Quis Monet Ipsos Monitores? Motivations, methodological issues and techniques for monitoring the controversy on surveillance as a topic in on-line scraped textual data

Alberto CAMMOZZO[a] and Andrea LORENZET[1*b]

[a] Observa Science in Society; [b] University of Padua and Observa Science in Society

This paper illustrates an automated computer program aimed at monitoring surveillance as a public technoscientific controversy within a big data perspective. The program collects on-line articles from selected newspapers, cleans them up, indexes, classifies and presents them according to a relevance criteria. This empirical work presents manifold challenges: (1) assessing the reasons for building such a tool; (2) defining some of the most relevant features of the selected theme; (3) tracing a methodologically sound path for the classification processes measuring the pertinence of a single newspaper article to the whole theme; (4) choosing sources; (5) selecting the textual analysis techniques and eventually building or assembling the necessary tools. Besides (a) describing the effort behind the production of the tool, this work (b) probes the methodological and technical issues of automated textual analysis of large (~700.000) items.

***Keywords**: surveillance; big data; media monitoring; public controversies*

Introduction

The current debate on surveillance is raging on the press following Wikileaks and Snowden-Datagate revelations, bringing back the central role of the press in monitoring government activities, as

1

*Corresponding author: **Alberto Cammozzo** | e-mail: ac+sts@zeromx.net

well as the interest of social scientists in interpreting these phenomena. But how does this monitoring happen? Does the word 'surveillance' mean the same for everybody? How does this theme get articulated in topics? Is there some topic that is covered more or less as time passes? Which are the most relevant keywords associated to the public representation of such controversy?
In order to answer to these questions, we are interested here in describing how sociologists and computer scientists can collaborate together in the attempt to create a materialization of what Bruno Latour calls *oligoptica* (2005, p. 181), sites seeing *very little* in Latour's words; that is, seeing a rather small portion of social reality, but seeing it very well, projected and trained as they are to observe in a very accurate and detailed way what they are looking at.
When describing oligoptica, the main interest of Latour has been the attempt to describe some of the elements sustaining the 'sociotechnical', both as a relevant concept for doing social research avoiding to refer to older dichotomies (such as humans/machine, science/technology, and nature/culture), and at the same time as a programmatic academic sound research path. To contribute to this latter effort, we thought that it was a good idea to reflect on how to create by ourselves an homemade oligopticon software, a program for observing the mass media, and in particular what digital online versions of newspapers do in relation to a specific issue, a public controversy (Venturini, 2010, 2012; Lorenzet, 2013) that is rather visible and present in the latter times, the issue of 'surveillance'.
While this does not mean to evaluate and monitor the actual controllers – being them public institutions or private companies - for sure by looking at the media by using text mining measures, we can understand and trace synthetic features regarding the spreading of a public discussion or controversy on the topic of surveillance, and in any case check how the media – one of the most powerful institutions in our society - cover the issue. At the same time, we are interested here in looking at relevant concepts and to the topics which are associated to the issue of surveillance in the social science literature, in order to understand connections and processes regarding the relevant issues identified by scholars

with media attention and public understandings of these phenomena.

In describing the reasons motivating us in choosing this topic and what for us this topic means, it is interesting to reflect here a little bit more on the concept of oligopticon software machine for several reasons, trying to critically connect this concept as Latour does, to the more traditional idea of *panopticon*, developed by another French scholar, Michel Foucault (Foucault, 1975).

The first reason is related to the collective idea, prejudice, and paranoia on the same topic of surveillance, as a potentially dark and ultimate political output of knowledge and digital societies. When speaking about surveillance there's obviously a dystopian strike and spin that we need to take into account in serious terms, not as the expression of a technophobia, but as one of the elements of the actual organization of public sphere at several levels, being them a local debate on the installation of CCTV systems in urban areas or the discussions on RFID and scanner systems at international and migration offices, affecting the crowds of travellers daily using airport services, or also the disrupting impact of technological innovations as Google Glass video-recording devices (Adey, 2003; Boyle and Haggerty, 2011; Fonio, 2011; Haggerty and Ericson, 2000; Lyon, 2003).

