Metilli: authors of DanteSources, of the tool for inserting reference knowledge, co-authors of the HDN ontology
- Cesare Concordia e Luca Trupiano: architects, developers and maintainers of the computing infrastructure
- Gaia Tomazzoli: co-author of the conceptualisation underlying the HDN and the SyntIt ontologies, reviewer/adviser of the GUI to the KG, author of the conceptualisation of the metaphors in the *Commedia* and provider of the set of such metaphors
- Chiara Paolini: reviewer of the conceptualisation underlying ORL

People

- Nicola Aloia: co-designer e co-developer of the GUI
- Oriana Mirabella, Martina Barbieri, Chiara Ciminello, Chiara Bertini who completed the knowledge on references.
- The HDN team, for inspiring and reviewing the conceptualisation underlying the HDN ontology.

Michelangelo Zaccarello who gave us the project and supported its development all along.

Thank you!

Il set di dati di DanteSearch

- Un insieme di **file XML** annotati secondo la codifica TEI (con una codifica *ad hoc*)
- Il dataset contiene la **lemmatizzazione** e l'**annotazione morfologica e sintattica** delle opere volgare e latine di Dante
- Strutture XML-TEI non omogenee
- Il dataset della Commedia: 3 file con le annotazioni morfologiche, uno per ciascuna cantica, 1 file con le annotazioni sintattiche (analisi del periodo) dell'intera Commedia

La realizzazione del KG della *Commedia*

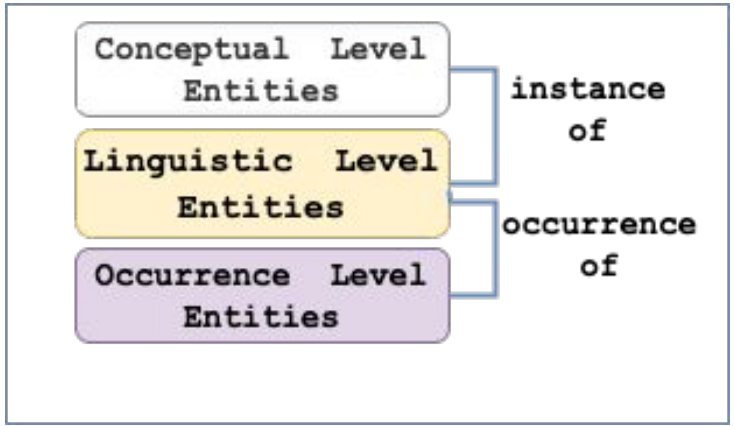
A. Vengono processati i file con le annotazioni morfologiche per creare il vocabolario

B. Si calcolano le occorrenze nel testo di ciascun simbolo del vocabolario e si assegnano a queste le caratteristiche morfologiche

C. Viene processato il file con le annotazioni sintattiche: vengono create le entità sintattiche

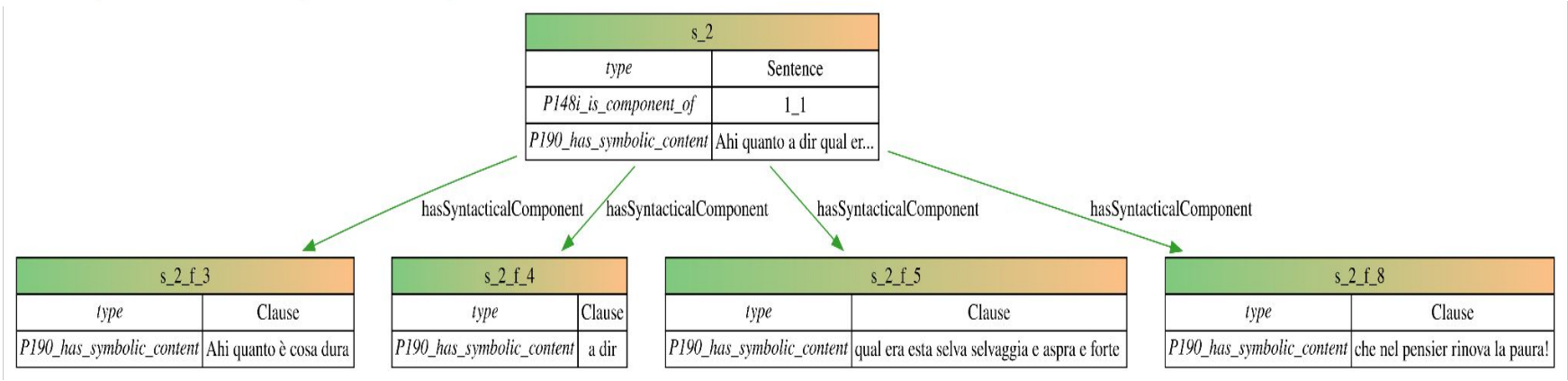
D. Si calcolano le occorrenze delle entità sintattiche assegnando loro la funzione e il tipo

