Computational Journalism

When journalism meets programming

Eirik Stavelin

Dissertation for the degree philosophiae doctor (PhD)
at the University of Bergen

2013

Scientific environment

This thesis was produced at the Department of Information Science and Media Studies at the University of Bergen. The affiliations include the institute's research groups for journalism studies and semantic and social information systems.

One paper was written in collaboration with Joakim Karlsen, Faculty of Computer Sciences, Østfold University College, Norway.

Acknowledgements

First of all I would like to thank my supervisors. Bjørnar Tessem has been my main supervisor and the technological anchor in my efforts while Martin Eide has provided the journalistic anchor as co-supervisor. I'm thankful for the interest you have shown and valuable, quick feedback on ideas, artifacts and written text you have given. About once every year Nicholas Diakopoulos, external supervisor and computational journalism scholar, have visited Bergen and given me excellent feedback and discussions. Combined, this feedback has been invaluable to my research.

Secondly I would like to thank the participants and interviewees for sharing their time and experiences with me.

Further, I owe a lot of people special thanks. Dag Elgesem, Helle Sjøvaag and Hallvard Moe invited me into their research project *NRKs nyhetstilbud på nett 2009*, and gave me a flying start as a researcher. Joakim Karlsen, co-author of *Computational journalism in Norwegian Newsrooms*, who has functioned as a research "sparring partner". I'm particularly grateful to Frode Guribye, Lars Nyre, and Anders Fagerjord for input and inspiration. My fellow PhD students, for creating a wonderful atmosphere for peer-review and social interruptions, and particularly my office cohabitant Torgeir Uberg Nærland for all the shared experiences and for making procrastination in the office time well spent. To colleagues in Fosswinckelsgate 6, lunch-time quizzers, hallway dwellers and coffee machine mechanics, for making my surroundings a inclusive an intellectually stimulating space. To friends: Linn, Chang, Ina, André, Lars Thomas, Fabia and Kjartan for the trips and hikes, discussions and shows, games and dinners. To family: Inger, Roar, Svein, Martin & Tine for your support and for teaching me the value of hard work. Most all of you have given me feedback on my work at some time or another, thanks.

Most of all, I owe my beloved Cathrine Sætre a special thanks for being awesome.

Abstract

Digital data sources and platforms allow journalists to produce news in new and different ways. The shift from an analog to digital workflow introduces computation as a central component of news production. This enables variability for end users, automation of tedious tasks for newsrooms, and allows journalists to tackle analysis of the increasingly large sets of data relevant to citizens. To journalism, computerization is a promising path for news production, particularly for those who are able to wield computers to their specific needs through programming as a journalistic method. Toolmakers and users, both internal in the newsrooms and external in academia and in the IT business, are putting effort into making computational journalism a reality.

While the hypothetical aspects of computational journalism are easy to find, this thesis provides studies of computational efforts in newsrooms as well as experimental prototyped suggestions in order to provide a better understanding of how practices in journalism intersect with computing as information science.

This thesis approaches software-oriented news production as (1) a socially situated practice in newsrooms and (2) a design science research problem. The newsroom approach includes an analysis of news applications; journalistic output that consists of software code as a part of news storytelling. The analysis focuses on what technical and visual elements these applications consists of and how they compare as journalistic products in relation to the core functions of the journalistic social contract. Further, authors of news applications as journalist-programmers are interviewed in order to give an account of how this practice is situated in the newsroom and how these practitioners view their efforts in relation to technical, social, and journalistic considerations. As a design science research problem, I have approached computational journalism as an effort to produce software *for* journalism by user testing a custom prototype for dealing with analysis of social media messages, and as an effort to produce software *as* journalism in creating a tool for

watchdogging the parliamentary data API, aided by expert parliamentary reporters to discuss how such an endeavor could be formulated and executed.

Results show that advanced technological work is used, both in creating news applications and in an array of other newsroom-internal workflows, to continue traditional journalistic functions and themes, under the premises of digital media logic where software creation can be used to gather, systematize, and analyze material as well as to publish code in digital journalism online. The practitioners that have these skills use them as a journalistic method and underline their positions as journalists not technologists. This view of technological work as journalistic is not universal in journalism, where technical work is often segregated from journalistic work. Creating software for journalism, as exemplified as a tool to aid analysis of user-generated content, requires solid understanding of what journalists do rather that what journalism is intended to do. Finding stories and sources in social media is a matter of negotiating limited resources and the authorship of messages counts heavily in favor of known persons over popular or alternative arguments. The types of stories the prototype was found to best aid were soft and human interest stories, findings in accordance with other studies of journalists' utilization of user-generated content. Creating software as journalism, taking a more user-centered design approach, created richer insight into how one subgroup of journalists (parliamentary reporters) relate to software in their beat. The possibilities for journalistic reinvention were clearly expressed, as was a stricter boundary between journalistic and technical work, where journalism is a function that transforms facts and data into journalism by adding context, interpretation, and explanations. The particularity of parliamentary reporters' workflow, that to a large extent depends on oral sources and traditional social networking, is mostly unsuited for computational aid based on the parliaments' API, but fact-checking and analysis of background information on members of parliament through a software-oriented approach is seen as complimentary and promising rather than threatening to the craft.

While computational journalism emerges from traditions of software-oriented news productions that to a large extent overlap as a merge of computer science and

journalism, some distinctive features distinguish and define this field. Both internally in the newsroom and as journalistic output, computational journalism is defined be a shift towards platforms, in creating spaces for finding, discussing and narrating stories. This can include the management of computable models, not merely collected sets of data. As a craft, creating software to solve journalistic problems, computational thinking becomes a key skill that defines both reasonable expectations and limitations, but also collaborations. The difference in technological sophistication between computational journalists as the newsrooms at large is under constant negotiation. Programming journalists strive for higher journalistic capital, while newsrooms adapt by both embracing computational efforts as possibilities for journalistic reinvention and keeping a distance by labeling the work as technical. Journalistic values and values of technology (or reasons for utilizing technology), can contradict each other. The gap that needs to be acknowledged in order to stay accountable in computational news production is above all an understanding of technology as a companion (and antagonist) of agency in news production.

List of publications

I. Stavelin, Eirik. 2012. “News applications – journalism meets programming.” Published in Norwegian with the title “Nyhetsapplikasjoner” in an anthology by Eide, Martin, Leif Ove Larsen, and Helle Sjøvaag. 2012. *Nytt På Nett Og Brett*. Oslo: Universitetsforlaget.

II. Karlsen, Joakim, and Eirik Stavelin. 2013. “Computational Journalism in Norwegian Newsrooms”. Published in Journalism Practice (July 23): 1–15. doi:10.1080/17512786.2013.813190.

III. Stavelin, Eirik. 2013. “The pursuit of newsworthiness on Twitter”. Submitted and accepted for presentation at Norsk Informatikkonferanse (NIK) in Stavanger 18.-20. November 2013 and publication in the NIK2013 proceedings.

IV. Stavelin, Eirik. 2013. “Watchdogging in code”. Submitted to Digital Journalism, and is currently in review. Presented in a shorter format at the Future of Journalism conference in Cardiff 2013.

Table of Contents

Part 1: Summary of research contribution

1. Introduction

Professional news production has throughout history always been technology oriented. From the printing press through the telegraph, from vacuum tubes to the current technologies such as mobile telecommunication and computing, the ways we produce and consume news media have followed the state of technological development. All through this development, some journalists have pioneered news production by utilizing new technology. The last 50–60 years of development in computing have had a significant impact on society, and journalism is no exception. Current efforts involve the combination of computer science and journalism into a hybrid craft called "computational journalism". It is this hybrid journalism that I aim to explore, describe, and analyze in this thesis.

Computational journalism is an emergent field, with high expectations and uncertain boundaries. My primary research objective is to answer the following research question: *How is computational journalism operationalized and how are computational methods perceived in Norwegian newsrooms?* In order to get closer to a reasonable answer to this question, I have approached the software-oriented form of news production from different angles. These represent two distinctly different approaches to research. One is in the newsroom studies tradition, with an analysis of journalistic output in the form of news applications, and an interview study with journalists who write code as a method of producing journalism. The other is an exploratory design science approach where I have designed software prototypes that let me explore what journalists would like software to do for them, and also allows me to inquire about how journalists perceive computational methods when presented as something very concrete and tangible in front of them.

1.1 Introduction to the articles

While appended at the very back of the thesis, the research papers are the center of a doctoral student's life and work. I will now briefly introduce the content of these articles and explain how they are connected.

Paper I, *News applications – journalism meets programming,* is an analysis of 79 news applications – journalistic web application where custom code is written to tell stories in a journalistic context. The material was exclusively gathered from traditional media institutions online so to capture how the established gatekeepers of information utilize the web in its richer end in terms of interactivity and multimedia. The paper accounts for the basic concepts that enable newsrooms to publish interactivity through code as "frozen labor", in addition to "frozen speech" in forms of traditional media content. As to whether these applications are journalistic, I categorize them using a traditional content scheme for online journalism, as well as align them in a triangle consisting of the three core functions – information, arena, and watchdogging – and find these applications fit the yardstick well. These applications are continuations of journalistic traditions, but are created with an untraditional skillset we do not expect to find in newsrooms or teach in current journalism classes.

