frontiers in Digital Humanities

SPECIALTY GRAND CHALLENGE
published: 02 December 2015
doi: 10.3389/fdigh.2015.00007

The Logic of the Big Data Turn in Digital Literary Studies

Jean-Gabriel Ganascia[1,2]*

[1] ACASA Team, Laboratoire d'Informatique de Paris 6 (LIP6), Pierre and Marie Curie University, Paris, France, [2] OBVIL Labex, Sorbonne University, Paris, France

Keywords: Digital Humanities, Digital Literary Studies, big data, induction, abduction, sciences of nature, sciences of culture

THE PROBLEM

The Digital Humanities, and especially the literary side of the Digital Humanities, i.e., Digital Literary Studies, propose systematic and technologically equipped methodologies in activities where, for centuries, intuition and intelligent handling had played a predominant role. The recent "big data" turn in the natural and social sciences has been particularly revealing of how these new approaches can be applied to traditional scholarly disciplines, such as literary studies. In so doing, big data can renew, with the use of computers, the Humanities, i.e., the disciplines rationally studying human works and cultural production. Digital Literary Studies are emblematic of these new approaches, certainly because they constitute the oldest subfield of the Digital Humanities, as some early projects like the *Trésor de la Langue Française* attest but also because they are the domain in which the intellectual stakes of mass digitization has already been extensively used and debated as demonstrated by Franco Moretti's *Graphs, Maps, Trees* (Moretti, 2005), for instance.

Some view this evolution enthusiastically as a shift toward the "hard" sciences. This is the case of Matthew Jockers who affirms in the chapter entitled "Revolution" of his book *Macroanalysis* (Jockers, 2013) that: "Now, slowly and surely, the same elements that have had such an impact on the sciences are revolutionizing the way that research in the humanities get done" (p. 10). Further on, he declares that literary methodology is "in essence no different from the scientific one" (p. 13).

Others assert that some questions cannot be dealt with using the same methods in the humanities and the natural sciences, like physics or biology. That is the case of Stephen Ramsay, who, in *Reading Machines* (Ramsay, 2011), assures us that, even if some problems in the Humanities, like authorship identification, can clearly find comfort with the methods developed by the natural sciences, for most literary critical endeavors, such as characterizing the subjectivity of Virginia Wolf in her novel *The Waves*, for instance, it is not possible to clearly identify a set of "falsifiable" facts.

Between these two extremes, many scholars provide convincing illustrations of what digitization allows and then discuss the nature and current evolution of the Humanities in general, and literary studies in particular. The *Companion to Digital Humanities* (Schreibman et al., 2004), the *Companion to Digital Literary Studies* (Siemens and Schreibman, 2008), and more recently an excellent online MLA Commons anthology dedicated to *Literary Studies in the Digital Age* (Price and Siemens, 2013) all provide various and enriching views on these topics.

We attempt here to conciliate the two above-mentioned and apparently antagonistic views with the help of a philosophical approach. More precisely, our Grand Challenge is in the service of establishing solid epistemological foundations for the Digital Humanities, which is necessitated by the increasingly important role attributed to digital tools in humanistic research. We also claim that employing a conceptual apparatus originally built by German neo-Kantian philosophers at the beginning of the twentieth century, in particular by Heinrich Rickert and Ernst Cassirer, seems particularly relevant today with the emergence of "big data," primarily because the logical nature of the possible inferences drawn from this sort of data needs to be clarified.

Edited by:
Frederic Kaplan, École Polytechnique Fédérale de Lausanne, Switzerland

Reviewed by:
Glenn Roe, Australian National University, Australia

***Correspondence:**
Jean-Gabriel Ganascia
jean-gabriel.ganascia@lip6.fr

Specialty section:
This article was submitted to Digital Literary Studies, a section of the journal Frontiers in Digital Humanities

Received: *04 May 2015*
Accepted: *02 November 2015*
Published: *02 December 2015*

Citation:
Ganascia J-G (2015) The Logic of the Big Data Turn in Digital Literary Studies. Front. Digit. Humanit. 2:7. doi: 10.3389/fdigh.2015.00007

