1. Content of Textual Analysis

In this section we set out to introduce the importance of content or textual analysis and the software tools as they are needed, placing them in the Web context. More specifically we discuss the benefits of using such methods, and suggest the ways by which the Web as a system of interlinked hypertext documents accessed via the Internet might strengthen current research techniques providing an alternative and supplemental information source to traditional statistics on social phenomena. In some cases, we argue, content or textual analysis can even act as a substitute for existing indicators and data collection methods (public opinion research, surveys, interviews, observation and participant observation), leading to reduced administrative burdens and consequently lower costs.

Content quantitative or qualitative analysis is accepted as a uniquely valuable methodological approach for generating knowledge, particularly in relation to social phenomena and human communication. According to Krippendorff the term *content analysis* defines "the analysis of the manifest and latent content of a body of communicated material through classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect". Within social sciences and humanities different kinds of content analysis have been widely adopted to systematically investigate a number of social issues. Historically, content analysis was developed for the analysis of textual material but the range of data has widened over the years to embrace nearly any cultural artifact (audio, video, image, multimedia, hypertext) .

In the 21st century, the textual fabric of contemporary society has undergone radical transformations, due to the ongoing information revolution. The proliferation of documents available on the Web and elsewhere is overwhelming. Citizens and enterprises, groups, organizations, institutions and governments do not only leave behind 'digital footprints' when using the internet. Millions of Web professional or amateur users increasingly create billions of web pages[1] and documents. Every day, a huge amount of online texts is produced for different purposes, on different issues, in different countries, languages and online environments, user generated contents in blogs and social network sites or social media, private conversations, emailing, blogging, news, academic work, etc. Moreover, all over the world, governments, institutions, libraries[2], museums, academic journals, public policy organizations etc try to digitize their content and make it available through the Internet, TVs and an ever-growing number of devices (smart-phones, tablets etc) for business, science, research, entertainment etc.

[1] The Web universe is constantly expanding, so its size is unknowable. In 2008 Google noted that it had identified (but not actually indexed) over a trillion (10^{12}) distinct URLs (Web addresses), and that several billion (10^{9}) new web pages appear daily (. For a daily count of web pages see http://www.worldwidewebsize.com/.

[2] A growing number of textual databases are available to researchers and could be used by content analysts. The US Library of Congress has made digitized versions of collection materials available online since 1994, concentrating on its most rare collections and those unavailable anywhere else. In December 2000, the Library of Congress initiated the *National Digital Information Infrastructure and Preservation Program*, a project to develop a comprehensive strategy to "save America's cultural and intellectual heritage in digital formats" providing one of the largest bodies of noncommercial high-quality content on the Internet (digitized photographs, manuscripts, historic maps, documents, video, sound recordings, motion pictures, books, as well as "born digital" materials such as Web sites (http://www.loc.gov/library/about-digital.html). See also, the *LexisNexis Group* the largest electronic database for legal and public-records related information, the *Educational Resources Information Center* (ERIC) on education-related literature, the *Oxford Text Archive* on humanities, the *Project Muse*, *JSTOR, Google Scholar,* etc.

Traditional media (newspaper and TV)[3] are quickly migrating to the Internet. News papers and other news providers are updating their on-line news feeds in near real-time, allowing interested parties to follow the news in near real-time. Search engines exacerbate the accessibility by making more and more documents available in a matter of a few key strokes. The Internet and the Web content have thereby become the most effective resources to research on contemporary economy, culture, politics, human communication, human behavior and human interaction.

The increasingly widespread availability of digital or digitized texts concerning virtually everything that matters to society and its members has moved content or textual analysis into the center of the researcher's interest. The Web has evolved into the resource of first resort for social researchers and scientists. However, although the volume of information explodes exponentially and the amount of data available to us is constantly increasing, our ability to produce genuine knowledge and statistics on social phenomena for public policy purposes from that information appears very limited.

