Between

Digital literary and cultural studies: the state of the art and perspectives

Fabio Ciotti

Literary Digital Humanities: the state of the art

Over the last decade Digital Humanities has ceased being a "niche discipline", becoming instead a major phenomenon in academic and cultural debates. According to numerous authorities, it represents one of the few points of resistance in the general decline of the humanities:

> Digital Humanities represents a major expansion of the purview of the humanities, precisely because it brings the values, representational and interpretive practices, meaning-making strategies, complexities, and ambiguities of being human into every realm of experience and knowledge of the world. It is a global, trans-historical, and transmedia approach to knowledge and meaning-making. (Burdick, Drucker, Lunenfeld, Presner, Schnapp 2012: vii)

The rapid spread of the term "Digital Humanities", rather than the more rigorous and older "Humanities Computing", indicates this success on a linguistic level, and shows the ambition of this vast and all-encompassing field of study, whose internal borders within the human sciences are becoming increasingly blurred.

The Digital Humanities today has a significant presence in university teaching[1], research and cultural heritage institutions (although their recognition in institutional contexts is far from adequate, especially in Europe). The discipline has organizations at national and international levels, which engage in scientific cooperation through bringing together a large number of scholars, organizing huge conferences and publishing authoritative monographs and periodicals.

In the past decades significant scientific results and outcomes have been achieved, and fundamental research infrastructures have been realized. These include:

1) The now widely shared theoretical and methodological awareness that the relationship between the humanities and the computational methodologies is epistemologically and theoretically relevant and not merely instrumental;

2) The development of the concept of *modeling* as the intellectual activity that characterizes the study of cultural objects and phenomena in the digital ecosystem, mediating between the level of theory and that of observation. Modeling is the method used in Digital Humanities,. It is relatively theory-independent but requires the formalization of theoretical entities and of the relationships between these entities, as well as the operationalization of the procedures to link those entities to observational data (ultimately to the textual and linguistic material or factual context);

3) The development of shared languages and standards for the modeling, representation and dissemination of high quality digital resources, which is an activity that comes out of strong cooperation with the information sciences community. This includes *Text Encoding*

[1] For the European area see for instance the Digital Humanities Course Registry set up in the context of the DARIAH, the European research infrastructure for the humanities at https://dariah.uni-koeln.de/.

Initiative (TEI) and related projects such as *Epidoc*, *Encoded Archival Description* (EAD) and other important metadata standards[2].

4) The extensive digitization campaigns of primary and secondary sources in textual and/or image facsimile formats, and the creation of vast online repositories that provide free access to an important part of the Western textual tradition.

5) The development of software frameworks and infrastructures for information retrieval, textual analysis, and online publication of textual resources, typically available freely as open source products or web services.

Despite these far-reaching outcomes in the theoretical and methodological rationales and in the development of general infrastructure for research, Digital Humanities still does not have a satisfactory influence in the individual traditional disciplinary fields. As John Unsworth (2003) said more than 10 years ago:

> We need (we still need) to demonstrate the usefulness of all the stuff we have digitized over the last decade and more – and usefulness not just in the form of increased access, but specifically, in what we can do with the stuff once we get it: what new questions we could ask, what old ones we could answer.

We could argue that the great influence of the post-structuralist or neo-idealist approaches in literary and cultural studies plays an important role in this distance between Mainstream and Digital Humanities: "Theory" without adjectives, as defined by J. Culler (1997), does not lend itself easily to interacting with the formalism of data structures and computational models. But it is also true that the computational methods for the analysis and the editing of texts and the

[2] TEI is the most widely used standard for the digital representation of textual data in the humanities (see http://www.tei-c.org), based on XML. Epidoc is a TEI customization for editing epigraphical inscription. EAD is an XML standard for the creation of digital finding aids in the archival context (see http://www.loc.gov/ead/).

results in terms of critics and scholarly editions have often fallen short of expectations, and where they have succeeded have rarely managed to acquire sufficient recognition in the context of the traditional disciplines.

Despite increasing theoretical awareness, the tools of representation and analysis produced so far have not satisfactorily addressed the problem of the specificity and complexity of the cultural and literary studies domain. In fact, the intellectual investment in the definition of new models and languages for the formal representation and processing of complex cultural objects has been rather low. Most commonly we have inherited and applied models and languages developed in computer science for different domains and necessities. The case of XML is a good example of this. For many good reasons, it has assumed a central role in the modeling of textual data. But it is well known that XML requires the adoption of a tree like data model that is not always suited to the structural nature of the objects to be represented, and that is unable to adequately represent the numerous and complex semantic levels that characterize a literary text (Ciotti 2011).