The second reason is to look at the development of another phenomenon, and precisely to how common users (being them individuals, companies or other organizations), are interacting with control devices in order not only to look and investigate at what others do, but to organize and manage their ordinary daily life. This is the function of the development and spread of data stored in digital and web companies that are managing, accessing and keeping for example email data, and it is also an issue regarding who has the right to own and use private data and for what reasons. In terms of public icons and myths, the outcome of this kind of discussion has been the development and growing interest on data journalism, and its most radical debate on issues such as the "Snowden affair" and the discussion on Julian Assange's Wikileaks (Ball and Wood, 2013; Landau, 2014, 2013).

As interested human beings in the fields of social and computer sciences, we want also to be part of this collective being formed on surveillance, and we are interested here in reflecting on the

opportunity to use web information within the frame of so-called *big data* research (Mayer-Schönberger and Cukier, 2013) to analyse and monitor what some digital media do in relation to the topic of surveillance. Our interest is not in becoming the controllers of the controlling, as our provocative question-title may be suggesting, but rather to understand how a specific public controversy, an issue such as surveillance, can be monitored in order to gain information about its spreading and diffusion in the public sphere.

This kind of work corresponds to the idea of needing a sound and reliable methodology for building social science oligoptica, an objective that was recognized by Bruno Latour himself when he coordinated and developed the EU-funded project Macospol (Mapping Controversies in Science for Politics), involving research teams based on the collaboration between social scientists and computer scientists focused on both the review and the *ad hoc* realization of digital tools to map and visualize data about public Technoscientific controversies, that is public debates in which science and technology have a relevant constitutive dimension within heterogeneous social settings (Latour, Camacho-Hubner and November, 2010; Beck and Kropp, 2011; Venturini, 2008; Latour, 2011; Mélard, 2009; Lorenzet, 2011, 2013; Yaneva, 2012).

For us this work is part of the effort in providing methodologies and techniques for this kind of study, the Mapping of Technoscientific Controversies, and at the same time corresponds to the need of understanding how debates relate to Public Communication of Science and Technology processes, that is how they are related to the generation of a technoscientific public sphere, and thus in the interactions between mass media, public opinion, and policy regulations, as suggested by Bauer in the Mapping the Cultural Authority of Science project (Bauer *et al.*, 2011; Bauer and Gaskell, 2002; Bauer *et al.* 2007).

Moreover, our effort is part of the development of digital methods, that seek to move in the analysis of web contents beyond the analysis of Internet cultures, and thus to use the Web as a repository of information in order to study not much media use, but instead to study relevant aspects of society itself thanks to information available on digital media (Rogers, 2013; Marres and Weltevrede, 2013).

Finally, the work here presented can be seen as one of the outcomes of the Science in the Media Monitoring project at

Observa Science in Society, a multidisciplinary team working on the analysis of newspapers and digital media coverage of S&T issues since 2007 (Neresini and Lorenzet, 2011, 2012, 2013; Giardullo and Lorenzet, 2013)

Collecting the press discourse

This section describes the method of acquisition, processing and archiving of newspaper articles, along with the means for their classification and presentation.
The Science in the Media Monitor (SMM) project is a collection of programs articulated in four modular steps (see Figure).

