Revista Latina de Comunicación Social # 070 – Pages 878 to 890
Research | DOI: 10.4185/RLCS-2015-1076en | ISSN 1138-5820 | Year 2015

How to cite this article in bibliographies / References

F López-Cantos (2015): "Communication research using BigData methodology". *Revista Latina de Comunicación Social*, 70, pp. 878 to 890.
http://www.revistalatinacs.org/070/paper/1076/46en.html
DOI: 10.4185/RLCS-2015-1076en

Communication research using *BigData* methodology

F López-Cantos [CV] [ORCID] [GS] Professor in Audiovisual Communication and Publicity. Department of Communication Sciences. Universitat Jaume I of Castellon (Spain) flopez@uji.es

Abstract
[ES] Introducción: Las tecnologías digitales están facilitando nuevas herramientas y metodologías de análisis en el ámbito de la comunicación, al igual que en otras áreas de investigación. En este trabajo abordamos los aspectos relacionados con la investigación en comunicación con la metodología *Big Data*, que está creando muchas expectativas y está dando buenos resultados en otros ámbitos científicos. **Metodología:** Tomamos como corpus de trabajo un conjunto de artículos y utilizamos sobre esa muestra las técnicas de análisis y representación de *Big Data* como metodología de análisis de contenido para valorar su pertinencia y efectividad. **Conclusiones:** Como resultado de nuestro estudio podemos concluir que esta metodología en nuestra área de conocimiento puede ser de utilidad pero tiene una efectividad limitada.

[EN] Introduction: Digital technologies are enabling new tools and analysis methodologies in the field of communication, as well as in other research areas. In this document, we approach aspects related to communication research using the *Big Data* methodology, which is generating much expectation and showing good results in other scientific areas. **Methodology:** A set of articles were used as corpus for this work and we applied the *Big Data* analysis and representation techniques as contents analysis methodology over this sample in order to evaluate its relevance and effectiveness. **Conclusions:** As a result of our study, we can conclude that this methodology can be useful in our field of knowledge, but it has a limited effectiveness though.

Keywords
[ES] Big Data; minería de datos; análisis de contenido; metodología; epistemología; representación científica; investigación en comunicación.

Revista Latina de Comunicación Social # 070 – Pages 878 to 890
Research | DOI: 10.4185/RLCS-2015-1076en | ISSN 1138-5820 | Year 2015

[EN] Big Data; data mining; contents analysis; methodology; epistemology; scientific representation; communication research.

Contents

Translated by **Yuhanny Henares.**

1. Introduction

We are submitted to numbers, data, both regarding adoption of standards that must be mandatorily met and that enable the objectivization and evaluation of our academic achievements in order to determine the course of our professional career, as well as when it comes to undertake our own research projects and promote their publication in national and international impact journals that, of course, must be preferably be based on empiric research and show quantifiable research results. "Numbers, numbers, numbers…", quoted Bruno Latour (2009) as a critic to this contemporary litany, questioning this current obsession for quantification which ends up impregnating every public and private areas as well as the contemporary culture itself.

And in this new contemporary digital environment of immediate and global response and highly interconnected and mediated by mathematical algorithms that impregnate all corners of our quotidian existence generating, as a logical consequence, an avalanche of data accumulating *ad infinitum* at an exponential rhythm. And in this senseless chaotic vortex there appears, as it usually does, a new creature of pure reason that, *mutatis mutandi*, promises to pacify the existing chaos and even contribute with new perspective of knowledge but that, above all and firstly, has turned into a new great business of the digital age, which is listed in NASDAQ with full rights: the *new* technologies of extraction, representation and analysis of *Big Data*.

There is an increasing number of Works that ask themselves about the usefulness of new techniques in different scientific areas (Leonelli, 2014; Taylor, Schroeder and Meyer, 2014), they treat epistemological (Kitchin, 2014; Raghavan, 2014), methodological (Burrows and Savage, 2014; Murthy and Bowman, 2014) or ethical aspects (Zwitter, 2014; Lyon, 2014; Gurevitch, 2014). And some works specifically talk about the impact of its use in the field of communication (Schroeder, 2014; Penney, 2014; D'Heer and Verdegem, 2014; Fischer, 2014), determining the noticeable significance the management of huge data bases using contents analysis techniques is acquiring.