Directions for the future: digital methods and tools

Given this main picture, which directions should be taken in the development of new digital methods and infrastructures for humanities research? How can such efforts help to fill the gaps and gain new insights into cultural and literary phenomenona?

No doubt, to consolidate and extend the results already attained is a mission to be pursued: text archives must be preserved and extended; transcriptions and editions of primary sources using the current formalisms must be promoted; standards must be maintained and their wide application fostered. These are the basic requirements and missions for a research infrastructure, such as the one promoted by the recently established European consortium DARIAH (Digital Research Infrastructure for the Arts and Humanities, http://www.dariah.eu). However, can a research infrastructure provide

the ability to enhance the overall level of research in its domain, providing common innovative methodological tools and resources? We acknowledge the fact that, contrarily to hard sciences or social sciences, in the Humanities and in Literary Studies it is very hard to find common methods. In fact, Humanities is the realm of individuality in analysis and interpretation. Anyway, the digital turn requires those methodological commons, and their implementation in computational tools and services.

Amongst the many emerging research fields in DH in this context, two present themselves as the most promising and interesting[3]:

1) Big Data and distant reading: the application of text mining, knowledge extraction or topic modeling algorithms and tools to the Humanities digital data (whatever they mean).

2) Semantic Web and Linked Open Data: the experimentation with new formalisms and data models for semantic annotation in the Literary and Cultural Heritage domain.

Big data and distant reading

Big Data is the hype of the moment. The term refers to the application of *data mining* and *machine learning* heuristics to search for implicit recurring patterns and regular schemes inside wide amounts of data (structured or not), usually not visible to the naked eye.

The search for those patterns is based upon complex statistical algorithms, the most known of which derive from bayesian probability theory, where probability is the measure of the *a priori* plausibility assigned to a state of knowledge or to a belief. When those algorithms are applied to textual data the more specific term *text mining* is used. In the DH context the most widespread method for textual corpora analysis is *topic modeling*, that is the research of patterns of lexical

[3] As confirmed by the trend analysis of the subjects of the forthcoming DH 2015 conference papers conducted by Scott Weingart in his blog at www.scottbot.net.

tokens pattern co-occurring with a noticeable frequency inside a text or a corpus (Jockers 2013, Underwood 2012)[4].

The groundbreaking steps in this direction are due to the researchers of the Stanford Literary Lab, founded and directed by Franco Moretti and Matthew Jocker. Moretti (2013) himself has attempted to give a literary theoretical rationale to these experiments, introducing the notion of "distant reading" (opposed to the traditional "close reading" method in literary criticism as defined by *New Criticism*). The basic idea of this approach is that there are synchronic or diachronic literary and cultural facts that are undetectable to the usual deep reading and local interpretation methods that require the scrutiny of hundreds or thousands of texts and documents (and millions of lexical tokens). In this way we can gain access to otherwise unknowable information that plays a significant explanatory role in understanding literary phenomena as the evolution of genres, the affirmation of a style and its reception, and the presence of recurrent content clusters in a given time span of literary history.

I cannot go deeper into this issue here. However, I observe that the enthusiasm showed by the now many practitioners of distant reading and topic modeling seems to overshadow some critical issues.

First, big data algorithms in general are completely independent from the context (they can be applied indifferently to stock exchange transactions or to very large textual corpora). They individuate similarities and recurring patterns independently from the semantics of the data. But in a sense, when you work with structured data semantics is fixed a priori in the data schema; if you work with non or low structured data (as is the case with large text only corpora), the characters (or the n-grams) are the atomic data, and they play a very limited semantic role. I am not saying that it is impossible to discover interesting phenomena also at this level, but many relevant facts concerning textual and literary objects are simply out of scope.

[4] Various tools implementing Latent Dirichlet Allocation (LDA), the most known variant of topic modeling algorithms, are available: Mallet, Stanford Topic Modeling Toolbox, Serendip.

Texts, as any other cultural artifact, are essentially intentional objects: to cite Daniel Dennet (1990), they are the products of the intentional stance of their producers and users. The meaning of a word, the usage of a metaphor, or the choice of a metric or rhythmic solution in a poetic text, are determined by the attribution of sense and meaning by the author and by the reader (I do not discuss here whether the former is more or less relevant than the latter). They can be, and often are, idiolectal or even unique. Purely quantitative and mass analysis can delineate the textual "degree zero", on which the secondary modeling system of culture and literature builds its significance (Lotman 1970).