Paper II, *Computational journalism in Norwegian newsrooms*, is an interview study with programming journalists. This paper is a work of collaboration with Ph.D. student Joakim Karlsen, who is interested in digital storytelling. The interview guide for this study is largely built up around questions that arose from Paper I. The basic aim for this thesis was to figure out who these journalist-programmers are, what they do, how they work, who they collaborate with, and the premises for doing this type of work. A semi-structured interview approach with quite open questions was used to allow as much as possible to be described from their perspectives. As the papers' backbone we used the concept of computational journalism as a rhetorical craft, a perspective that underlines both how computational journalism is similar and different from journalism at large. We found the differentiating key skill (programming) to be indistinguishable from the problem-solving solutions they apply – a computational thinking that favors computational methods. We also found a strong focus on finding stories in data, and more traits of data journalism than computational journalism.

Paper III, *Newsworthiness on Twitter*, has a very different point of departure. One of the promising democratic aspects of the web is that it lets anyone express themselves in online debates. Through Twitter, a micro-blogging service, such debates accumulate over topics for those who have an interest in analyzing them. Topics that generate interest in the audience are by default topics that media institutions care about, and I wanted to explore the possibilities for facilitating analysis of such material. My approach was to cluster Twitter messages by grooming the language (applying stemming, removing stop words, giving key linguistic and media-specific elements greater weight) to automatically create subsections of a Twitter corpus with similar topics based on the words in use. This application was given a graphical user interface and evaluated by journalists with special responsibility or interest in social media. The evaluation focused on how the system was perceived in terms of utility and areas of improvement, but also how these kinds of methods were seen in relation to the participators' work responsibilities. The evaluated application was found to be interesting, but with some key flaws and good features both in design and requirements. Among the methodological shortcomings were the (still) quite noisy output and the lack of possibility to exclude material in the user interface, and among the requirements was the lack of focus on identifying who the authors of Twitter messages were.

Paper IV, *Watchdogging in code*, is another design approach that picks up the trail from papers I and II. A variable I initially coded[1] for Paper I was whether the applications' data were updated after publication. None were. When discussed in paper II with journalists who had programmed some of the applications from Paper I, it became clear that this was often an intended goal for the applications, but for various reasons this never happened. In *Watchdogging in code* I built a web application not too unlike some of the ones from Paper I, but I built it on top of a data API instead of an isolated data collection. This created a continuously running news application that updates as new data are exposed in the API. This solved the problem

[1] Coded - as in assigned variables from a coding scheme in the content analysis tradition, not as in programming.

of the application lagging behind, as data were in sync with the API. The data source I used was the Norwegian parliament API, and the designed prototype can be found online at wwww.samstemmer.net. The problems presented now are slightly different from those of the news applications described in Paper I. Now the journalistic angle of the application cannot be decided once and for all, and the potential unpredictability of live data must be given different frames. My approach was to let the parliamentary reporters explain the outline of a basic requirement specification by pointing out what works well and what does not, and through this dialogue try to capture how a parliamentary reporter would imagine such an information system. The "test" session focused more on exposing the data to the reporters than evaluating the currently implemented array of different reports/visualizations/hypothesis the application consists of. The results include the imagined features of a future system that watchdogs the parliament though code, but also a discussion of the neutrality/biases of a tool such as samstemmer.net. The question of who the journalist is and how they can verify their facts, becomes an issue when software takes the role of a watchdog.

1.2 Structure of the thesis

Writing an article-based thesis allows for small dives into different aspects of a phenomenon, but the article format demands a strong focus on presenting the studies' results. This creates distance between the papers as they approach the field quite differently methodologically. The subprojects also gathered data on wider aspects of the problems at hand than the papers present. The composite form of this first part of the thesis, the summary of the research contribution[2], contains a model of software-oriented news production that is not explicitly discussed in the papers, but that is a result of working with the material from different angles.

[2] This part of an article-based Ph.D. thesis is often referred to as the *final contribution*, but the contribution offered in this thesis is likely not the final word on computational journalism, hence the alternative terminology.

In chapter 2 I present how the overarching research question is broken into smaller subprojects. Chapter 3 contains a review of the field, with particular focus on the history of software-oriented news production. The fuzzy terminology used to describe how journalists create stories through software and software through journalistic needs creates an uncertainty in whether computational journalism represents a continuation, revitalization, or a theoretical proposition for a potential journalistic practice. I build on this literature and emphasize the differences in semantics used and journalistic foci and contexts to differentiate computational journalism from its predecessors when I propose a definition of computational journalism and a model of the field in chapter 4. This model is both a summary of relevant theory and a result of my own work, and is created (iteratively adjusted) alongside the work with this thesis. Methodological considerations and choices are explained and discussed in chapter 5. The papers results are summarized in chapter 6, before the results are discussed in chapter 7.

Interdisciplinary work, as research on the intersection of computing and journalism unavoidably is, challenges the fields it is intersected by. It will never be a "pure" version of its parent fields, and readers are thus warned: this is not a work on journalism or a work on information science, it is a work on computational journalism, which consists of both.

2. Research questions

In this project I aim to explore the intersection between information science and journalism studies, in particular the potential for computational journalism in this field. My overarching research question reads as follows: *How is computational journalism operationalized and how are computational methods perceived in Norwegian newsrooms?*

"Operationalized" in this context means "put into operation or use", as in "implemented" or "effectuated". This question is composed of two different, but assumed related, aspects of computing in the newsroom: 1) What kind of work it is and how it is situated in newsrooms, and 2) how other journalists see this kind of work. The assumption is that to understand computational journalism in a newsroom, one needs to have some understanding of how this newsroom understands computational journalism. This question has been approached from various angles, and has been broken down into smaller areas of focus in the different subprojects. I want to describe how computational journalism is effectuated or practiced and how this is understood by journalists – both those who program and those who do not.

News applications are one example of journalistic output that requires some more advanced technical skill, and usually some computer programming knowledge. *What are news applications, and how do they compare as journalistic products?* is the question raised in Paper I. In Paper II the questions aim to capture an understanding of computing in the newsroom from a programming journalist perspective. The opening question in this study was: *what is computational journalism to you*? The next approach was initiated by a need expressed by a journalist: the need to understand large collections of social media messages. In Paper III I asked by design, and produced a tool that clusters similar Twitter messages together and breaks down large messy collections into smaller more coherent subgroups. I asked professional journalists: *how does this approach align with the journalistic work of finding stories, sources, and arguments in social media messages*? This is one way of breaking down the question of how computational methods are perceived. Another is found in Paper

IV, where I asked parliamentary reporters: *what should a tool that monitors the parliament API be like*? Inherent in this question is a discussion around performing the watchdogging function of journalism through software, a concept that also implies that journalistic values and standards should be included in the software. How can we ensure that this happens?

3. Theories and concepts

On what basis should computational journalism be measured or interpreted? Is it a process, an occupation, a public service, a boundary object, a set of methods, a mindset, or perhaps all, some, or none of these things?

Computational journalism can presumably be understood as all of these things. As a proposed intersection of journalism and computer science it is a part of information science by both containing a social component (human actors/organizations/social structures) and a technological component of creating and using technological tools. Therefore, theoretical considerations should involve theories that incorporate both aspects.

As a practice in or a function of journalism, computational journalism also positions itself in a long tradition of journalism research. The production of news, or journalism's professional practices, is in this context usually found under the sociology of news. Schudson's four approaches to the sociology of news stand central in defining research perspectives in this field, divided into the political context of news-making, cultural approaches, and economic and social organization (Schudson in Curran and Gurevitch 2005, 172–190). In general this perspective is focused on "how journalism matters" (Zelizer 2004, 206). Alongside production we often find the political economy of news and journalistic ideology (Sjøvaag 2011, 10) and how journalism is produced operate within and strongly relate to these factors. In terms of ideology, Breed notes, "every newspaper has a policy, admitted or not", in his contribution to understanding how such policies are learned and why they are followed (Breed 1954). The functions journalism performs, such as gatekeeping, deciding how and what gets through to an audience (cf. White or Bleske in Berkowitz 1997, 63–80, or Schudson in Curran and Gurevitch 2005, 174), and establishing ideals such as objectivity (Tuchman 1972) and a notion of a social contract or journalistic responsibility in regards to its position in democracy (cf. Roppen and Allern 2013; Østbye 2009; Sjøvaag 2010), create the frames to understand how journalism matters.

Research into the sociology of journalism changed in the 1960s and 70s from a general focus on media effects to newsroom studies of production. This turn represented a shift in focus from actors (journalists, editors, etc.) to structures (that provide boundaries to actors such as a dominant consensus in political, economic, geographical, ethical, cultural, etc., questions). Later a focus was given to actors within a cultural perspective in social systems, that both influence and are influenced by the actors they consist of (Eide 1992). The emphasis on this double hermeneutic, as explained by Anthony Giddens, is used in many fields in the social sciences, including information systems and research into online interactivity (A. O. Larsson 2012, 57–71). As a new social practice, it is reasonable to consider computational journalism as a particularly "negotiable" aspect in journalism, and it is uncertain how a "stable for now" structure or actor of this trade will settle into, or become part of, Norwegian newsrooms.