The following essay is divided into four parts. The first recalls the distinction between the "sciences of nature" and the "sciences of culture," which is at the heart of the Rickert and Cassirer conceptual apparatus. The second analyzes the status of the Digital Humanities with respect to this distinction. The next part shows that the use of big data does not necessarily restrict one to making purely inductive inferences, in the logical sense, from the data. It also explains why the logic of the Digital Humanities is closer to the logic of the traditional Humanities, even if, by making use of large digital datasets, they at first sight seem incompatible. Lastly, the final part concludes on the role of theory in Digital Humanities and gives some examples of the new and exiting areas of investigation that Digital Literary Studies opens.

THE LOGIC OF THE HUMANITIES

At the beginning of the twentieth century, a German neo-Kantian philosopher, Heinrich Rickert – who influenced many important intellectuals, among them the sociologist Max Weber and the young Martin Heidegger – attempted to base the Humanities on a rigorous foundation. More precisely, he wanted to scientifically characterize culture understood as the result of goal-oriented activities. The notion of the "Sciences of Culture" (*Kulturwissenschaften*[1] in German, which designates "Humanities" in American English) (Rickert, 1921) was introduced to epistemologically ground the Humanities as empirical sciences that interpret human achievements and activities as the results of mental processes. Rickert clearly distinguishes the scientific characterization of the mind enacted by the Humanities from that of the psychological sciences, which deal with mental phenomena using the methods of the physical sciences. He affirms that spiritual phenomena have a specificity that cannot be reduced to their physicality alone, even if they can be submitted to a rational and empirical inquiry.

According to him, and to his student Ernst Cassirer (Cassirer, 1923, 1942), the underlying logic of the "sciences of culture" totally differs from the logic of what they call the "sciences of nature" (*Naturwissenschaft*), i.e., the natural sciences.[2]

Briefly speaking, Rickert and Cassirer first differentiate the theoretical sciences like mathematics, which deal with abstract and perfect entities, such as numbers, figures, or functions, from the empirical sciences that are confronted with the material reality of the world. Then, among the empirical sciences, they further differentiate the "sciences of nature," which deal with physical perceptions, and the "sciences of culture" that give meaning to human works. According to them, the "sciences of nature" proceed by generalizing cases: they extract general properties of objects and they determine laws, i.e., constant relations between observations. As a consequence, the logic of the "sciences of nature" is mainly inductive, in the logical sense of the word, i.e., this logic goes from the observation of many particular cases to the construction of general laws that cover and summarize the observations, even if the practical modalities of reasoning for researchers may be deductive or abductive. The important point is that the particular cases have to be forgotten; they have to be abstracted and analyzed in general terms as composed of well-defined objects that make no reference to the context of the situation. The validity of this scientific activity relies on the constancy and the generality of the extracted laws.

By contrast, the "sciences of culture" do not proceed by generalizing multiple cases. They do not extract laws, i.e., relations between observations; they do not even work with physical perceptions, but with meaningful objects that have to be understood. In brief, their main function is to give sense to the works of humans, i.e., our shared cultural record. Their means of investigation is to understand particulars, and their general methodology is to observe individual instances and give meaning to them. However, they often have to choose, among the particulars, instances that are paradigmatic, i.e., that can teach general lessons that may be reused in other circumstances. In other words, the "sciences of culture" are not properly interested in the singularity of cases, which should be ignored, but in the overall understandability of the individual instances under study. Their methods help to give meaning to observations of complex individual cases.

ARE DIGITAL HUMANITIES "SCIENCES OF NATURE" OR "SCIENCES OF CULTURE"?