Moreover, although on a large scale quantitative methods can give some insights into lexical meanings and their distribution, we must observe that the meaning in literary texts is multi-layered, and that some layers do not have direct lexicalization or have a very complex and dispersed one (think about aspects of a narrative text at different abstraction levels like anaphors, themes, plot and fabula, actants).

Besides these theoretical pitfalls, there are some methodological and pragmatic ones as well. Firstly, to do use the Big Data methods you must have Big Data. The dimension of the data sets used in hard sciences, economics and (partially) social sciences are bigger by many orders of magnitude than the largest textual datasets that we can have in literary studies. The efficacy and adequacy of the probabilistic algorithms in this context is not so certain.

Secondly, if a so called very large textual set is composed of documents spread over a long period of time, diachronic variation of the form and usage of the language (both on the syntactic and semantic levels) can invalidate purely quantitative and statistic measures.

Finally, yet importantly, there is the problem of data quality and of the assessment of the protocols followed to build the data sets, a problem recently raised also in the context of hard sciences[5]. If in social

[5] See for examples what David Crotty (2014), senior editor at Oxford University Press, observes in a recent blog post: "Detailed methodologies

sciences a given level of statistical error in the data is "acceptable", it is hardly questionable that Humanities and Literary studies have a much lower threshold.

Semantic technologies and ontologies: towards Rich (linked) Data

Big Data methods rely on the application of quantitative algorithms to large sets of simple and possibly unstructured data. The semantic oriented approach is instead based on the modeling of complex human interpretations of data through formal languages: we can say that in this case we are creating and processing Rich Data. This approach has similarities with the humanities tradition of annotation and comment, and has informed the text encoding field in the Digital Humanities.

The first and most ambitious formulation of this idea in the context of modern digital and networked technologies is the seminal vision of the Semantic Web by Tim Berners-Lee in the late 90s (Berners-Lee, Hendler, Lassila 2001). He proposed that the information resources on the Web should be associated with a set of semantic metadata, so that their intended semantics could be accessed and processed by software agents.

Semantic Web has become an official W3C initiative, which has developed a number of languages and data models. The most basic one it is *Resource Description Framework* (RDF), a simple data model that allows for binary predicates to be stated (subject – predicate – object). RDF as such does not specify the content of those statements. That role

---

would be of tremendous value across the spectrum of scientific research. The validity of many types of sociological studies, for example, depends greatly on how those studies were conducted. Why not offer all the gory details to better help readers understand whether the experiments were well conducted so we know whether the data is worth reusing? Beyond reproducibility, increased availability of trusted protocols would be a boon to scientific progress simply because more people would have more access to more techniques".

is reserved to the formal ontology level, using RDFS and OWL formalisms (Antoniou, Van Harmelen, 2008; Di Noia, De Virgilio, Di Sciascio, Donini, 2013).

The original general vision of Semantic Web has proved to be unfeasible for many technical and social reasons. Nonetheless, Semantic Web methods and technologies have had successful application in many restricted and controlled domains, and in the context of the *Linked Data.*

The term Linked Data (introduced again by Tim Berners-Lee) refers to a set of guidelines for publishing and interlinking structured data on the Web (Bizer, Heath 2011). These principles are the following:

1. Use URIs as names for things;
2. Use HTTP URIs, so that people can look up those names;
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL);
4. Include links to other URIs, so that they can discover more things.

Linked Data movement has grown quickly and Cultural Heritage initiatives play a relevant role in this[6]. My point is that a virtuous convergence between cultural and literary digital resources, ontologies and linked data practices represents a big opportunity 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, and it must not be considered in contrast with those tools and methods.

[6] See the Web site of the "Linked Open Data in Library Archives and Museum" (LODLAM) network, http://lodlam.net.

The case for ontology and why humanists should care

In this overall picture I want to stress the centrality of formal ontologies and of ontology building for the humanities.

The term has been inherited from classic and medieval metaphysics (since the Aristotelian system which denoted the theory of being and its categories). Qualified by the adjective "formal", it now refers to the idea of giving a formalized account of a conceptual description of (a portion of) the world:

> 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. (Gruber 2009)

The relevance of formal ontologies for literary and cultural objects digital processing are both theoretical and operational.

First, in the DH community a great relevance has been given to the notion of the model and modeling. But the problem with the model/modeling notions is that they are umbrella terms, relating to a wide and diverse collection of conceptual objects and practices. In general, we can sort the roles assigned to modeling in scientific activity into three areas:

- representation/communication: models ensure that a community of practice shares the fundamental concepts of a domain;
- explanation/prediction: models relate facts and concepts providing explanations and possibly predictions of the behavior of a system;

- multiple views/perspectives mediation: models mediate between the different perspectives that can arise within a single community of practice and between different but proximal communities of practice.