Combining qualitative and quantitative content or textual analysis and official statistics could add value to the information provided by official statistics to public policy makers. Moreover, the Web content analysis could become an efficient alternative to the methods of tracking markets, political leaning and emerging ideas and a reliable and more *unobtrusive*[4] substitute source of information compared to surveys, observation and participant observation, public opinion research, questionnaires, interviews, focus groups etc and therefore could contribute to a reduction in the cost and the administrative workload.

The dual impact of new information technology both on what kind of content corpora can be created and on what kind of analysis it makes possible is an ongoing debate to the present day. Most researchers recognize that the use of information technology affects both the ways of collecting content and texts and building analysis corpora and the ways of undertaking the analysis. Many researchers, engage with what is termed "virtual methods" that import standard methodologies from the social sciences and the humanities and try to deal with the huge amount of available texts, the complexity of forms and the diversity of formats with the assistance of content analysis software (CAQDAS). Within this research rationale the Web content is perceived as a *static corpus* of information.

Web epistemologist Richard Rogers, on the other hand, argues that any static corpus is cut off at the moment of its compilation of the *dynamic* and constantly changing environment of the Internet. He also asserts that there is an ontological distinction between the natively digital and the digitized, that is, the objects, content, devices and environments that are "born" in the Internet, as opposed to those that have "migrated" to it. He also insists that the current methods of study should change, however slightly or wholesale, given the focus on objects and content of the *medium*. He proposes what he terms, the *digital methods program* introducing the term *online groundedness*[5], in an effort to conceptualize research that follows the medium, captures

[3] The BBC Archive made available via the internet the entire television and radio program archive (the worlds largest), documents and photographs from as far back as the 1930s. http://www.bbc.co.uk/archive/.

[4] Hence, there is little danger that the act of measurement itself will act as a force for change that confounds the data .

[5] Rogers , relying on Daniel Miller's and Don Slater's ethnographic work on Internet usage in cybercafés in Trinidad and Tobago, that found that the Internet was a space where Trinis performed their own culture, introduced the term *digital groundedness*, or *online groundedness*, in order to assert that people appropriate the Internet in their own

its dynamics and makes grounded claims about cultural and societal change. As he puts its "Indeed, the broader theoretical goal of digital methods is to think through anew the relationship between the Web and the ground" .

In what follows we discuss very briefly two of the major available strategies, the virtual and digital methods as far as the Web content is concerned.

2. Virtual methods - Software packages for qualitative content analysis (CAQDAS)

Content or textual analysis qualitative research involves many epistemological and methodological choices; there are very many different ways of doing it, and they require different sorts of data and different research tools and strategies. In general, the content analysis procedure consists of six steps[6]:

1. Research design: Identify a research problem, review the literature, specify a purpose for research, and choose the analytic approach and software.
2. Collecting texts (data): define the set of database/ actors/ sites/ networks/ countries etc. Define the time span for collecting documents. Make the choices about different types of data (simple or multi-media file types and formats including rich and plain text files, PDF, image files, audio and video files, spreadsheets, etc).
3. Corpus construction: Decide the way of systematically collecting the data either following a statistical sampling (quantitative), a semantic (qualitative), or a mixed method rationale.
4. Indexing – coding: In general, coding, or data ordering, is a major task in all qualitative analysis because this is the process by which the researcher identifies which of the data appears to be important. Codes become the building blocks for the analysis and interpretation of the data. The researcher must choose between different approaches to coding (e.g., inductive, deductive or a combination of the two), types of coding (e.g., open, axial and selective coding), principles of coding, coding schemes and frames. He must also decide between three common ways of creating codes: creating apriori codes (which can be based on available theoretical, empirical and/or clinical knowledge), creating and naming codes as they emerge from the data, and creating in-vivo codes (i.e., identify a direct quote in the data as a code).
5. Analysis of data
6. Interpretation and evaluation of the research.

ways making it fit their own national or cultural practices.

