Big Data, Mechanical Turks, and the Excess of Textual Circulation

Didier Girard

▶ **To cite this version:**

Didier Girard. Big Data, Mechanical Turks, and the Excess of Textual Circulation. Crossways Journal, 2017, 1 (1), https://crossways.lib.uoguelph.ca/index.php/crossways. hal-02610402

HAL Id: hal-02610402
https://hal.science/hal-02610402

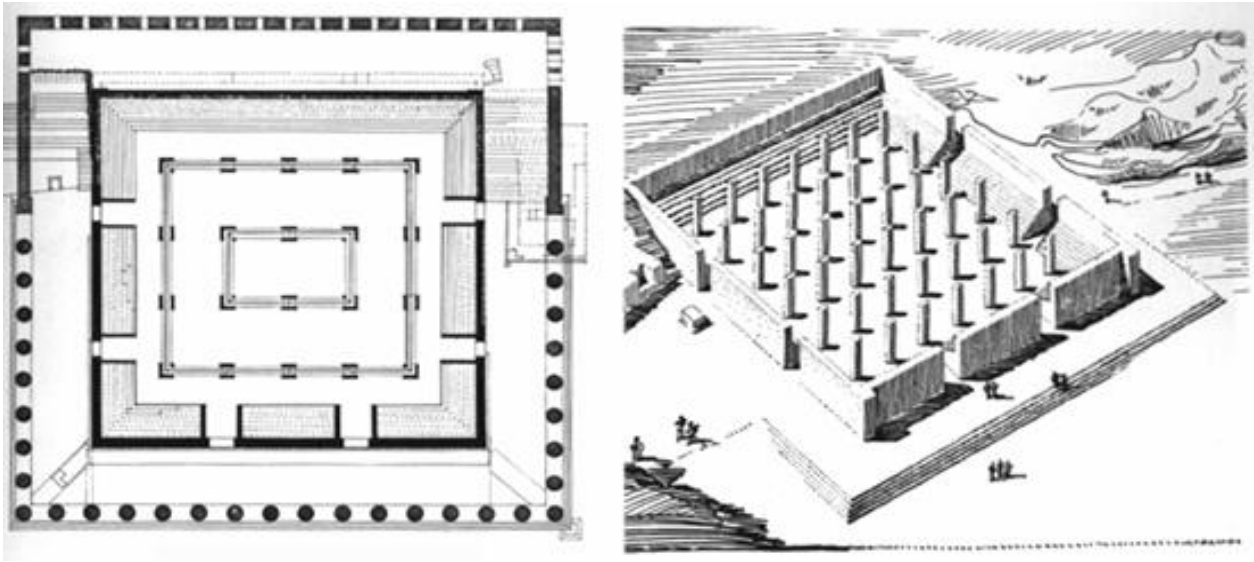
Submitted on 16 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Big Data, Mechanical Turks, and the Excess of Textual Circulation

Didier Girard
François Rabelais University
France



This paper will not exactly deal with literary issues but with Intelligent Technology and cultural sociology, although the very thing known as a book probably remains the *epitome* of what intelligent technology can be at its best: a compact device opening up to unlimited possibilities. Ten years of intense intercultural and transcultural activities across five continents with three post-graduate or doctoral programmes[1] give us today the opportunity to take some critical distance and a broader view on what we have been doing and witnessing in cultural and research practices. After all, we now have the necessary amount of expertise with just over 300 postgraduate students (between 20 and 38 years of age), about 150 co-supervisors around the world with every possible academic background, 2 dozen international universities on every continent except for Asia and Africa, and their countless research centres. So we are in a position to assess, in relatively objective terms, what global textual mobility means in the academic sphere today, and how it might affect academic research in the next twenty years.

Our international students (composed of just under 50 different nationalities, roughly one third European, one fourth South American, one sixth Asian and Indian, just as many from Russia and the West Balkans area, 8% North American and a more limited number of students from Africa, the Middle East or the Pacific) have produced academic papers in 6 different languages, and in just as many fields of study: comparative literature (39%, including 6% in classical literature), cultural anthropology (25%), visual cultures (14%, mostly cinema studies but also the Fine Arts), contemporary philosophy (8%), translation studies and socio-linguistics (7%), and finally “Other”, including musicology in particular (7%).