Ontological modeling formalizes the common sense concept of model, giving it a precise logical semantics and a definite functional role in each of these areas.

Creating formal models based on explicit conceptualization and logical foundation assumes that all the discourses are firmly grounded in a common "setting" of the domain.

Formal ontologies permit the application of computational inferences and reasoning methods to explain and to make predictions. Their grounding in *description logic* has made possible the development of efficient automatic reasoners and inference engines:

> Logic reasoning is one possible application for ontologies. It is probably helpful (i) to check consistency during ontology development, (ii) to enable semi-automatic merging of (domain) ontologies as well as (iii) to deduce hidden information contained in the ontology. (Zöllner-Weber 2009)

Finally, Semantic Web modeling provides methods to compare and eventually merge different ontologies and, being based on the Open World Assumption, ensures the functionality of the model even if it is incomplete or conceived as a work in progress.

In the Humanities and Literary Studies, conceptual formalization must face the deep problem of the indeterminacy of theories and of theoretical terms. We can concede that indeterminacy is a characteristic of the object domain. Nevertheless, as long as we want to use computing we need to reduce that which is implicit and, with the consciousness that formal modeling is inside the hermeneutic process and that we are expected to modify and adapt it, must formalize it *ad infinitum*. Nonetheless, at a given synchronic moment the model must be determined, isomorphic to the domain and at the same time dependent on the perspective of the community of practice who has

responsibility for it. To recall Willard McCarty, an ontology is an account of what the community knows, as much as it is an account of how it knows what it knows.

Toward an infrastructure for a Literary Semantic Web

Given this theoretical context, I propose a sort of Literary and Cultural Semantic Web, a digital environment and infrastructure incorporating semantic methods and practices of digital interaction and cooperation already available and tested in the Digital Humanities community. The components of this networked infrastructure of resources, tools and services are:

1. large, high-quality document archives belonging to different linguistic traditions / cultures in standard encoding formats;
2. a set of methods and computational tools for the distributed and cooperative annotation of digital resources;
3. a set of domain specific shared ontologies to ground the annotations, organized in a multilayered way, each dedicated to a particular aspect of the intratextual, extratextual and intertextual structure:
    - real places and spaces chronologically adapted
    - real people (including authors)
    - works and literary history categories
    - historical events
    - fictional places and worlds
    - fictional characters and entities
    - themes and motives
    - rhetorical figures
    - genres and stylistic features
4. tools able to visualize and process "semantic" levels of digital information, which allow knowledge transfer and sharing within

the digital environment as linked data.

Although these technologies and the relevant underlying methodologies are now widely used, there are critical aspects to highlight: formal ontology has the dual capacity of fixing prior knowledge of what is in the domain, and, simultaneously, promoting the development of new knowledge; multiple ontological analysis can be connected with the same (passage of) text, creating multiple knowledge and cultural content layers that overlap with the textual layer, and thus uncovering its complexity; these stratified texts can be re-used in different contexts, and by different kinds of users: from "professional scholars" to culturally aware users who are attracted by the potential text mash-ups. As a result, it is possible to use data for processing and activities such as: visual representations of the texts' content, the integration of text and maps, their re-use for tourist-oriented services, etc.

Building such an infrastructure is obviously a very demanding task. But many of the building blocks are already there. Above all the history and evolution of the Web has shown that it is possible to build complex systems through a public, incremental and cooperative process, and that this strategy proves to be much more efficient and effective than private, monolithic and centralized ones.

The infrastructure we are envisioning is intrinsically cooperative and driven by crowd sourcing. This infrastructural model opens a space into which "experts" - professional scholars - and "non-experts" can enter, to read, visualize and analyze the resources at different levels of complexity, and, in so doing, enrich them. The traditional experts' literary, aesthetic, historical-critical reading and interpretation are no longer exclusive or dominant.

Open data, collaborative annotation, ontologies, relations with the context and connection to other network resources all contribute towards defining a new form of digital "cultural literacy" (Hirsch 1987).

Works cited

Antoniou, G., Van Harmelen, F., *A semantic Web primer*, Cambridge, Mass., MIT Press, 2008.

Berners-Lee, T., Hendler, J., Lassila, O. «The Semantic Web», *Scientific American*, 284, 5, 2001, pp. 34–43.

Bizer, C., Heath, T., *Linked Data: Evolving the Web into a Global Data Space*, «Synthesis Lectures on the Semantic Web: Theory and Technology», I, 1, 2011.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., Schnapp, J. *Digital_Humanities*, Cambridge, MIT Press, 2012.