[6] In qualitative content research, like grounded theory building, these phases and steps of analysis must be evaluated against four research quality criteria: construct validity, internal validity, external validity and reliability. Briefly, according to Pandit , *construct validity* is enhanced by establishing clearly specified operational procedures. *Internal validity* is enhanced by establishing causal relationships whereby certain conditions are shown to lead to other conditions, as distinguished from false relationships. In this sense, internal validity addresses the credibility or "truth value" of the study's findings. *External validity* requires establishing clearly the domain to which the study's findings can be generalized. Here, reference is made to analytic and not statistical generalization and requires generalizing a particular set of findings to some broader theory and not broader population. Finally, *reliability* requires demonstrating that the operations of a study - such as data collection procedures - can be repeated with the same results.

Many researchers in order to solve the information overload problem and the differences that the online environment brings with it use Computer Assisted Qualitative Data Analysis Software (CAQDAS) to assist with organization and analysis of non-numerical or unstructured qualitative data.

CAQDAS packages are very common among qualitative researchers nowadays. The first programs have been available since the 60s, but it took until the late 1980s, when personal computers entered the desktops of qualitative researchers, that the first programs[7] began to be widely recognized in the field of social research and humanities. In past decades, advances in CAQDAS technologies have been applied throughout the social sciences, enabling researchers to address questions of greater scale and complexity with better accuracy, consistency, and transparency. Similarly, linguists have found ways to utilize the processing and analytical power of computer technology to assess large quantities of text in ways not feasible manually. Since then, a growing variety of software packages that tend to be more specifically focused on a given task, with relatively sophisticated functionalities, became available.

With the new-generation qualitative software packages[8], it is possible to manage, access and analyze qualitative data in many different formats and to keep a perspective on all of the data, while supporting rigorous data analysis without losing its richness or the closeness to data that is critical for qualitative research. Some can work only with text, others can handle images, sound and video.

Researchers must, of course, tailor their methods to the requirements of the research by selecting specific techniques and integrating them with other methods, substantive considerations, and theories. There is a wide variety of techniques for analyzing text that researchers may use (word-frequency counts, key-word-in-context (KWIC) listings, concordances, classification of words into content categories, content category counts, and retrievals based on content categories and co-occurrences). Some packages build hierarchical trees of categories, others let the researcher build their own "trees", and others simply list the categories alphabetically. Most of them can create reports according to the analyst's needs.

The advantages of using this software include being freed from manual and clerical tasks, saving precious time, having increased flexibility, and having improved validity and auditability of qualitative research. According to Weber , compared with human-coded or interpretive modes of text analysis, one of the most important advantages of computer-aided content analysis is that the rules for coding text are made explicit and when applied to a variety of texts, generate formally comparable results,

[7] Nowadays there are at least 40 commercial and a few open source CAQDAS packages. For practical support and training for users of a range of software programs designed to assist qualitative data analysis (CAQDAS packages: ATLAS.ti, MAXQDA, NVivo and QDA Miner), as well as summary of strengths and weaknesses of each package in specific research, discussion of qualitative and quantitative analysis strategies in the context of each CAQDAS package etc, see *CAQDAS Networking Project* at University of Surrey: http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/. On CAQDAS methods and applications see also, *Forum: Qualitative Social Research* http://www.qualitative-research.net/index.php/fqs/index. See, also for in-depth analysis step by step of three popular packages: ATLAS.ti 5, MAXqda 2, and QSR NVivo7. See also on the process of analyzing the media files in this data set using *Transana*, a qualitative software package designed for the transcription and qualitative analysis of video and audio data.

[8] A CAQDAS program usually has: Content searching tools, Coding tools, Linking tools, Mapping or networking tools, Query tools, and Writing and annotation tools. Software supports tagging, coding and indexing of texts thereby supporting segmentation, linking, ordering and reordering, structuring and the search of retrieval of texts for analytic purposes.

process which can lead to the accumulation of research findings. A second major advantage of computer-aided content analysis is that, once formalized either by computer programs and/or content-coding schemes, the computer provides perfect coder reliability[9] in the application of coding rules to text. High coder reliability then frees the investigator to concentrate on other aspects of inquiry, such as validity, interpretation, and explanation. However, as Weber, notes, even with the assistance of computers a remaining difficulty is that there is too much information in texts. Their richness and detail exclude analysis without some form of data reduction. Thus for Weber, the key to content analysis is choosing a strategy for information loss that yields substantially interesting and theoretically useful generalizations while reducing the amount of information analyzed and reported by the investigator .

