Digital Digging

Towards a methodology for online grounded investigative reporting

Master thesis for the New Media course

Name: Jelle Kamsma

Student number: 5645662

Address: Gevleweg 102, 1013AJ Amsterdam

Phone number: 06 38483324

Email: jellekamsma@gmail.com

Supervisor: Dr. P.L.M. Vasterman

Second reader: Dr. T. Poell

Date: August 16, 2010

UNIVERSITEIT VAN AMSTERDAM

Abstract

This thesis introduces the term online grounded investigative reporting (OGIR) for the implementation of new online methods in the field of investigative journalism. My claim is that by looking at online environments cultural change and societal conditions can be diagnosed. The Web and the 'real world' should not be seen as separate realms but impact each other in numerous ways. Technological innovations have always affected the work practices of investigative reporters in significant ways. Therefore, I argue that a clear understanding of the way information is organized on the Web is vital for conducting solid online investigations. Online research often works with incredibly huge datasets and it is very easy to find false correlations or patterns. Instead of following the quantity I propose to follow the methods of the Web. The objects through which the Web is organized, such as hyperlinks and search engines, should be a part of any investigative reporting that takes place online. These object are not simple tools that can be used without consequences but instead actively influence the results of the investigation. In the analytical part of this thesis I discuss three elements of the Web that might prove useful for journalists. Besides describing a couple examples of online grounded investigative reporting, I will focus on the tools and techniques that are available online. Many of these have been developed by the Digital Methods Initiative which shares many methodological concerns with OGIR. In the last part of the thesis these concerns are further addressed as I try to design a methodology for OGIR. A number of strategies, derived from the examples in the second part, are described. One of the most productive ways of doing online grounded investigations is by examining trends over a period of time. With this strategy it becomes more justifiable to make correlations between certain online trends and major events in the real world. Also a number of unique responsibilities for conducting OGIR are discussed. The reporter has to be completely transparent about the way the investigation was conducted while at the same time not obscuring the public with methodological considerations. Finally it is important to note that it is up to the practitioners in the field to define what should constitute a shared methodology for online grounded investigative reporting. The scope of this thesis is merely to start a debate on how such reporting should be done.

Keywords:

Online grounded investigative reporting, methodology, journalism, online groundedness, Digital Methods

Table of contents

INTRODUCTION

The news industry is in the midst of a 'perfect storm': the economic downturn causes an increasing decline in advertising revenues, there is a structural and irreversible shift of advertising to the Internet, and 'fragmented audiences [are] moving increasingly to non-linear consumption which is less susceptible to commercial impact and therefore less valuable to advertisers.' (Barnett, 2009: p. 1) As consumers and advertisers move to the Web it becomes apparent that the commercial model that underpins the gathering, production and distribution of news faces significant economic pressures. Media enterprises are desperately seeking a new model to make money of the Web. The basic challenge is two-fold: the consumers expect news content to be free, whilst advertisers are expecting much lower rates around the news. The digital revolution forces publishers to distribute the news into 'bite-size, multi-media friendly packages that can attract enough click to sustain the interest of advertisers.' (Currah, 2009: p. 1) This is a worrying development since newsrooms are being sucked into the 'clickstream' and determine their editorial agenda by what proves successful on the Web. The underlying logic seems to be that news has to be fast, cheap and easy to comprehend. Journalist Loretta Tofani explains: 'The dominant message, amid buyouts and pink slips, is produce, produce, produce! The result is that reporters tend to produce more good and mediocre stories at the expense of the great and vital stories, which are still out there.' (Tofani, 2001: p. 64)

This might also explain why investigative reporting is the main victim of the digital revolution in the news industry. Besides being inextricably linked to the general crisis in journalism, investigative reporting has some unique vulnerabilities of its own. Edward Wasserman explains:

> It's expensive, offers uncertain payback, ties up resources that could be used in more conventionally productive ways, fans staff jealousies, offends powerful constituencies (including touchy readers), invites litigation, and usually comes from the most endangered class in the newsroom, the senior reporters whose ranks are being thinned aggressively through forced retirement. (Wasserman, 2008: p. 7)

However, the aim of this thesis is not to discuss the negative influences that the Internet is supposed to have on journalism in general and on investigative reporting in particular. What I suggest on the other hand is to look at new ways in which the Web can be used for investigative reporting. That is to move away from the more common approach of looking at the Web as solely a place of distribution of news content. Instead my interest lies in examining how Web-based content can be used in the gathering of information and the production of investigative news stories. The central question being: how can online objects such as search engines or hyperlinks be appropriated for investigative journalism? The overall purpose of this thesis is to describe a methodology for online

investigative reporting that incorporates the concepts, strategies and techniques linked to the Web. Methodology here refers broadly to the logic of analysis and the justification for the methods that are in use within a given field. By building upon the work done by the Digital Methods Initiative I hope to make a beginning with the methodology of, what I call, online grounded investigative reporting. In the end it is up to the practitioners of the field, the investigative journalists, to determine the requirements, procedures, and rules of analysis in conducting research on the Web. The intention of this thesis is not to prescribe what is right and wrong but to start a debate on what should constitute a shared methodology for online grounded investigative reporting.

This thesis is divided into three parts. The first part will provide a historical and a theoretical framework within which my methodology can be understood. I will begin this thesis by discussing the development and nature of investigative reporting. Even though online grounded investigative reporting requires specific techniques and strategies, it retains the need for creative and critical thinking. At the same time technological innovations have always had a major influence on how the practice of investigative reporting evolved. It is therefore important to look at investigative reporting from a historical perspective. I will describe the implementation of computers, databases and the Internet in the newsroom. In the following chapter I will discuss how the Internet differs from other media and propose a medium-specific approach. The structural features of the Internet have to be taken into consideration when taking the Web as an object of study. The way information is organized online is vital in this respect. Investigative reporters conducting online grounded investigations must be aware of this.

In the second part of my thesis, I will analyze three distinct elements of the Web that might prove useful to investigative reporters: the Website, hyperlinks and search engines. After discussing the ways in which investigative reporters already made use of these elements in their work I will describe some further possibilities. Tools as well as the techniques are being discussed and looked at from a critical perspective.

The third part of my thesis will try to identify the lessons learned in the second part in order to make the beginning of a methodology for this kind of reporting. Some general strategies will be proposed for conducting online grounded investigations. I will also draw attention to a number of responsibilities of which reporters should be aware. Besides the general responsibilities that apply to every investigative reporter, the online investigations bring along a number of additional considerations. I will conclude my final chapter with some guidelines for presenting the finding in a comprehensible way to the public. In the end it is always up to the public whether online grounded investigative reporting can be a succes.

PART I

1
THE CHANGING NATURE OF INVESTIGATIVE REPORTING

It is often argued that freedom of the press is essential for a flourishing society. It is an argument that extends back to the position set forth by John Milton in the late 1600s that freedom of the press is an acknowledgement of the 'peoples' right to know' and the need for a 'marketplace of ideas'. (Altschull, 1990) Investigative journalism can be seen as a logical consequence of these political theories. A strong press can investigate and unveil the actions, policies and performances of those who are in power. However, the exact nature of investigative reporting has always been the topic of vigorous debate among both academics and investigative journalists themselves. Especially since reporters are constantly changing the ways in which they conduct their investigations. Technological innovations, most notably the introduction of the computer in the news room, can often be seen as the main catalysts in these processes. As I argue in this thesis, the widespread use of the Internet marks a new phase for investigative journalism in which reporters have to appropriate new strategies and become aware of new responsibilities. This does not mean that traditional strategies by which reporters investigate situations are replaced. New tools can be used with great effectiveness by an investigative journalist but must always be supplemented with a traditional approach. In this chapter I will place the practice of investigative journalism in a historical perspective to show how technological innovations influenced the work of the reporters.

1.1 Defining investigative journalism

For exploring the issue of what constitutes investigative reporting we best begin by looking at the United States. Many definitions have their roots in the American culture of investigative reporting and although I will argue that some of these definitions are outdated, they provide a good starting point. Many views on investigative journalism derive from the muckraking era that started at the beginning of the twentieth century in the US. Pieces like "History of Standard Oil" by Ida Tarbell and "The Treason of the Senate" by Graham Phillips were characterized by the drive to expose abuses of the system. Their stories annoyed the men in power to that extend that president Theodore Roosevelt used the term 'muckrakers' for reporters who addressed wrongdoings. While it was meant as an invective, reporter quickly saw it as an honor to be called a muckraker. (Van Eijk, 2005: p. 14) In the beginning of the twentieth century investigative stories mainly focused on the missteps of commercial companies but during the sixties muckraking reemerged by focusing on corruption in government agencies. The Watergate scandal as exposed in *The Wall Street Journal* by Bob Woodward and Carl Bernstein can be seen as the culmination of this trend.

The muckraking tradition and famous stories like the Watergate scandal have had tremendous influence on how investigative journalism is defined. The foreword in *The Reporters Handbook*, a publication by the American association of investigative reporters gave the following, widely cited, definition:

> It is the reporting, through one's own work product and initiative, matters of importance which some persons or organizations wish to keep secret. The three basic elements are that the investigation be the work of the reporter, not a report of an investigation made by someone else; that the subject of the story involves something of reasonable importance to the reader or viewer; and that others are attempting to hide these matters from the public. (Ullmann & Colbert, 1991: p. xvi)

Although this definition does list several aspects that the major stories in the American history of investigative journalism have in common, it has also been criticized for being too limited. Especially the last element, the attempt of others to hide matters from the public, is subject to vigorous debate. Some stress the role of the investigative reporter as someone who exposes deliberately concealed information, while others deny that secrecy is necessary. The authors of the book *Investigative Journalism* (1976), David Anderson and Peter Benjaminson stated that 'investigative reporting is simply the reporting of concealed facts.' (As cited in Protess, 1991) The suggestion that some people are deliberately trying to conceal information that is important to the public also brings focus to the moral aspect of investigative journalism. Especially in the early muckraking stories but throughout the history of investigative journalism, reporters tried to make a clear distinction between the good guys (the public) and the bad guys (often big corporations and government agencies). According to Protess the motive of investigative reporters is to advocate changes in society. 'They seek to improve the American system by pointing out its shortcomings.' (Protess, 1991) This clear moral stance is hardly ever made explicit in investigative stories but is often hidden in the narrative through the use of explicit terms like 'corrupt', 'wasteful' and 'greedy'.

Although this focus on deliberately concealed information does account for some of the most exciting investigative stories, some would say this definition is too narrow. Eugene Roberts, editor of *The Philadelphia Inquirer*, denounces this definition because of its connotation with scandals. 'One of the reasons I don't much use the term 'investigative reporting' is that it misleads and confuses. To many people, investigative reporting means nailing a crook or catching a politician with his pants down. This, I think, is too narrow a definition.' (Roberts, 1988: p. 12) The focus on scandals does indeed shift the attention from some of the other aspects of investigative journalism. I would therefore rather turn to the definition as put forward by Everette Dennis and Arnold Ismach. They defined investigative reporting by stating that it should focus on 'all sectors of society that require examination, explanation or airing, but that are hidden from the public view.' (Dennis & Ismach, 1981: p. 81)

This definition is also consistent with the three kinds of investigative journalism proposed by the Dutch association of investigative reporters, the Vereniging van Onderzoeksjournalisten (VVOJ). Besides revealing scandals they see it as the task of the investigative reporter to examine how governments, companies and other organizations function. The third kind of investigative journalism the VVOJ describes is the tracing of social, economic, political and cultural trends in society. (Van Eijk, 2005: p. 22) Describing changes in society is not usually associated with investigative reporting but is becoming increasingly important. This thesis will argue that the Web opens up a whole new field for investigative reporters to analyze societal trends.