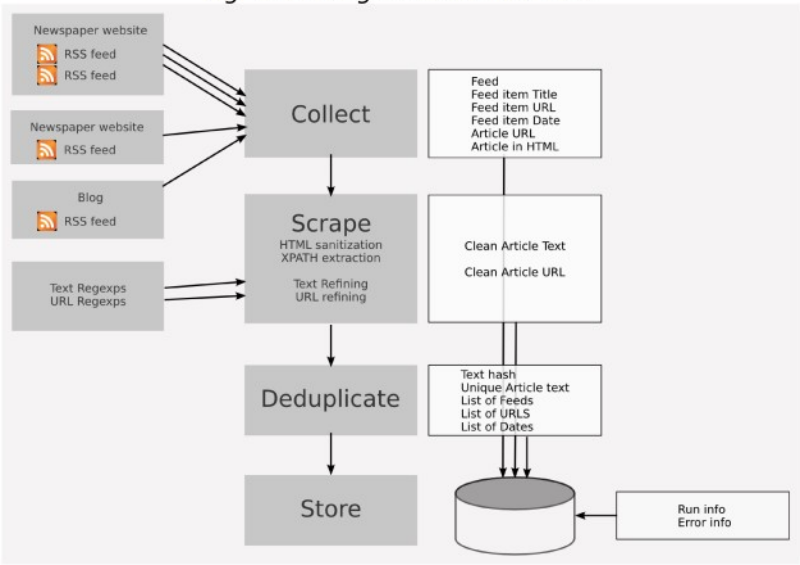
1) source selection: SMM has monitored since 2007 six newspapers, covering the most of Italian national daily paper readers (excluding those that offer only a regional coverage). For each of these, the a mix of automatic and manual methods have been used to identify their active RSS feeds of national interest from their online editions;

2) corpus building: The RSS feed items are collected and stored. Each item points to a web page through its URL, which is retrieved in HTML format. The page is then sanitized and cleaned-up to avoid processing errors. Significant textual data (the body of the article) is automatically extracted from the page ("scraped") with XPATH extraction and regular expressions discarding non-relevant parts as advertisings, fixed parts of newspaper page, images, scripts, etc. Each article is then de-duplicated, that is checked for its uniqueness, trying to spot duplicate items published at the same or different URLS (articles with the same content may be published at different URLs) or from different RSS feeds (same content, different feed); to identify updates of the same news article over time and keeping the more recent one (same URL, different date); and to spot articles that are shared between different newspapers. Following the de-duplication check, only "unique" articles are stored. Alongside the process, metadata is collected, providing information on the process itself;

3) data and metadata management and classification: each stage uses a no-SQL MongoDB database to store its data. The final corpus consisting in unique (de-duplicated) newspaper article text data is stored in yearly collections, along with the metadata regarding the process of its collection. Reporting modules build a daily report informing SMM staff on the harvesting process metrics. Corpora collections are then indexed using an Apache Foundation Solr platform. After indexing, scores are calculated according to classifiers. A classifier is a list of terms (a lexicon) where each term has a value weighting the importance of that term in the semantic field of the given theme. For each document and each classifier a score is calculated as the sum of the values of each term in the lexicon present in the document. Some classifiers may have special terms, called multipliers, used to enhance the value of other co-occurring terms whose meaning may be too broad or polysemic for the topic;

4) data analysis, presentation and query interface. A web interface allows to peruse documents sorted according to their classifiers score, or to search them by content, origin, type of source, time span, etc. Articles whose score is above a given threshold (named the *salience* threshold) are considered “relevant” for the given theme according to one of the classifiers. The use of multiple classifiers with different lexicons and weights is necessary to fine-tune the classification process and to test better classifiers that fit the evolution of the semantic field of the theme. Building and testing the lexicon is a delicate process described in the next section. Only relevant articles are further classified in four relevance classes (low, medium, high, very high) following their quartile distribution: articles in the top 25% rank are considered very highly relevant. The graphic interface shows also the “daily salience trend” in time, an indicator of the relevance of the documents collected in each day for each source type (newspapers, blogs). A demo of the SMM system is accessible from this URL:
http://www.observa.it/science-in-the-media-monitor/?lang=en.

Figure 1: Program data workflow



Construction of the lexicon for the surveillance topic

In order to understand how a lexicon can be useful in order to categorize the press articles related to a specific theme, in our case surveillance, it is useful to reflect on the relationship between events, media discourses, and the selection of specific keywords to be inserted in the lexicon. After that we will describe some techniques that can be used in order to give weights to keywords inside a classifier's lexicon.
We devised a three-layered model for the analysis of the relationship between public issues and a keyword classifier (see

Figure 2); the three levels of the model correspond to the development of the conceptual path that allows us to go from facts concerning a theme, to media items covering a public issue, and finally come to a lexicon, that is a series of keywords that we can use in order to detect the relevant newspaper articles covering selected topics.