Journalism serves the function of enlightening and correcting the public through information and exposure to diverse views and standpoints, which a democratic community at large benefits from. What news is, as a key question in the sociology of news, has been found to be a shared understanding across newsrooms around the world. "The primary purpose of journalism is to provide citizens with the information they need to be free and self-governing", write Kovach and Rosenstiel, following with nine principles to fulfill this task:

> *Journalism's first obligation is to the truth.*
> *Its first loyalty is to citizens.*
> *Its essence is a discipline of verification.*
> *Its practitioners must maintain an independence from those they cover.*
> *It must serve as an independent monitor of power.*
> *It must provide a forum for public criticism and compromise.*
> *It must strive to make the significant interesting and relevant.*
> *It must keep the news comprehensive and proportional.*
> *Its practitioners must be allowed to exercise their personal conscience.*
>
> *(Kovach and Rosenstiel 2007)*

Through journalism news is created, curated, and disseminated to a public with the need to fill gaps in their information. I find these principles useful in relation to

computational journalism, as they do not depend on the form of the output or particular contexts to work. They simply outline what journalism should be for us to treat something as news or journalism. Journalism claims a special position in the information society as it promises to be truthful and loyal to the public before any other interests.

Latent in these understandings of journalism is the media as a central component of a Habermasian understanding of a public sphere, where the media provides functions for reaching good decisions for a collective through exposing arguments to public scrutiny and deliberation. This makes journalism important to democracy. This is a given in the journalism community, but is not necessarily so evident to the spectator watching from the outside, who sees a lot of sport and entertainment and few high-impact Watergate-type stories. In addition to the public sphere function, journalism has given itself the mission to expose injustice. This function is brittle, culturally dependent, and in the eye of the beholder, and the possibility that it works as intended is clearer when looking at societies that do not have a functioning free press. Issues such as journalism's position between its political and economic dependence on various entities in society (such as the state or corporations) and its simultaneous need to stay critical and independent, are among the parameters for defining what type of media a country or state has, such as in the framework provided by Hallin and Mancini (2004).

Technology is hard to find in the classic sociology of news. That is, technology is often mentioned, but rarely discussed in detail and rarely given any significant position in relation to journalism. It is observed from a distance and with a self-evident naturalness, "these technologies [personal computers, online and database research, remote transmission, digital photography] are generally introduced to reduce labour costs and to provide the technical capability to make the newspaper more 'user-friendly', with more interesting and attractive page design" notes Schudson (in Curran and Gurevitch 2005, 178). Or as Zelizer states: "As journalism has expanded into new technological frames, the set of practices involved in doing news work has changed. For instance, typesetting skills of the print room have given

way to a demand for computer literacy" (2004, 42). The focus it is given is that it exists in the newsroom and that journalism happens around it, not how it works or how it is potentially a part of journalism itself. Bruno Latour has suggested that technology is the "missing masses" in sociology (Latour 1992), in the sociology of news it is at least taken for granted in much of the classic literature. This means that most of this theory can only function as a backdrop in computational journalism, as it does provide neither frameworks nor terminology or empirical evidence to how technology is a part of news production.

Research into the ideology of journalism continues to keep technology at a distance. In *What is journalism? Professional identity and ideology of journalists reconsidered,* Deuze put focus on how new media and multiculturalism interface with contemporary journalism. He argues:

> *[T]his approach is inspiring because it helps us to look beyond infrastructures (as in computer hardware and software) or representationalism (as in the number of minority journalists in a newsroom) when assessing what journalism as a profession is (or can be) in a context of fast-changing techno-logy and society. (Deuze 2005, 443)*

When later looking at journalism and technology, he focuses on multimedia as a possible umbrella term for "digital media, new media, information and communication technologies, internet, interactivity, virtuality and cyberspace" (ibid). The intersection of all this creates a convergent media, where "multi-skilling" (the mastering of newsgathering and storytelling techniques in all media formats) becomes a necessity.

In studying the production of online news, Klingenberg concludes that "[d]igital technologies have changed journalistic production in newsrooms, but not according to journalists' preferences" but instead in favor of "productivity, efficiency and profitability of news businesses" (2005, 62). Another way digital technology has changed journalism concerns how it is used "to learn about the stories that competitors and other players are working on" (Boczkowski 2009, 40). The web has not only offered news organizations a new platform for dissemination of news, it has

also given the user a chance to be a producer, through social media sharing sites[3] or services of media companies. While this has been theorized as a notion of a public sphere, it is also noted that "most news organizations are not enthusiastic about allowing audience members to become co-authors of content" (Mitchelstein and Boczkowski 2009, 573). Research into online news has kept focus on the new or promising aspects of the new platform, such as interactivity and multimedia (see Steensen 2010 for an overview). Still, online news is quite similar to news in general, and particularly to news on paper. A term for repurposing news for the web, noted by Boczkowski, is shovelware – "the taking of information generated originally for a paper's print edition and deploying it virtually unchanged onto its web site" (Boczkowski 2005, 55). While one particular case is described in the quote above, I think this illustrates how technology is seen to be insignificant and somehow detached from the message, which may indicate why the transformation into digital journalism is a slow process. Newspapers, radio, and television can all present journalism in forms such as news bulletins but also as documentaries, debates, and commentaries. Journalism is independent of, or at least adaptable to, the different media channels. This is, perhaps, one reason why technology is so subdued in the older literature. The shift to a fully digital platform creates at least one fundamental shift in the production of news: numerical representation. Both data coming in and going out to an audience are now (mostly) digital and thus programmable. A logical reply to this change would be to emphasize programming as a basic journalistic skill. This reasoning seems to be becoming more common now, and programming is becoming a more frequently used word in journalism research and education.

More recent sociology of online news has identified the "multilayered dynamics of journalistic work in the digital age" (Powers 2012, 25), where computer technology and programing get more attention. That technological work and journalism seem to blend poorly is one observation in this field. In the paper *In forms that are familiar*

[3] E.g. blogger.com, twitter.com or wordpress.com for text, flickr.com or instagram.com for images, youtube.com or vimeo.com for video. New services for online expression have arrived regularly over the last few years, and this trend is likely to continue as some of these services both become massively popular among the public and valuable on the stock market.

and yet-to-be invented, Powers (ibid) accounts for how technological work is presented in 939 articles in journalism trade industry publications between 1975 and 2011. The literature Powers uses are search results for queries containing "computer" and "news", or "programmer". He finds three distinct ways in which technological work is discussed: (1) as exemplars of continuity; (2) as threats to be subordinated; and (3) as possibilities for journalistic reinvention.

If we quickly jump to a theory in information science, Powers' finding overlaps nicely with Orlikowski's theory of *technology-in-action* as structural consequences of technological use as related to the enactment types (1) inertia, (2) application, and (3) change (2000). Orlikowski intends to provide a structuration theory that includes treatment of technology, as Giddens' theory does not directly address this. Information systems constitute parts of, and are used in, structures. The technology facilitates (arguably) some forms of use, but does not dictate how an artifact will in the end be used. Technology use in relation to facilities (hardware, software, etc.), norms, and interpretive schemes (assumptions, knowledge, etc.) creates structures (or an instance of technology-in-practice, where Orlikowski allows multiple parallel use-structures). Technology, as part of the structure, partakes in its own re-enactment by providing a specific constituent materiality inscribed by designers and previous users. While people through general use change the structures that can consist of technologies, programmers have a particularly central role as they can change not (necessarily) how technology is used, but what kind of functions it can perform. Software as rules or even laws (Lessig 2006) of social spaces partakes in shaping social action, and computational journalism can be imagined as such an action.

A different way of relating to journalistic values is by creating maps of the field through empirical variables of preferential data in a Bourdieuian tradition. Hovden (2012) offers such a map, or a space to map, journalistic traits in the Norwegian journalism field. His analysis outlines four different types of journalists, based on clustered ontological views on journalism as well as demographic variables and merits. These journalist types can be used to understand and explain how journalists relate to what journalism-internal power structures define as important or "good journalism".

On a practical level journalism is often described as a process, an understanding that is frequently noted in technology-oriented journalism (e.g. European Journalism Centre, 2010; Gynnild 2013; Meyer 1973). This process that consists of "information gathering, organization and sensemaking, communication and presentation, and dissemination and public interaction" (Nicholas Diakopoulos 2010). On a macro-level the process perspective opens for a discussion if computational journalism represents a favorable outcome in treating journalism businesses with a business process reengineering methodology (cf. Al-Mashari and Zairi 1999), to transform journalism into better version of itself. On a micro level, this understanding aligns well with the Heideggerian perspective of the Aristotelian description of *techne* – craftsmanship, a process of creation (Heidegger 2001). This perspective does provide good space for human or individual creativity and expressivity to form an object with a given goal, purpose, and context. Computational journalism as a method, occupation, or process makes good sense in this perspective.

Theories that provide artifacts with functional expressivity, such as Latours' actor-network theory or activity theory, can be applied if looking at concepts such as bias, or to understand what the technological impacts on journalism are. These theories underline human-computer interplay or cooperation as crucial to any actions performed by machines and grant non-humans some agency and acknowledge latent capacity for action in objects. These perspectives hold great promise for future research on computational journalism[4], and also steer the debate in the direction of describing computational journalism as boundary objects (Star and Griesemer 1989) as spaces for collaboration across social worlds (such as the hacks and hacker worldviews[5]). Theoretic approaches from science and technology studies represent a

[4] I have used actor-network theory in the formal requirements for the PhD work, in a non-published philosophy of science essay. The theoretical apparatuses presented in actor-network theory offer ample concepts to cope with journalistic technology, but demand empirical data with a certain contextual richness (e.g. detailed data from observations) that my studies have not emphasized.