1.2 From documents to databases

Documents have always played an important role in investigative journalism. Dick van Eijk explains the importance of documents by claiming they contain 'the ultimate facts'. 'They speak for themselves, literally, without a journalist between them and the audience. They are the textual equivalent of photographs, which in the eyes of many have truthfulness of their own.' (Van Eijk, 2005: p. 15) One of the muckrakers, Ida Tarbell, spend much of her time researching archives. She examined the records of government investigations both meticulously and relentlessly. As historian C.C. Regier described her approach:

> Three years passed before she was ready to begin writing; five years elapsed before her investigations were completed. In that time she had mastered the history of the company, penetrated the technical intricacies of her material, pieced together the evidence, and prepared herself to tell the complicated story in such a way that the average man and woman would understand it. (Regier, 1932: p. 123)

With the arrival of the computer some of Tarbell's methods might not be duplicated today. Governments at all levels have turned to storing their records on electronic media. New skills are required to access and understand all the information available.

The introduction of computers into the newsroom took place in the early days of television. One of the earliest examples of computer-assisted news occurred during the presidential election of 1952 between Dwight D. Eisenhower and Adlai E. Stevenson. A computer was programmed to predict the outcome of the election on basis of the early returns. With only seven percent of the votes counted it predicted a landslide victory for Eisenhower which no one believed since it was supposed to be a very close contest. Eventually the final count was unbelievably close to the early predictions made by the computer. (Cox, 2000) Around the 1980s newsrooms also adopted online databases. When these developments came together, reporters began to have access to 'resources that would change the nature of news reporting considerably.' (DeFleur, 1997: p. 37) Besides increasing the

amount of information available to develop a news story, it also made reporters more familiar with software for spreadsheet analysis, database construction and tools to discover statistical manipulation. Skills that would become incredibly important since governments and other institutions also began to use computers for record keeping. (DeFleur, 1997: p. 33)

All these technological innovations paved the way for a new kind of journalism. Philip Meyer coined the term *precision journalism* for a new approach to fact gathering, analysis and reporting by adapting social science research methods in the practice of journalism. Meyer's book *Precision Journalism: A Reporter's Introduction to Social Science Methods* suggested that quantitative techniques from the social sciences such as surveys and public opinion polls could become useful tools by which reporters could more adequately perform their roles. Meyer believed in pushing journalism toward science by 'incorporating both the powerful data-gathering and –analysis tools of science and its disciplined search for verifiable truth.' (Meyer, 1991: p. 5) According to Meyer both social sciences and journalism attempt to develop accurate descriptions of reality. However, journalists depend too much on qualitative techniques founded in personal intuitions and suspicions which make their accounts of the 'world outside' inaccurate. (Lippmann, 1922) Methods of the social sciences on the other hand were founded on 'theory development, measurement, quantitative data analysis, formal research designs, and a probabilistic epistemology' (DeFleur, 1997: p. 37) which provide more accurate tools for describing reality.

Meyer's main focus is on qualitative techniques which would be very difficult to apply for journalists without the help of computers. Hence is precision journalism closely related to computer-assisted investigative reporting; a form of journalism that applies computers to the practice of investigative reporting. For example, computers were used to analyze electronic records from governments in order to find newsworthy situations. One of the pioneers in this field, David Burnham, developed a conceptual framework to use when investigating the records of a government or public agency. When conducting a data analysis Burnham asked two key questions: (a) What is the stated purpose or mission of the agency? (b) What problems or procedures prevent the agency from achieving its stated purpose or goal? (Walker, 1990) Philip Meyer carried this concept on to the second step by stating that many stories derived from databases can be developed by comparing groups or subgroups as a simple first step. (Meyer, 1991) With these methods journalists were able to test popular theories. For example, following the civil riots in Detroit in 1967 survey data from African Americans who lived in the riot area were analyzed by computer. Popular believe was that those who participated in the riots did so because they were at the bottom of economic ladder and poorly educated. The analysis, however, showed that people who attended college were just as likely to participate in the riots as those who failed to finish high school. (DeFleur, 1997) The number of stories that have been uncovered through database analysis has since only increased and is to this day an important part of investigative journalism.

The introduction of the Internet marked another major development in the journalistic profession, especially in investigative reporting. While at first the Internet was mainly used among journalists for communicating with each other, it quickly became more common to use the Internet for conducting research. (Cox, 2000) Barry Sussman called the Web the 'investigative reporter's tool.' As he explains:

> What the Web does incomparably well is to provide information —instantly— on just about anything. Want to know about where there have been cases of bird flu? Or what can go wrong with voting machines? […] Googling not only provides answers, but it connects reporters and anyone else with possible places and sources to go to find out more. (Sussman, 2008: p. 45)

As the pressure on journalists to produce only increases, the Web can proof to be an incredibly useful tool. Naturally Sussman understands that information on the Web can be unreliable as he for example points out that a number of edits on Wikipedia that have been traced back to the CIA. Yet he argues that it is up to the reporter to determine how trustworthy the information is. 'There are plenty of reliable, dedicated groups and individuals responsibly sharing important information through the Web.' (Sussman, 2008: p. 45) Therefore Sussman sees using information on the Web (or the wisdom of the crowds) for journalistic purposes as one of the few ways in which news producers can continue and maintain their essential watchdog role. However, he also points out that few investigative assignments should be completed online and he underlines the importance of working with actual sources; people who have stories to tell and documents to back up what they know.

1.3 Towards online grounded investigative reporting

In the previous paragraphs I basically tried to sketch the two main fields in which investigative reporters operate nowadays. First there is the traditional approach which consists of finding sources, persons or documents, in the material world often with the aim to expose some kind of scandal. The traditional approach and its associated strategies still form the baseline against which investigative reporting is understood. The introduction of electronic and later on online databases opened up a whole new field for investigative journalists. Databases supplemented with analytical software gave investigative reporters the opportunity to closely examine official records from government and public agencies. This approach is especially useful to examine governments', companies', and other organizations' policies or functions. These two approaches are well-established in investigative reporting and the associated strategies and responsibilities have been documented in diverse literature. What I, however, propose in this thesis is a whole new approach, one closely related to the last kind of investigative journalism proposed by the VVOJ. I suggest to turn to the Web in order to describe social, economic, political and cultural trends in

society. Of course, as described in the previous paragraph, the Web is already widely used within journalism. However, my approach to the Web is bit more ambitious than to simply check sources and find information online. I suggest seeing the Web itself as a field for investigative reporting with unique possibilities to examine society through computational means. This means not just *looking* at the Web for relevant information but *using* the Web for investigative purposes. As I will further explain in the course of this thesis this means using the way information if organized online to make statements about society.

I use the term *online grounded investigative reporting* (OGIR) for the type of journalism that tries to examine society and culture on the Web using computational tools. The type of journalism I propose basically sets itself apart with two presuppositions. First of all it moves beyond the real/virtual dichotomy in which the Web is seen as a world separated from the 'real world'. Online grounded investigative reporting is grounded in the online but tells us something about our society and our culture. Since more and more of our social and cultural activities migrate to online environments, I suggest that with the proper strategies the Web can be appropriated as a new field for investigative reporters. Although I want to move past the distinct separation between the online and offline, I also believe we have to take into account some of the fundamental differences between the two. We have to be aware of how information on the Web is organized, structured, interrelated, presented and retrieved. The second presupposition of online grounded investigative reporting is therefore that not only the content on the Web can tell us something about the 'real world' but also the way information is organized. Information, knowledge and sociality on the Web are organized by certain recommender systems such as Google. By utilizing the distinct features of these systems we can for example say something about the popularity of a certain topic. We look at Google but see society.

In the next chapter I will put these two presuppositions in a more theoretical framework and further explore the way information is organized on the Web. The historical framework provided in this chapter should make clear that investigative reporters should not be afraid of new technologies but fully embrace their potential. On the other hand, journalists should be aware of lessons from the past. With the introduction of the Internet the need for creative and analytical thinking has not become outdated. In contrary, as the next chapter will show, the Internet has complicated things in many ways. The need for skilful reporters to interpret these new technologies is greater than ever.

2
THE ORGANIZATION OF INFORMATION ON THE WEB

Digital technologies, like the Internet, present us with new opportunities that in many cases ease the constraints previously experienced by investigative reporters using traditional methods. In the previous chapter I already gave some examples of how computers transformed the newsroom. Especially the Internet, with its 24-hour accessibility and instantaneous communication, provides an enormous data pool from which journalists can draw. Huge quantities of data are not something of which journalists should be afraid. Philip Meyer argued in his book *The New Precision Journalism* in relation to using computers to investigate public records that '[t]he quantitative change in the amount of time and effort to search and link such records has led to a qualitative shift in the things that journalists can discover.' (Meyer, 1991: p. 238) I argue that the same qualitative shift will occur if investigative journalists would better utilize the possibilities of the Internet. However, these promising attributes of the Internet must also be looked at with caution. When conducting online investigations journalists should be careful in choosing an effective method rather than simply applying offline methods to online environments. Christians and Chen suggest that the 'direct application of traditional research methods to cyberspace often generates difficulties in execution or unsuccessful results. The Internet presents a new technology for research, and the formal features of the Web need to be taken into consideration when implementing it as a research strategy.' (Christians and Chen, 2004: p. 19) Exactly these features I want to discuss in this chapter by looking at the ways information is organized on the Web. This means not just looking at content of the Web but also taking into consideration some of its more structural elements.

2.1 The real/virtual dichotomy

The Internet has persistently been seen as a virtual realm apart from the real world. This is also the reason why people ascribe the Internet its revolutionary potential. It is claimed to offer new opportunities for redefining identity, citizenship, community or even democracy. (Barlow, 1996; McNair, 2006; Turkle, 1995) This notion of the Internet as utopian space separated from the 'real world' has been widely critiqued by numerous scholars. Communications scholar Steve Jones was one of the first to propose to move beyond the perspective of the Internet as a realm apart. (Jones, 1999) *The Virtual Society?* (1997-2002) further critiqued the digital divide model with a number of empirical studies. They argued in relation with the real/virtual dichotomy that virtual interactions not substitute but rather supplement the 'real'. Identities are grounded in both the offline and the online. (Woolgar, 2002) A scientific approach to study culture through computational means, closely related to the kind of

journalism I propose, is the Digital Methods Initiative, supervised by Richard Rogers. He proposes a research practice which 'grounds claims about cultural change and societal conditions in online dynamics.' (Rogers, 2009: p. 5) He does not consider the Web a virtual space separated from the 'real world'.

> It concerns a shift in the kinds of questions put to the study of the Internet. The Internet is employed as a site of research for far more than just online culture. The issue no longer is how much of society and culture is online, but rather how to diagnose cultural change and societal conditions using the Internet. The conceptual point of departure for the research program is the recognition that the Internet is not only an object of study, but also a source. (Rogers, 2009: p. 8)

The goal of Digital Methods is therefore to re-conceptualize the relation between online and offline culture. Rogers introduces the term online groundedness as the effort to make claims about cultural and societal change grounded in the online. In the same way I propose a type of investigative journalism that is grounded in the online. Hence, I will use the term online grounded investigative reporting for the kind of journalism described in this thesis.