While CAQDAS packages are very useful when the content analysis is carried out on large amount of documents and when there is an imperative need for statistical data and generalizations translated into quantifiable charts and graphs, they have technical limitations as far as the in-depth and in-context analysis in methods like discourse analysis is concerned, where the material to be analyzed has to be understood in relation to its particular discursive, interactional or rhetorical context . Moreover, software use may lead to increasingly deterministic and rigid processes, privileging of coding, and retrieval methods, reification of data, increased pressure on researchers to focus on volume and breadth rather than on depth and meaning.

3. **Digital methods**

The turn to *digital methods* in social and especially cultural research has been based on what one of the leading new media theorists, Lev Manovich (2007) has termed “cultural analytics”[10], the necessity to develop the appropriate methods and to build new tools for collecting, storing and analysing large data sets which would give us systematic insight into personal and group behaviour in the social and cultural field. Web epistemologist Richard Rogers claims that the Web is a knowledge culture distinct from other media that gives tremendous research opportunities that would have been improbable or impossible without the Internet. He also insists that *virtual methods* , imported from the social sciences, can no longer capture and analyze the *dynamic* status of data and content that proliferate in the Web in different forms and formats. Hence, he proposes that researchers should concentrates on new methodological strategies which adjust better to the digital culture features and the imperatives of the Internet. Practically, this means that digital methods seek to learn from the methods built into the dominant devices online and repurpose them for social and cultural research. Hence, according to the new methodological turn, the challenge is to study both the info-web as well as the social web with the tools and natively digital objects (for example, the hyperlink, the website, the url, the thread*,* the spheres and the tag) that organize them

[9] The central problems of content analysis originate mainly in the data-reduction process by which the many words of texts are classified into much fewer content categories. Reliability problems usually grow out of the ambiguity of word meanings, category definitions, or other coding rules. See .

[10] The term is borrowed from *Google Analytics*. According to Rogers, *Cultural Analytics* is a “big science type of idea” aiming to build quite large scale data collection facilities to take advantage of all the digital traces online and analyze them to think about culture production, state of culture etc .

and to find how devices themselves, such as search engines, are part of the analysis[11].

Briefly, the wide array of methods and tools which have developed in the last decade under the novel rubric *digital methods* is characterized by the following aspects:

1. The conception, that the Web is a *content circulation space* and not as a set of single Web sites (Roger, 2010).
2. The recognition, as Rogers (2009) has noted that an ever increasing part of social interaction was actually taking place in the online world, especially for the growing part of the population which has been described as "digital natives".
3. The attempt to develop data sets "native" to social science inquiries, moving away from reliance on limited access to commercial data sets generated by Search Engines like Google, Yahoo etc.
4. The realization that older, more conventional methods of social research may be inappropriate or lacking for research in the digital world primarily because of their inability to capture and analyse the *dynamic* nature of the online world, both as a *site* of social interaction (constantly changing sites, interfaces and online infrastructure and environments) and in terms of the actual social interactions taking place within it (constant and ever changing production of texts, fluidity and instability of content as compared to traditional content outputs of mass mediated communication and so on).
5. The ambition, therefore, to develop data, tools and research practices which would be characterized by "online groundedness" (Rogers 2009) and aiming to introduce to new ways of doing Internet research.

In the past decade, within the field of digital methods, considerable amount of research has been done in academia, as far as the content or textual analysis is concerned, applying a variety of techniques based on automated data mining algorithms. In what follows we present very briefly, T*ext Mining*, *Social Network Analysis* and *Issue Analysis*[12].