The main question here concerns the epistemological status of the Digital Humanities. On the one hand, their objects of study, i.e., human works and cultural records, bring them close to the "sciences of culture"; on the other hand, their method of investigation, and especially the use of computers and huge datasets, seems to bring them close to the "sciences of nature." Therefore, at first sight, Digital Humanities in general and Digital Literary Studies in particular, belong to both the "sciences of nature" and the "sciences of culture." However, this dual membership does not answer the initial question about the specificity of the Digital Humanities and their status compared to that of the "sciences of the nature." Clearly, we must pursue our investigation further. To do this, let us consider the three following points:

First, Digital Humanities and Digital Literary Studies are empirical sciences, as are the traditional Humanities, since they are based on facts. Even if, as Ramsay claims (Ramsay, 2011), it is difficult to objectively characterize the subjectivity of an author such as Virginia Wolf, it is absolutely necessary to give facts that support any hypothesis.

Second, as Ramsay also claims (Ramsay, 2011), humorously quoting Jarry's Dr. Faustroll (Jarry, 1911), Digital Literary Studies do not function as purely inductive sciences: even if they are based on facts and even though some questions, like the authorship identification problem, look similar in their formulation to investigations in the "sciences of the nature," nobody really aims in this context to establish general laws. As part of the Humanities, Digital Literary Studies examines the human record and considers particulars – e.g., a novel, the work of an author, a generation of writers, a genre, a culture, etc. – in order to understand these works as goal-oriented activities and to characterize their specificity. But, unlike the traditional Humanities, Digital Literary Studies also

[1]For instance, the German title of Cassirer's book *Zur Logik der Kulturwissenschaften*. has been translated in English "The Logic of the Humanities" (Cassirer, 1942).

[2]For the sake of clarity, we use here the term "sciences of nature" to refer to the concept of *naturwissenschaft*, as used by Rickert and Cassirer in their works, even if it looks similar to the common notion of natural sciences.

makes use of massive datasets that are automatically processed. In so doing, they propose new digital hermeneutic operators that give meaning to these human records, without necessarily aiming to delineate – and still less to discover – general laws.

Finally, the modality of reasoning in the Humanities is essentially abductive, in the sense that Charles Pierce gives to this word, which means that humanists are looking for provisional explanations, i.e., for facts that enforce an explanatory hypothesis within a theoretical framework. For instance, in literary criticism, intertextuality (Bloom, 1973; Compagnon, 1979; Genette, 1982), interdiscursivity (Adam, 2006), or textual genetics (Grésillon, 1994; Hay, 2002) are theoretical frameworks to which scholars refer when they search for explanations that make literary works more understandable. As mentioned in Murray-Jones (2011), the use of computers in the Humanities does not necessarily lead one to abandon theory. On the contrary, programs need to refer to well-defined theoretical frameworks on which they can bring pieces of material evidence to bear. This does not mean that each program need be a theory, or that each individually encodes a theory, but rather, a program, e.g., a visualization tool, that has not made an explicit reference to the theoretical framework on which it is built is useless and has no real scientific value whatever the facts that it seems to generate.

In summary, point one does not provide any clear evidence in favor of the Digital Humanities belonging either to the "sciences of nature" or to the "sciences of culture"; point two seems, at first sight, to turn the scales toward the "sciences of nature"; while point three seems to favor the "sciences of culture." Point two is of key importance here, because it is through the use of huge datasets that the Digital Humanities are clearly distinguished from the traditional Humanities. Does this, however, as Moretti (2005) and Jockers (2013) suggest, necessarily lead to a change of logic in the "sciences of the culture," which become inductive in the same manner as the "sciences of the nature"? We will investigate this question further in the next section by detailing the nature of data-based reasoning.

THE LOGIC OF *BIG DATA*

Taken literally, the locution "big data" refers to the size of data. But, what does "big" mean for the Digital Humanities? A million, a billion, and a trillion bytes are small compared to the Terabytes and Petabytes that are usually considered as the standard for "big data." In the case of Digital Literary Studies, the total number of texts that can be characterized as literary works, including novels, poetry, and theater, does not exceed a few million books, which has been seen characterized as a delimiting horizon by Gregory Crane in his famous paper, "What do you do with a million books?" (Crane, 2006). If we consider an average upper size of 1 million characters per book, the overall digital library corresponds at most to a few Terabytes, which is quite small compared to the current magnitude of scientific big data. Nevertheless, the Digital Literary Studies need not restrict itself to investigations of digitized literary texts. In a recent paper, Kaplan (2015) clearly expresses three levels of big data for the Digital Humanities:

- the level of human records, which corresponds in our case to literary works.
- the level of social interactions, which, in the case of literary studies, could include scientific theories that influenced novelists, newspapers to which authors contributed or that related current world events, and many others. This level corresponds to the intellectual landscape at time of writing, and while the idea of digitizing the integrality of the intellectual context for any given author may seem unrealistic; furthermore, this is perhaps also a case of confusing the map and the territory.
- the third level gathers material exchanges with technical devices, such as e-books, which gives an idea of the way people are reading, or with computers, which will allow us to keep track of different writers' drafts or search queries of authors or readers on the web. In the future, this will certainly be a key source of information that will allow us to evaluate the ways in which works are produced and received.

However, even if "big data" are often characterized by the famous "3Vs" acronym – Volume, Variety, and Velocity – neither the volume of the datasets, nor their variability and "velocity," i.e., their constant evolution, can fully encapsulate the logic of "big data." As mentioned in many publications (Aiden and Michel, 2013; Mayer-Schonberger and Cukier, 2013), one of the key characteristics of big data is the absence of sampling. The totality of data is used during the exploitation, without restriction to a random selection, like in a survey or a poll, as was the case with classical statistical studies in the past. In the case of literature, almost all the published literary texts, scholarly books, and newspapers will be digitized in the coming years. This means that it will not only be possible to detect specificities of an author that distinguish him/her from others or that characterize his/her work or the generation of writers to which he/she belongs, etc., but it will also be possible to identify citations, influences, plagiarism, pastiches, or reuses on a truly massive scale as most of the possible sources of inspiration, i.e., most of the writings to which the authors could have had access, will be available in digital form.

Besides the absence of sampling, we should also note that the algorithmic exploitation of big data does not extract solely causal relations, but rather empirically observed correlations, among which only some may correspond to actual causal relationships. Therefore, contrary to intuition, the inferences drawn from big data are not necessarily inductive: either the reasoning starts without any theory and generates correlations that need to be proved to constitute a true body of knowledge, in which case it corresponds to actual inductive inferences, or it starts from a theory that is used to find possible explanations of the data, which corresponds to abductive inferences.

THE LOGIC OF DIGITAL LITERARY STUDIES

It follows from what we have said that the logic of the Digital Humanities equipped with big data techniques is definitely a continuation of the logic of the Humanities, i.e., it may be either inductive or abductive, even when using very large datasets. More precisely, even if inductive inferences play a role in the digital humanists' investigations, their main modalities of reasoning are essentially abductive, which means that digital humanists as

humanists are looking for explanations, i.e., they are seeking facts that strengthen new hypotheses within a theoretical framework.

This point echoes similar debates that shook the Digital Humanities community (Fitzpatrick, 2011; Gold, 2012; Presner and Schnapp, 2013) a few years ago, underscoring the antagonism between those who envisage the Digital Humanities as a new theoretical approach to the Humanities, which would constitute a paradigm shift in the Kuhnian sense, and those who think that it is now time to focus on methods and, more precisely, on tangible implementations of software that procure empirical evidence (Berry, 2011; Cecire, 2011). According to this latter view, it would mean that the Digital Humanities are moving us toward a "post-theoretical age" in which interpretation becomes less important than the making of tools, archives or other digital methods. As Porsdam (2011) claims, this antagonism conceals a deeper opposition between classical humanist culture, for which rational thinking is a discursive activity, and techno-scientific culture, which asserts that the reaching of certainty today requires formalization.

Our claim here is that, despite these debates within the larger Digital Humanities community, for Digital Literary Studies, there is no real antagonism between the logic of the "sciences of culture," as described by Rickert and Cassirer, and the making of tools that help to interpret huge databases with respect to existing theories. In other words, computer-aided methods can be seen as a continuation of traditional humanistic approaches. As such, they can afford many opportunities to renew humanistic methods and to make them more accurate, by helping to empirically confront working hypotheses with datasets that now approach the entirety of our printed record, taking into consideration not only literary works themselves but also the intellectual landscapes surrounding the authors of these works.