Although I argue to move beyond the absolute distinction between real and virtual in investigative journalism, this does not mean that offline journalistic practices should simply be applied to online environments. Rogers makes a distinction between the natively digital and the digitized: 'that is, between the objects, content, devices and environments 'born' in the new medium, as opposed to those which have 'migrated' to it.' (Rogers, 2009: p. 5) One example of how the application of offline methods to online environments can fail is the election of Dutchman of the year 2005 organized by the Dutch newspaper *de Volkskrant*. The paper used an online survey which was subsequently hijacked by the followers of the popular weblog Geenstijl.nl who massively voted for Rachel Hazes, the widow of a famous Dutch singer. (Geelen, 2005) The results of the survey were clearly untrustworthy which shows how easy it is to manipulate online surveys. The approach I therefore propose is to look at the natively digital, the elements 'born' in the Internet such as recommendation systems, hyperlinks and search engines. In this chapter I will look at how the Internet is organized through these kind of objects and how these can be used for investigative journalism. Without a clear understanding of how the Web is organized it is impossible to conduct solid investigations.

2.2 Some unique properties of the Web

Marshall McLuhan argues in his most well-known article "The Medium Is the Message" that to understand media we should look past the content and focus on how the medium itself influences our actions.

> It is the medium that shapes and controls the scale and form of human association and action. The content or uses of such media are as diverse as they are ineffectual in shaping the form of human association. Indeed, it is

> only too typical that the "content" of any medium blinds us to the character of the medium. (McLuhan 2003, p. 203)

Although McLuhan's ideas have been widely criticized for its techno-deterministic tendencies, it does bring focus to the ways in which communication technologies shape our way of thinking. Technologies are never neutral and therefore all research using communication technologies like the Internet must be familiar with concerns of technology as a whole. The main task while using Internet technologies in research is to discover their unique properties. As Christians and Chen argue: '[T]he intellectual challenge is to identify the distinguishing properties of particular media technologies such as books, cinema, satellites, and the Internet. Regarding television or radio of fiber optics, communications scholars must work deeply into their fundamental properties in order to know them distinctly as their own. (Christians and Chen, 2004: p. 20) I therefore propose that reporters doing online grounded investigations should use a medium-specific approach which accounts for these distinguishing properties of the Web.

One fundamental properties of the Internet with special relevance to online grounded investigative journalism is its 'ephemerality' or, in other words, the transient nature of the Web. Harold Innis, McLuhan's precursor at the University of Toronto, studied the introduction of communication technologies like papyrus, the printing press and radio and noticed a bias regarding space and time. He argued that oral communication systems are biased towards time while print systems are biased towards space. (Innis, 1952) In this respect, Steven Schneider and Kirsten Foot talk about the sense of permanence that clearly distinguishes the Web from oral media such as live television and radio. Just like print media, Web content has to exist in a permanent form in order to be transmitted. However, the Web also differs from other permanent media.

> [T]he permanence of the web is somewhat fleeting. Unlike any other permanent media, a website may destroy its predecessor regularly and procedurally each time it is updated by its producer; that is, absent specific arrangements to the contrary, each previous edition of a website may be erased as a new version is produced. By analogy, it would be as if each day's newspaper was printed on the same piece of paper, obliterating yesterday's news in order to produce today's. (Schneider and Foot, 2004: p. 115)

It can therefore be argued that, while oral cultures privilege time over space and print systems are biased towards space, electronic culture dislocates us from both space and history. 'It ruptures historical consciousness and disconnects us from our geographical home in mediating structures.' (Christians and Chen, 2004: p. 21)

Another fundamental property of the Web is its traceabillity. The Internet stands out from other media as a site where social life is made traceable. (Thrift, 2004) The Web basically is a vast archive containing

information on countless of social interactions. Whether it be the comment panels of a popular weblog or the formation of networked content on Wikipedia, it al becomes visible on the Web. These traces of social live can be tremendously valuable for investigative journalism. Especially since the Web as an archive full of traces of social live is not only huge in size but also searchable. Navigating through huge amounts of information becomes possible thanks to search engines. However, as I will argue in the next paragraph, working with very large datasets also poses some risks.

2.3 The computational turn

New approaches and methodologies are being developed for the study of culture and science with huge datasets as provided by the Internet. Chief-editor of *Wired Magazine* Chris Anderson argued in 2008 for a methodological turn in science. In the article "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" he claims that since it is possible to analyze information at petabyte scale, scientific models have become obsolete.

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (Anderson, 2008)

However, these statements should be looked at with great caution. Unlike Anderson claims, numbers do not always speak for themselves. As is argued in a response to Andersons bold claims, it is very easy to find false correlations when dealing with petabytes of data. 'In a huge amount of data, you *expect* to find tons of patterns. Many of those patterns will be statistical coincidences. And it can be very difficult to identify the ones that are.'(Chu-Carroll, 2008) How then should we deal with these huge quantities of data? New media scholar Lev Manovich has announced Cultural Analytics as a way to analyze the huge quantities of user-generated content on the Internet as well as other cultural artifacts. He claims that these huge quantities of data will turn 'culture into data' and make possible the study of 'culture as data.' (Manovich, 2007: p. 4) Instead of a qualitative content analysis, Manovich' focus lies on finding patterns. By visualizing the results it becomes possible to see 'real-time cultural flows around the world.' (Manovich, 2009: p. 3)

Another way of doing social research on the Web is the previously discussed Digital Methods Initiative. In a previous paragraph I already talked about how Digital Methods aims to re-conceptualize the relation between offline and online culture. Furthermore, Digital Methods is an attempt at social research on the Web where the

devices, such as Google, Twitter or Wikipedia, themselves are a part of the analysis. As Rogers puts is: 'We look at Google results and see society, instead of Google. That is a shorthand way of saying that we see institutions and issues in the ranked lists that are returned in the search results.' (Rogers, Stevenson and Weltevrede, 2009: p. 1) Both Digital Methods and Cultural Analytics deal with large quantities of information. This is an aspect of both methodologies that has to be critically examined. Especially Cultural Analytics is vulnerable to the critique posed by Chu-Carrol since with this program datasets are approached in a very open way without formulating a research question or hypothesis. It is all about the patterns. Digital Methods is different in this respect because it is not following the quantity but the methods of the Web. Digital Methods focuses on how 'information, knowledge and sociality [on the Web] are organized by recommender systems – algorithms and scripts that prepare and serve up orders of URLs, media files, friends, etc.' (Rogers, 2009: p. 12) This requires examining how digital objects like the permalink, IP address, URL, etc. are handled on the Web. The devices used by large popularities on the Web are the main focus of the research. I propose the same approach in online grounded investigative journalism. The medium once again becomes the message but not in the techno-deterministic fashion of McLuhan. Instead, as Rogers proposes, we should follow the medium to ultimately find society.

2.4 The nature of objects

When following the methods of the Web one should be very precise in examining the nature of the objects at play. Different studies have documented how new communication technologies, like email and the Internet, have altered the practice of journalism. (Cochran, 1997; Dumlao and Duke, 2003) Some authors claim that these shifts in the newsroom can best be understood using the actor-network theory (ANT) described by Bruno Latour. (Plesner, 2009; Turner 2005) Their claim is underpinned by one of the fundamental characteristics of ANT; objects too have agency. Journalists doing research on the Web should be aware that the tools they use and the objects they study are not simple intermediaries but instead active mediators. The interaction between humans and non-human objects in the newsroom is not a new trend. The use of the telephone or even earlier the telegraph is an example of this. What is however new, is the seamlessness with which they are now part of newsroom. Plesner concludes her article as follows:

> [O]ne thing seems fair to identify as a new or changing condition of newswork in relation to new ICTs, namely their seamlessness. Journalists – and other actors involved in newswork – seem to approach them with little suspicion and little critical distance as compared to the situation a decade ago. Humans and their communication technologies have become integrated more tightly. (Plesner, 2009: p. 624)

A critical perspective on the relation between new media and online grounded investigative reporting is therefore essential. I will introduce some of the central concepts of ANT to get a better understanding of this complex relationship. As a social methodology for mapping controversies ANT can, I believe, offer some valuable insights for investigative journalists using the Web. Bruno Latour explained using ANT once as 'just look at controversies and tell what you see.' (Venturini, 2009: p. 2) This statement could just as easily be applied to investigative journalism.

ANT can be seen as an attempt to rethink sociology by moving away from social explanations. Latour criticizes 'the sociology of the social' for using social mechanisms as an explanation for all kinds of activities. Instead he is more interested in how different, heterogeneous actors create associations. One of the most important contributions in this light is the proposition that the type of agencies participating in interaction remains wide open. Objects should be considered mediators in the sense that 'their input is never a good predictor of their output; their specificity has to be taken into account every time.' (Latour: 2005: p. 39) Agency is not restricted to humans since the only question one needs to ask about any agent is the following: 'Does it make a difference in the course of some other agent's action or not?' (Latour, 2005: p. 71) Technological devices such as computers and mobile phones but also search engines and recommender systems influence certain practices.

> ANT is not the empty claim that objects do things 'instead' of human actors: it simply says that no science of the social can even begin if the question of who and what participates in the action is not first of all thoroughly explored, even though it might mean letting elements in which, for lack of a better term, we would call non-humans. (Latour, 2005: p. 72)

When a group of heterogeneous actors, humans and non-humans, are associated they make up 'actor-networks'. These actor-networks are not fixed entities but are always in the making. Venturini uses magma as a metaphor to show how collective life is constantly melted and forged. (Venturini, 2009: p. 7) Connections and associations are formed all the time and ANT is interested in tracing these concrete links. This results in a practice that is quit common in journalism but not as obvious in the social sciences: follow the actor. The goal is to create accounts based on how actors themselves feel they relate to other actors. As Latour puts it:

> …either we follow social theorists and begin our travel by setting up at the start which kind of group and level of analysis we will focus on, or we follow the actors' own ways and begin our travels by the traces left behind by their activity of forming and dismantling groups. (Latour, 2005: p. 29)

Because of its call for a focus on heterogeneous actors (humans and non-humans), I feel ANT offers some useful insights for journalists. As Plesner showed with her research, journalists approach new media with little critical

distance. On the one hand this means that most journalists are so comfortable with using search engines, email etc. that it goes without saying. However, in the online grounded investigative reporting that I propose, is critical distance towards the tools used in the investigation of great importance.

In this chapter I tried to explain the importance of a medium-specific approach when dealing with the Web. A clear understanding of the way the Internet generates, organizes and manipulates information as well as the role online objects play in this process, is vital for conducting solid online investigations. The considerations put forward in this chapter will certainly come back in the following chapters in the second part of my thesis. In these chapters I will discuss how three fundamental elements of the Web, the Website, the link and the search engine, can be used for online grounded investigative reporting. This more analytical part will provide an inventory of possible techniques and methods that can be used by investigative reporters interested in the online.