[5] The organization named hacks/hackers (http://hackshackers.com) is based on the view that different worlds needs to collide and reorient: "Journalists sometimes call themselves 'hacks', a tongue-in-cheek term for someone who can churn out words in any situation. Hackers use the digital equivalent of duct tape to whip out code. Hacks/Hackers tries to bridge those two worlds. […] to invent the future of media and journalism" (Hacks/Hackers 2010).

different view than what the sociology on news has focused on, from the study of how journalism matters to how people and artifacts matter in journalism. It does not capture journalism in all its forms and from all angles, but it creates a space where technology and humans alike become important for understanding how news comes into existence.

In exploring new opportunities, such as computational journalism, it makes sense to keep the theoretical scaffolding to a minimum to avoid inhibition of creativity. The understanding of journalism I promote in this regard is a "back to basics" idea of accurate information as a necessity to make good personal and collective decisions. For computing, I suggest a broad understanding of the application of algorithmic treatments of data though a computer. What aspects of computing will provide fruitful interaction with journalism remains largely unknown and opening up the possibilities makes more sense for innovation and exploration than narrowing them down. For an example of how this can be applied as a framework, see Diakopoulos (2012).

While the theoretical sociological accounts of journalism give technology little space, journalism also has a history of software-oriented production. These practices create a space where computational journalism is less alien and new.

3.1 Software-oriented production of journalism

In order to position computational journalism in the tradition of utilizing computing in journalism, other waves of computer journalism efforts need to be accounted for. The nomenclature for computing in journalism is fuzzy, and also changes over time. In the academic literature and in online forums the same projects and efforts are frequently labeled under different names. "Computational exploration in journalism" is one label given to this development (Gynnild 2013) – a name that underlines the fact that we do not yet know how and what a sustainable stable merge of computing and journalism will be. A "final" or truly stable merge will never occur, as both

technology and journalism are changing all the time.[6] But as the various names for software-oriented journalism currently found in the literature contain semantic variation that suggests differences in skills and application, I will describe the most frequently used names before suggesting a model that underlines the subtle differences in the historical background.

3.1.1 Computer-assisted reporting & precision journalism

"Computer-assisted news reporting refers to anything that uses computers to aid in the news-gathering process" states Melisma Cox in the opening lines of her paper *The development of computer-assisted reporting* (Cox 2000). The name computer-assisted journalism is also sometimes used, but CAR, short for computer-assisted reporting, is used most often. Cox starts her narrative in 1952, when CBS used a computer to predict the election results in the American presidential election. According to Cox, this practice was pioneered by a handful of individuals, with Philip Meyer being central. "Philip Meyer can be credited as one of the innovators of computer-assisted reporting […] with his coverage of the Detroit riots in 1967" (ibid, 7). A few years later, Meyer published the landmark book *Precision Journalism* (Meyer 1973), which has been updated several times, but even from the first edition included insight into how computers can be applied to problems in journalism. "In this book [the 1991 edition], Meyer explains that beginning in the 1970s, journalism started to become scientific, a journalism which he labels as precision journalism" (Cox 2000, 8). Precision journalism is an effort to make journalism more accountable and scientific by applying methods from the social sciences (mainly statistical methods in Meyers' book); computers merely made this more practical. The fact that the computer became a defining factor of what CAR is, Meyer later writes to be an "embarrassing reminder" that journalism does not take technology for granted compared to other professions (Poynter Institute 1999).

[6] A stable or "stabilized for now" status (Orlikowski 2000) would in this context mean a readily identifiable practice that can be said to be similar enough across social contexts to be captured with the same term.

Following Cox's narrative through the 1970s and 80s we come to the introduction of databases as a journalistic tool. A key methodological trick that lead to several Pulitzer Prizes is the ability to join two datasets (e.g. persons driving school busses vs. persons convicted of traffic violations or who are drug dealers) to find intersecting rows, or to narrow the scope of large datasets to fewer candidates for hypothesis testing.

The basic tools of CAR are described as spreadsheets, database managers, and on-line resources. Cox also includes web access and e-mail as important technological advances in the CAR tradition. The tools included in the early days of CAR delude the significance of the name today, as e-mail, web searches, and word processing are no longer technological substitutes that distinguish the technologically advanced journalists from others – they are now standard tools used by everybody. Today these tools that became common property are usually not referred to as CAR tools or methods. Usage of technological tools still typifies the CAR tradition today.
CAR has also been studied as a practice in line with the tradition of newsroom studies, with methods such as qualitative interviews and content analysis (Parasie and Dagiral 2012) identifying a particular epistemology of CAR reporters, and surveys and questionnaires (Garrison 1998a) finding that larger newsrooms hold an advantage over smaller ones in the use of computer-supported methods.

The CAR tradition is still relatively strong today, with its own annual conference and teaching institution (National Institute for Computer-Assisted Reporting, NICAR), a wealth of reading material (cf. DeFleur 1997; Garrison 1998b; Houston 1996; Houston et al. 2002), and active mailing lists for collegial discussion and problem solving.[7] In Scandinavia the most successful CAR initiative was the Danish International Center for Analytical Reporting (DICAR), co-founded by Tommy Kaas and Nils Mulvad. Mulvad also authored a few books on the subject in Danish

[7] In particular, the NICAR-L mailing list from IRE (http://www.ire.org/resource-center/listservs) is a well-used and active channel.

(Mulvad and Svith 1998; Mulvad, Swithun helgen, and Svith 2002). DICAR was closed at the end of 2006.

Earlier this year, Espen Andersen (journalist and developer at the Norwegian Broadcasting Corporation, NRK) published a book titled *Datastøttet journalistikk* (Andersen 2013), a Norwegian phrase Andersen uses explicitly synonymously with CAR. The techniques and example projects mentioned in this book exceed the basic tools summarized by Cox when it comes to technological sophistication, but Andersen follows the same historical path from the 1950s, with pioneers such as Philip Meyer, and into the current world of seemingly abundant data with programming and databases as key tools.

3.1.2 Data journalism

In this context the word data describes digital structured or unstructured raw material that journalists use to investigate, argue, and explain facts. Typical examples of data are public data such as tax records, budgets, census data, etc., and private data such as social media messages (tweets, images, videos) and transaction logs (e.g. Netflix usage or cellular phone usage), or leaked data such as in the case of Wikileaks.

Working with data (public or otherwise) has been a part of journalism since its beginning (Rogers 2011), but the digitization of data has made this an increasingly more interesting path for newsrooms. Journalism's need to explain complex data to the man on the street has given a certain boost to data visualization and storytelling (cf. McGhee 2010; Segel and Heer 2010; Weber and Rall 2013). Data journalism is described as a growing trend in Europe, inhibited by lack of knowledge about how to work with data (Sirkkunen, Aitamurto, and Lehtonen 2011; Nygren, Appelgren, and Hüttenrauch 2012).

More recent books on computerized methods and data use in journalism include: *Facts are sacred: The power of data* (Rogers 2013) and *The data journalism handbook* (Gray, Chambers, and Bounegru 2012). The name "data journalism" might suggest a specialized form of journalism devoted to the collection and analysis of

data in line with the "analyst", "researcher", or the more recent "data scientist" roles – which use math, statistics, and more advanced forms for computing as central tools, but this is not the case in these books.

The term "data journalism" is found on awards such as the international *Data Journalism Awards* (Burn-Murdoch 2012) and the Norwegian *Prisen for årets datajournalistikk* [data journalism of the year] (NxtMedia 2013), but working with data is a central part of most computerized angles in the production of news. Working with data offers challenges to journalism beyond the technical (Sarah Cohen 2011), and is also included in the explanations for both precision journalism and computer-supported reporting.

Rogers' book offers the term "data journalism", synonymous with "computer-assisted reporting":

> *'Data journalism' or 'computer-assisted reporting'? […] These are just two terms for the latest trend, a field combining spreadsheets, graphics, data analysis and the biggest news stories to dominate reporting in the last two years. (Rogers 2013)*

Paul Bradshaw of Birmingham City University explains in *The data journalism handbook* that the difference between data journalism and "the rest of journalism" is perhaps the possibility to combine the traditional "nose for the news" with large amounts of digital data. "And those possibilities can come at any stage of the journalist's process: using programming to automate the process of gathering and combining information from local government, police, and other civic sources, as Adrian Holovaty did with ChicagoCrime and then EveryBlock" (Bradshaw in Gray, Chambers, and Bounegru 2012, 2).

Holovaty and his projects are cited in several of the above-mentioned works. His insight on the name and relevance matter can be seen in this short blog post:

> *It's a hot topic among journalists right now: Is data journalism? Is it journalism to publish a raw database? Here, at last, is the definitive, two-part answer:*
>
> *1. Who cares?*

> *2. I hope my competitors waste their time arguing about this as long as possible.*
>
> *(Holovaty 2009)*

One could argue though, if works such as *EveryBlock* need a label, database journalism might fit better than data journalism.