1. ***Text mining***

The problem of text mining has gained increasing attention in recent years because of the large amounts of text data, which are created in a variety of social network, web, and other information-centric applications. Prado& Ferneda offer a broad definition of text mining as "the application of computational methods and techniques over textual data in order to find relevant and intrinsic information and previously unknown knowledge".

According to Ronen Feldman text mining tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management. In a manner

[11] See *Digital Methods Initiative. DMI*, is a collaboration of the New Media program, Media Studies, University of Amsterdam and the Govcom.org Foundation. DMI is dedicated to reworking method for Internet research, and in particular to learning and developing techniques for studying societal conditions and cultural change with the Web. Its director is Richard Rogers, Chair, New Media & Digital Culture, University of Amsterdam, https://wiki.digitalmethods.net/Dmi/MoreIntro.

[12] Nowadays, these applications are becoming common in industry, as well as defense and law enforcement. They are also increasingly used in the sciences - particularly bioscience - and social sciences, where researchers frequently deal with very large volumes of data. The humanities, however, are still only just beginning to explore the use of such tools. As an example of text mining in literary interpretation see .

analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections. Hence, for text mining systems, preprocessing operations center on the identification and extraction of representative features for natural language documents. These preprocessing operations are responsible for transforming unstructured data stored in document collections into a more explicitly structured intermediate format, which is a concern that is not relevant for most data mining systems. Moreover, because of the centrality of natural language, text mining exploits techniques and methodologies from the areas of information retrieval, information extraction, and corpus-based computational linguistics.

According to Prado& Ferneda text mining techniques can be organized into four categories:

1. *Classification* techniques consist of the allocation of objects into predefined classes or categories.
2. *Association analysis* can be applied to help the identification of words or concepts that occur together and to understand the content of a document or a set of documents.
3. *Information extraction* techniques are able to find relevant data or expressions inside documents
4. *Clustering* is applied to discover underlying structures in a set of documents.

A key element of text mining is its focus on the *document collection*. At its simplest, a document collection can be any grouping of text-based documents. Practically speaking, however, most text mining solutions are aimed at discovering patterns across very large document collections. The number of documents in such collections can range from the many thousands to the tens of millions. Document collections can be either *static*, in which case the initial complement of documents remains unchanged, or *dynamic*, which is a term applied to document collections characterized by their inclusion of new or updated documents over time. Extremely large document collections, as well as document collections with very high rates of document change, can pose performance optimization challenges for various components of a text mining system.

As a typical real-world document collection suitable for text mining Feldman mentions the example of *PubMed*, the National Library of Medicine's online repository of citation-related information for biomedical research papers[13]. The real size of document collections like that represented by PubMed makes manual attempts to correlate data across documents, map complex relationships, or identify trends at best extremely labor-intensive and at worst nearly impossible to achieve. Automatic methods for

[13] PubMed contains text-based document abstracts for more than 12 million research papers published in the English and other languages on topics in the life sciences. It also represents the most comprehensive online collection of biomedical research papers. The publication dates for the main body of PubMed's collected papers stretch from 1966 to the present. The collection is dynamic and growing, for an estimated 40,000 new biomedical abstracts are added every month. PubMed's data repository can represent substantial document collections for specific text mining applications. For instance, a relatively recent PubMed search for only those abstracts that contain the words *protein* or *gene* returned a result set of more than 2,800,000 documents, and more than 66 percent of these documents were published within the last decade.

identifying and exploring interdocument data relationships dramatically enhance the speed and efficiency of research activities. Indeed, in some cases, automated exploration techniques like those found in text mining are not just a helpful adjunct but a baseline requirement for researchers to be able, in a practicable way, to recognize subtle patterns across large numbers of natural language documents.

For more examples on text mining or Text analytics especially on social media and multimedia see .

2. *Social network analysis – issue network analysis*

Sustained attempts towards network theorizations of social activity developed well before the rise of computer mediated communication (CMC), let alone the Web 2.0. As a consequence, developing appropriate analytical tools for SNA in the CMC era has been a dynamic process, testing already established notions and practices in a new technological and communicative environment.