PART II

3
THE WEBSITE

Investigations into Websites by journalists have long been dominated by rhetorical analyses. Journalists focus often solely on the content with little or no attention for its structuring elements. An investigation by the Dutch radio program *Argos* used for example the Websites of local newspapers in the US, archives and military sources to reconstruct how many casualties there were among American Special Forces in Afghanistan and Iraq. (Jaspers, 2004) Since there was no official number of casualties known, the Internet provided valuable information that was used to reconstruct the story. The way the Internet is used in this example is basically no different from how paper sources are used. In fact the investigation into the number of casualties among Special Forces could have been conducted by looking at the paper editions of the local newspapers. The only difference is that, since all the information was digitized, the investigation could be done in way less time. However, as I have argued in the previous chapter, the way information is structured on the Web differs in some fundamental ways from other media. The ephemeral nature of Websites has to be taken into consideration when using the Web as a site of investigation. Although rhetorical analyses of Websites by investigative reporters can be very useful, I propose to involve both the content and the structuring elements of the Web in the investigation. Because of the transient nature of the Web, saving Websites is not an easy task. There are, however, a couple of initiatives that aim to archive the Web, most notably the Internet Archive and its Wayback Machine. With the Wayback Machine investigative reporters can use the ephemeral qualities of the Web to come up with new stories. In this chapter I will begin with discussing the difficulties of archiving Websites before turning to the ways Internet archives can be used in investigative journalism.

3.1 The Internet Archive and the Wayback Machine

The Internet Archive was created in 1996 as a digital library of Webpages and other digitized cultural artifacts with free access for both researchers and the general public. Five years after its creation the Internet Archive introduced the Wayback Machine, a service that allows people to visit archived versions of Websites. The Internet Archive is collecting the Webpages with search engine technology called Alexa Crawl. This technology crawls the estimated 16 million archived Websites every two months. The archive currently contains almost two petabytes of data and is growing at a rate of 20 terrabytes per month. ("Internet Archive Frequently Asked Questions", 2010) Since its introduction the Internet Archive has become an important source of information not only to researchers but also in legal cases ranging from copyright and trademark infringements to business torts and defamation. In

the United States the court ruled in 2004 that Webpages taken from the Wayback Machine are admissible as evidence. (Gelman, 2004)

However, before using the Wayback Machine as a resource for information one should realize that the Internet Archive as an object is formed by the archiving process. Web archiving scholar Niels Brügger does indeed argue that '[U]nlike other well-known media, the Internet does not simply exist in a form suited to being archived, but rather is first formed as an object of study in the archiving, and it is formed differently depending on who does the archiving, when, and for what purpose.' (Brügger, 2005: p. 1) In case of the Wayback Machine this is relevant when one looks at how the archived pages are retrieved.

> [W]hat stands out for everyday Web users accustomed to search engines is not so much the achievement of the existence of an archived Internet. Rather, the user is struck by how the Internet is archived [by the Internet Archive], and, particularly, how it is queried. One queries a URL, as opposed to keywords, and receives a list of stored pages associated with the URL from the past. In effect, the Internet Archive, through the interface of the Wayback Machine, has organized the story of the Web into the histories of individual websites. (Rogers, 2009: p. 16)

The Wayback Machine makes it possible to investigate the evolution of a single or multiple pages over a period of time. This evolution may give clues about both broader developments both on- and offline. By looking at what Niels Brügger calls a site's 'textual grammar', from its structure and features to metaphors used for navigation, (Brügger, 2009) one can find stories that tell something about both online culture and offline society.

One of the clearest examples of this is a sample project by the Digital Methods Initiative. In "Google and the politics of tabs" they used the Wayback Machine to look at the evolution of the Google front-page from 1998 until late 2007. They took snapshots of the different versions of the front-page which were turned into a movie. It became clear that the Google front-page remained remarkable stable with a search box and two buttons: 'Google web search' and 'I'm feeling lucky'. However, in the upper left corner there are several tabs that refer to different services provided by Google. By looking at which tabs have risen to a prominent place at the front-page and which tabs have been relegated to the 'more' or 'even more' buttons, one can say something about which services are privileged by Google. 'It tells the story of the demise of the directory, and how the back-end algorithm has taken over the organization of Web information at the expense of the human editors and the librarians.' ('Google and the politics of tabs', 2010) The Google directory is a service provided by Google that organizes the Web by topic via human editors. The demise of the directory does not only tell us something about a shift of focus in Google's business strategy. There is a larger issue at stake, namely the demise of the human editor. This is something that not only implicates online environments but also reflects in the real world. We rely

in everyday life more and more on computerized processes. The clearest example probably being the library, instead of asking a librarian for a book we just use Google's algorithm to find it. What this sample project by the Digital Methods Initiative showed us is that it is possible to make statements about society by looking at the evolution of a Website.

3.2 The tracing of societal trends on the Web

Also in investigative journalism the opportunities the Wayback Machine offers have been recognized. One example is an article in *NRC Handelsblad* on the rise of extremist language on sites in the Netherlands. (Dohmen, 2007) Instead of analyzing pamphlets or interviewing extremists they analyzed 150 right-extremist sites over the course of a decade. They used the Wayback Machine to find and compare different versions of the sites. Unfortunately the article gives little or no clarity about how he conducted this research. However, one possible method would be to scrape the pages of the rightwing and extremist sites from the Internet Archive, place the text and/or images in a database and systematically query it for the presence of certain keywords over a period of time. The journalist used keywords like 'hoernalisten' (derogative term for journalists), 'multikul' (a sneer towards the multicultural society) and 'linkse kerk' (referring to leftwing parties). The newspaper found that right-wing sites were increasingly employing more extremist language.

Besides analyzing the language used on the sites the article also did more quantitative research. Interesting in this respect is the correlation they found between certain major events, like the attacks of 9/11, the rise of Pim Fortuyn, the political murders of both Fortuyn and Theo van Gogh and the rise of Geert Wilders, and the number of extremist sites. They found that there was and explosion in the number of rightwing and extremist sites after these events. Thus, by looking at the online, through the analysis of particular sets of archived sites, they were able to make statements about cultural changes. Their claims were online grounded as the Websites became the baseline against which society was judged. In the article the relation between the on- and offline was put as following: 'The Internet seems to be impacted by the social hardening in the Netherlands. But possibly the Internet also spurs this social trend, namely because of the anonymity.' (Dohmen, 2007) (Own translation) Naturally it is difficult to be concrete about the exact relation between on- and offline cultures. However, one solution I would propose is to stay away from absolute numbers and instead to focus on trends. It is rather meaningless to talk about the number of extremist sites that are in the air, especially as many of these are very short-lived. As the article notes, half of the researched 150 sites have already been abandoned. It would therefore make more sense to look at these sites over a larger period of time. Whether it is about the increase in the number or sites or about the increasing extremist language, only when looking at it from a historically perspective you are able to make statements, as the article does, about the relation between on- and offline events. The

Internet Archive and the Wayback Machine are therefore crucial tools for investigations into Websites. This kind of investigative reporting on the Web fits perfectly in the third kind of investigative journalism as defined by the Vereniging voor Onderzoekjournalisten: the tracing of social, economic, political and cultural trends in society.

I do, however, also want to point to yet another way the Wayback Machine is used in the article. The found out that the person behind the only Dutch right-extremist Web hosting service, eigenvolkeerst.com gave a false name during the registration process. But in the archived version of 2004 of the site the URL robbierob.nl was visible in the title bar. A domain that is registered to Rob Copier, a prominent figure in the extreme-right movement. Even though this URL was quickly removed from the title bar, with the help of the Wayback Machine the initiator of eigenvolkeerst.com could still be traced. Here we see how the Wayback Machine can be used in a more traditional form of investigative journalism, namely to expose something that is being kept secret. This shows also one of the main advantages of research on the Internet, every action leaves traces. Something that also the Dutch royal family found out the hard way. The *NRC Handelsblad* used the Wikiscanner software, a technology that connect edits made on Wikipedia to IP addresses, to show that Dutch Prince Friso and his wife Mabel Wisse-Smit edited the Wikipedia entry about a scandal that had forced the prince to renounce his claim to the throne. With the Wikiscanner this edit was traced back to the Royal Palace of Queen Beatrix. The couple later confirmed that they were visiting the queen at the moment when the edit was made. (Verkade, 2007)

While online grounded investigative reporting is still in its infancy, the article in *NRC Handelsblad* about the rise of right-extremist sites in the Netherlands already showed its potential. I would like to argue that the Internet Archive makes an excellent source for investigative reporters wanting to make claims about society grounded in online environments. The next chapter will focus on the relations between Websites. Through the use of hyperlinks it becomes possible to map connections between different sites.

4
THE LINKS

Networks have always been tremendously interesting to investigative reporters. Questions like who knows who and where is the power located, have intrigued journalists since the beginning of the trade. The Dutch newspaper *de Volkskrant* has compiled since 2006 every year a Top 200 of the most influential people in Dutch society. The Top 200 consists of a group of people who won their spurs in politics, public administration or business and who are now active in boards of directors, committees and councils. The list ranks the most influential administrators of a database of eight thousand people through a network analysis developed by the Dutch communication scientist Wouter de Nooy. The starting point being that someone high on the list has more chance to influence government policy. In the article the used method is described as follows: 'A person's influence score is calculated on basis of his or her proximity to the chairmen of a number of organizations that influence government policy the most. This is measured in the number of intermediaries in the network.' ("Volkskrant Top 200, editie 2009", 2009) A series of confidential interviews with well-informed administrators were held to determine the number of intermediaries in the network.

Another example of network analysis in investigative journalism was published in the Flemish weekly *Trends*. In their attempt to map the Belgian economical elite, this magazine used a similar approach as *de Volkskrant*. They formed a database with 2000 administrators from 200 Belgian businesses. With a computer program they conducted a matrix analysis to determine interrelations between businesses and administrators. *Trends* used their research to conclude that 'Belgium no longer has a cluster of connected, strong businesses. The financial fragmentation is complete'. (Carbonez & Brockmans, 2004) What is interesting is the fact that both *de Volkskrant* and *Trends* postulate that the number of connections in the network determines how influential or powerful a person or company is. Although this presumption can be criticized on a number of grounds, someone can still exercise influence without having a lot of additional positions, it might be interesting to take this approach to the Web. Instead of using a predetermined database, I argue that this type of research could also be done with Web-based content. By adapting the described network analysis to the specificities of the Web we can look into how different entities may be characterized by the types of hyperlinks given and received. The Digital Methods Initiative already found certain linking tendencies among domain types.

> '[G]overnments tend to link to other governmental sites only; non-governmental sites tend to link to a variety of sites, occasionally including critics. Corporate websites tend not to link, with the exception of collectives of them – industry trade sites and industry 'front groups' do link, though. Academic and educational sites typically link to

partners and initiatives they have created. Taken together, these linking proclivities of organizational types display an everyday 'politics of association'.' (Rogers, 2009: p. 14)

These 'everyday politics of association' might proof very interesting material for news stories. In this chapter I want to discuss in what ways online network analysis can be used for investigative journalism.