A simple search of the term *Big Data* on *Google* shows more than 765 million links. However, and despite these numbers, it seems that this new research trend in Human and Social Sciences has scarce impact in our area and it seems rather alien for us at the same time we are suffering a progressive demand of *quantification* of our work paradoxically.

A priori, undoubtedly, the appeal of being able to work with *Big Data* is being potentiated with the

new possibilities current digital technologies offer for their management and obtaining of profitable results in a fast and economic manner, and the improvement derived from the analyses of studied phenomena.

Anyway, every new era brings new mythologies with it and hence, we have included in the title of this work, the possibility that a new era opens, the *Big Data Era*, because as we already know and quoting Latour again (2009): "Change the instruments, and you will change the entire social theory that goes with them".

We haven't found a single article on greatest impact journals in our area of knowledge, collected on the IN-RECS 2011 directory, where said term appears; a situation that contrasts, as we mentioned before, with the increase of presence of this area of knowledge in international academic journals on communication.

Therefore, it seems relevant to analyze where these promised improvements get to regarding its application in our field of knowledge, Communication Sciences, and determine to what extent this research trend might be useful as analysis methodology on communication studies [1], and evaluate whether we can talk about a supposedly upcoming *New Big Data Age.*

2. Methodology

In order to carry out this research, the method we use should first meet a set of premises so that it is valid and, undoubtedly, will position us far from orthodoxy and usual practices.

We start from a simple hypothesis: there is a research trend emerging in Human and Social Sciences represented in a group of 117 texts we have found through the search of the term *Big Data* on *Sage Journals*.

We have obviously got this far because at some point in our readings we have found the term, and we have even taken a glance at some article about it in a first approach to our object of study, and we want to perform a more exhaustive research; hence, we carried out an initial search in several data bases, Sage Journal in our study. We could have used other data bases (JCR, Scopus, etc.) but we selected *Sage* only because it ensures enough sample for testing the analysis of BigData analysis techniques and what we try to do, is to get conclusions regarding the validity of application of data mining methodology in our area, not analyze the impact of BigData in the whole of communication researches, an aspect that should be managed in a specific manner for its adequate analysis and, besides, it is not this research's intentions.

In our search, the results are a total of 117 references that encompass our corpus to test BigData methodology. And now, it is when the research method we are going to use differs compared to usual ones.