The overall aim of all SNA is to provide a structural analysis that focuses on connections, which can provide insight into how one person, group, or event can and does influence another. These people or groups or events cannot, and do not, act in an autonomous fashion; rather, their actions are constrained by their position in the overall network, which is in turn constrained by the other networks and institutions in which they are embedded (the overall ecology).

Depending on the particular research aim, we may distinguish SNA research into the following themes

1. *Exploratory data analysis (EDA)* for networks. Such methods include descriptive measures and analyses that assist in summarizing and visualizing properties of networks and in investigating the dependence of such measures on other network characteristics.

2. *Model development and estimation*. The second guiding theme is the value of developing plausible models for social networks whose parameters can be estimated from network data.

3. *Impact of network change on network properties*. A third theme is the importance of understanding how different measures of network structure change following “node removal” or “node failure.”

4. *Processes or flows on networks*. A fourth theme is the importance of distinguishing the structure of a network from the different types of dynamic processes or flows that the network might support

An interesting example of the potential of the methodological turn towards digital tools for the analysis of Web content and web social networks is the *Digital Methods Initiativ*e based in Amsterdam.[14] New Media expert Richard Rogers and his colleagues, exploring and deepening the SNA rationale develop methods, techniques, software applications and info-tools since 1999. Rogers studies what he terms ‘adjudicative’ or 'recommender' cultures of the Web that help to determine the reputation of information as well as organizations. He emphasizes that, instead of

[14] https://wiki.digitalmethods.net/Dmi/MoreIntro.

building “an all-purpose” standard social scientific tool, the teams’ goal is to make “situated software” for specific kinds of research questions . He also insists on the “online groundedness” of the Internet and tries to develop metrics that will aid in analysis of particular countries. The most well-known tool Rogers has developed with his colleagues is th*e Issue Crawler*[15], a server-side Web crawler, co-link machine and graph visualizer. It locates what Rogers and colleagues have named "issue networks" on the Web - densely interlinked clutches of NGOs, funders, governmental agencies, think tanks and lone scientists or scientific groups, working in the same issue area. According to their theoretical approach, unlike social networks, *issue networks* do not privilege individuals and groups, as the networks also may be made up of a news story, document, leak, database, image or other such items. Taken together these actors and 'argument objects', serve as a means to understand the state of an issue either in snapshots or over time. Rogers and colleagues also developed the *Election Issue Tracker*, a pre-RSS newspaper query machine employed in the Netherlands to understand whether media aided the rise of populism. Other tools Rogers and colleagues have developed include the *Web Issue Index of Civil Society,* also known as the *Issue Ticker*, where the campaigning behavior of NGOs is monitored. The Index is a novel form of attention cycle research, showing whether attention to issues is rising or falling, not according to newspaper coverage, but rather according to civil society campaigning. The tools form the infrastructure of the *Digital Methods Initiative*, which specialises in repurposing online devices (and 'methods of the medium') for research that goes beyond the study of online culture only[16]. Since then a set of allied tools and independent modules have been made to extend the research into the blogosphere, online newssphere, discussion lists and forums, folksonomies[17] as well as search engine behavior. These tools include scripts to scrape web, blog, news, image and social bookmarking search engines, as well as simple analytical machines that output data sets as well as graphical visualizations.

4. **Conclusion**

Bauer argues that an adequate understanding of social phenomena requires a multitude of methods and data: “methodological pluralism arises as a methodological necessity” . We believe that both virtual and digital methods in content or textual analysis can coexist and would add quality and depth to research and IaD methods for policy makers. While large documents collections containing valuable information can be mined with text mining techniques, since texts and documents are more than collections of words, phrases and paragraphs, CAQDAS techniques can be helpful when working with very rich text-based and/or multimedia unstructured information and where deep levels of analysis on small or large volumes of data are required. On the other hand, the analysis of the medium digital objects can respond better to the instability and *ephemerality* of websites and other digital media, and the complexities associated with

[15] A crawler is a program that retrieves and stores pages from the Web, commonly for a Web search engine. A crawler often has to download hundreds of millions of pages in a short period of time and has to constantly monitor and refresh the downloaded pages.