3.1.3 Database journalism

Analyzing a database or utilizing one for reseach are activities that are already claimed as precision journalism, data journalism, and CAR. What Holovaty suggests ("Newspapers need to stop the story-centric worldview" (Holovaty 2006)), and later does with *EveryBlock*, is to turn online news sites into more granular databases and produce structured information that can be resused at a granular level. An online news story should not be a "blob" or a "text", but a combination of the elements the story consists of (persons, places, events, dates, etc.) also on the database level, so that the individual pieces can be recombined for multiple and/or future-use contexts.

A different operationalization of this concept is found on *Homicide Watch D.C.*, where Laura and Chris Amico do crime reporting at a very granular level (Amico and Amico 2011). "Homicide Watch D.C. is built around 'objects'-incident, victim, suspect, case-and uses structured information about location, age and race to build a very detailed picture of this one type of crime in one city" explain Anderson et al. (2012, 30). As with *EveryBlock, Homicide Watch* allows for the reuse of story elements as structured data. One could call it "structured journalism" as suggested by Chua, who uses *politifact.com* as an example (Chua 2010). All these sites, to a cetrain degree, expose the structure of the database and make content avaliable through a URL structure that clearly maps to queries (e.g. homicidewatch.org/victims/method/shooting/ lists victims that were shot, and homicidewatch.org/suspects/gender/f/ lists suspects that are female).

"Database editor" occationally apears as a title in some newsrooms, but other than that the database journalism name has not seemed to stick. *EveryBlock* is now closed and Homicide Watch struggles to find a buiness model (Carr 2012), but the lessons

learned from applying a strict database logic to news content might prove to holde lasting value[8].

3.1.4 Data-driven journalism

Yet another more recent name for doing journalism with computers is "data-driven journalism". If we look at the categories from the international data journalism awards we find:

- *Data-driven investigative journalism: using data to uncover facts*
- *Data storytelling (text, visualisation, video…)*
- *Data-driven applications (mobile or web): serving data to your public*
- *Data journalism website or section*

None of these, though they are at times hard to separate, would fall outside of the scope of what computer-supported reporting, data journalism, and database journalism are described as doing.

The European Journalism Centre runs a project called *datadrivenjournalism.net* (#DDJ for short in other online contexts), which "is aimed at enabling more journalists to use data-sets as a source for reporting" (from the "about" section on the webside, European Journalism Centre, 2013). In the project's explanation of what data-driven journalism is they quote Jonathan Stray: "Data journalism is obtaining, reporting on, curating and publishing data in the public interest". Again, the terms are used synonymously. The organization's report from a 2010 symposium offers an "overview on what data-driven journalism might mean and how it can provide a new perspective for journalists" (European Journalism Centre, 2010, 5) and presents the

[8] Laura Amico of HomicideWatch is taking the concept into education, and to the next generation of journalist this might be a normal format. http://www.niemanlab.org/2013/09/laura-amico-from-homicide-watch-to-education-testing-a-new-kind-of-structured-journalism/

topic as data production, usage, integration, data visualization, storytelling with data, and new formats for presenting information and stories.

In *A new style of news reporting: Wikileaks and data-driven journalism* (Baack 2011), the terms are also used interchangeably with each other. Baack quotes The Guardian's data blog editor Simon Rogers on the issue of Wikileaks: "Wikileaks didn't invent data journalism. But it did give newsrooms a reason to adopt it".

As with database journalism, one could draw from the name "data-driven" that we describe a subcategory of (technological?) journalism here. In cases such as Wikileaks, large datasets arrive before any journalistic hypothesis or story idea is in place, and the process of analyzing the data drives the journalists towards a story they had no chance of knowing of before the data arrived. It becomes a "follow the money" or "follow the evidence" kind of game through datasets. This too seems not to be the case; these terms are used interchangeably.

Under the title *Data-driven journalism and the public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago*, Parasie and Dagiral describe two different ways of thinking between "old" computer-supported reporters and a newer wave of "programmer-journalists" (Parasie and Dagiral 2012). Beyond the differences in epistemologies that Parasie and Dagiral find, the programmer-journalists differentiate themselves in that they do write software code as journalists, not as engineers with a contract and a requirement specification. This sets the programmer-journalists apart from other journalists as toolmakers, not only tool users. It creates a slight shift towards computing/programming as a creative, contextually dependent craft that can be used not only in journalism, but also as journalism, and underlines computing as something more than a tool to manage and analyze data and databases. This element of professional orientation among differently skilled newsroom workers (designers, animators, programmers, etc.) is found to be a success criteria for the New York Times' newsroom (Weber and Rall 2013). It also creates new occupational titles, such as the aforementioned

"programmer-journalists", but also "news apps developer", "editorial programmer", and "hacker-journalist", labels not always easy to decide upon (Pilhofer 2010).

3.1.5 Computational journalism

"One thing machines do better is create value from large amounts of data at high speed. Automation of process and content is the most under-explored territory for reducing costs of journalism and improving editorial output", Anderson et al note in the report *Post-industrial journalism: Adapting to the present* (C.W. Anderson, Bell, and Shirky 2012, 25). This is what computational journalism aims to do: create value for journalism by applying computing to tasks journalists elsewise would do manually (or not do at all).

After a 2009 summer workshop entitled *Developing the field of computational journalism,* a provisional definition of computational journalism was given in an end-of-workshop-report:

> *For now though, we define computational journalism as the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism. In some ways computational journalism builds on two familiar approaches, computer-assisted reporting (CAR) and the use of social science tools in journalism championed by Phil Meyer in Precision Journalism: A Reporter' s Introduction to Social Science Methods (Rowman and Littlefield, 2002). Like these models, computational journalism aims to enable reporters to explore increasingly large amounts of structured and unstructured information as they search for stories (Hamilton and Turner 2009, 4).*

This definition is largely an updated version of Philip Meyers' precision journalism, but explicitly includes algorithms and focus on accountability. It is updated to fit a world with an abundance of important data, where keeping up with the scale is a problem.

A more process-oriented definition is offered by Diakopoulos in *A functional roadmap for innovation in computational journalism:*

> *I define Computational Journalism as the application of computing to the activities of journalism including information gathering, organization and sensemaking, communication and presentation, and dissemination and public interaction with news information, all while upholding values of journalism such as balance, accuracy, and objectivity (Nicholas Diakopoulos 2010, 1).*

The activities, the journalistic process, are emphasized here. It is the step-by-step process found in most introductory journalism books that is to be exposed to computing, while the values of journalism are to be upheld. This suggests that the introduction of computing might distort, obscure, hide, or affect elements of the process in a way a non-computer-supported process does not. The computation must be applied in accordance with the established values of the traditional journalistic profession. As such, the definition includes stronger non-functional requirements, or quality requirements, that demand computing incorporate – or align to – journalistic values than the above descriptions. It also ties computing and journalism together, as something more than just the combination of the two; it is a true meld, a new entity.

Variation of these definitions exists, but the general idea of "upgrading" the journalistic process with digital, computerized, algorithmic means and upholding the means and end of traditional journalism is established.

A hypothetical field?

In contrast to the forms of technology-oriented journalism I have mentioned above as practices that are performed in media production, it is not clear from the definitions whether computational journalism is something that happens in the world, or something we hope will happen in the world and therefore should put research efforts into. While cases in real-world media institutions can be pointed to, much of the literature that uses the term computational journalism is hypothetical. Flew et al underline this in their paper titled *The promise of computational journalism*. They explain what computational journalism is good for:

> *Ultimately the utility value of computational journalism comes when it frees journalists from the low-level work of discovering and obtaining facts, thereby enabling greater focus on the verification, explanation and communication of news. Such an understanding serves to dissolve the illusion that news providers employing computational journalism can automatically deliver better news to their readers simply because they are able to move more information about at faster speeds, and from more remote locations. In other words, computational journalism has less to do with systems that transmit data and information only as a commodity. Computational journalism, like journalism per se, is a constructive, meaning-making enterprise. (Flew et al. 2011, 167)*

This is a supposition and does not clarify whether this should happen or if it actually happens. In exploring computer games as an interface to news, Bogost, Ferrari, and Schweizer note that "[t]hese future *computational journalists* will spin code the way yesterday's journalists rattled off prose" (2010, 178). Further hints of a hypothetical field are found in papers such as *Computational journalism: A call to arms to database researchers* (Cohen et al. 2011) and books such as *Understanding digital humanities*, where different tools and formats are imagined:

> *For example, one could imagine a form of computational journalism that enables the public sphere function of the media to make sense of the large amount of data which governments, among others, are generating, perhaps through increasing use of 'charticles', or journalistic articles that combine text, image, video, computational applications and interactivity (Berry 2012, 15).*

In a speech at the 2013 symposium *computation + journalism,* initiator Irfan Essa summarized that there is "No need to define 'Computational Journalism' or 'Journalism'" and "Let's stop defining things, but building/doing". If computational journalism is a matter of creating tools for journalism, the current literature on what journalism is should suffice to define the non-functional requirements for such information systems. But in order to study computational journalism as a potential social creative craft performed in newsrooms, some defining limitations are useful in identifying and discussing the practice. An intersection of computing and journalism suggest both the creation of technological tools, but also the use of such tools. Definitions can describe how and what elements of journalism need to be incorporated into computer systems to ensure successful tools and a meaningful

practice. For a craft practiced in newsrooms, the "what" is equally important as the "how" in bridging the gap between two fields that traditionally have attracted people with quite different mindsets, skillsets, and values. While perhaps rare, computational journalism can now and again be observed as an operationalized practice, but this is not exclusive to newsrooms.