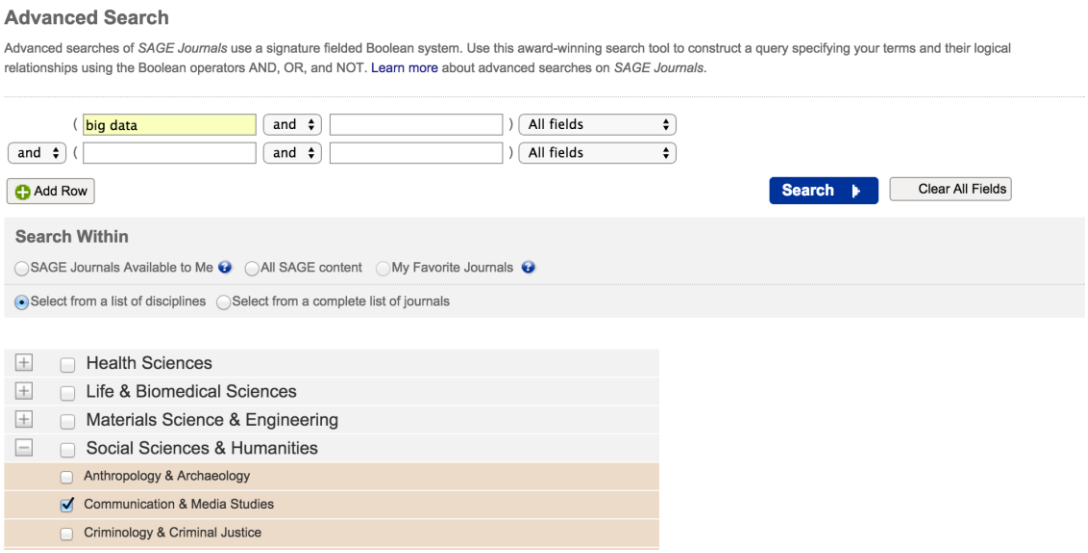


Illustration 1 Search on Sage Journal

If we performed a research using the conventional methodology, the next step would be to select from the list of references, the ones that seemed more adequate and give them a first reading. From this point, we would collect other bibliography from insert in each one of the articles we are reading and, manually or using a text editor, we would make notes for their potential use, in a kind of contents analysis that would allow us to dig deeper and know more about our object of study.

After several readings and re-readings, some discards and extensions, and the consequent new notes and concept updating, we could start elaborating an argumentative discourse where we could introduce our own particular analysis of the corpus that encompasses our object of study. After many drafts and re-elaborations, we would finally get a text that is suitable for its dissemination and that, in an organized manner, we could present our discourse complying with stablished standards and following the style and composition indications of the chosen journal so that, luckily, it could be considered positive for its publication.

The working method we are about to use is similar in some aspects, nevertheless, it is radically different in others. We won't read anything; we won't even read a single text line. Those tasks will be left at the hands of the so promising BigData technology. And we will observe whether expectations are met.

We are going to use a software we have selected among those that are available for free, *QDAMiner*. In order to work with structured data, we might use, *RapidMiner* for example or the much more sophisticated software *R*, with a variety of analyses and graphs possibilities, but our bet is to work on data mining as such, that is, with data without any structure whatsoever in order to see the real possibilities of *Big Data* technologies.

Working with already elaborated data bases offer more analysis possibilities, but in general we find non-structured data when we are about to perform a contents analysis in any research in

communication, for example, for analyzing press, and hence it is more interesting for us to work with non-structured data. This is why we chose *QDAMiner* software.

Let's analyze the functionality of this software, and determine whether all informatics Engineering behind the term *Big Data* is helpful or is not helpful at all in most of research projects we carry out: we could read and analyze one hundred or two hundred texts, but not one hundred thousand, which is what the *Big Data* software promises.

The questions we want to answer are the following:

Firstly, is data extraction by brute force without previous selection and filter from our part, helpful in any way? And given that this usefulness is only partial, can specific techniques be applied in order to improve quality of data? And, to what extent do they involve our intervention anyway?

Secondly, once the data are obtained, with what criteria we consider them adequate for our study?

And finally and once they are validated, what form of presentation will be more adequate for its analysis?

And further, from this first analysis, can we go back from the start and walk over the path again in order to deepen in the research and /or get new perspectives?

In order to answer these questions, we are going to evaluate results obtained in the different stages of the methodological analysis procedure using the BigData techniques, following the guidelines we introduce below.

