Thematic annotation and ontologies: application to digital epigraphy

(Daniele Silvi, University of Rome “Tor Vergata”)
(Fabio Ciotti, University of Rome “Tor Vergata”)

Thanks to the diffusion of new digital technologies in literary studies, much of the various national textual traditions are now available in digital format; a good percentage of them are available in advanced encoding formats such as XML/TEI, although the level of coding of these resources is mostly 'limited' to the representation of the textual and editorial structure. These high quality digital collections of textual resources represent an optimal starting point to carry out a systematic program of thematic analysis and annotation.

Our team, in Rome, University of Tor Vergata, is part of the PRIN national research project “Poetry of the memory, memory of the poetry. Word and theme occurrences in inscriptions and in the literary system”. Project is aimed to the development of formal models and computational tools for the morpho-syntactic and semantic analysis of Greek, Latin and Arabic texts. This project has a front-end on-line application, named Memorata Poetis. The website is actually under development and is provided with an account-protected access.

Our research is aimed to the thematic tagging of the early poetic texts, like sonnet, of the italian vernacular literature (epigraphic and epigrammatic tradition; and we’ll see how digital Epigraphy brings interesting results. Vernacular literature is literature written in the speech of the "common people". In the European tradition, this effectively means literature not written in Latin) and this job rise several theoretical and methodological issues we are currently heading, such as the selection (and definition) of the Canon and the relationship between epigraphy of the Classic Age and the Vernacular one (if present).

The suggested list of themes, derived from the classic latin literature, has shown its ineffectiveness from the beginning, due to the original and peculiar structure of the vernacular poetry, to wich it has to be now applied.

From the experience in thematic annotation, held on short texts of Vernacular poetry of the early centuries, some problems came out, first practical and, consequently, theoretical. The list follows:

1. The List of Themes is not really suitable to the content of vernacular poetry and, consequently, to a developmental and trans-cultural dimension. In other words, lacking of themes / motifs born with the vernacular poetry itself (Morgana, as an ideal of beauty);
2. Inadequacy of the concept of "temario" as a static entity;
3. Generalization of themes and motifs, which leads to a '"subjectivist impressionism";
4. Disconnection between discursive and non-discursive elements, between text and themes / motifs, like suggested by foucaldian dispositive analysis;
5. Multiple and sometimes fleeting variations of the same theme / motif, due to its evolution over time and across cultures (ex. the theme of "foreigners", which changes with the advent of Christianity. Hospes / Hostes).

The first two above mentioned problems, were partially (and temporarily) solved by adding new entries to the list of themes (temario) and crossing, on a same textual segment, several annotations in order to generate new ones.

Thematic annotations, in fact, rests on the basic themes or motifs, in itself generic. Just the dynamics of relations between different elements, seen in previous slides, makes us understand the importance of the application of ontologies to thematic annotation of literary texts. Despite the expansion of the theme's list may seem a simple solution, really it is not. An

important attempt to overcome this deception was, in fact, already started by Fahad Khan and Federico Boschetti, from the Zampolli Institute of Computational Linguistics of Pisa (IT). Their work is oriented to rewrite the current taxonomy, enriching it and turn it into an ontology. We intend to continue in this direction, already traced, flanking the ontology of the themes with an upper thematic ontology, that distinguishes between themes, motifs, imagery, etc. in short, between a variety of thematic items, not only linguistic.

Given the purpose of our proposal of a digital tematology, we consider more efficient the potential of Semantic Web technologies (WS) and Linked Data. The basic idea of the Semantic Web, is to associate a formalized description of certain parts of their meaning (intensional or extensional) to informative resources on the Web.

At the base there is the Resource Description Framework (RDF) that allows us to express semantic properties through simple predicative assertions. Each assertion has a "subject - predicate – object" structure. An assertion specifies a predicative relationship between subject and object (in RDF are allowed only binary relations).

RDF does not provide a default vocabulary of properties and relationships, that subsuming and organizing resources. It is a simple and rigorous data model for specifying properties of the resources, whatever they may be. In such a context, heterogenic and differentiated as the web is, we can theorize the existence of numerous diagrams and semantic vocabularies, based on different conceptualizations of particular domains, on different terms and languages. In general it can be assumed that there are also conceptualizations mutually contradictory, and/or changing over time. In order to make these conceptualizations usable in a computational way (at least partially) it is necessary to obtain an additional level of modeling: the formal ontologies.

The term ontology is now adopted to designate a large and differentiated class of objects, ranging from controlled vocabularies, to thesaura to the formal ontologies themselves.