The mere issue of transporting, carrying, and “crossing” borders with texts and documents, either virtually or physically, is in itself a giant metaphor for what this conference is to tackle; I often have this mental image of young researchers smuggling, peddling, or hawking texts from one part of the world to the next, from one semester to another, and in itself this is going back to the origins of what mædieval universities and *sodalitates litterarie* were, in other words what I would coin as *le colportage de la connaissance*. Linguistically speaking, and given the worldwide structure of the programme and the selection of our candidates, it is remarkable that only 60% of the theses defended were written in English, whereas 18% were written in French, 10% in Spanish, 5% in Portuguese, 4% in German, and 3% in Italian.

Impressive and innovative as they might be for the still prevailing campus-based academic workshop, those statistics should not overshadow the great lesson of global multi-site further education. Academics of all calibres heavily rely on e-technology to carry out their research, find their primary and secondary sources and exchange their thoughts with fellow-students and supervisors alike. The traditional sanctuary

represented by the supervisor’s office, the library, or the conference room is no longer physical, the place of initiation is no longer *here and there* but *every- or anywhere*, a formless and ubiquitous cloud of archives and documents. The fear that this virtual space – all those i-clouds hovering above (why “above”?) our heads – is one of textual evanescence, corruption and degeneration as the call for papers for this conference so elegantly put it. And yet, should we accept and embrace this radical change, this transition to a state where everything might be in a state of “permanent” mutation and reflection?

Telestrion was actually a place in Eleusis, Greece, precisely the place of initiation into what is known as the Eleusian mysteries. Despite the fact that we know that the place could house 1000 participants, it is crucial to notice is that there is NO written text or testimony to describe what happened inside and what the initiation consisted of. Telestrion can alternatively be read as a sword word, a *mot valise*, formed by the chance encounter of tele/telos (the idea of distance and remote controls) and histrion, an *histrio*, a wandering minstrel and something of a showman – as we all live in a post-Debord world in which documents have turned into screens and words into images.

© Dider Girard

© Dider Girard

The *hic et nunc* has most probably become an obsolete notion, lacking in *ubique* and *omnia* and this is precisely the core of the matter. Our students are no exception to the new way of the world: they do cope with extremely elaborate visa procedures, shifting time zones, mutating identities, existential fluxes, and institutional multiversity, but the major academic issue supervisors around the globe have had to face – and struggle with – in the context of *Crossways* and *Interzones,* was questioning their own investigation tools and goals, and not letting themselves be submerged by the accumulation of data. The complexity obviously dwells on the preconceived idea that today everybody can potentially know everything (the most recent estimates say that we have just passed the 50% watershed of digitalization of any written source or document of any period). We also know for sure that, paradoxically, not everybody is able to make use of this apparently unlimited knowledge and circulation of ideas – the main risk being an extremely flimsy and superficial understanding of major texts that are overshadowed by masses of secondary literature that do not really take into consideration (or far worse, that take for granted) the epistemological context of the works under scrutiny. Research today means something vastly different from what it was, let us say, 15 years ago. Pessimists will say that going back to cultural hierarchization is what is crucially and urgently needed (especially in the endangered field of the Humanities), whereas optimists will see Big Data as the most exciting conceptual, political and sociological challenge of the millennium to come. Willingly or not, Big Data is a technological parameter, a piece of *tekhnè* (*τέχνη*) that should be put first on the agenda of every research centre to redefine their priorities, recommended methodologies, and ultimate *raison d’être* in a dematerialized open-source information age.

The European Commission, of all governmental or non-governmental institutions, is very much aware of this key issue. Its branch for Education and Culture has funded a 2-year-long *Big Data Public Private Forum* (*BIG – Big Data Public Private Forum*) through their Seventh Framework Program to engage companies, academics and other stakeholders in discussing Big Data issues. The project aims at defining a strategy in terms of research and innovation to guide supporting actions from the European Commission in the successful implementation of the Big Data economy. In Great Britain, the government announced the founding of the Alan Turing Institute in 2015 (*The Alan Turing Institute*); it was named after the famous WWII computer pioneer, code-breaker and social martyr, and will focus on new ways to collect and analyze large data sets. Projects of the same kind and scope abound everywhere around the globe, as the irresistible cocktail of money, power and thirst for communication is everywhere to be found in Big Data politics. Space is traditionally associated with processes of location and localization, identification and distinction (I am here and you are there); time tends toward bonds and relationships, a way of connecting opposites (I was young and I am old) and a way of picturing the common deployment of phenomena in nature. But as philosopher Elie During makes it clear: “It is because it endures (and not because it stretches out in space) that the universe can keep itself together. But is it still in one piece? What new ways of connecting (or disconnecting) things, what modes of coexistence are appropriate for the articulation of space and time in Space-Time.”