Figure 1: ACTA issue network on the Web, March 2010 (Generated with the Issue Crawler)

4.1 Finding networks

In order to map an online network via hyperlinks, the Digital Methods Initiative has developed a piece of software called the Issue Crawler. The Issue Crawler follows hyperlinks to map a network of sites. Its built-in analytical approach is co-link analysis, whereby a site must receive at least two inlinks from the other sites in the network to be included. In order to use the Issue Crawler you first need a starting point in the form of a list of seed URLs. One has to define the key actors around a certain issue. These can be defined with rather traditional journalistic methods. In another research paper I tried to map the network around the ACTA (Anti-counterfeiting

Trade Agreement) negotiations. My starting points consisted of the Websites of the organizations who replied to the request by the US Trade Office to comment on the ACTA negotiations since they were clearly affected by the issue. This list was provided by the Public Knowledge website and consisted of both the websites of public interest groups and industry lobby groups. (Kamsma, 2010: p. 7) Another method to find your starting points could be to rip the outlinks from three key websites, triangulate these with each other to determine which sites are in at least two of the networks. Once triangulated, you have got a list of the top actors around a certain issue. The Issue Crawler can be used to crawl these specified starting points. It will capture the starting points' outlinks, and performs co-link analysis to determine which outlinks have at least two starting points in common. The crawler repeats this step two times and will visualize the results. Figure 1 shows an example of such a network. The figure shows the issue network around the ACTA negotiations I used in my research. However, before turning to the ways these networks can utilized in investigative journalism, we first need to clearly define what kind of network we are exactly dealing with.

In the article "Recipe for Tracing the Fate of Issues and Their Publics on the Web" Noortje Marres and Richard Rogers describe how public debate plays out on the Internet. By following the hyperlinks between Websites related to a certain issue they construct, what they call, issue-networks.

> Following hyperlinks among pages dealing with a given issue, we found that these links provided a means to demarcate the network that could be said to be staging the controversy in the new medium. Those Webpages which treated the issue, and which received a significant number of links from other pages presenting the affair, we decided, disclosed the controversy on the Web. Thus, we came to focus on sets of inter-linked pages that treated the affair in question, dubbing them issue-networks, as our most useful unit of analysis. (Marres & Rogers, 2005: p. 1)

It is very tempting to characterize the online activity around certain issues as public debate. However, as Marres and Rogers point out, the links between the different sites in the issue network do not so much represent a constructive dialogue between the different actors. 'Acknowledgements of other sites, by way of hyperlinks, characteristically are one-way recognitions, whereby the sender of the link 'frames' the site of the receiver.' (Marres & Rogers, 2005: pp. 2-3) Instead of talking to each other, the different websites are rather trying to define the issue by building from or countering the issue-definitions by other actors in the network. Issue-networks can therefore best be defined as a set of heterogeneous actors (government agency's, NGO's, industry lobby groups etc.) mobilized around a certain issue. This is also apparent in figure 1, the issue-network around the ACTA negotiations. You see government agency's (for example the US Trade Office and the site of the White House), public interest groups (for example Public Knowledge and the Electronic Frontier Foundation) and industry lobby

groups (for example the Motion Picture Association of America) all represented in the network. The Issue Crawler is in this respect a great indicator for which groups or organizations have something at stake in a certain controversy.

4.2 Mapping controversies

As also became clear in the previous paragraph, the Issue Crawler and its associated methods are very useful for mapping controversies. It therefore offers a lot of potential for investigative journalism. Investigative reporters are always looking for conflicting interests and stories since those are an indication that there is something going on. The social researcher Tommaso Venturini defines a controversy as follows:

> [T]he definition of controversy is pretty straightforward: controversies are situations where actors disagree (or better, agree on their disagreement). The notion of disagreement is to be taken in the widest sense: controversies begin when actors discover that they cannot ignore each other and controversies end when actors manage to work out a solid compromise to live together. Anything between these two extremes can be called a controversy. (Venturini, 2009: p. 4)

In his article "Diving in magma: how to explore controversies with actor-network theory" Venturini looks at how controversies can be used for social research. He warns the reader from the start that mapping controversies is no piece of cake. One of the reasons for this is that '[c]ontroversies display the social in its most dynamic form'. (Venturini, 2009: p. 5) New actors come and go and often it is not quite clear where these actors stand in the conflict. Definitions of the particular issues are constantly shifting and the actors often don't seem to agree on anything. In other words, to tresh out a controversy is not an easy task and often requires a lot of time and effort. As many handbooks on investigative journalism will tell, the first step in an investigation should be to demarcate the network. (Kussendrager, 2007) Which people are involved and how do they relate to another? It's especially in this first phase of the investigation that the Issue Crawler can come in handy.

There are, however, some limitations to the use of the Issue Crawler in mapping controversies. First of all, the links should be seen for what they are. As said before they don't necessarily represent a constructive dialogue or a fixed relation but rather are one-way recognitions. I therefore would like to point to the fact that the Issue Crawler can only provide a partial perspective on controversies. As Marres points out in points out in her dissertation "No Issue, No Politics": 'The findings of Issue Crawler are constrained by the presence of hyperlinks among Websites, and they are limited to the sources disclosed in this way.' (Marres, 2005: p. 113) For the purpose of the investigative reporting these limitations should be seen from a different perspective. The Issue

Crawler provides the means to clearly demarcate the research object and to disclose the controversy in a way that is methodologically justifiable. Secondly, the Issue Crawler shows how actors are mobilized around an issue online, but controversies are never acted out only in online environments. According to Marres 'we should certainly not make the mistake of defining the Web a priori as a key site for the enactment of public controversies today.' (Marres, 2005: p. 112) Instead, we should always notice the relation between on- and off-Web practices. Online practices do, however, offer one big advantage. The Web is largely still a text-plus-still-image type of medium which makes doing the investigation much easier. As Marres puts it: '[A]s issue networks on the Web present us with tangles of text, image, documents, claims and organizations, as opposed to problems, hypotheses and claims as they arise in situated practices, here we encounter issues in a relatively reified state.' (Marres 2005: p. 113) This makes it easier to trace the controversy and to come to clearly underpinned conclusions in the news story. Furthermore, in this stage of the research you are not yet dependent of human sources, which can be unreliable and do not always show the tip of their tongue.

Taking into consideration the above described limitations and advantages, one example of an investigation that could be conducted would be the 9/11 conspiracy controversy. In the aftermath of the 9/11 attacks many conspiracy theories circulated via email and the Web. This was a controversy that mainly played out on the Web since mainstream media gave little or no attention to the phenomenon, especially in the early stages. It would be interesting to investigate how the different actors mobilized around this particular issue and to see how the different conspiracy theories came into being. One would begin by creating an issue network with the Issue Crawler as explained above. The 911truth.org site would provide an excellent starting point since many groups and persons advocating 9/11 conspiracy theories identify themselves as part of the 9/11 Truth Movement. The issue network that would thus be created would provide clues about which actors are defining the issue. The nodes of the visualized network are clickable and make accessible the pages that make the controversy traceable on the Web. By switching back and forth between the nodes of the networks and the particular pages one can see how the different sites are building upon each others definition of the conspiracy. Especially now that the conspiracy theories have become much more influential (five years after the attacks an article in *Time Magazine* stated that: This is not a fringe phenomenon. It is a mainstream political reality.") (Grossman, 2006) it is useful to investigate how some of these theories came into being online. The Issue Crawler is an excellent tool to do this.

Although there are not yet clear examples of online network analysis's conducted by investigative reporters, I think this chapter showed that with the use of the Issue Crawler there are certainly possibilities. Journalists, however, who want to use this method must very well understand that hyperlinks between Websites do not necessarily indicate close relations between the sites. More interesting would it be to study how the different site build upon

each others characterizations of a certain issue. The Issue Crawler used in this way is an excellent indicator for which organizations are affected by the issue. Reporters could use more traditional investigative techniques to find out the exact nature of the conflict. Besides hyperlinks, another important organizing mechanism of the Internet is the search engine. In the next chapter I will look at the possibilities search engines and especially Google offer to investigative reporting.

5
THE ENGINE

The engine has become one of the most prominent features of the Internet. Without ordering devices such as Google it would be much more difficult to efficiently navigate the Web. Before, I have talked about how the demise of the human editor in favor of the back-end algorithm was visible in the evolution of the Google front-page. This development was dubbed 'googlization' and has been widely critiqued for different reasons. People are becoming increasingly aware of the power Google has over their Internet use and online presence. Google's business model and aesthetics largely determine how information and knowledge get circulated online. Library science scholars have expressed their concern about the changing locus of access to information from public shelves to commercial servers. (Rogers, 2009: p. 7) For some this is frightening thought but in this chapter I would like to argue that the way Google handles information and knowledge also offers opportunities for investigative reporters.

As said before the reason that the Internet is navigable is because of search engines and especially Google. The genius of Google lies in the fact that it does not use a great organizational scheme for the Web. Instead, they got everybody else to do that for them. Google's defining feature is PageRank, a system that looks at who links to whom online. Google does not only look at the content of sites but also at the links pointing to these sites. The more links you get, the higher you will end up in the search results. Furthermore, pages with a lot of inlinks are presumably more trusted and influential. Thus, Google counts their outlinks as being 'worth' more. (Grimmelmann, 2009: p. 942) However, it must be said that the exact algorithm used be Google is kept secret. It is therefore very hard to make any definite conclusions about the hierarchy of search results.

Although this system is brilliant in its simplicity it is very hard to compute. That's why Google has a huge server farm, 450.000 servers in 2006 costing millions of dollars in electricity each month, to process al the searches. (Carr, 2006) This is also the reason why the search engine market is dominated by only a few companies namely Google, Yahoo and Microsoft. Of these three Google is the absolute champion with market share of over 62 percent in the United States. Yahoo and Microsoft are far behind with a market share of respectively 19 and 13 percent. (comScore, 2010) In this chapter I will mainly focus on Google since, due its sheer size, it has tremendous impact on how people use the Internet. What can Google, by looking at the hierarchy of search results and search behavior of large groups, tell us about societal or cultural trends in the offline world? In this chapter I will provide a critical perspective at the ways in which Google can make a contribution to investigative news stories. Not only will I be looking at tools explicitly provided by Google but also at how Google and other search engines can be utilized for investigative reporting through the use of other open source tools.

5.1 Predicting trends

Journalists have steadily become more aware of the potential of search engines in the prediction of societal trends. While Google and other search engines previously were more reserved in disclosing information about their users' search activities, they have become more somewhat more open. With the introduction of Google Trends in 2004 and its more sophisticated successor Google Insights for Search in 2008 it provided ample opportunities for news stories. While the tools described in the previous two chapters have not yet really been adopted in the news room, Google Trends has not gone by unnoticed. Especially the results of Google Flu Trends were picked up by different mainstream media. With this service Google claimed to be able to predict outbreaks of the seasonal flu based on current and historical search data in conjunction with official medical statistics. A study presented at the American Thoracic Society concluded that Google Flu Trends is indeed good at suggesting flu-like illnesses although not at pinpointing actual cases of the flu. Still, researchers found that Google's tool was about 72% accurate in predicting confirmed cases of the flu over the 2003-2008 study period. (Ortiz, 2010)

Also in other fields Google made some successful predictions. On basis of the search queries Google predicted the winner of the Eurovision Song Contest. The group that was searched the most via Google also turned out to be the winner during the final. Yet another tool based on Google Insights is the Google Barometer, which is used to identify economical trends. The hypothesis is that if many people are looking for luxurious vacations and expensive fashion, this is a good sign for the economy. Hal Varian, chief economist at Google, is in a interview with the Dutch newspaper *de Volkskrant* optimistic about the potential of Google Insights:

> It shows that unemployment rates, at least in the United States, almost exactly match the search for topics about unemployment. The beauty of Insights for Search is that it uses actual searches, while official data on unemployment, car sales, vacation destinations and so on, is always dated. To be clear, we don't say that we can predict everything, but our tools can be used to make better predictions. (Hal Varian, as cited in Keuning, 2010) (Own translation)

Mainstream media, however, are still hesitant to incorporate the data provided by Google in their news stories. Rather they designate a separate article to the issue which is more about Google than about the trends it tries to predict. This hesitation is not ungrounded as there are still a lot of unanswered questions. It is often uncertain what specific search terms indicate. Further in this chapter I will address these issues but first the way Google operates will be discussed.