Qualified by the adjective "formal", ontology now refers to the idea of giving a formalized account of a conceptual description of (portion of) the world. The relevances of formal ontologies for digitally processing literary and cultural objects, are both theoretical and operational.

The formal ontologies, in addition to the establishment a structured terminology for entities in a given domain, also defines the semantic shared by a given community, in formal-logical terms:

> In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application.

Our point is that a virtuous convergence between cultural and literary digital resources, ontologies and linked data practices represents a big chance for the future development of Literary and Digital Humanities. Building this kind of Rich Data for humanities research can also enhance the efficacy of text mining technologies, so we must not see it in contrast with those tools and methods.

Formal ontologies permit the application of computational inferences and reasoning methods to express explanation and make predictions. Their grounding in description logic has made possible the development of efficient Artificial Intelligence and inference engines.

Making ontologies and linking them to texts is not only a way for ensuring already existent knowledge exchange, but also a way to build knowledge.

In Humanities and Literary Studies conceptual formalization must face the deep problem of the traditional indeterminacy of theories, vagueness of terms, intrinsic ambiguity of the domain. Nevertheless, as long as we want to use computing we need to reduce the implicit and formalize, with the consciousness that formal modeling is inside the hermeneutic process and that we are expected to modify and adapt it, ad infinitum. The model must be determined at any given synchronic moment isomorphic to the domain and at the same time dependent to the perspective of the community who has the responsibility on it. An ontology, in the end, is an account of what the community knows as much as of how it knows what it knows, to recall Willard McCarty.

The purpose of a formal tematological domain ontology is (1) the definition of fundamental concepts: theme, motif and related concepts such as stereotype, locus communis or topos, image, character-type; (2) the identification of their mutual relations (hierarchy, similarity, instantiation, etc.); (3) the specification of their further properties, and (4) the various relationship (manifestation, exemplification, instantiation) between themes and motifs and their manifestations or discursive expression in general (in fact, such an ontology could be easily generalized to non-linguistic objects), as epigraphic dispositive of some some epigrams illustrates with heterogeneous elements (Sperlonga Cave).

Coming to the distinction between the two main concepts of the theme and motif, it is still valuable the suggestion of C. Segre:

> Si chiameranno temi quegli elementi stereotipi che sottendono tutto un testo o una parte ampia di esso; i motivi sono invece elementi minori, e possono essere presenti in numero anche elevato. Molte volte un tema risulta dall'insistenza di più motivi. I motivi hanno maggior facilità a rivelarsi sul piano del discorso linguistico, tanto che, se ripetuti, possono operare in modo simile a dei ritornelli; i temi sono perlopiù di carattere metadiscorsivo. I motivi costituiscono di solito risonanze discorsive della metadiscorsività del tema.

We can then extrapolate the following reports: a theme usually has no immediate discursive manifestation, and applies at the whole text, or at very large portions of this (think of the coincidence theme/chapter in *La Coscienza di Zeno*). One motif is a content unit smaller that can compose a theme (but not necessarily all the grounds of a text are functional to the theme or themes that characterize it) and has an immediate related discursive.

We believe that the ontological technologies of the Semantic Web provide a suitable instrumental apparatus for the creation of a thematic repertory that is not preeminently an enumeration - as are those recently published in Italy - but rather it organizes itself into multiple typological layers while allowing a rich network of horizontal relationships between classes and between instances of themes and motifs. The pursuit of this program, however, requires addressing several theoretical, methodological and instrumental issues.

1. First, we need to get to a formal and strict definition of the concepts of of the domain of thematics.
2. Secondly, we must design and implement a systematic repertoire of themes and motifs.
3. And finally, it's necessary to proceed to a systematic correlation between themes/motifs and texts.

Each of these points presents different problematic aspects and would require a lengthy treatment. In this paper we will restrict ourselves to consider only the first point: that is, the definition of a formal ontology for the thematic analysis of literary texts.

To partially overcome the objection about the proteic and elusive nature of these concepts it should be noted that the logical foundations of formal ontology languages such as OWL (for example, the assumption of the criterion of the open world) allows for some flexibility in the definition of concepts and in the association of properties to classes and individuals. In other words the modeling formalism is capable of at least partially represent the complexity and even the plurivocity of the domain objects.
The availability of this thematic ontology would permit to annotate the discursive manifestation of themes in encoded digital texts, producing progressively a thematic mapping of the literary tradition; this would form an invaluable empirical basis on which to extend systematically the literary analysis with computational and formal tools.