E. Le occorrenze delle entità sintattiche sono associate alle occorrenze dei simboli di vocabolario

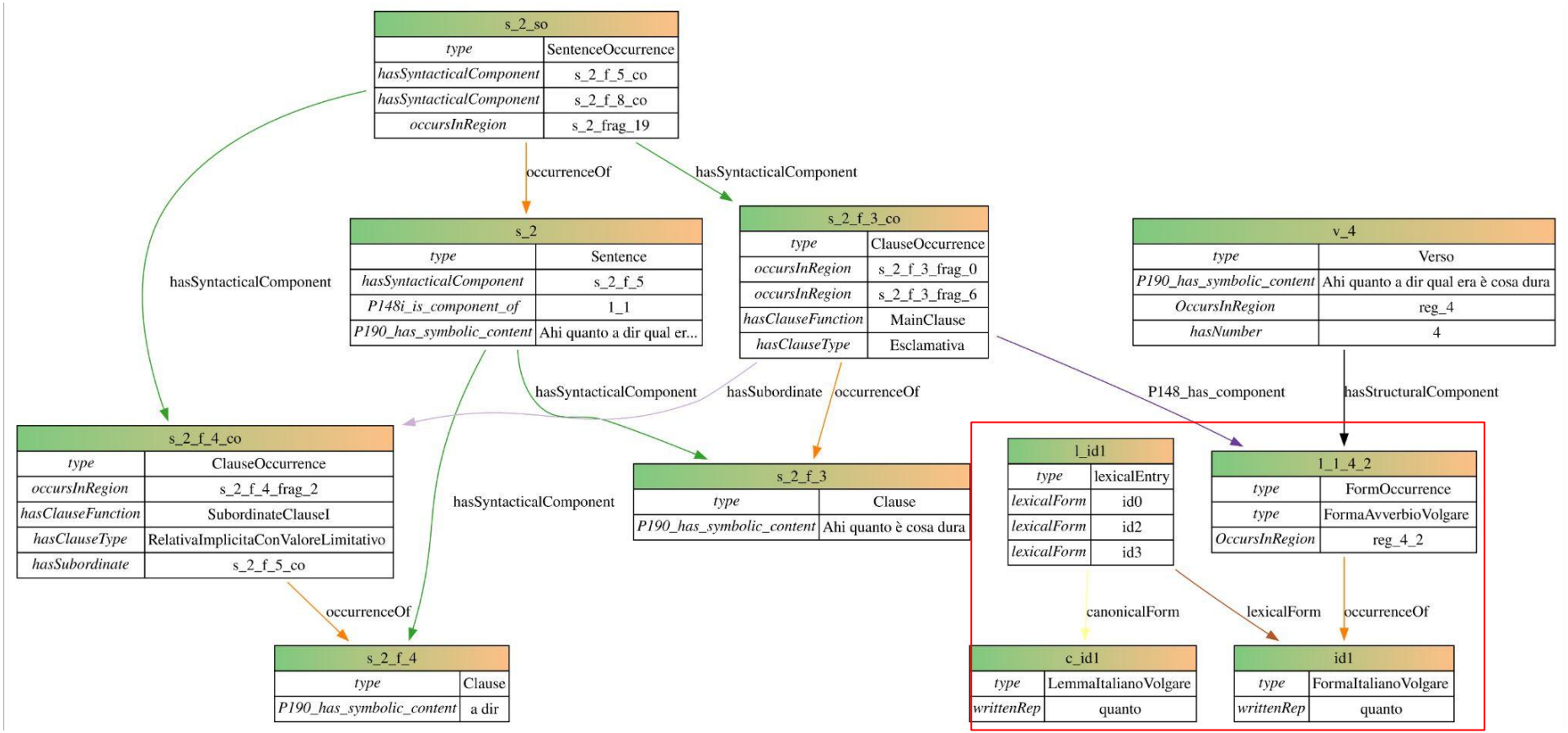


Esempio: entità sintattiche

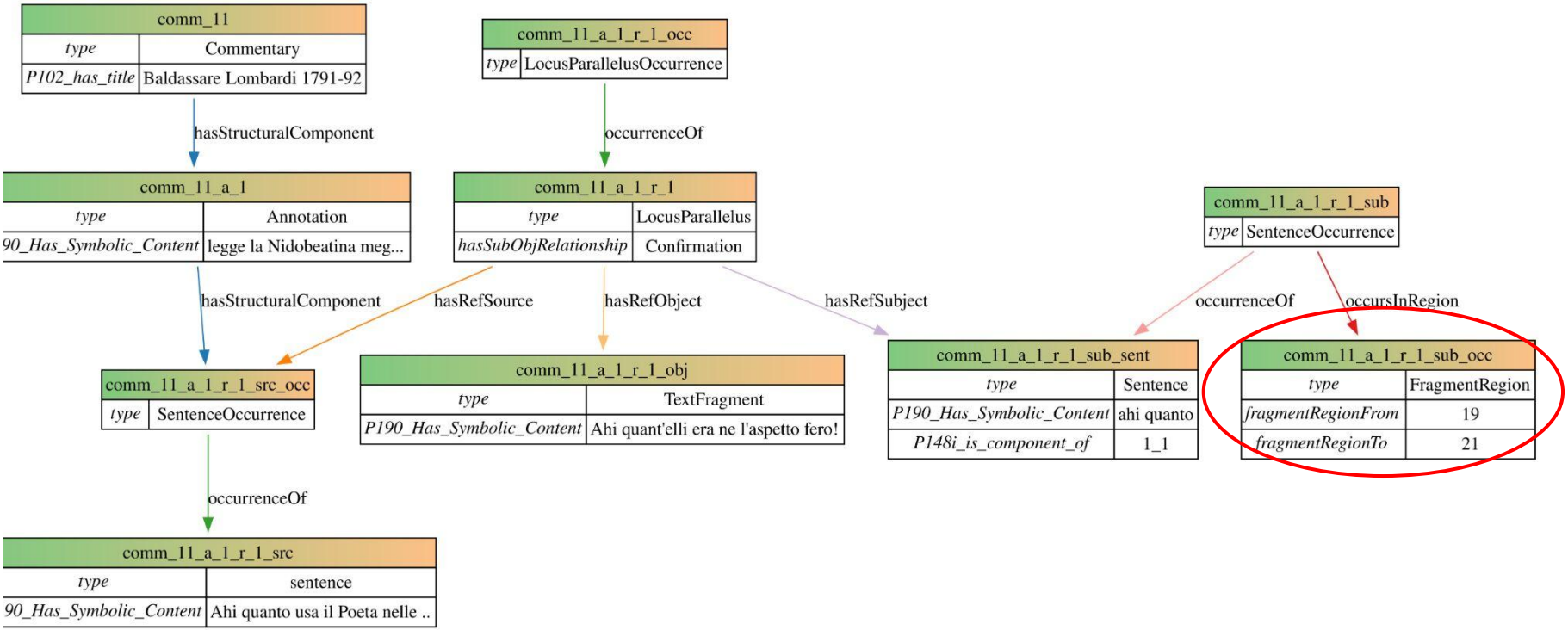
```
@prefix ecrm: <https://erlangen-crm.org/current/> .
@prefix hdn: <http://dantenetwork.it/data/commedia/> .
@prefix olires: <https://dantenetwork.it/ontology/olires/current/> .
@prefix syntit: <https://dantenetwork.it/ontology/syntit/current/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .


hdn:s_2 a syntit:Sentence ;
    syntit:hasSyntacticalComponent hdn:s_2_f_3,
        hdn:s_2_f_4,
        hdn:s_2_f_5,
        hdn:s_2_f_8 ;
    ecrm:P148i_is_component_of hdn:1_1 ;
    ecrm:P190_has_symbolic_content "Ahi quanto a dir qual era è cosa dura esta selva selvaggia e aspra
                                    e forte che nel pensier rinova la paura!"@it .
hdn:s_2_f_3 a syntit:Clause ;
    ecrm:P190_has_symbolic_content "Ahi quanto è cosa dura"@it .

hdn:s_2_f_4 a syntit:Clause ;
    ecrm:P190_has_symbolic_content "a dir"@it .

hdn:s_2_f_5 a syntit:Clause ;
    ecrm:P190_has_symbolic_content "qual era esta selva selvaggia e aspra e forte"@it .

hdn:s_2_f_8 a syntit:Clause ;
    ecrm:P190_has_symbolic_content "che nel pensier rinova la paura!"@it .
```