5.2 The politics of Google

Because of the way Google organizes information unexpected results may occur at certain times. This became apparent in 2003 when many opponents of the former president of United States G.W. Bush linked to his official biography Website with the phrase 'miserable failure'. Within no time the page had zoomed to the number-one hit if you searched for 'miserable failure'. This form of manipulating the Google search results has been dubbed Googlebombing. It is an effective online strategy that can be used in political context. As Grimmelmann points out 'This is a significant new form of politicking. Land a bomb like this and you can convince the world that Google agrees with your position. A successful Googlebomb doesn't just *reflect* the consensus of web users; it can help *construct* that consensus.' (Grimmelmann, 2009: pp. 942-943) The hierarchy of search results is so important since people will often follow the first link provided by Google and do not look further than the first ten results. It could therefore be argued that the first fifty or so links tell a story that might be very interesting for investigative reporters to uncover, especially when the Google search hierarchy is being appropriated by political activists to make a statement. The 'miserable failure' Googlebomb shows how political battles are nowadays being fought in online environments.

Another, more serious, example of how the hierarchy of search results can have political implications took place in 2004. In this year the number-one Google hit for a search on 'jew' was the anti-Semitic site jewwatch.com, which describes itself as 'An Oasis of News for Americans Who Presently Endure the Hateful Censorship of Zionist Occupation'. (Jew Watch, 2010) Besides showing how search results can be manipulated, this example also showed something else. Unlike other Googlebombs this one drew a response from Google. In a box next to the search results they wrote a note: 'Offensive Search Results. We're disturbed about these results as well.' Google also explains why this site ends up so high in the results: 'A site's ranking in Google's search results relies heavily on computer algorithms using thousands of factors to calculate a page's relevance to a given query. Sometimes subtleties of language cause anomalies to appear that cannot be predicted. A search for "Jew" brings up one such unexpected result.' (Google, 2010) However, Google did not change the results which would be a relatively simple thing for them to do. In fact they kept relying fully on the algorithm. This neutral stance by Google makes the search results a rather interesting object for investigative

Figure 2: Issue cloud showing the relative importance of different climate change skeptics, august 3, 2010. (Generated with the Google Scraper)

stories since they are generated (and thus demarcated) solely by the engine and not by the personal beliefs or morals of Google. The search results are generated by the input of large groups of Internet users (through the use of hyperlinks) so then the question arises in what sense does the hierarchy in search results reflect or differs from hierarchical relationships in society.

One way to investigate this is through the use of the Google Scraper, a software tool provided by the Digital Methods Initiative. With this tool one is able to capture the top 1000 results from a Google query. Subsequently one is able to study the source distance, the distance of a source from the top of the Web. One can search within the results for a certain organization but also for particular keywords. In a sample project by the Digital Methods Initiative the researchers looked at to what extent climate change 'skeptics' were present in the climate change spaces on the Web. The central question being: 'Does Google grant the skeptics voice in its returns for a "climate change" query?' ("The engine", 2010) The researchers made a list of the eleven foremost climate change skeptics and harvested the top-100 results for the query 'climate change'. This data was put into the Google Scraper which showed the results in a tag cloud. The tag clouds make it for example fairly easy to see how many times a specific climate change skeptic is mentioned in the first hundred Google returns for the query 'climate change'. (See figure 2)

5.3 Using engines for investigative stories

The previous paragraphs show that there are basically two ways in which search engines can be utilized for investigative news stories. The first way is to look at the search behavior of large groups of people in order to predict certain trends. Google makes this possible with tools like Google Insights for Search and Google Trends. The problem with these tools however is that all the data is provided by Google which leaves little room for creative and analytical thinking on part of the journalist. The tools provided by Google also have the problem that, because they are so easy to use, it is also very easy to find false correlations. For example, if during an election a lot of people are searching for a certain politician this does not mean they will also vote for him or her. This example seems rather obvious but also some of the search terms used by Google in their economical Barometer can be criticized in the same way. For example, Google Insights defines search terms like 'good job' and 'lease car' as positive for the job index. However, what do these terms exactly indicate? That people have a good job or are looking for one? These search terms can be interpreted as both positive and negative for the job market. Journalists should not be tempted by first impressions and thus be very careful in using the tools provided by Google. That is not to say that the tools cannot be used in more productive ways for investigative reporting. I would only like to point to the fact that the reporters should look very careful at the exact terms that have been used and what these terms would indicate.

A second way search engines can be used for investigations is by looking at the source distance, the distance of a source from the top of the search results, with the help of tools like Google Scraper. It can be very useful to analyze which actors are granted a voice by Google. This is also a phenomenon that should be looked at over time. Search engine returns can be very volatile towards Website rankings over time. One example of this is the site 911truth.org which ranked steadily in the top-10 results of Google for the query '9/11' until September 2007. For three weeks, the site suddenly dropped out of the first 200 results, and subsequently the top-1000. These type of anomalies raise questions that should be examined further. As argued before the influence Google has over our search behavior is tremendous. Other information outlets, like newspapers and television stations, can be held accountable for the information they provide us with since they are held by journalistic codes. Google and other search engines are accountable to no one but are becoming more and more powerful. That is why I claim that search engine returns open up a new field for investigative journalist to explore.

The second way I described for using search engine for investigative purposes is still very much underused in journalism. I do, however, want to claim that this method shows the most potential. Search engine results can tell stories currently overlooked by investigative reporters. In the previous three chapters I discussed several methods that would constitute what I call online grounded investigative reporting. In the third part of my thesis I will try to identify the lessons that can be learned from these examples to make the beginning of a methodology for this kind of reporting.

PART III

6
TOWARDS A METHODOLOGY

The purpose of this chapter is to describe a methodology for conducting online grounded investigative reporting. In the previous chapters I already discussed some of the possibilities the online holds for investigative journalists. The concepts, techniques and principles that were described in these earlier chapters will also be incorporated in this methodology. Thus in short, this chapter will bring together all of the issues, examples and discussions around OGIR of the previous chapters. However, before designing a methodology it is important to clarify what is meant by that term. In the philosophy of science a methodology is defined as a set of constructs about how research in a particular field should be conducted. It relates to the logic of analysis used in a given field as well as the justification for the appropriated methods. That is to move beyond simply describing the 'tool kit' available for journalists conducting OGIR. Although the specific procedures or techniques used in OGIR constitute an important part of a methodology, the concept itself is broader. It refers to a set of constructs, shared within a given field, about how its research should be conducted and reported in order to be regarded as legitimate by its practitioners. The philosopher Maurice Natanson explained this as follows: 'By "methodology," I understand the underlying conceptual framework in terms of which concrete studies in history, sociology, economics and the like are carried out, and in terms of which they receive a general rationale.' (Natanson, 1963: p. 271) The aim of this thesis is to describe this shared rationale by discussing the strategies, ethical responsibilities and standards for presenting findings to the public. The word 'shared' is important is this context since it is never up to one person to design the methodology for a particular field. The last word is ultimately up to the practitioners. There must be considerable consensus within the field about the ways investigations are conducted. This consideration signals that I must be modest in my attempts to describe a formal methodology for OGIR, especially since this discipline in journalism has barely even proven itself. Still I argue that by describing a number of strategies and responsibilities this thesis can be seen as the start of a debate about what should constitute a shared methodology for OGIR.

6.1 Strategies

When conducting an online grounded investigation one should begin with preparing a carefully devised plan or strategy. Although often tempting it might not be such a good idea to blindly dive into the big pools of data the Internet provides. As I argued in the second chapter, it is not difficult to find false correlations. A methodology therefore should include strategies that can be used to achieve a desired goal. The strategy used in a particular investigation serves as a plan that guides the project to its goal. A strategy should begin by ensuring a complete understanding of the nature of the online data. The first step in each Web-based research should be to gain a clear understanding of the object of study and its methods. The way information, knowledge and sociality is organized on the Web can be enormously complex. One way to

start might be to identify all the actors around a certain issue while realizing that objects or technologies also have agency. They too can be agents. This requires knowledge on how certain Web devices organize their content. For example, when trying to provide an account about the evolution of a certain Web page through use of the Internet Archive, one should not take the Wayback Machine for granted. The Wayback Machine is not a time capsule that can take you back to the early days of the Internet (although the name might suggest otherwise). As I tried to explain in the third chapter, the output of the Wayback Machine is largely defined by the archiving process. The Wayback Machine should not be seen as a passive intermediary but as an actor that actively mediates the results it provides. The same goes for many other Web objects such as search engines or hyperlinks as well as the tools used to analyze these objects such as the Issue Crawler or the Google Scraper. When conducting online grounded investigations the reporter should always realize and account for how the objects studied and the tools used influence the results. After the reporter has a thorough understanding of the nature of the objects used in the investigation, specific goals can be formulated. This can be in the form of either a research question or a hypothesis. For example the article in *NRC Handelsblad* about the rise of right-wing sites may have started with the hypothesis that the Internet spurs the use of extremist language. The reason why it is important to formulate a clear research question or hypothesis is that when dealing with huge quantities of data you do not get lost. Patterns can pop up anywhere in the analysis of large datasets and it might prove very difficult to distinguish between coincidental and meaningful ones. For this reason I would claim that journalists should be very careful with using the tools provided by Google such as Google Insights for Search. Without a proper research question it might be very tempting to follow the results provided by Google blindly. But as I tried to explain in chapter 5, more searches for 'good job' can be explained as both positive and negative for the economy.

After the goals are determined, the reporter can begin with collecting and analyzing the data. This might be collection of snapshots harvested with the Wayback Machine, a network of interlinked sites generated with the Issue Crawler or a list of search engine returns. These datasets should be strategically examined which can be done in a number of ways. I will describe some of these possibilities while also discussing the benefits and any possible disadvantages.

> *Examining trends over time* – One of the most obvious but at the same time most productive strategies might be to examine data over a period of time find out if significant trends have occurred over these years. If the results show a clear upward or downward trend, a newsworthy issue may have been discovered. This strategy was used in the article in *NRC Handelsblad* were they discovered an increase of extremist language at right-wing sites over the course of a decade. Thanks to the ephemeral qualities of the Web this strategy is especially useful for online grounded reporting. The Web is constantly evolving while at the same time these changes leave traces that can be tracked down through the use of for instance the Wayback Machine. An additional advantage stems from the fact that is remains very difficult to account for offline phenomena through the use of online data. It is, for example,

difficult to say whether 150 extremist sites is a lot for a country like the Netherlands. However, when looking at these kinds of data over a period of time it becomes more justifiable to make the connection between certain online trends and offline major events. After the attacks of 9/11 there was for example an increase of the number of ring-wing sites.