Today, there are 4.6 billion mobile-phone subscriptions worldwide, and up to two billion people accessing the Internet on a daily basis via their smart phones or other electronic devices. If you take into consideration the additional fact that for 25 years, since the advent of e-technology (in the early 90’s), 1.5 billion people around the world have entered the middle class, the growth in digital information exchange is bound to increase exponentially. As a reminder, the world's effective capacity to exchange information through telecommunication networks has soared from 281 petabytes a year in 1986, to 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007, and up to around 700 exabytes in 2014. The zettabyte indeed is looming large, and near.

Multiples of bytes

Decimal			Binary				
Value	Metric		Value	IEC		JEDEC	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte		–
1000^5	PB	petabyte	1024^5	PiB	pebibyte		–
1000^6	EB	exabyte	1024^6	EiB	exbibyte		–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte		–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte		–

Such figures are somewhat cryptic, if not esoteric – but they also do reveal something about us in the twenty-first century. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes of data were created. On the one hand, in academic terms, this means an unprecedented and unfathomable goldmine of archives and documents to consult easily, sometimes free of charge. But on the other, this represents a trap, a black hole, into which postgraduate or doctoral students are only too prone to fall by expanding their primary and secondary sources *ad libitum* – and alas, sometimes *ad nauseam*, turning their academic investigations into fireworks of disconnected comments and remarks. The personal computer industry has indeed paved the way for such an evolution in everyday practice: IBM was reported to be the first company to have built the largest storage array, with a capacity of 120 petabytes, and what we can call the Internet Archive surpassed 15 petabytes as early as 2014.

Moreover, the evolution of Big Data is not just to be measured by size, but also by speed. Today’s researchers, whatever their field of study, need to update or refresh their “data bank” more often and more regularly than in the past as the span of time that separates the writing of an article, the delivery of some speech, or any artistic/cultural happening (the event) and its publication tends to shrink as years, and months, go by. The alertness required from young researchers is another academic issue to be considered

because it either misleads them into a browsing frenzy and into a wider and wider scope of data that takes them astray or far from the initial intellectual target, or results in a relentless postponing of the moment when the urge to give a form to their own thinking is stronger than the passive absorption of ever new data.

The accelerated dissemination of data also leads to the mutability and versatility of information exchanged or gathered. For the doctoral student, the necessity to devote more time to cross references and multiple checking methods becomes vital. A critical (preferably clinical!) turn of mind is more indispensable than ever. In addition, the use of multivariate methods that probe for the latent structure of the data, such as factor analysis and cluster analysis, have proven useful as analytic approaches that go well beyond the bi-variate approaches (cross-tabs) typically employed with smaller data sets. In other, more practical words, these new research parameters tend to develop the schizoid components in the young researcher’s psyche. S/he must now welcome, exploit and let him/herself be inspired by the limitless potentialities offered by today’s I.T. which are great (this paper is NOT suggesting that Intelligent Technology is nipping 21st century academic research in the bud), but also at the same time analyze and be clear-sighted about the intellectual routes taken as s/he moves on. Hence, the need and urgency of sharply monitored supervision mechanisms to challenge and encourage such awareness.[2] Because, having said all that, one must never forget that (according to neurologists) the human brain's ability to store memories is equivalent to about 2.5 petabytes of binary data, a long way from the skyrocketing standards of the zettabyte age we are entering.

In the current race for innovation and hypermodernity at all costs in academic circles across the world, it seems urgent to destroy the myth according to which speed and acceleration preserve youth. When Bergson introduced the philosophical concept of duration at the beginning of the 20th century, he wittily remarked: “We’ll have to find other ways of not aging.” Slowing down and acting up could be tomorrow’s alternative.