Computational journalism operationalized

Computational journalism, as one understanding of "computational exploration in journalism", exists in academic, entrepreneurial, and newsroom contexts (Gynnild 2013). A closer look at how it is operationalized in the different contexts underlines the distinguishing elements in computational journalism compared to earlier efforts.

Entrepreneurial efforts

Entrepreneurs push computing in journalism forward from the outside. Much innovation in technology is pushing journalism without being tailored or adapted. That we today can record video and create a full multimedia story on a mobile phone is quite remarkable in a journalistic context, but it is not by design a journalist tool. The same can be said about countless useful inventions and technologies. Software examples in use in journalist storytelling include *Storify* (storify.com), research tools such as *Openrefine* (github.com/OpenRefine), charting libraries such as IBM's *Many Eyes* (www-958.ibm.com), *Tableau Public* (tableausoftware.com), and *Highcharts JS* (highcharts.com), and crowdsourcing tools for gathering information such as *Crowdmap* (ushahidi.com), and numerous others. They are useful and have an impact on journalism, but are not designed for journalism as a primary field, similar to other general software tools such as spreadsheets and word processors.

Adrian Holovaty's *Everyblock* is different. It was built to fulfill the information function of journalism, and designed within those frames. It was also funded as a winner in the Knight News Challenge 2007 (newschallenge.org). Similar is Jonathan Stray's *Overview* (overview.ap.org), a document-clustering tool for journalists to categorize unstructured text documents. It was designed to solve a problem journalists face and need good trusted tools for. It is integrated with *DocumentCloud* (documentcloud.org), a tool also designed for journalists, which helps reporters

manage, analyze, and publish documents. These tools were created to solve journalistic problems on journalistic premises.

Another entrepreneurial effort, Narrative Science (narrativescience.com), has taken an interesting approach to journalism. They produce text stories from structured data. Statistics from a children's baseball match can be computationally analyzed and written as a textual story – the kind of story that is rarely covered. Financial data has also turned out to function as input for these story-writing machines. For some journalists this might seem like a doomsday device, while others see it as a future that must be adapted to (Morozov 2012; Farr 2013; Fassler 2012). What this "automated journalism" turns out to be in the end is not clear yet, but its origin is. Narrative Science grew out of an academic research project called *Stats Monkey* at Northwestern University ("Intelligent Information Laboratory at Northwestern University - Projects - Stats Monkey" 2013).

Academic efforts

In research and higher education, computational journalism exists both as a field of research and as a field of study for students.

The use of journalism as a field of teaching computing has proved fruitful for both younger (Wolz et al. 2010) and older students (Pulimood, Shaw, and Lounsberry 2011). In recent years more specialized computational journalism classes have been offered at several teaching institutions. By looking at the content of these classes we get an understanding of what these schools suggest as important methods and theories. The topics included in such classes include web programming; SQL; text/data mining (NLP); social computing; development/deployment for the cloud; journalistic practices in the digital age; visualization; structured journalism & knowledge representation; network analysis; computer security, surveillance & censorship; web design; database design; data journalism and investigative reporting. The curricula naturally vary a bit from school to school, but the general idea of computational journalism being an intersection of computing and journalism that

requires both technical and journalistic skills is rooted in all of them.[9] A blog post from 2011 by Interactive Technology Editor at the Associated Press Jonathan Stray (who later came to teach computational journalism at Columbia University) titled *A computational journalism reading list,* suggests literature that largely overlaps with the above-mentioned school curricula. He also states that "'Computational journalism' has no textbooks yet", but provides a good outline of what one could include in the list (Stray 2011a).

The academic research literature on computational journalism mainly falls into one of two categories: computer/information science or journalism studies. The reason for this probably has more to do with the academic traditions of publishing than the subject matter. The label "computational journalism" in academic research requires the work to be relatively recent, and that the researchers choose to frame work in this way. Work that is highly relevant to journalism in computer/information science is often published without touching the semantics and references I have outlined here, but does not use the computational journalism tag or key word. Works such as *Weaving a safe web of news* (Kiscuitwala et al. 2013), where a platform for safe communication for citizen reporters is build and discussed, and *Information credibility on Twitter* (Castillo, Mendoza, and Poblete 2011), that explores metrics for identifying the credibility of news on Twitter, can serve as examples. Computational journalism has yet to become an advantageous tag to label computer/information science. This is not an absolute though, as exemplified by Diakopoulos (2010) and Diakopoulos, De Choudhury, and Naaman (2012). It is also under the initiative of the mainly technologically oriented Georgia Institute of Technology where a series of symposiums on computational journalism were initiated and held.[10] Not only does academic research on computational journalism result in knowledge and papers, but

[9] The topics mentioned are found in curricula from 1) The Tow Center for Digital Journalism, Columbia Universirty http://www.compjournalism.com/?p=84 2), Georgia Institute of Technology http://compjournalism.wordpress.com/ 3) Duke University http://www.cs.duke.edu/courses/spring12/cps296.1/ and 4) Arthur L. Carter Journalism Institute at New York University http://journalism.nyu.edu/undergraduate/concentrations/computational-and-digital-journalism/

[10] In 2008, http://www.computation-and-journalism.com/symposium2008/, and 2013, http://computation-and-journalism.com/symposium2013

also in tools. Examples include *Jigsaw* (Stasko, Görg, and Liu 2008), *Timeflow* (Viegas, Wattenberg, and Cohen 2010), *SRSR* (N. Diakopoulos, De Choudhury, and Naaman 2012), and *NewsCube* (Park et al. 2011). The research community at Northwestern University has developed a bundle of applications in this niche with examples such as *TimelineJS* for visualizing timelines and *SoundCite* for citing audio on the web.[11]

In journalism studies the introduction of new technologies and production techniques is noticeable. The focus here is rarely on tools or technologies, but on the sociological aspects of computing in a newsroom. The fact that American journalism is in a major economic crisis is perhaps also a driving force for exploration not only in business models, but also in the creation and management of news production. Computation can be used to speed up work and increase efficiency in almost any field, and journalism is assumed to not be an exemption. Jacobson's content analysis of the New York Times' multimedia output includes several exotic categories indicating re-thinking of what online news can be (Jacobson 2012). The creation of these kinds of products is also studied and new ideals are found, such as an "open-source or hacker culture" (Royal et al. 2012, 5-24) and that "[w]hat is new is that even programmers and designers belong to the journalistic team of the newsroom and define their task as a journalistic one" (Weber and Rall 2013, 164). As computing affords a technology-focused approach to journalism innovation, journalism is being studied in alignment with other (digital) cultures, such as the open source and hacker cultures (Lewis and Usher 2013).

Much research in journalism has focused on hypertext, interactivity, and multimedia in online journalism (see Steensen 2010 for an overview). While online journalism is, perhaps, the place we expect new things to happen, given the digital platform, computing is rarely mentioned at all. The most focused and recent literature in journalism that explores journalism in a "computational light" is already mentioned

[11] See http://knightlab.northwestern.edu/projects/ for details and more examples.

above (e.g. Flew et al. 2011; Chris W. Anderson 2012; Gynnild 2013; Parasie and Dagiral 2012; Royal et al. 2012; Weber and Rall 2013).

Newsroom efforts

Producing content, matters more than exploring technological possibilities in newsrooms. Newsrooms have a strong internal culture focused on story creation, and we do not expect to see big technological innovations come from this environment. We might expect newsrooms to utilize new technology (such as the infamous CNN hologram from the American 2008 election, a technology provided by the company Vizrt), but not create it themselves. Newsrooms are traditionally technology users, not producers. People who create technologies have accordingly traditionally found jobs in other places than newsrooms. As digitization increasingly makes journalistic work a matter of manipulating computers, the matter of how computers partake in news production becomes increasingly more relevant. The perspective that bias and ideology exists in algorithms is no longer a thought experiment for wine drinking computer enthusiasts, it is a matter of fact that newsrooms need to include both as creators of computer-supported journalism, but also as supervisors of other actors that create digital media content (such as governments, corporations, and individuals).

A captivating, if not to say overly optimistic, example of computational journalism is found at the Washington Post, and their *Truth Teller* prototype. “The goal of Truth Teller is to fact check speeches in as close to real time as possible” (Haik 2013). This was executed by applying speech-to-text algorithms alongside lookups against a database of known facts. How successful *Truth Teller* was is so far unanswered, but the idea shows that the journalism community has problems they would like to solve using computers.

A smaller, single-story example of utilizing computation is the example of how the NYT story *How Mariano Rivera dominates hitters* (Roberts, Carter, and Ward 2013) was produced. An application was written in *Processing*[12] to investigate data on the

[12] Processing is a programming language “initially created to serve as a software sketchbook and to teach computer programming fundamentals within a visual context”, see http://processing.org for more info.

successful athlete Mariano Rivera's baseball pitches, and create visualizations for a video describing his technique (González Veira 2013; found in Weber and Rall 2013). Here, custom code was written and applied to a single story (but perhaps reusable for future analysis), which exemplifies how a larger dataset (1300 ball throws in multiple x/y/z coordinates) can be dealt with by computational means that what we normally expect news media to analyze in detail. While programming was used in the project, it is tempting to label it as an advanced form of data journalism that has similarities with the CAR tradition. The matter of aligning computational journalism is a matter of aligning overlapping entities.