The upper level of our model corresponds to the “pragmatic plane of events”, where we have several issues occurring in the real world. These events are basically facts that can be taken up by the media following the mechanisms and processes that are part of the “agenda setting” phenomenon, according to which the mass media have the specific function of selecting and deciding the hierarchy of relevance of real world events and thus impacting in the long run on readers’ reality perception (McCombs and Shaw, 1972).

The middle plane of our conceptualization corresponds exactly to the outcome of the agenda setting. Around the selected events the mass media generate not only a hierarchy of relevance, but also specific discourses. Media discourses are here intended as narratives that frame events and give to them some specific meanings and not others that resonate and can be understood by the public, becoming thus significant for most of them. Both quantitative coverage and the features of media discourses may vary depending on the specific media source, so what is relevant in one media arena (i.e. blogs), ma be not in another (i.e. newspapers), and vice versa; at the same time some issue may generate similar discourses and framing in different sources.

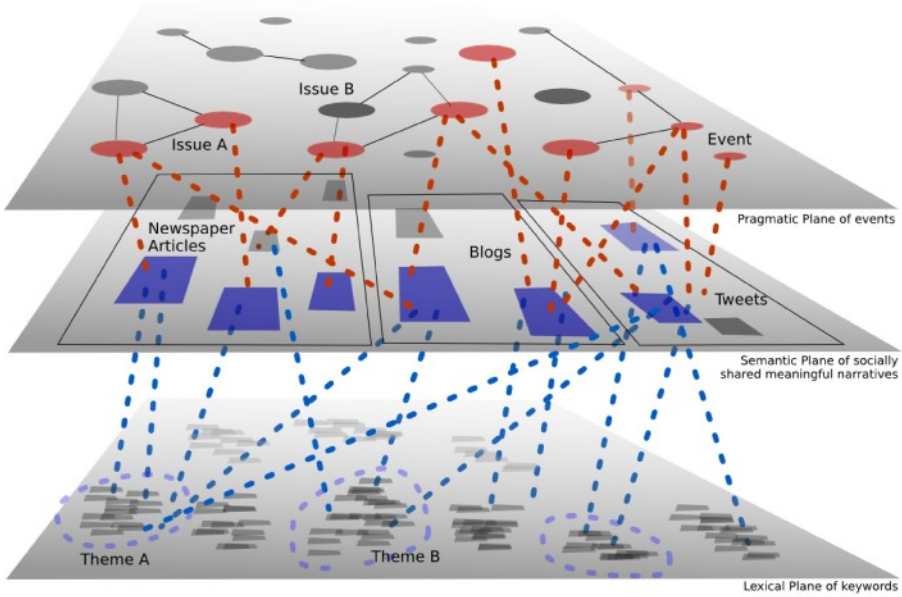


Figure 2 – Our three layered conceptual model for thesaurus construction

The lower part of our scheme is representing the attempt of our analysis to condensate issues reported by the media in a series of keywords, a lexicon, that is able to describe the articulation of the issue in the different medias, attempting to be as *neutral* as possible, that is avoiding noise and biases of sorts. Our aim is thus to build around each theme a classification list of these terms acting as a content selector: on the basis of this list of keywords we can obtain a selection of the relevant mass media coverage, and at the same time describe the relevant features of the discourses and framings occurring in relation to those events. Attempting to build a method for constructing an unbiased lexicon to capture the media discourse on a given topic means that the keywords composing the lexicon are chosen: 1) in a way that is as much as possible reproducible, 2) relying on a broad base of common

understanding of the topic, and 3) in a way that can be tested upon.