To conclude, the Grand Challenge that we support in this paper is to build tools that are able to analyze literary works through the prism of such or such theory, and that can furthermore provide evidence of the fecundity of the theory under consideration. These tools automate hermeneutic operators, among which some are traditional and others new, and which are made possible by the mass digitization of our shared cultural record. We have followed this approach in the development of MEDITE (Ganascia et al., 2004), a text aligner for textual genetics and the comparative publishing of textual variants, and also with PHŒBUS (Ganascia et al., 2014), a program that detects textual reuses and other forms of intertextuality. This is also what we are currently doing in terms of theories of "interdiscursivity," which we explore by detecting semantic patterns in passages of texts, both for stylistic analysis (Boukhaled and Ganascia, 2015; Frontini et al., 2015), by extracting syntactic patterns, and for semantic analysis (Mpouli and Ganascia, 2015), by extracting similes using comparative markers. Our hope is that these approaches will allow us, in the near future, to generate new theories of interpretation inspired by the making and the use of programs operating on "big data," and to open new areas of intellectual investigation in the field of Digital Literary Studies.

In addition to this conclusion and its possible contributions to Digital Literary Studies, our Grand Challenge also seeks to contribute more generally in laying of the epistemological groundwork for the Digital Humanities. To do this, we are drawing not only on technology and on the epistemology of technology, but also on a classical philosophical tradition that likewise attempted, a century ago, to lay the epistemological groundwork for the Humanities. Today, we believe that such an epistemological reflection is both necessary and urgent due to the increasing impact of digitization on all humanistic endeavors.

FUNDING

OBVIL Labex, programme investissement d'avenir.

REFERENCES

Adam, Jean-Michel. (2006). Intertextualité et interdiscours: filiations et contextualisation de concepts hétérogènes. *Revue Tranel (Travaux neuchâtelois de linguistique)* 44: 3–26.

Aiden, Erez, and Michel, Jean-Baptiste. (2013). *Uncharted: Big Data as a Lens on Human Culture*. New York, NY: Riverhead Books, Penguin Group.

Berry, David. (2011). The computational turn: thinking about the digital humanities. *Cult. Mach.* 12. Available at: http://www.culturemachine.net/index.php/cm/article/download/440/470

Bloom, Harold. (1973). *The Anxiety of Influence: A Theory of Poetry*. New-York, NY: Oxford University Press.

Boukhaled, Mohamed Amine, and Ganascia, Jean-Gabriel. (2015). Computational study of stylistics: a clustering-based interestingness measure for extracting relevant syntactic patterns. In *Proceedings of the 16th* International Conference on Intelligent Text Processing and Computational Linguistics, Cairo.

Cassirer, Ernst. (1923). *Substance and Function*. Chicago, IL: Open Court.

Cassirer, Ernst. (1942). *Zur Logik der Kulturwissenschaften*. Vol. 47. Göteborg: Göteborgs Högskolas Årsskrift. [Translated in English under the title *The Logic of the Humanities*. New Haven, CT: Yale University Press (1961)].

Cecire, Natalia. (2011). "Introduction: theory and the virtues of digital humanities," in conversations. *J. Digit. Humanit.* 1:1. Available at: http://journalofdigitalhumanities.org/1-1/introduction-theory-and-the-virtues-of-digital-humanities-by-natalia-cecire/

Compagnon, Antoine. (1979). *La Seconde main ou le travail de la citation*. Paris: Seuil.

Crane, Gregory. (2006). *What Do You Do with a Million Books?*. Vol. 12. D-Lib Magazine. Available at: http://dx.doi.org/10.1045/march2006-crane (accessed October 5, 2015)

Fitzpatrick, Kathleen. (2011). *"The Humanities, Done Digitally" – The Digital Campus 2011 – The Chronicle of Higher Education*. Available at: http://chronicle.com/article/The-Humanities-Done-Digitally/127382/ (accessed May 3, 2015)

Frontini, Francesca, Boukhaled, Mohamed Amine, and Ganascia, Jean-Gabriel. (2015). *"Linguistic Pattern Extraction and Analysis for Classic French Plays"*. Sydney, NSW: Digital Humanities Conference.