[16] See http://www.uva.nl/over-de-uva/organisatie/medewerkers/content/r/o/r.a.rogers/r.a.rogers.html .

[17] According to Wikipedia, a *folksonomy*, a term coined by Thomas Vander Wal, (from the coupling of *folk* and *taxonomy),* is a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorize content. Also known as, *collaborative tagging, social classification, social indexing*, and *social tagging*.

their dynamic character (Rogers, 2009).

The best way to verify if Web content analysis can be useful for developing new and alternative indicators for governmental and public policy research and management is to **build case studies and perform some small-scale experiments, especially in areas where statistics are not available or data impossible to acquire with the traditional methods and techniques or where collecting data has a very high cost** (i.e. conducting interviews, public opinion researches, surveys etc). Moreover, the experiments will be a good opportunity to discuss the methodological perspectives and the epistemological issues and to explore the merits of different forms of software and scrutinize the features and their adaptations. Finally, it will be a great occasion for the ethical problems embedded in such research practices to be addressed (the right to privacy, the right to not participate to the ongoing research, the copyright issues etc).

5. **Bibliography**

Aggarwal, C. C. & C. Zhai (eds.) (2012). *Mining Text Data.* Science+Business Media, New York, Dordrecht, Heidelberg, London, Springer

Bauer, M. W. & G. Gaskell (eds.) ([2000] 2005). *Qualitative Researching with Text, Image and Sound. A Practical Handbook.* London, Thousand Oaks, New Delhi, Sage Publications.

Breiger, R., K. Carley & P. Pattison (eds.) (2003). *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. Committee on Human Factors, National Research Council* Washington, The National Academies Press.

Dempster, P. G. & D. K. Woods (2011). "The Economic Crisis Though the Eyes of Transana". *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research,* 12(1), Art. 16.

Feldman, R. & J. Sanger (2007). *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data.* Cambridge, New York, Melbourne, Cambridge University Press.

Fletcher, W. H. (2011). "Corpus Analysis of the World Wide Web". In Chapelle, C. A. (ed.) *Encyclopedia of Applied Linguistics.* Wiley-Blackwell.

Hine, C. (ed.) (2005). *Virtual methods: issues in social research on the Internet.* Oxford, Berg.

Kellehear, A. (1993). *The Unobtrusive Researcher: A Guide to Methods* St Leonards, NSW, Allen & Unwin.

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology. 2d edition.* Beverly Hills, London, Sage Publications.

Lewins, A. & C. Silver (2007). *Using Software in Qualitative Research: A Step-by-Step Guide.* Los Angeles, Sage.

Macmillan, K. (2005). "More Than Just Coding? Evaluating CAQDAS in a Discourse Analysis of News Texts ". *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research,* 6(3).

Miller, D. & D. Slater (2000). *The Internet: An Ethnographic Approach.* Oxford, Berg.

Pandit, N. R. (1996). "The Creation of Theory: A Recent Application of the Grounded Theory Method ". *The Qualitative Report,* 2(4).

Prado, H. a. D. & E. Ferneda (eds.) (2008). *Emerging Technologies of Text Mining:*

Techniques and Applications. Hershey, New York, Information science reference.

Rogers, R. (2009). *The End of the Virtual: Digital Methods.* Amsterdam, Amsterdam University Press.

Rogers, R. (2010). "Internet Research: The Question of Method". *Journal of Information Technology & Politics*, 7, pp.241-260.

Sculley, D. & B. M. Pasanek (2008). "Meaning and mining: the impact of implicit assumptions in data mining for the humanities". *Literary and Linguistic Computing,* 23(4), pp.409-424.

Webb, E. J., D. T. Campbell, R. D. Schwartz & L. Sechrest (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences* Chicago,, IL: Rand McNally.

Weber, R. P. (1990). *Basic Content Analysis* Newbury Park, CA, Sage Publications, Inc.