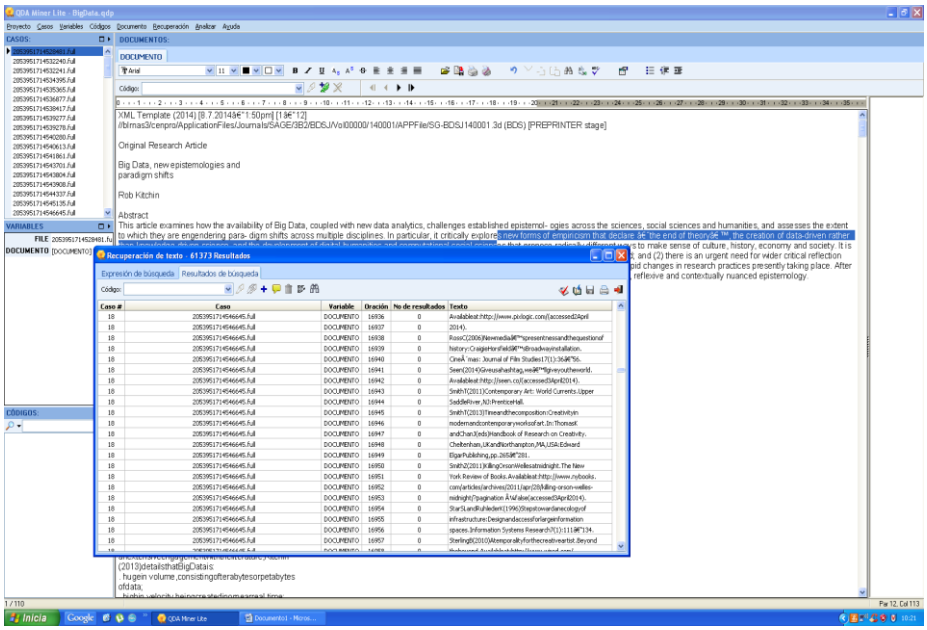
3. Results

Firstly, and once all articles are stored in *pdf* format in our computer, we execute a file import to *QDAMiner*. A total of 110 files are stored, the ones accessible from our subscription to *Sage Journal*, and the import is made on lots and without further complication in less than one minute. A *QDAMiner* project is generated from the folder they are located in, where all original documents are included and transformed into plain text.

Afterwards, and once inside *QDAMiner*, we perform the contents analysis. This stage is rather controversial and we started having trouble and questioning the helpfulness of the software and the promises of the *Big Data Age*. If we carry out a search of raw data in order to segment contents, we can use data recovery at the level of what this software calls *sentence*. On the contrary, if we carry out a search using key words that could be of interest, we find the typical problems of semantic categorization and besides, we will be reading the text necessarily with the bias that would inevitably introduce our particular reading.

We proceed to evaluating the efficacy of the first option, because we have decided to avoid reading and categorization, and explored the text recovery tool. We got a total of 74.894 segments at minimal unit level, which is a sentence, for the total of 110 documents (or *cases* as they are called). We are

unaware about the algorithm the software uses to determine these units and according to results, raw segmentation doesn't seem to be helpful at all.



Therefore, let's further explore more specific searches that have some usefulness. But then the question we immediate formulate is: what text are we looking for? in the end, whatever it is, it will constitute a *Kantian a priori* and it won't provide more results from those we already know previously they could exist and this, undoubtedly, will make us loose the remaining information we might extract and that, at least hypothetically, could be also of some interest.

But there is nothing we can do, it is necessary to structure information in a more efficient manner. So, what to do? Is this everything this software offers and what the *Big Data* technology promises? We understand it isn't, we are only at the first stages of our analyses and really, up until now, we haven't dedicated more than a few minutes of our time for all this. There should be something more, but, what else can we do?

We are using the Lite version, which is free, and there are many software options that are not accessible but, are we really going to need these functions or are they more of the same thing? We reviewed the Help section and, effectively, it is more of the same thing: greater versatility when it comes to crossing data and exploiting already existing data, but we are not going to be able to get any useful information unless we decode the texts manually, at least minimally, through a Thesaurus with its hierarchical structure and correspondences that, likewise we need to create ourselves.

And the question is what use is the software and *Big Data* then? Very little actually, only to computerize a manual process that we can later exploit but that, in the end, we would need to do it as it has always been done: by ourselves. It doesn't seem that there is nothing more than smoke behind

the *Big Data* label because nothing can be done if there is the need to label manually when we face huge volume of data, or can we?

And yes, something more can be done by using a technique that combines manual and automatic work and that consists in codifying one or several texts manually so that afterward, and in an automatized manner, the coding can be extended to the remaining of the corpus. That is, we need to select one or more representative articles, analyze and codify them thoroughly according to our interests in order to explore the rest of the corpus afterwards.

This entails and extra work of reading we wanted to spare and the complex and hard work of elaborating a thesaurus that keeps a minimal semantic coherence and, in any case, the inconvenient aforementioned: selection of representative articles and their categorization, will irremediably introduce bias in our research. But there is no other choice.