To reach this objective for the “surveillance” theme, we first devised two techniques to create a lexicon for categorizing and selecting contents: first automatic topic detection applied to a selection of academic articles and second a “snowball” method applied on newspaper articles selected from a keyword search of “core” terms for the topic; both techniques are based on the analysis of a sample of articles that are somehow strongly related to the issue and constitute in different ways a “ground truth” archive.

Automatic Topic Modeling

The first technique is based on the application of the LDA (Latent Dirichelet Allocation) algorithm for automated topic modeling (Blei *et al.*, 2003; Blei and McAuliffe, 2007).
The aim is to explore thoroughly the semantic field of the theme, in this case “surveillance” issues, relying on textual sources devoted to that theme, assuming that those sources will cover all the possible topics that articulate that matter or theme, at least for “experts” from various disciplines. The second assumption is that an automated topic detection will uncover *all* these topics and present their specific keywords. On the on hand we can expect that topics covered in an interdisciplinary review of academic literature on the theme of “surveillance issues” will be as broad as possible, ranging from CCTV cameras acceptance to government surveillance. On the other hand we must acknowledge that the process of selection of the academic articles remains biased by an “academic” understanding of the research field and the knowledge of literature sources of the reviewer.
Topic modeling algorithms are statistical methods allowing to automatically detect the most significant *topics* within a given set of documents. During the last decade several topic modeling algorithms have been proposed, differing mainly on their assumptions (for example on the basis of the relationships among the topics to be extracted). The method we used for this analysis is based on the algorithm *Latent Dirichelet Allocation*, on the basis of which we find the assumption that documents are characterized by

a given set of *topics*, where a *topic* is defined as a distribution on a fixed set of words: for example within the topic "biomedical research and stem cells", the words regarding biomedical research and stem cells will be present with a high probability. Topics manifest within documents in different proportions: to do the analysis here described, we used the open source software called "Mallet" (Mccallum, 2002), allowing to apply LDA to a set of documents, specifying the number of topics to be extracted and the number of keywords to be visualized (the sum of which can be considered those that best represent each topic).

Regarding the first method we built the corpus for Topic Detection with a two-pronged approach: on the one hand we retrieved 390 articles from 2002 to 2013 published in the journal "Surveillance & Society", which is an interdisciplinary peer-reviewed and open access journal devoted to surveillance studies (Lyon, 2002; Marx and Muschert, 2007). On the other hand we built a review of special issues on surveillance in peer-reviewed academic journals other than "Surveillance and Society". The initial core of articles found with traditional scientific literature exploration methods has been complemented with a crowdsourced approach, publishing the provisional review in two specialized mailing lists (*surveillance listserv* and *liberationtech*), and asking for integrations. The resulting list (Cammozzo, 2013) consists of items from 27 journal special issues from a broad disciplinary spectrum, of which 67 were suitable for download and automated text extraction.

Table 1 shows a sample of LDA output from the Mallet program trained on versions of both corpora modified with Treetagger (Schmid, 1995) to include only nouns. Please note that the topic title in square brackets has been added by the authors and does not come from the program output, that consists only in the first 10 keywords identifying a topic.