Ganascia, Jean-Gabriel, Fenoglio, Irène, and Lebrave, Jean-Louis. (2004). «Manuscrits, genèse et documents numérisés. EDITE: une étude informatisée du travail de l'écrivain». *Document numérique* 8: 91–110. doi:10.3166/dn.8.4.91-110

Ganascia, Jean-Gabriel, Glaudes, Pierre, and Del, Lungo Andrea. (2014). Automatic detection of reuses and citations in literary texts. *Lit. Linguist. Comput.* 29: 412–21. doi:10.1093/llc/fqu020

Genette, Gérard. (1982). *Palimpsestes: La Littérature au second degré*. Paris: Seuil, coll. «Essais».

Gold, Matthew. (2012). *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press.

Grésillon, Almuth. (1994). *Eléments de critique génétique*. Paris: PUF.

Hay, Louis. (2002). *la littérature des écrivains. Etudes de critique génétique*. Paris: Corti.

Jarry, Alfred. (1911). *Gestes et opinions du docteur Faustroll, pataphysicien*. Paris: Fasquelle. Available at: http://fr.wikisource.org/wiki/Gestes_et_opinions_du_docteur_Faustroll/Texte_entier (accessed May 3, 2015)

Jockers, Matthew. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana-Champaign, IL: University of Illinois Press.

Kaplan, Frédéric. (2015). A map for big data research in digital humanities. *Front. Digit. Humanit.* 2:1. doi:10.3389/fdigh.2015.00001

Mayer-Schonberger, Viktor, and Cukier, Kenneth Niel. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Eamon Dolan/Houghton Mifflin Harcourt.

Moretti, Franco. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Mpouli, Suzanne, and Ganascia, Jean-Gabriel. (2015). Broadening the horizons of computational stylistics: an experiment on automatic simile description. In *2015 Conference of the Poetics and Linguistics Association (PALA)*, Canterbury.

Murray-Jones, Patrick. (2011). Theory, digital humanities, and noticing, in conversations. *J. Digit. Humanit.* 1. Available at: http://journalofdigitalhumanities.org/1-1/theory-digital-humanities-and-noticing-by-patrick-murray-john/

Porsdam, Helle. (2011). Too much 'digital', too little 'humanities'? An attempt to explain why many humanities scholars are reluctant converts to digital humanities. In *Arcadia Project*, Cambridge University Laboratory, University of Cambridge. Available at: http://arcadiaproject.lib.cam.ac.uk/docs/DigitalHumanities.pdf (accessed October 23, 2013)

Presner, Todd, and Schnapp, Jeffrey. (2013). *Digital Humanities Manifesto 2.0*. Available at: www.humanitiesblast.com/manifesto/Manifesto_V2.pdf (accessed May 3, 2015)

Price, Kenneth, and Siemens, Ray eds. (2013). *Literary Studies in the Digital Age: An Evolving Anthology*. New York, NY: MLA Commons, Modern Language Association. Available at: https://dlsanthology.commons.mla.org/ (accessed October 5, 2015).

Ramsay, Stephen. (2011). *Reading Machines: Towards an Algorithmic Criticism*. Urbana, IL: University of Illinois Press.

Rickert, Heinrich. (1921). *Kulturwissenschaft und Naturwissenschaft*. 5th ed. Tubingen: J.C.B. Mohr (Paul Siebeck).

Schreibman, Susan, Siemens, Ray, and Unsworth, John eds. (2004). *A Companion to Digital Humanities*. Oxford: Blackwell. Available at: http://www.digitalhumanities.org/companion/ (last access on October 5, 2015)

Siemens, Ray, and Schreibman, Susan eds. (2008). *A Companion to Digital Literary Studies*. Oxford: Blackwell. Available at: http://www.digitalhumanities.org/companionDLS/ (accessed October 5, 2015)

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.