The selection of sample documents for their coding can be solved in two manners. Either trying to determine representative texts from a quick look at the articles' titles and summaries, in which case we would need to determine what would the criteria be. Or either by extracting randomly a statistically significant sample. The first option wouldn't correspond to the use of *Big Data* because it will only be valid for a small sized corpus and it would be biased by the Kantian a priori we have been mentioning anyway; and the second one, suffers from the bias of the statistical selection of sample.

Moreover, from this moment, there are endless possibilities because we have to think about the theoretical framework from where we will start elaborating the thesaurus in order to codify the selected articles. And every time we get farther away from our raw data, we introduce a greater bias on data we get and loose more original information on the way.

Keeping this in mind, we carry out a simple coding operation from a selection of fifteen articles from a randomized sample, we track them in a fast title and summary reading and from these, we selected five titles that seemed more adequate. We analyzed the summaries in them and built a small thesaurus using a combination of down-up / up-down technique, that is, we will establish a priori categories according to our research interests and we will establish others from the reading of summaries from that sample of articles.

Selected sample texts to be codified are the following:

- Complementary Social Science? Quali-quantitative experiments in Big Data World.
- Big Data and The Brave New World of Social Media Research.
- Big Data, new epistemologies and paradigm shifts.
- Emerging practices and perspectives on Big Data analysis in economics: Bigger and Better of more the same?
- Theses on the Philosophy of History: The Work of Research in the Age of Digital Searchability and Distributability.

In order to continue with the analysis of their summaries, we previously established the next categories in our brief thesaurus, which we will be building *ad hoc*. We are interested about finding

out everything related to the *Big Data working methodologies*, the *application to communication research*, everything that might affect the current *communication theories*, texts related to *epistemology* and, finally, texts talking about *ethical issues*.

From these and the analysis of sample texts, we will create codes with the significant terms found and analyze searches in the corpus in order to extend our analysis to the totality of documents talking about different aspects regarding *Big Data*. For example, for our purposes we will include all searches of terms such as *methodology, empirical, ethics or theory*, in English because there is not a single journal in other language within the corpus.

We performed the search of these terms and coded the segments in which they appear in their corresponding categories. What we obtain from all this, is the apparition frequency table that follows and that we exported to an Excel file:

Category	Code	Description	Count	% Codes	Cases	% CASES
metodologías	methodology		23	3,70%	19	17,30%
teorías	theory		213	34,60%	64	58,20%
ética	ethics		78	12,70%	29	26,40%
aplicaciones	empirical		108	17,50%	42	38,20%
BigData	Big Data		194	31,50%	62	56,40%

Obviously, this list could be much longer, but it is not necessary for our purposes and works as a sample.

Everything we've got after applying all possibilities that the data mining software offers, was a frequency table of apparitions. For its elaboration, we could establish filters, variables during searches, crossing data and even generate multidimensional tables using sophisticated algorithms. Our table could have all the complexity we would want however, in the end, it wouldn't be more than that, a frequency table of apparitions.

Finally, and with results obtained using all methodological tools offered by Big Data for contents analysis, which are rather poor and quite biased, what we can do is elaborating graphic representations of data that allow analysis in a simple manner by researchers or, for its presentation to general public.

At a greater level of concretion, we could even further develop our texts analysis and deepen on different theories that appear in our corpus of study for example:

Category	Code	Description	count	% Codes	Cases	% CASES
teorías	theory		213	25,40%	64	58,20%
teorías	semiotics		5	0,60%	4	3,60%
teorías	cultural studies		28	3,30%	17	15,50%
teorías	meme		191	22,70%	5	4,50%
BigData	Big Data		194	23,10%	62	56,40%

Anyway and as we can see, in order to get minimum results, even if they can become more sophisticated graphically and even appealing, our intervention is needed and, despite it being more or less efficient with small sets of texts, as corpus increases in size and complexity, it stops being appealing and there imposes a bias which makes results obtained invalid.

Although in this graphic, the frequency of apparition of word meme is 191, this value obviously only indicates that; that the word appears a certain number of times, but there is no analysis of the theoretical aspects associated to *Big Data* in the texts of our corpus from this epistemological framework, whereas there is no deep analyses made in only five of them, while there is an analysis from the semiotic perspective in four of them even though the apparition of the term is significantly lower. But in order to get these results, we need to read all texts, a reading that is impossible to do when we face a series of texts that exceed some tens.