To create a single story, as in the case of NYT and Mariano Rivera, takes a lot of resources. Data collection, analysis, animation, programming, and video creation all take time and resources. For the coding part, a natural goal would be to reuse it as new data arrives, or on larger datasets. A model for this is observable at the Chicago Tribune with their News Applications Team.[13] By building news applications instead of single stories, the Chicago Tribune creates databases the audience can browse, visualizations to aid narration of complex data, and interactive news experiences in general.

Through Gynnild's idea of computational exploration in journalism it is clear that computational journalism is not a field that exists solely inside newsrooms. Still, newsrooms are a likely place to find this as a practice. A definition of computational journalism should thusly encompass computational journalism as a practice both inside and outside of media institutions.

A note on crowdsourcing

"When you aggregate enough individual participants, you get a crowd. One thing that crowds do better than journalists is collect data" (Anderson, Bell, and Shirky 2012, 24).

[13] They keep a blog at http://blog.apps.chicagotribune.com that gives some insight into what they are doing.

A number of the texts mentioned, including the above-quoted Anderson et al., that deal with computing and journalism also deal with crowdsourcing. The topic has also been given specific attention as a key method.

The world's most famous crowdsourcing project is perhaps Wikipedia, but in journalism The Guardian's work on the 2008 UK MP expense scandal is likely to come close (see Daniel and Flew 2010; M. Andersen 2009). This was a true landmark project that stimulated a growing faith in this method, and that demonstrated the potential of letting the audience work/participate. In Norwegian newsrooms this is noticeable as crowdsourcing was mentioned in many of the interviews as an area where they wanted to do more work in the future. Crowdsourcing is indeed an example of computational thinking, and is demonstrated to be utterly useful and successful in some areas of journalism. In an information perspective this has also proven useful, e.g. in 2009 Verdens Gang created the web portal vaksineguiden.no where readers could contribute local instructions about how the mass vaccination program for H1N1 (swine flu) was organized in the 429 Norwegian municipalities, an information problem the central government struggled with. From a watchdogging perspective, the MP expense scandal serves as an example of success, while the same procedure used in Bergens Tidende in 2013 turned into a more toothless endeavor.[14] This method does not make magic alone; some useful or scandalous data needs to be involved.

In relation to computation, crowdsourcing represents an inverted mode: usually human input is computed by software to produce output, but in crowdsourcing data are exposed to humans as processors to process (compute/collect/improve/assess/categorize/pin-point, etc.). It is about managing and aligning data and the audience for interaction. It is creating platforms for co-

[14] A Bergens Tidende research team asked the audience to help them go through receipts from public bodies' expenses (http://www.bt.no/nyheter/innenriks/I_verdens_rikeste_land/Hjelp-oss-a-sjekke-2952534.html#.Uk1bl2SpZJU) with results such as *Full fest på statens regning* [Party at the Governments' expense] (http://www.bt.no/nyheter/innenriks/I_verdens_rikeste_land/Full-fest-pa-statens-regning-2954892.html#.Uk1bKGSpZJU), where it was exposed that public employees sometimes drink more alcohol at official dinners than the guidelines prescribe. The consequences were small if not absent. The same method was applied by Verdens Gang in 2012, also without major scandals such as the British MP expense case (http://www.vg.no/spesial/2012/depdok/).

production or co-investigation, and thus is a child of social computing. While the "computing" part of computational journalism is toned down to a minimum in crowdsourcing, (the software development part consists of creating a tool for exposing data or collecting data, or in some lucky cases just using such a tool), it still fits my criteria for computational journalism. Originally, computers were indeed humans, and "computer" a job description for one who computes (performs calculations). While crowdsourcing is an interesting and new way of producing journalism, perhaps even a field of its own, the computing (done by machines) part is rather meager.

4. Aligning computational journalism

The following alignment of computational journalism is partly based on the literature review in the previous chapter, but is also formed by the work with and results from the articles. Figure 1 shows the latest version of the alignment between computational journalism and other often used terms for software-oriented news production – a model that has been reorganized and reconfigured multiple times during the last few years.

Are precision journalism, CAR, data journalism, database journalism, data-driven journalism, and computational journalism just different names for the same thing?

They all have in common a computer-oriented approach to journalism and the branding of this activity; they all also separate the practitioners from "regular" journalists. They all require specialized skills in more advanced use of computers. To argue that these things are the same, rebranded every few years in order to stay new, fresh, and interesting is not totally wrong. Philip Meyer, one of the men accredited as a pioneer of CAR, argued over 10 years ago that we should stop using the term CAR, as working with computers "no longer defines us", and that we needed to "move on to a fresher, more ambitious concept" (Meyer in Poynter Institute 1999, 5). Staying fresh is one reason for the plethora of names for this concept.

But there are differences. In essence, precision journalism emphasizes the use of scientific methods, CAR emphasizes digital tool use, database journalism emphasizes structure of information storage and retrieval, data and data-driven journalism emphasizes finding stories in data sets, while computational journalism emphasizes the merging of computing and journalistic values in tool creation and method application. There are subtle differences in the semantics, as well as the journalistic foci.

Table 1 Comparing software-oriented modes of news production.

	Precision Journalism	*CAR*	*Data Journalism*	*Database Journalism*	*Data-driven Journalism*	*Computational journalism*
Focus	Make journalism scientific	Utilizing computer tools to produce journalism	Finding, analyzing and presenting data as/in journalism	Adding and exploiting the advantages of structure in data journalism	Pursue unknown or presumed stories by following the "data trail"	Creating, adapting or using computational tools and method in/as journalism
Distinctive skills	Social science methods	Advanced computer tool-use	Data wrangling, data storytelling	Database theory & practice	Analytical, investigative research	Computational thinking, programming

Input/output

All these share fundamental foci and skills, such as producing news by means of computers, providing citizens with important information and a general "nose for news" and the need to balance and explain results of analysis in disseminating news items to audiences. While the similarities perhaps are easier to pinpoint than the differences (many of these names are indeed used interchangeably by both scholars and practitioners), the names, as descriptions of practices in journalism, suggest variations as shown in table 1.

In input all of these names suggest that data (structured or unstructured, digital or analogue datasets) are to be transformed or treated in order to become journalistic output. Especially the names "data journalism", "database journalism" and "data-driven journalism" suggest this. A consequence of making journalism scientific, and in examples in Meyers book, the collection and analysis of data also requires precision journalism to have data collections and input. Computational journalism shares this with all the others, but as trade that also creates software it also allows computable models to function as input.

The output from none of these are defining for the practices, and can be in traditional forms such as textual stories in newspapers, manuscripts for anchormen in studios for

radio or TV or new forms such as interactive multimedia products on digital platforms. Computational journalism is different in regards to output as it potentially produces software as news (e.g. as news application) or for newsrooms (e.g. DocumentCloud).

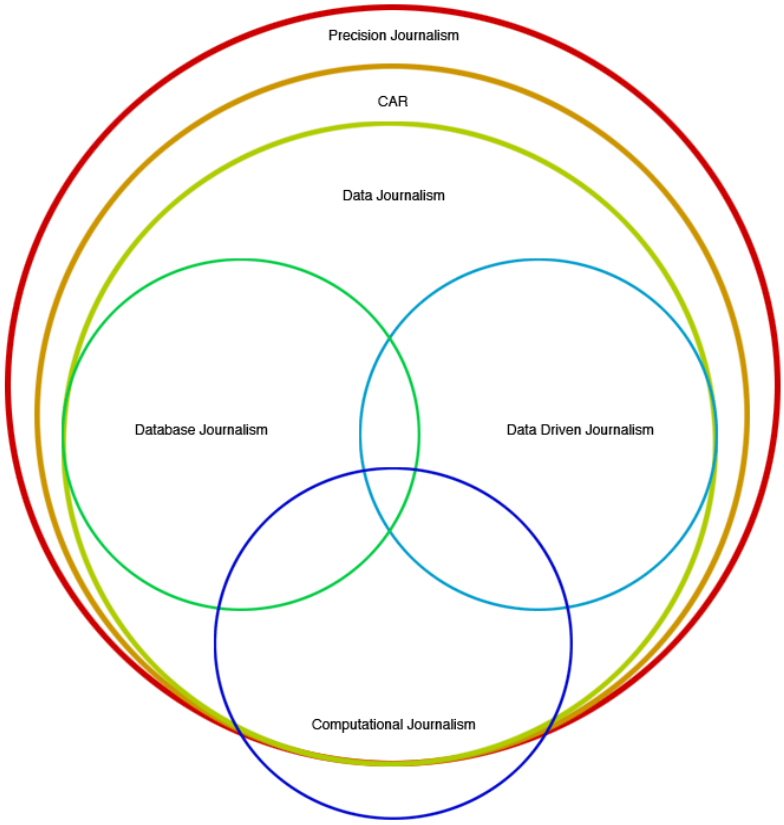


Figure 1: Computational journalism positioned with other types of computer-supported journalistic efforts. The rings bear solid borders in this illustration, but the borders between the practices are actually quite fuzzy. The amount of overlap between the different journalistic types is also made for illustrative purposes.