Topics detected from Surveillance & Society articles

Topic number and [title]	Topic keywords
1. [Surveillance Studies]	surveillance study technology practice form resistance issue relation work lyon
2. [health & medicine]	health hiv surveillance disease individual sex testing medicine practice population
3. [children]	child parent school teacher family care mother home risk mobility
4. [military]	state police intelligence surveillance citizen war security control germany year
5. [risk]	power part order time risk case fact sense question person
6. [prisons]	home space people city wall prisoner street water prison toilet
7. [gender issues]	woman body discourse identity man gender narrative practice violence experience
8. [data protection law]	privacy protection information law datum data court case individual act
9. [data protection systems]	system datum technology information data project device network design collection
10. [government policies]	policy government crime process approach control strategy issue network agency
11. [privacy and drugs]	drug study research hair cent survey result community level sample
12. [identity cards]	system identification card identity technology database individual recognition biometrics id
13. [Internet]	individual life people information user internet site world network participant
14. [urban security]	space city security centre control system operator mall shopping room
15. [surveillance studies]	medium student event university animal time article interview paper research
16. [school]	year people area school community girl group time offender life
17. [video surveillance]	camera image surveillance space film technology television art work video
18. [consumer data protection]	consumer information datum market user company marketing advertising form product
19. [book reviews]	book work world author text chapter analysis reader page argument
20. [communication]	information government service internet communication law response phone agency canada
21. [Foucault theory]	power foucault control society life body subject form individual space
22. [workplace]	organization employee information staff company monitoring management patient control work
23. [immigration control]	security border airport mobility state risk people movement immigration passenger
24. [crime control]	police camera crime system surveillance officer evidence safety driver area
25. [government]	welfare people agency service program state population case government technology

Table 1 – topics detected with Mallet from 390 articles retrieved from Surveillance & Society journal

Some topics are common to both sources: data protection laws, Foucaultian theory, consumer data exploitation and protection, video surveillance, crime control, immigration control, data protection law.
Other topics are specific to one source or the other: surveillance in schools and child security, surveillance and gender issues, healthcare, drugs and patient privacy, surveillance in sport events, surveillance and identification technologies.
Taken together, the topics from both sources illustrate the way the research debate around surveillance is articulated. An ideally unbiased lexicon should take into account keywords coming from all these topics in order to cover the "surveillance" theme.
We have found it useful to make multiple runs changing the number of iterations and the number of topics, as LDA results greatly depend on program parameters: the results shown in the tables should be taken as an example of a larger set of outputs.
Even if topic detection is not intended to directly provide us with keywords for the lexicon but rather to map the semantic field of the "surveillance" theme, some of the keywords from the topic analysis were used to enrich the classifier.

Seeding from the press discourse

Regarding the second point, that is the "snowball" method, in order to analyse the press discourse we collected a sample of data from news search engine Google News, by using three *seed* keywords – "surveillance", "privacy", and "data".
From this operation, we obtained a list of articles that we analysed with text mining software Rapidminer (Mierswa /et al./, 2006), obtaining the list of key terms with the highest score in terms of the measure tf-idf (term frequency-inverse document frequency) (see Table 2).

From this list, a manual selection of the keywords relating to the topic of surveillance has been made, and for each we obtained three different metrics, that is: the keyword frequency within the whole corpus; the IDF, measured as the logarithm of the total number of articles and frequency ratio, and the keywords salience,

that is the percent of articles on the total number of articles that include at least one time the keyword.

The evaluation of the values of these measures allows us to define a scale of values for the keywords, thus weighting them and at the same time to assess their presence within the corpus.

Keywords obtained with "snowball" method

Keyterms	Frequency	(1/log (tot.articles/ frequency))	Keyterms salience (% of articles on total including keyword)
Society	18915	1,06	11,29
System	17753	1,03	10,60
Datum	17089	1,01	10,20
Citizen	16699	1,00	9,97
Personal	16081	0,98	9,60
Service	12077	0,88	7,21
Police	10879	0,84	6,50
Safety	10344	0,83	6,18
Services	10172	0,82	6,07
Data	9455	0,80	5,65
Web	8754	0,78	5,23
Information	8696	0,78	5,19
Control	8693	0,78	5,19
Internet	4973	0,65	2,97
Computer	3265	0,58	1,95
Protection	3221	0,58	1,92
Code	2852	0,57	1,70
Digital	2704	0,56	1,61
Mobile	2683	0,56	1,60
Web	2444	0,54	1,46
Processing	2329	0,54	1,39
Device	2085	0,52	1,24
Users	2066	0,52	1,23
Online	1839	0,51	1,10
Surveillance	1070	0,46	0,64
Social	1059	0,45	0,63
Authority	900	0,44	0,54
Privacy	830	0,43	0,50
App	156	0,33	0,09
Cloud	141	0,33	0,08
Nsa	21	0,26	0,01

Table 2 – List of keyterms with the highest specificity in the considered corpus; IDF scores have been obtained has been tested on a corpus of 167472 documents (words have been translated from the Italian).