Examine deviant results – This strategy can reveal unusual situations that can provide information for good news stories. Certain dramatic and unusual information can come to light that may provide the foundation for an unusual news story. This strategy fits in with one of the more traditional types of investigative journalism, namely to reveal scandals. The Wikiscanner software could be used to reveal who is behind certain Wikipedia edits. It turned out that the Dutch Royal family was behind some censoring Wikipedia edits in the entry on Princess Mabel's relationship with the Dutch drug lord Klaas Bruinsma.

Serendipity as search concept – Anomalies that are coincidently stumbled upon may be the start on an interesting story. One example of this is the high ranking of the anti-Semitic site Jewwatch.com in the Google search results for the query 'jew' (Grimmelman, 2009). Obviously, these examples should be followed up and developed into a story. This approach shows how the Web can function as a place to find ideas for news stories. A relatively easy overlooked anomaly in the Google returns for the query 'jew' may provide the lead to an intriguing online struggle between anti-Semitic groups and Jewish interest groups in their attempts to manipulate the Google search results.

Make a before and after comparison – This approach could shed light on a suspected causality. By examining some occurrence at a certain point in time and then looking at the same situation at a later date, after a set of circumstances have taken place between the two, you could find interesting correlations. As said before, certain major event in the real world may trigger responses in online environments. Analyzing these online environments may provide clues about how people are dealing with certain events. One example could be to study the site of the White House before and after an election. How does election of a new president influence the layout and content of this site? Also the formation of online networks could be analyzed in this fashion. By using the Issue Crawler one could analyze how the publication of the secret documents about the wars in Iraq and Afghanistan by Wikileaks circulates the Web.

It must be noted that the approaches described above show only a few of the possibilities of OGIR. All of these strategies serve as a general outline to achieve different kinds of goals, whether it is to show a suspected causal sequence, to expose a scandal or making visible some kind of trend. Obviously, a good news story requires more than this. Before your findings can

be reported to the public there are a number of responsibilities that the investigative reporter must assume. In the next paragraph I will describe some of the considerations that become important when conducting online grounded investigations.

6.2 Responsibilities

Before discussing the responsibilities that are unique for online grounded investigative reporting I would first like to attend to some considerations that derive from the general responsibilities of any form of investigative journalism. The fundamental responsibilities off investigative reporters are described by DeFleur as follows: 'First they have an ethical responsibility to get the facts straight. Second, they have a similar responsibility to reach conclusions from those facts that represent an accurate picture of reality. Third, they have an obligation to report what they have found to the public in such a way that it can be readily understood.' (DeFleur, 1997: 229) These very basic responsibilities apply to all forms of investigative reporting or journalism for that matter.

However, online grounded research implies a set of new concerns and responsibilities. New technologies almost always raise new ethical concerns for both reporters and editors but the advent of the Internet showed this in particular. Especially the privacy question remains a big issue. People carelessly put private information on the Web that, because of the already discussed nature of the Internet, becomes very difficult to take off. Fierce discussions take place among journalists how they should handle sensitive information like this. But the online privacy question goes much deeper. Also search queries may provide very specific information about a particular person. A now infamous search engine data release in 2006 made the searches of a half million people over the course of three months publicly available. The data release sparked, as Rogers puts it, 'frightening and often salacious press accounts about the level of intimate detail revealed about searchers, even if their histories are made anonymous and decoupled from geography (no IP address).' (Rogers, 2009: p. 18) Numerous trials were run against Google for violating the privacy of their users. In regard to Google Trends and Google Insights for Search, Google gets around the privacy concerns by pointing to the fact that their graphs are based on aggregated data from millions of searches done on Google over time.

The same goes for the practices described in this thesis. For the most part, they work with large datasets which makes it impossible to trace your findings back to a specific person. However, in some instances specific persons were implicated in the story. The scandal around the Wikipedia edits by prince Friso and his wife Mabel is one example, the person accused by the *NRC Handelsblad* of providing a hosting service to right-extremist sites is another. The royal couple was granted the right to reply but the person that was accused of enabling right-extremist sites was not. The audi alteram partem rule should always be implemented in journalism, especially when the conclusions in the news story are based on online grounded investigations. The Web is invested with disinformation so the reporter should always verify the data used in the analysis. Barry Sussman describes in his article on digital journalism an entire disinformation industry 'consisting of corporate funded think tanks, phony grassroots groups, co-opted citizens organizations, experts for hire, fake news and fake

reporters, government officials on all levels who lie and promote disinformation.' (Susman, 2008: p. 46) He notes, however, that the Internet provides journalists also with reliable sources to help sort out what is real and solid from what is fake and disingenuous. It is up to the reporter to make responsible decisions in this respect.

Another unique responsibility for reporters doing online investigations derives from the fact that for many people the Internet remains a black box. They see it as a clever piece of technology without knowing exactly how it works. Since the online grounded investigations discussed in this thesis are conducted by reporters rather than by independent outside experts it might prove difficult to satisfy the public that the data have been used responsible. The problem here is that there is no way for the reader to know whether the reporter has done the investigation correctly or drawn proper conclusions. One way of dealing with this problem is making the data used for a news story publicly available for people to recheck. However, due to the ephemeral nature of the Web it is not possible to simple redirect them to the sources used. The hierarchy in Google search results for example depends, besides on the time of the search, also on the location and the preferences of the user. It is therefore important to keep records of the analyzed data to ensure that the conclusions derived from this data can always be (re)checked. These records, for example a collection snapshots of sites collected with the Wayback Machine, can be made available online so that critics or others can examine the data themselves. Furthermore, it is vital that the reporters always respond to readers who have questions, people who request additional information or critics who raise objections. Especially since most of the public are born before the introduction of the Internet and are digital immigrants at best, it is important to maintain an open relation with the public. Maybe one of the most difficult elements of online grounded investigative reporting is to present your finding in an understandable manner to your public. The next paragraph will discuss this element further.

6.3 Presenting your finding to the public

After the investigation is completed, the results should be interpreted and made comprehensible for the public. This might prove to be a difficult task since your findings often do not speak so much to the imagination of the public as for example governmental corruption would do. The only way to work around this is to provide a understandable report explaining the way the investigation was done. The major problem with the article in *NCR Handelsblad* about the rise of right-extremist sites is that it remains very vague about how the study was conducted. The source of the data, the Wayback Machine, was provided but a brief description of the overall approach was absent. This makes it very difficult for the readers of the article to value the conclusions. On the other hand, the public should not be obscured with extensive digressions about the exact methodological considerations. In other words, there is delicate balance between what public needs to know and what can be omitted. It is up to the reporter to cover the minimal amount of information that needs to be disclosed to the public so that the nature and limitations of the investigation can be readily understood.

In the second chapter I already tried to shatter Chris Anderson's argument that numbers speak for themselves. I argued that this was not the case with Internet research but this is especially untrue in a news story. The numbers should never obscure the human side of the story. Reporters should realize that the research process is not finished after the online investigation has been completed; it is only the start. Online grounded investigative reporting cannot replace good writing or effective interviewing. This requires creative and analytical thinking not unlike in traditional investigative reporting. The findings should be supplemented with interviews with key figures or experts. The article about right-extremist site was, for example supplemented with quotes from the director of the discrimination hotline. These types of experts are necessary to interpret the results provided by the investigation by putting them into context. Furthermore the *NRC* article made extensive use of quotes from the analyzed sites. Although these quotes were anonymous they did provide a more human side to the story. So, good stories are always about people and not just about numbers and facts.

The facts and numbers that are necessary should be presented in such a way that they are easy to understand for the public. This could be achieved by visualizing the results. Putting big numbers into graphs and charts can help people to grasp the large quantities while at the same time showing correlations and patterns. In regard to data visualization there are increasingly more possibilities. News organizations are more and more seeing the opportunities of data visualization. *The Washington Post* made extensive use of data visualization during their investigative project 'Top Secret America'. ("Top Secret America", 2010) The stories in the paper were supplemented with online info graphics that showed the huge national security buildup in the United States after the attacks of 9/11. This method, were stories in the paper are further clarified with online graphics, could also be very useful to give the public more insight in online grounded investigative reporting.

The methodology for online grounded investigative reporting is still in its infancy. It should therefore be noted that when the methodology further evolves, its practitioners have less to worry about accounting for fundamental epistemological concerns since these issues will be accounted for in the methods themselves as well as in the interpretation of the results. For this reason I predict a slow start for the kind of journalism I propose in this thesis. The public has to become more familiar with the objects by which the Web is organized in order to fully comprehend the scope of OGIR. However, as the number of digital natives, referring to those growing up in online environments, is increasing by the minute I also predict a bright future for news stories grounded in the online. The Web is full of stories waiting to be uncovered.

CONCLUSION

Let me begin my conclusion by stating that online grounded investigative reporting should not be seen as a replacement of traditional investigative journalism but as an extension. The characteristics of investigative reporting like creative and analytical thinking are still of vital importance for the success of this practice. But, as was explained in the introduction, the profession of journalism is under enormous pressure of the market, and OGIR might be one possible solution. News organizations have to deal with budget cuts which lead to journalists losing their job. The same amount of news has to be made with fewer contributors. Especially investigative reporting suffers from this tendency. It is expensive, ties up resources that could be used in more productive ways and has an uncertain outcome. Because of these reasons, investigative reporters are often the first to go. New media and especially the Internet are regularly blamed for the demise of journalism. My claim, however, is that the Internet opens up new, promising and cost-efficient ways of doing journalism and especially investigative reporting. By analyzing a number of tools, techniques and possibilities, I have sketched the outlines of a methodology for what I call online grounded investigative reporting.

The proposed methodology in this thesis is not meant to be seen as a definite way of doing online grounded investigative reporting. It is merely the start of a debate on how such reporting should be done. In the end it is up to the practitioners in the field to decide on which methods are acceptable. Important to remember in this respect is the fact that OGIR is an extension of the traditional means of investigative reporting. Old values and responsibilities will and have to remain intact. Examples of OGIR are still scarce but already show what is possible when using the Web as object of study. Although I am aware of the limitations of this study, both the concepts and the methodology should be further examined, I do believe that this thesis clears the path for exiting new ways of doing journalism. On the other hand, I also tried to be critical about using the Web for investigative purposes. I stressed the importance of not relying on first impressions. The way the Internet is structured makes it easy to find all kinds of patterns and correlations but without a thorough understanding of the medium and its methods, these finding become meaningless.

In this thesis I limited myself to discussing three elements of the Web and their potential uses for investigative reporting. It should be noted that each of these elements, the Website, the link and the search engine, can provide enough material to devote an entire thesis to. Furthermore, besides the elements I discussed, there are numerous other opportunities for investigative reporters on the Web. One could think of networked content on Wikipedia. The collaborative efforts to create knowledge are thanks to Wikipedia's transparent nature open for

investigation. Another promising field, not mentioned in this thesis, are social network sites and how these sites organize sociality. People put tremendous amounts of personal information online which offer new opportunities for demographic research. In other words, much work is still to be done.

References

Altschull, Herman. *From Milton to McLuhan: The Ideas Behind American Journalism*. White Plains: Longman, 1990.