What strikes me as much more seminal, though, is the role played by what is known as crowdsourcing in Big Data phenomena because it raises a number of issues that are reshuffling and shaping up our own cultural sociology. Until recent years, in most countries where education and research are supported by State, the added value to any researcher’s expertise, publications, and findings tended to be rewarded by job promotions, patents or other forms of professional recognition which could in turn become profitable in “real life.” With crowdsourcing practices, the line between amateurs and professionals becomes very difficult to draw and this consequently challenges the cultural fabric of our universities.

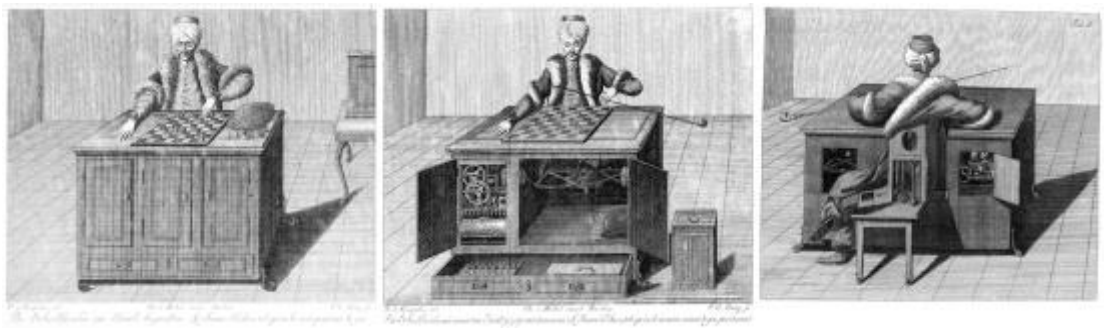
By definition, crowdsourcing combines the efforts of numerous self-identified volunteers or part-time workers, where each contributor, acting on their own initiative, adds a small contribution that combines with those of others to achieve a greater result; hence, it is distinguished from outsourcing in that the work comes from an undefined public, rather than being commissioned from a specific, named group. Jeff Howe and Mark Robinson were the first journalists in 2005/2006 to use the term in *Wired* magazine (Jeff Howe revealingly wrote “distributed labour networks are using the Internet to exploit the spare processing power of millions of human brains” in the 14.06.2006 issue of *Wired*). A couple of years later, Daren C. Brabham was the first academic to publish a scholarly article using the word crowdsourcing, defining it as an "online, distributed problem-solving and production model.” More recently, communication analysts (including a well-known team of researchers from the University of Valencia in Spain) tend to focus on the psychological drives that seem to irresistibly lead Mr or Ms Everybody to volunteer for such collaborations without being paid (or so little paid!), even when the organizer is a perfectly well-known strictly commercial venture or firm.

> Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken (Estellés-Arolas and González-Ladrón-de-Guevara).