4. Conclusions and discussion

In this sense and finally, methodologies and technologies this *new Big Data Age* offers are not innovative at all and they are usual for a long time in the field of documental sciences and biblioteconomy since the computerized technologies have been available for text management and the building of data bases. Nothing new, only faster computers and distributed computing, but this is not enough for a change of paradigm in research.

With the term Big Data, what is really promoted are technologies that increase the storing, recovery and data visualization capacity that enable facing huge amounts of information faster and at a reasonable cost. A huge rapidity is promised in data analysis and, specially, the increase of synthesis capacity with the use of graphic tools that show results in a fast and visually attractive manner. But this is not innovative at all, only that, historically, the data collecting and analysis, together with visual presentation of analysis results was a burdensome task that entailed much resources and time, and today it is done in a more efficient manner and there are more spectacular graphic results as well, but there is no research paradigm being developed at all.

The problem these improvement promises generate is that they are still based on a neopositivism impregnated by outdated attributes associated to objectivity. And this new mythology that is building around the *Big Data* term and the supposed improvements it introduces in the generation and progress of knowledge seems to forget the solid critics the neopositivism received from Kuhn (1962), with its proposal of relativism of scientific paradigms, or the most radical Feyerabend (1975), with its epistemological anarchism, among others.

It is very likely that the dazzle the supposedly *New Age* produces, is promoted by the graphic potency of the tools used, which enables elaborating spectacular static or dynamic visuals that multiply, at several levels, the penetration capacity of this new mythology in all ambits. The applications that computing distributed for the fast management of structured data, for example, in medical and pharmacological research, economic or network analyses, among many other fields, are quite noticeable, and however the problems inherent to the representation of knowledge have not yet disappeared.

The difficulties of setting boundaries to semantic fields are well known, as well as the impossibility to avoid biases in the selection of contents since they are inherent to a verbal language for the elaboration of data bases in the fields of Documentation Sciences, see López Yepes (1995) for example.

Regarding the audiovisual language, the limitations in data representations have already been questioned in the renown work of Anscombe (1973), but the current problem is that the new promises go beyond the sole improvement of the efficacy in representation and they intend to be the culmination of an old dream those navigators-geographers that promoted cartography of new territories started to gestate since maps started to be printed (Ford, 1992: 131). And in the last century, there has been a transit from the illustration of the worldly to the representation of the intangible, with all the problems related thereto (Lynch, 2006; self-reference 2013), in what

Heidegger called a few decades ago *The Era of the World Picture* where the final purpose of modernity ended with the conquering of the world as an image (cfr. Gross, 2006), and which implications and limitations have been explored deeply and with critical analysis for the last decades, as in Manovich (2002) for example, among others.

Therefore, it is true, and to conclude, that the promise of an improvement in the representation of knowledge, of even a change of paradigm in this supposedly *New Big Data Era*, is nothing more but another version of the already known contemporary techno-neopositivist mythology that has been gestating over the last decades around Silicon Valley (Gregg, M, 2015), which is now promoted with the invaluable impulse of the spectacular representation of data that the current computing capacities allow today.

In any case, techno-economic interests promoting the supposedly new *Big Data Era* sums up to the questionable proposals the contemporary digital culture brings, together with the questionable imposition of quantification in all fields with new false old promises that only nurture ancestral mythologies.

Regarding the field of Communications Sciences, we understand that a data mining software such as the one evaluated in this document or similar, can be somewhat helpful to enable contents analysis in bibliographic reviews such as this one and, in general, in empiric researches where the analysis of the corpus of textual data with a considerable size is needed. But it is no panacea at all and we must keep in mind that the unavoidable biases introduced through the whole process become greater in relation to the size of the corpus analyzed, and despite the carefulness in methodological issues, described data obtained are supra necessarily "multifaceted and multitrued" (Focault, B. & Meirelles, I, 2015).

5. Note

1 For a compilation about different analyses techniques and methodologies in communication please consult Vilches, L. (Coord.) (2011).