Testing the classifier

The classifier is made of keywords and weights. The presence of a keyword in a document adds its weight to the document score. Some words may be polysemous: for instance /police/ may be a relevant term in the surveillance discourse, but may be present in lots of newspaper articles that are not relevant in the "surveillance" theme.

Some keywords may be highly *specific* to the theme (their presence states the relevance to it), others may be more vague, but still relevant when the theme pertinence has been ensured by the presence of other highly specific keywords.
Testing the efficacy of the classifier means testing for 1) the effect of the presence or absence of certain keywords; and 2) the role of the certain keyword's weight value in the overall scoring effect. The score efficacy has to be measured against the relevance threshold value.
In the classifier testing process the documents whose score is near to the threshold value have a very important role, as they allow to assess the classifier *sensitivity* and *selectivity.* That is, measuring the number of relevant articles that were not identified as such, and the number of non-relevant documents that were mistakenly considered relevant.
This test, at the moment, is performed “by hand”, examining the documents whose score is in a certain range around the threshold value, and weighting the keywords and their values accordingly.

Conclusions

In order to automatically observe a phenomenon like surveillance in the press by building an automated software for classification of articles entails within the methodology of SMM – Science in The Media Monitor – different approaches. In this paper we described topic detection on scholarly articles on surveillance and a snow-ball analysis of a sample of the press discourse as tools to (1) explore some of the most relevant linguistic features of the articles, and (2) build a thesaurus classifier to be tested on a sample of articles and then applied to the whole corpus. Both techniques provide means to understand how quantitative analysis of rather large corpora of texts can be relevant for sociological analyses of technoscientific public issues; in particular these processes can be useful in view of automatic text classification as part of both a selection and interpretive processes, whose implications will be deepened in the course of future research.

References

Adey, P. (2003).Secured and Sorted Mobilities: Examples from the Airport. *Surveillance & Society* 1 (4) 500-514.

Ball, K.S., Wood, D.M. (2013). Editorial. Political Economies of Surveillance. *Surveillance & Society,* 11, 1–3.

Bauer, M. W., and Gaskell, G. (ed) (2002). *Biotechnology-the making of a global controversy*. Cambridge: Cambridge University Press.

Bauer, M. W., Allum, N., and Miller, S. (2007). What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda. *Public understanding of science*, 16(1), 79-95.

Bauer, M., Shukla, R., & Allum, N. (2011). *The Culture of Science: How does the Public relate to Science across the Globe*. London: Rutledge.

Beck, G., & Kropp, C.,(2011). Infrastructures of risk: a mapping approach towards controversies on risks. *Journal of risk research*, 14(1), 1-16.

Blei, D.M., McAuliffe, J.D. (2007). Supervised Topic Models., Neural Information Processing Systems 121–128.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning* , 3, 993–1022.

Boyle, P., Haggerty, K.D. (2011). Civil Cities and Urban Governance Regulating Disorder for the Vancouver Winter Olympics. *Urban Studies* 48, 3185–3201.

Cammozzo, A. (2013). Special issues on Surveillance (working paper). [Online] Available from: http://cammozzo.com/Papers/SurveillanceSpecialIssues.pdf. [Accessed 19th November 2014].

Fonio, C. (2011). The silent growth of video surveillance in Italy. Information Polit 16, 379–388.

Foucault, M. (1975). *Surveiller et punir* . Paris: Gallimard.

Giardullo P. e Lorenzet A. (2013). La ricerca emergente nei media: nanotecnologie, neuroscienze, biologia sintetica e proteomica, in Neresini F. e Lorenzet A. (a cura di) *Annuario Scienza e Società 2013*, Bologna: Il Mulino, 39-54.

Haggerty, K.D., Ericson, R.V. (2000). The surveillant assemblage. The British Journal of Sociology, 51, 605–622.