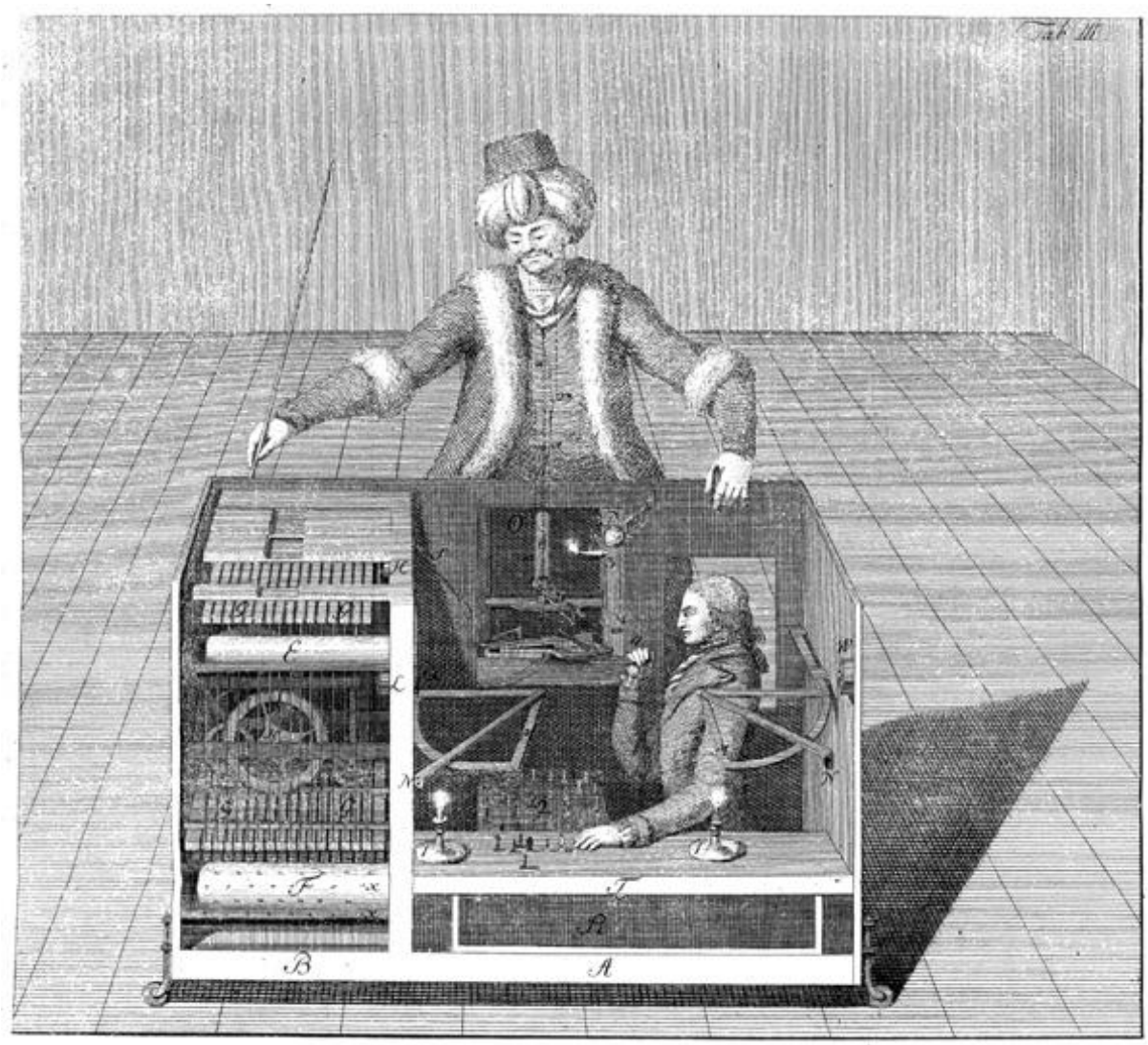
Indeed, the fact that these collaborations are rewarded symbolically or paid for is a crucial issue. Open source information has changed the market forces at play and the professional sectors of education and culture. Customers are ready to pay (or participate financially through subscriptions of all sorts) only if the data bank sought for is exhaustive or of an extraordinary size. Idealists would even object to this by insisting on the idea that crowdsourcing is based on giving the world’s research and results back to the

people, or on the idea that crowdsourcing is a way of sharing the burden of many tasks difficult to carry out by one person. In other words, this would be a democratic version of the *ancien régime* concept of the Republic of Letters initiated by enlightened intellectuals and other Encyclopaedists. Examples abound: Wesay (software to collect linguistic data from all over the world), Wikipedia, the Oxford English Dictionary, weather or astronomy observers, open food, crowdjournalism, photos, medicine, genealogy, politics, design... Crowdsourcing has always existed of course (various prizes and contests, crowdfunded art works or events, charities of all sorts...) but the Internet has made it so that people can participate without being scrutinized physically while doing it, and that could also explain the enthusiasm with which individuals reach out to the rest of the world with observations resulting from hours of solitary investigation via impersonal screens. The real mystery is why they would be so generous with the crowdsourcing institution that sucks their blood and brains (what is known in the trade as a digitally updated version of “piggybacking”) and why they ignore the commercial use crowdsourcers might make of them.