6. Bibliography

FJ Anscombe (1973): "Graphs in Statistical Analysis", in *American Statistician*, 27, pp. 17-21.

R Burrows and M. Savage (2014): "After the crisis? Big Data and the methodological challenges of empirical sociology", in *Big Data & Society*, April–June, pp. 1–6.

E D'heer and P. Verdegem (2014): "Conversations about the elections on Twitter: Towards a structural understanding of Twitter's relation with the political and the media field", in *European Journal of Communication*, Vol. 29(6), pp. 720–734.

PK Feyerabend (1975): *Tratado contra el método. Esquema de una teoría anarquista del conocimiento.* Madrid: Tecnos, ed. 1981.

E Fisher (2015): "'You Media': audiencing as marketing in social media", in *Media, Culture & Society*, Vol. 37(1) 50–67.

B Focault & I Meirelles. (2015): "Visualizing Computational Social Science: The Multiple Lives of a Complex Image", in *Science Communication*, Vol. 37(1), pp. 34-58.

BJ Ford (1992): *Images of Science: A History of Scientific Illustration*. London: British Library.

Gregg, M. (2015): "Inside the Data Spectacle", in *Television & New Media*, Vol. 16(1), pp. 37–51.

A Gross (2006): "The Verbal and the Visual in Science: A Heideggerian Perspective", in *Science in Context*, 19 (4), pp. 443-474.

L Gurevitch (2014): "Google Warming: Google Earth as eco-machinima", en *Convergence*, Vol. 20(1), pp. 85–107

B Kitchin (2014): "Big Data, new epistemologies and paradigm shifts", in *Big Data & Society*, April–June, pp. 1–12.

TS Kuhn (1962): *The Structure of Scientific Revolutions*. Chicago: Chicago University Press

B Latour (2009): "Tarde's idea of quantification", en *The Social after Gabriel Tarde: Debates and Assessments*. London: ed. M. Candea, Routledge, pp. 145–162.

S Leonelli (2014): "What difference does quantity make? On the epistemology of Big Data in biology, in *Big Data & Society*, April–June pp. 1–11.

M Lynch (2006): "The Production of Scientific Images. Vision and Re-Vision in the History, Philosophy, and Sociology of Science", *Visual Cultures of Science: rethinking representational practices in knowledge building and science communication*. Hanover, N.H.: Dartmouth Collegue Press., p. 26 and ss.

D Lyon (2014): "Surveillance, Snowden, and Big Data: Capacities, consequences, critique", en Big *Data & Society*, July–December, pp. 1–13.

L Manovich (2002): "The Anti-Sublime Ideal in Data Art", disponible en http://manovich.net/index.php/projects/data-visualisation-as-new-abstraction-and-anti-sublime.

D Murthy y S Bowman (2014): "Big Data solutions on a small scale: Evaluating accessible high-performance computing for social research", in *Big Data & Society*, July–December, pp. 1–12.

J Penney (2014): "Motivations for participating in 'viral politics': A qualitative case study of Twitter users and the 2012 US presidential election", in *Convergence*.

P Raghavan (2014): "It's time to scale the science in the social sciences", in *Big Data & Society*, April–June, pp. 1–4.

R Schroede. (2014): "Big Data and the brave new world of social media research", in *Big Data & Society*, July–December, pp. 1–11.

L Taylor; R Schroeder y E Meyer (2014): "Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?", in *Big Data & Society*, July–December, pp. 1–10.

L Vilches (coord.) (2011): *La investigación en comunicación. Métodos y técnicas en la era digital.* Barcelona: Gedisa.

A Zwitter (2014): "Big Data ethics", in *Big Data & Society*, July–December, pp. 1–6.

How to cite this article in bibliographies / References

F López-Cantos (2015): "Communication research using BigData methodology". *Revista Latina de Comunicación Social*, 70, pp. 878 to 890.
http://www.revistalatinacs.org/070/paper/1076/46en.html
DOI: 10.4185/RLCS-2015-1076en

Article received on 20 October 2015. Accepted on 27 November.
Published on 21 December 2015