Landau, S. (2013). Making Sense from Snowden: What's Significant in the NSA Surveillance Revelations. *I*IEEE Security and Privacy, 11, 54–63.

Landau, S. (2014). Highlights from Making Sense of Snowden, Part II: What's Significant in the NSA Revelations. *I*IEEE Security and Privacy,12, 62–64.

Latour, B. (2005). *Reassembling the Social-An Introduction to Actor-Network-Theory.* Oxford:Oxford University Press.

Latour, B. (2011). Network Theory| Networks, Societies, Spheres: Reflections of an Actor-network Theorist. *International Journal of Communication*, 5 (2011): 15.

Latour, B., Camacho-Hübner, E., & November, V. (2010). Entering a risky territory: space in the age of digital navigation. *Environment and Planning D: Society and Space*, 28, 581-599.

Lyon, D. (2002). Surveillance Studies: understanding visibility, mobility and the phenetic fix. Surveillance and Society, 1(1), 1–7.

Lyon, D. (2003). Airports as data filters: Converging surveillance systems after September 11th. Information, Communication and Ethics in Society 1(1) 13–20.

Lorenzet, A. (2011). Using the World Wide Web for the cartography of technoscientific controversies. *TECNOSCIENZA: Italian Journal of Science & Technology Studies*, 1(2), 185-193.

Lorenzet, A. (2013). *Il lato controverso della tecnoscienza. Biotecnologie, nanotecnologie e grandi opere nella sfera pubblica*. Bologna: Il Mulino.

McCombs, M. E., and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2), 176-187.

Marres, N., & Weltevrede, E. (2013). Scraping the Social? Issues in live social research. Journal of Cultural Economy, 6(3), 313-335.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt.

Marx, G.T., Muschert, G.W. (2007). Personal Information, Borders, and the New Surveillance Studies. Annual Review of Law and Social Science,3 (1), 375–395.

Mccallum, A., 2002. *MALLET: A Machine Learning for Language Toolkit.*http://mallet.cs.umass.edu

Melard, F. (2009). Analysis of the participative set up and its ouputs-the Debategraph and the bee controversies. [Online] Available from:L http://orbi.ulg.ac.be/bitstream/2268/62810/1/Deliverable%20D6a%20EN.pdf. [Accessed: 15th May 2014]

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006). *YALE: Rapid Prototyping for Complex Data Mining Tasks*. In Eliassi-Rad, T., Ungar, L. H., Craven, M., and Gunopulos, D., editors. Paper Presented at the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), pages 935–940, New York ACM Press.

Neresini F. e Lorenzet A. (ed.) (2013). *Annuario Scienza e Società 2013*, Bologna: Il Mulino.

Neresini F. e Lorenzet A. (2012), *Il dibattito sull'energia nei media italiani*, in Neresini e Pellegrini (a cura di) *Annuario Scienza e Società 2012*, Bologna: Il Mulino, pp. 43-61.

Neresini F. e Lorenzet A. (2011), *La scienza nei media italiani: tendenze e temi emergenti*, in Bucchi e Pellegrini (a cura di) *Annuario Scienza e Società 2011*, il Mulino, Bologna, pp.39-53.

Rogers, R. (2013). *Digital methods*. MIT Press.

Schmid, H., 1995. TreeTagger: a Language Independent Part-of-speech Tagger. Inst. Für Maschinelle Sprachverarbeitung Univ. Stuttg. 43.

Venturini, T. (2008). Introducing the cartography of controversies. *Etnografia e ricerca qualitativa*, 1(3), 369-394.

Venturini, T. (2010). Diving in magma: How to explore controversies with actor-network theory. *Public understanding of science*, 19(3), 258-273.

Venturini, T. (2012). Building on faults: how to represent controversies with digital methods. *Public Understanding of Science*, 21(7), 796-812.

Yaneva, A. (2012). *Mapping controversies in architecture*. Ashgate Publishing, Ltd..