The phrase “Mechanical Turks” has been applied to describe such amateur enthusiasts who are paid very little for providing information to capitalistic companies, but also self-styled fab labs or think tanks which take advantage of the precarious social status of their collaborators allowing them to collect large sets of complex information at minimum cost. The term was invented by Amazon.com who has had its own pools of AMTs since 2005. It is only to be hoped that academics of the next generations will not become fellow-UnivMTs around the world, and that the institutions and politicians understand the real and long-term value of culture and higher education in a global economy! An allegorical digression is maybe needed here to make my point clear: The Mechanical Turk is also a famous automaton built in Vienna in 1770 by Wolfgang von Kempelen (who was also the author of the illustrations reproduced below). Contrary to the popular fallacy that tends to convince viewers that an automaton is “animate” and truer than life, this one was touted as a genuine machine that could play chess at the highest level. It consisted of a large pedestal, housing intricate machinery on top of which stood a chessboard. To this box was attached the upper half of a man dressed in oriental robes and a turban (a rather puzzling aesthetic choice when one thinks about its proclaimed high-tech ambitions since each performance began with a lengthy introduction to prove to the audience that the Turk was a complex piece of sophisticated machinery, not a hybrid dummy). The automaton would then face a challenger. The Mechanical Turk first dazzled the court of the empress Maria Theresa in Vienna. The machine moved its own pieces, and would instantly recognize illegal moves by its opponent. The automaton soon became a sensation, toured Europe (including London at No. 9 Savile-Row, Burlington Gardens) and North America, and was matched against some of the best chess players of the time. It lost some games, but won surprisingly many. The Turk remained popular after its inventor’s death, even playing games against Napoleon Bonaparte and Benjamin Franklin.



The secret of the Mechanical Turk was kept for over 50 years but as early as 1783, Karl Gottlieb von Windisch (1725-1793), a Hungarian German writer and the first publisher of an academic Journal in Eastern Europe, produced a series of letters published as *Briefe über den Schachspieler des Hrn. von Kempelen: nebst drey Kupferstichen die diese berühmte Maschine vorstellen,* and translated as *Inanimate Reason; or a Circumstantial Account of That Astonishing Piece of Mechanism, M. de Kempelen's Chess-Player.* The machine was in fact an elaborate illusion, and contained an ingeniously hidden compartment that housed a human operator. This hidden chess master could observe the position on the chessboard above, and manipulate the movements of the Turk. The original Turk was destroyed in a fire, but some of its parts survived. It was reconstructed in 1984 but a hidden human operator was no longer necessary. The present incarnation of the Turk is truly autonomous; a chess-playing computer guides its moves! The machine is a machine and looks like a machine. The circulation of texts and iconographic documents is not only geographical, or a matter of 2-D translations in a rational Space/Time continuum, but morphs according to trans-historical layers as well. *Rien n’est historique.*



Notes

[1] *Crossways in European Humanities* (2005-2011) Masters’ Course, *Crossways in Cultural Narratives* (2012-2018) Masters’ Course. Both have been coordinated by the University of Perpignan – Via Domitia, http://www.munduscrossways.eu. *Cultural Studies in Literary Interzones* (2010-2017) Joint Doctorate coordinated by the Università degli Studi di Bergamo (Italy), http://www.mundusphd-interzones.eu .

[2] Research labs of all types could use the so-called 3V’s model (Volume, Velocity and Variety), operational in the information industry, and fine-tune it to their own specific purposes. Latest developments now consider an additional set of features, namely Variety (the inconsistency of data has to be regularly assessed) Veracity (the intrinsic quality of data sources) and what I would call the Vortex (connections and correlations between data must be established to extract something useful from a complex ramification of apparently unrelated sources). See Hilbert and López, and Billings.

Works Cited

BIG – Big Data Public Private Forum, big-project.eu. Web.

Billings, Stephen A. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains.* Wiley, 2013.

Brabham, Daren C. “Crowdsourcing as a Model for Problem Solving an Introduction and Cases.” *Convergence: The International Journal of Research into New Media Technologies*, 14.1 (2008): 75-90.

---. *Crowdsourcing.* Massachusetts: The MIT Press, 2013.

During, Élie. ‘’Relativity as an Accelerator of Metaphysics.” *Palais / Le nouveau magazine du Palais de Tokyo*, 1 (2006), reproduced in *LOST (IN LA).* Los Angeles: Public Fiction, 2012.

Estellés-Arolas, Enrique, and Fernando González-Ladrón-de-Guevara. "Towards an Integrated Crowdsourcing Definition." *Journal of Information Science* 38.2 (2012): 189–200.

Hilbert, Marin, and Priscila López. "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* 332.6025 (2011): 60–65.

The Alan Turing Institute, turing.ac.uk. Web.

Wired Magazine, vol. 14, no. 06, 2006.