Ciotti, F., "La rappresentazione digitale del testo: il paradigma del markup e i suoi sviluppi", *La Macchina nel Tempo: Studi di informatica umanistica in onore di Tito Orlandi*, Perilli, L., Fiormonte D. (eds), Firenze, Le Lettere, 2011.

Crotty, D., "Nevermind the Data, Where are the Protocols?", *The Scholarly Kitchen*, November 8, 2014, http://scholarlykitchen.sspnet.org/2014/11/18/nevermind-the-data-where-are-the-protocols (last access 14/12/2014).

Culler, J., *Literary Theory: A Very Short Introduction*, Oxford, Oxford University Press, 1997.

Dennet, D., "The Interpretation of Texts, People and Other Artifacts", *Philosophy and Phenomenological Research*, L, Supplement, 177-194, 1990, reprinted in Losonsky, M., (ed.), *Language and Mind: Contemporary, Readings in Philosopohy and Cognitive Science*, Oxford, Blackwells, 1995.

Di Noia, T., De Virgilio, R., Di Sciascio, E., Donini, F. M., *Semantic Web. Tra Ontologie e Open Data*, Milano, Apogeo, 2013.

Gruber, T. R., "Ontology", *Encyclopedia of Database Systems*, Springer-Verlag, 2009.

Hirsch, E.D., *Cultural Literacy: What Every American Needs to Know*. Boston, Houghton Mifflin, 1987.

Jockers, Matthew L., *Macroanalysis: Digital Methods and Literary History*. Champaign, University of Illinois Press, 2013.

Lotman, J., *Struttura del testo poetico*, II ed., Milan, Mursia, 1990 (*Struktura chudozestvennogo teksta*, Moskva, Iskusstvo, 1970).

Moretti, F., *Distant Reading*, London, Verso, 2013.

Underwood, T. "Topic Modeling Made Just Simple Enough", *The Stone and the Shell*, April 7, 2012, http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough (last access 14/12/2014).

Unsworth, J., "Tool-Time, or 'Haven't We Been Here Already?' Ten Years in Humanities Computing", paper presented at the conference *Transforming Disciplines: The Humanities and Computer Science*, Washington, DC, 17-18 January 2003, http://people.brandeis.edu/~unsworth/carnegie-ninch.03.html (last access 14/12/2014).

Zöllner-Weber, A., «Ontologies and Logic Reasoning as Tools in Humanities?», *Digital Humanities Quarterly*, III, 4, 2009, http://www.digitalhumanities.org/dhq/vol/3/4/000068/000068.html (last access 14/12/2014).

The author

Fabio Ciotti

Fabio Ciotti is Assistant Professor at the University of Roma Tor Vergata, where he teaches Digital Literary Studies and Theory of Literature. He is President of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD, the Italian digital humanities association); elected member in the TEI Consortium Technical Council and in the EADH (European Association of Digital Humanities) Executive Board.

His scientific and research work covers various aspects and themes of Digital Humanities and Literary Studies, both from the theoretical and the practical point of view: the applications of computational methods to the analysis of narrative texts; digital text encoding and representation; applications of XML and TEI

technologies to literary computing; modeling and creation of digital libraries; applications of new media and computer mediated communication to Humanities research and teaching. Recently his research interests concern the application of Semantic Web/Linked data principles and technologies to humanities digital libraries and textual corpora. He is interested in particular in ontologies for the semantic analysis of literary texts and for the semantics of markup languages.

Fabio Ciotti has been Chair of the Local Organization Committee for TEI Conference and Members Meeting 2013; a member of the Program Committee for the TEI Conference 2014 as well as for AIUCD conferences. He has been scientific consultant for text encoding and technological infrastructures in several digital libraries and archives projects, most notably Biblioteca Italiana (Italian literary tradition, http://www.bibliotecaitaliana.it) and DigilibLT (Late Latin tradition, http://www.digiliblt.unipmn.it/). He is currently involved in the Geolat project - aimed at building an ontology for the geographical knowledge of the ancient world and funded by Fondazione San Paolo - and in an Italian Ministry of University funded project (PRIN) for the thematic annotation of Latin and early Italian texts.

Email: fabio.ciotti@uniroma2.it

The paper

Data invio: 30/08/2014
Data accettazione: 30/10/2014
Data pubblicazione: 30/11/2014

How to quote this paper

Ciotti, Fabio, "Digital literary and cultural studies: the state of the art and perspectives", *Tecnologia, immaginazione e forme del narrare*, Ed. L.

Esposito, E. Piga, A. Ruggiero,, *Between*, IV.8 (2014), http://www.Between-journal.it/