Anderson, Chris. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired Magazine* 16.07, 2008, n.d. Web. <http://www.wired.com/science/discoveries/magazine/16-07/pb_theory > (Accessed May 14, 2010)

Barlow, J. P. "A Declaration of the Independence of Cyberspace." *Projects.eff.org*, Electronic Freedom Foundation, 1996, n.d. Web. <http://homes.eff.org/~barlow/Declaration-Final.html> (Accessed August 8, 2010)

Barnett, Steven. "Journalism, Democracy and the Public Interest: rethinking media pluralism for the Digital Age." RISJ Working Paper, 2009, n.d. Web. <http://reutersinstitute.politics.ox.ac.uk/fileadmin/documents/Publications/Navigating_the_Crisis_in_Local___Regional_News_Final.pdf> (Accessed June 24, 2010)

Brügger, N. *Archiving Websites: General Considerations and Strategies.* Centre for Internet Research, Aarhus, 2005.

Brügger, Niels. "Website history and the Website as an object of study." *New Media & Society* 11 (2009): 115-132.

Carbonez, Xavier. Hans Brockmans. "Wie vervangt de oude Belgische elite." *Trends.* Februari 26, 2004.

Carr, David F. "How Google Works." *Baselinemag.com*. Baseline, July 6, 2006, n.d. Web. <http://www.baselinemag.com/c/a/Infrastructure/How-Google-Works-1/> (Accessed on August 2, 2010)

Christians, Clifford G.. Shing-Ling Sarina Chen. "Introduction: Technological Environments and the Evolution of Social Research Methods." In: Mark D. Johns, Shing-Ling Sarina Chen, G. Jon Hall (ed.), *Online Social Research: Methods, Issues, & Ethics.* New York: Peter Lang Publishing, 2004.

Chu-Carroll, Mark. "Petabyte Scale Data-Analysis and the Scientific Method." *Scienceblogs.com,* Science Blogs, July 4, 2008, n.d. Web. <http://scienceblogs.com/goodmath/2008/07/petabyte_scale_dataanalysis_an.php> (Accessed May 15, 2010)

Cochran, W.. "Journalism's New Geography: How Electronic Tools Alter the Culture and Practice of Newsgathering." *Electronic Journal of Communication* 7 (1997):1–10.

"comScore Releases June 2010 U.S. Search Engine Rankings." *ComScore.com*. comScore, July 13, 2010 n.d. Web. <http://www.comscore.com/Press_Events/Press_Releases/2010/7/comScore_Releases_June_2010_U.S._Search_Engine_Rankings> (Accessed on August 2, 2010)

Cox, Melisma. "The development of computer-assisted reporting." Unpublished paper, 2000, n.d. Web. <http://com.miami.edu/car/cox00.pdf> (Accessed May 14, 2010).

Currah, Andrew. "Navigating the Crisis in Local and Regional News: A Critical Review of Solutions." RISJ Working Paper, 2009, n.d. Web. <http://reutersinstitute.politics.ox.ac.uk/fileadmin/documents/Publications/Navigating_the_Crisis_in_Local___Regional_News_Final.pdf> (Accessed June 24, 2010)

DeFleur, Margaret. *Computer-assisted investigative reporting: development and methodology.* Mahwah: Lawrence Erlbaum Associates, 1997.

Dennis, Everette. Arnold Ismach. *Reporting Processes and Practices.* Belmont: Wadsworth Publishing, 1981.

Dohmen, Joep. "Opkomst en ondergang van extreemrechtse sites." *NRC Handelsblad.* August 25, 2007, n.d. Web. <http://www.nrc.nl/binnenland/article1831689.ece/Opkomst_en_ondergang_van_extreemrechtse_sites > (Accessed June 30, 2010)

Dumlao, R. and S. Duke. "The Web and E-Mail in Science Communication." *Science Communication* 24 (2003): 283-308.

Geelen, Jean-Pierre. "Hallo, het was maar een spelletje!" *de Volkskrant.* December 23, 2005.

Gelman, L. “Internet archive's web page snapshots held admissible as evidence.” *Packets* 2 (2004).

“Google: An explanation of our search results.” *Google.com.* Google, n.d. Web. <http://www.google.com/explanation.html> (Accessed on August 3, 2010)

“Google and the politics of tabs.” *Digitalmethods.net.* The Digital Methods Initiative, n.d. Web. <http://www.digitalmethods.net/Digitalmethods/TheWebsite> (Accessed 20 July, 2010)

Grimmelmann, J. "The Google Dilemma," *New York Law School Law Review* 53 (2009): 939-950.

Grossman, Lev. "Why The 9/11 Conspiracies Won't Go Away". *Time Magazine*. September 3, 2006, n.d. Web. <http://www.time.com/time/magazine/article/0,9171,1531304-1,00.html.> (Accessed august 2, 2010)

Innis, Harold. *The Bias of Communication.* Toronto: University of Toronto Press, 1951.

"Internet Archive Frequently Asked Questions." *Archive.org*. The Internet Archive, n.d. Web. <http://www.archive.org/about/faqs.php#The_Wayback_Machine> (Accessed 24 June 2010.)

Jaspers, Huub. “Special Forces in Afghanistan and Iraq.” *Argos.* VPRO/VARA. Radio 1, Hilversum. 14 May 2004. Radio.

“Jew Watch News.” *Jewwatch.com.* Jew Watch, n.d. Web. <http://www.jewwatch.com/> (Accessed on August 3, 2010)

Jones, S. “Studying the Net: Intricacies and Issues.” In: S. Jones (ed.), *Doing Internet Research: Critical Issues and Methods for Examining the Net.* London: Sage, 1999: 1-28.

Kamsma, Jelle. “The leaking of official documents and democratic politics” Unpublished paper. March 31, 2010.

Keuning, Wouter. “Zoekopdrachten kunnen voorspellen; Interview Hal Varian, hoofdeconoom bij Google.” *de Volkskrant.* August 1, 2010.

Kussendrager, Nico. *Onderzoeksjournalistiek.* Groningen: Noordhoff Uitgevers, 2007.

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

Lippmann, Walter. *Public Opinion.* New York: Macmillan, 1922.

Manovich, Lev. "Cultural Analytics: Visualizing Cultural Patterns in the Era of More Media." *DOMUS* , 2009, n.d. Web. <http://softwarestudies.com/cultural_analytics/Manovich_DOMUS.doc> (Accessed May 14, 2010)

Manovich, Lev. "White paper: Cultural Analytics: Analysis and Visualizations of Large Cultural Data Sets." 2007, n.d. Web. <http://softwarestudies.com/cultural_analytics/cultural_analytics_2008.doc> (Accessed May 14, 2010)

Marres, Noortje. "No Issue, No Politics." Dissertation, University of Amsterdam, 2005.

McLuhan, Marshall. "The Medium is the Message." In: Noah Wardrip - Fruin and Nick Montfort (ed.), *The New Media Reader.* Cambridge: MIT Press, 2003.

McNair, Brian. *Cultural Chaos: Media in the Age of Dissolutions*. New York: Routledge, 2006.

Meyer, Philip. "Reporting in the 21st Century." Presentation at AEJMC in Montreal, 1992. In: DeFleur, Margaret. *Computer-assisted investigative reporting: development and methodology.* Mahwah: Lawrence Erlbaum Associates, 1997.

Meyer, Philip. *The New Precision Journalism.* Bloomington: Indiana University Press, 1991.

Natanson, Maurice. *Philosophy of the Social Sciences*. New York: Random House, 1963.

Ortiz, Justin R. Hong Zhou. David K. Shay. Kathleen M. Neuzil. Christopher H. Goss. "Does Google Influenza Tracking Correlate With Laboratory Tests Positive For Influenza?" Am. J. Respir. Crit. Care Med. 181 (2010): A2626.

Plesner, Ursula. "An actor-network perspective on changing work practices: Communication technologies as actants in newswork." *Journalism* 10 (2009): 604-626.

Protess, David L. (ed.). *The Journalism of Outrage – Investigative Reporting and Agenda in America.* New York: The Guildford Press, 1991.

Regier, C.C. *The Era of the Muckrakers*. Chapel Hill, N.C.: The University of North Carolina Press, 1932.

Roberts, Eugene. "The Finest Reporting is always investigative." *IRE Journal* vol. 11, no. 1 (1988): 12-14.

Rogers, R. *The end of the virtual – Digital Methods.* Amsterdam: Amsterdam University Press, 2009.

Rogers, R., M. Stevenson and E. Weltevrede, "Social Research with the Web," *Global Informaton Society Watch 2009*, Association for Progressive Communications and Hivos, 2009.

Schneider, Steven. Kirsten Foot. "The web as an object of study." *New Media and Society* 6 (2004): 114-122.

Sussman, Barry. "Digital Journalism: Will It Work for Investigative Journalism?" Nieman Reports 62.1 (2008): 45-47.

"The engine." *Digitalmethods.net.* The Digital Methods Initiative, n.d. Web. <http://www.digitalmethods.net/Digitalmethods/TheOrderingDevice> (Accessed August 4, 2010)

"The link." *Digitalmethods.net.* The Digital Methods Initiative, n.d. Web. <http://www.digitalmethods.net/Digitalmethods/TheLink> (Accessed 23 July, 2010)

Thrift, Nigel. "Remembering the technological unconscious by foregrounding knowledges of position." *Environment and Planning D: Society and Space* 22 (2004):175-190.

Tofani, Loretta. "Investigative Journalism Can Still Thrive at Newspapers." Nieman Reports 55.2 (2001): 64.

"Top Secret America." *Washingtonpost.com.* The Washington Post, n.d. Web. <http://projects.washingtonpost.com/top-secret-america/> (Accessed August 12, 2010)

Turkle, S. *Life on the Screen: Identity in the Age of the Internet.* Simon & Schuster, New York, 1995.

Turner, Fred. "Actor-Networking the News." *Social Epistemology* 19 (2005): 321-324.

Ullman, John. Jan Colbert. *The Reporter's Handbook: An Investigator's Guide to Documents and Techniques.* New York: St. Martin's Press, 1991.

Van Eijk. Dick. *Investigative Journalism in Europe.* Amsterdam: Vereniging van Onderzoeksjournalisten (VVOJ), 2005.

Venturini, Tommaso. "Diving in magma: how to explore controversies with actor-network theory." *Public Understanding of Science* 20 (2009): 1-16.

Verkade, Thalia. "Mabel en Friso pasten lemma aan." *NRC Handelsblad.* August 29, 2007, n.d. Web. <http://www.nrc.nl/binnenland/article1832920.ece/Mabel_en_Friso_pasten_lemma_aan> (Accessed July 21, 2010)

"Volkskrant Top 200, editie 2009." *de Volkskrant.* November 5, 2009, n.d. Web. <http://www.volkskrant.nl/binnenland/article1089671.ece/Volkskrant_Top_200%2C_editie_2009> (Accessed July 22, 2010)

Walker, Ruth. "Computer Databases Can Be Valuable Sources." *Christian Science Monitor*, September 25 (1990): 14.

Wasserman, Edward. "Investigative Reporting: Strategies for Its Survival." Nieman Reports 62.3 (2008): 7-10.

Woolgar, S. "Five Rules of Virtuality." In: S. Woolgar (ed.), *Virtual Society? Technology, Cyberbole, Reality.* Oxford: Oxford University Press, 2002: 1-22.