Esempio: entità sintattiche e grammaticali



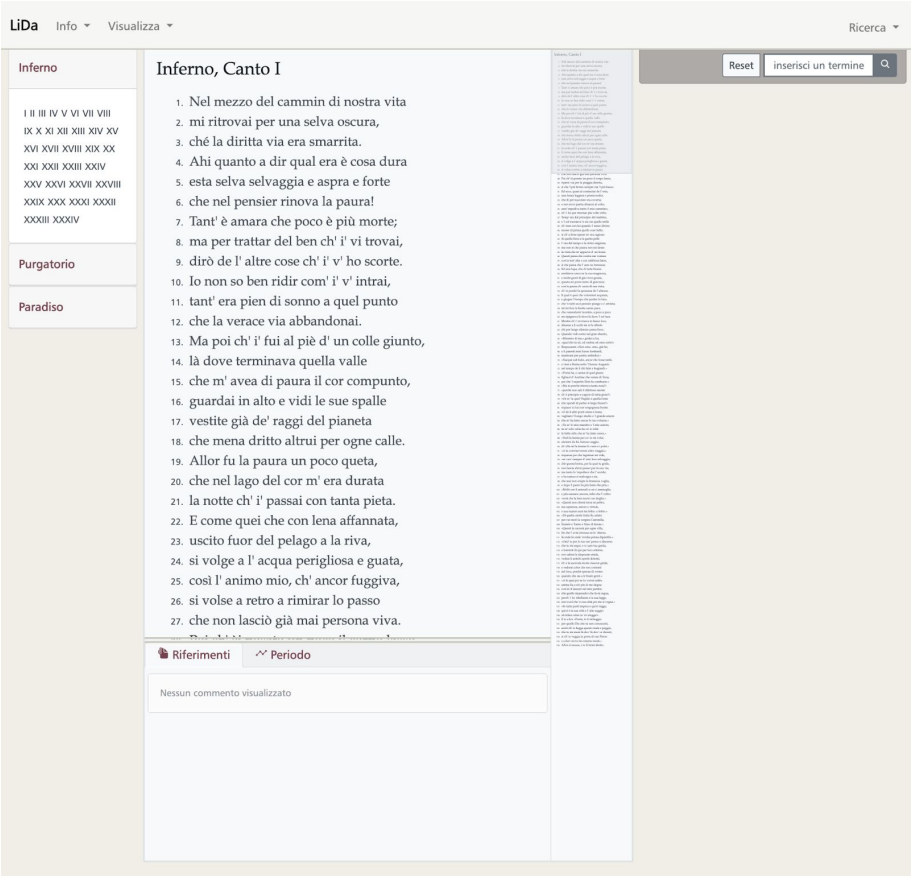
Esempio: riferimenti



I dati di LiDa

- Il dataset è rappresentato in RDF:
  - struttura flessibile
  - le entità sono identificate da un IRI e possono essere referenziate e riutilizzate come (L)LOD
  - è possibile applicare ragionatori automatici e inferire nuova conoscenza
- Ricerche semantiche usando SPARQL

La GUI: demo


LiDa
Info
Visualizza
Ricerca
Inferno
I II III IV V VI VII VIII
IX X XI XII XIII XIV XV
XVI XVII XVIII XIX XX
XXI XXII XXIII XXIV
XXV XXVI XXVII XXVIII
XXIX XXX XXXI XXXII
XXXIII XXXIV
Purgatorio
Paradiso
Inferno, Canto I
1. Nel mezzo del cammin di nostra vita
2. mi ritrovai per una selva oscura,
3. ché la diritta via era smarrita.
4. Ahi quanto a dir qual era è cosa dura
5. esta selva selvaggia e aspra e forte
6. che nel pensier rinova la paura!
7. Tant' è amara che poco è più morte;
8. ma per trattar del ben ch' i' vi trovai,
9. dirò de l' altre cose ch' i' v' ho scorte.
10. Io non so ben ridir com' i' v' intrai,
11. tant' era pien di sonno a quel punto
12. che la verace via abbandonai.
13. Ma poi ch' i' fui al piè d' un colle giunto,
14. là dove terminava quella valle
15. che m' avea di paura il cor compunto,
16. guardai in alto e vidi le sue spalle
17. vestite già de' raggi del pianeta
18. che mena dritto altrui per ogne calle.
19. Allor fu la paura un poco queta,
20. che nel lago del cor m' era durata
21. la notte ch' i' passai con tanta pieta.
22. E come quei che con lena affannata,
23. uscito fuor del pelago a la riva,
24. si volge a l' acqua perigliosa e guata,
25. così l' animo mio, ch' ancor fuggiva,
26. si volse a retro a rimirar lo passo
27. che non lasciò già mai persona viva.
Reset
inserisci un termine
Riferimenti
Periodo
Nessun commento visualizzato

Attività in corso

- Revisione delle ontologie
- Mapping ed integrazione di tutte le opere pubblicate in DanteSearch
- Realizzazione del KG per le metafore
- Estensione della UI di visualizzazione e ricerca

Risultati

Il grafo è pensato per essere esteso

- con la conoscenza sulle opere dantesche che per limiti di risorse non sono state considerate dal progetto
- con l'accesso ad altre risorse linguistiche (Vocabolario Dantesco)

o per essere linkato da altri grafi

- da un analogo grafo sulle opere di un altro autore
- da un grafo sulla letteratura italiana

o per essere usato da altri tool di accesso

- da un tool per gli studenti
- da un tool per i lettori

o per essere consumato da tool più sofisticati

- da un tool di AI

Acknowledgments

- Il lavoro è stato supportato in parte dal progetto Hypermedia Dante Network (HDN), finanziato nell’ambito del programma PRIN 2017 (2020)

- Il gruppo di lavoro di LiDa è composto da Carlo Meghini, Gaia Tomazzoli, Nicola Aloia, Luca Trupiano, Cesare Concordia, Daniele Metilli, Chiara Paolini