This chart can be used to plot journalistic output, but also to read the skillsets necessary to produce the various journalistic outputs. Read from top to bottom, this figure positions computational journalism in relation to other names for doing journalism with computers. The whole precision journalism tradition is story-centric, and so are the sub-elements in my illustration. Computers in general, and the CAR movement specifically, make journalism more scientific and fit within the precision

journalism it came from (one could arguably do precision journalism without a computer, but modern day CAR falls inside Philip Meyer's precision journalism). Data journalism, if not completely synonymous with CAR, falls inside this tradition. Data-driven journalism, if interperated as different from data journalism, falls inside it and overlaps with database journalism. Computational journalism overlaps with all of these, but also covers a field outside the story-centric tradtition of doing journalism with computers. Computational journalism is also initiated from outside of newsrooms and is described as the intersection between computer science and journalism. Computing can be applied to journalism without being story-centric, but still be very important to journalism. Creating a general tool such as a clustering algorithm or a database engine falls outside of this scope, but creating or tailoring such tools for journalism falls inside. Indeed, Christopher Groskopf did develop *PANDA*, a database/data management tool tailored for newsrooms (Coulter 2012) and Jonathan Stray did create *Overview* a "general-purpose document set exploration system for journalists" based on clustering (Stray 2011b).

The fact that computational journalism does not fully overlap with the other data and story-centric efforts allows for the explanation of efforts where, for example, models are presented or games and other forms of computer-supported layers are applied in journalism. It also allows for the inclusion of work that is independent of journalistic institutions and traditions, but still incorporates the goals and criteria for what we normally describe as journalism. This could be NGOs, bloggers, citizen journalists, etc.

4.2 Computational journalism defined

I summarize based on my interpretation of the historical background and argue that computational journalism differs from the other mentioned computer efforts in several distinctive ways:

1) Platform-centric instead of story-centric.

Computational journalism is initiated from outside of the newsrooms, and is so far anchored in academia rather than in the media industry. This leads to a shift from the story-centric way of thinking that dominates the newsrooms to a more platform- or product-centric thinking that goes with the tradition of information systems. By platform I mean spaces or opportunities for expression of opinion and spaces or opportunities for analysis and interpretations. As opposed to facilitating the narration or exploration of one story, it facilitates the narration or exploration of multiple stories or aspects of stories. For computing or software development to make sense in a newsroom beyond CAR or data journalism, the systems that get produced need to run over time, longer than the spotlight time a typical news story gets. This is an underlying assumption from a computing perspective. One single story will not weigh up for all the hours of work software writing takes, so the software must handle more than one headline. The Mariano Rivera baseball-story from NYT serves as an example; analyzing and visualizing larger data sets through custom code is extremely resource intensive, but if the code can be run every time new data arrives, or for all players in the league, we have transformed a story into a platform for finding and telling stories. We need to create and allow systems to run continuously as new data arrives, or support frequently repeated tasks to achieve this. This is a way to exploit that work done in software scales much better than other forms of journalistic work.

2) Can add computable models.

Another difference computing represents is adding models as a base for stories rather than data collections. A model in this context is a set of assumptions or definitions that define aspects of the world, rather than measured records of individual data. Examples of models can be the tax system for a country, distances/transport speeds to assess feasibility of movement on a schedule, the economic structures surrounding piracy in Somalia (Bogost, Ferrari, and Schweizer 2010), or the anticipated growth of the population and housing prices, etc., in an area to discuss city planning.

3) Applies computational thinking.

Computational thinking is a take on problem solving that emphasizes the delegation of tasks between man and machines as a key point:

> *Computational thinking builds on the power and limits of computing processes, whether they are executed by a human or by a machine. […] Computational thinking confronts the riddle of machine intelligence: What can humans do better than computers? And what can computers do better than humans? […] Computational thinking involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science. Computational thinking includes a range of mental tools that reflect the breadth of the field of computer science (Wing 2006, 33).*

While it is common sense from a computer science perspective to exploit computers' capabilities where possible, it requires insight to know when and how to apply computation to successfully solve a problem. The perspective that allows for efficient utilization of computation is unevenly distributed in society and is often clustered in pure technology businesses or departments. Computational journalism requires application of computational thinking in journalism.

In order to account for the goal and direction for computational journalism in multiple environments, I define computational journalism as the overlap between computing and the purpose and goals of journalism as summarized by Kovach and Rosenstiel (2007). This includes efforts in non-editorial spaces such as entrepreneurial and academic and does not limit the field through the established practices in newsroom cultures. As long as technology is created and adapted in alignment with the reasons for championing journalism as a democratic boon, computational thinking is applied to solve information problems important to society, and the activity has a public audience in mind, I consider it computational journalism.

My definition contains a strong normative notion, as opposed to a purely descriptive account. The purpose and goal of journalism, as described by Kovach and Rosenstiel, are normative; journalism is a trade based on ideals and ideas claiming that enlightened people are capable of making better individual and collective decisions. As a part of journalism at large, witch often is defined in normative terms; a normative definition makes sense also for computational journalism. The normative foundations are implied in the concept of the Forth Estate (Eide 2012), a concept computational journalism, as any other serious journalistic endeavor should aim to

fulfill. Kovach and Rosenstiels' principles are also technologically neutral, and does not depend on a particular organizational form (e.g. a traditional newsroom), an element that allows computational journalism to be performed by anyone *or anything* that aims for fulfill these principles.

My definition also deviates from the other mentioned definitions of computational journalism. Hamilton and Turners definition includes one particular aim to I find too narrowing " …aims to enable reporters to explore increasingly large amounts of structured and unstructured information as they search for stories". This puts computational journalism in place as a function of speeding up journalistic research, but exempts many other known and yet-to-be-invented use-cases for computing in newsrooms. Diakopoulos' definition is sufficiently agile, as it is quite general in relation to what the "activities of journalism" and "values of journalism" are, but exemplifies values with "balance, accuracy and objectivity" and objectivity is still a highly problematic value many journalists have abandoned. I find the values of journalism in Kovach and Rosenstiels' principles more concrete and applicable. They are also derived from a wide range of newsrooms and practitioners in different locations and contexts, and provide a more both explicit and universal model for how journalism can "…provide citizens with the information they need…".

5. Methodology

Before we begin, a note on nomenclature. When I write “research design” I describe a framework a researcher creates to conduct research – a map or plan of how the research shall be done. When I write “design science research” or “design science” I refer to the research paradigm, not to be confused with “design research”, which is “the study of design itself and designers - their methods, cognition and education” (Vaishnavi and Kuechler 2004). I will not be discussing the latter in this thesis.

5.1 How can we study computational efforts in journalism production?

The overarching research design for this thesis consists of a set of methods and contexts that seem promising in providing understanding of what computational journalism is and how this is perceived by both programming journalists and more traditional journalists. In retrospect, I note that the research design shares a lot with that of a collective case study, where multiple case studies are selected to illustrate an issue (Creswell 2009, 74). While the project as a whole follows this structure of inquiry, the subprojects utilize different methodological foci of data collection. The largest distinction concerning this is the utilization of design science research in two of the projects, and a more traditional media studies approach with text analysis and interviews in the other two. The contexts are online news, newsrooms, and experimental settings.

The main research questions raised in this thesis are so general that they can, at best, be answered indirectly. They also span what can reasonably be answered with one single method. The methodological tools that I have utilized in this study still fall within the traditions of the social sciences, and information science in particular. The approach I have taken includes the study of artifacts produced by journalists through computational means, interviews with practitioners of software-oriented news production, and the design and evaluation of artifacts specially crafted to fit the scope of computational journalism.

The methodological considerations are mainly done to frame computational journalism from two different angles: 1) as a social practice in newsrooms and 2) as an explorative field of design science. Papers I and II take the newsroom approach and papers III and IV apply design science methodology.

On the qualitative – quantitative spectrum all the utilized methods fall on the qualitative side of the scale. A flexible research design allows for exploration and has the advantage of letting the data guide the outcome of the studies to a stronger degree. As relatively little research exists on computational journalism, it is non-trivial to point to good quantitative measures that capture it neither as a performance by journalists nor as journalism performed by machines. Computational journalism can be studied through quantitative means, but I have chosen qualitative methods as I want to let relevant actors (journalists) partake in defining what computational journalism is, and how it can be understood. Comparative efforts could also be applied. This can and should be done in the future; for instance, a cross-Atlantic comparative analysis of newsrooms would be very interesting under the hypothesis that North American newsrooms lead the way in journalistic innovation in this field, but I have focused on trying to initially provide an account and understanding of the phenomenon.

5.2 The products

Paper I offers an analysis of news applications – journalistic products written in code. These are not the only form of products produced with the aid of custom code for newsrooms, but they represent the most visual cue of such practices to the audience and thus a reasonable point of departure for the study of this as a practice.

All journalism is sooner or later about creating a product – a story or a piece of information for an audience. While not all products of computation in journalism can be identified as a particular product of computation (it can, for instance, be fact-obtaining and validation as the basis for a traditional story on television or radio or in a newspaper), some can. Journalistic products are regularly studied and measured, in

content analysis (e.g. Neuendorf 2002; Sjøvaag and Stavelin 2012), in framing analysis (e.g. Entman 1993), and in critical readings of various kinds. Journalistic stories and services that are the result of computing can be analyzed through these means. Variations in classic content analysis have been used to illuminate aspects of computer-enabled journalistic stories (e.g. Parasie and Dagiral 2012) and to identify likely candidates for such stories under labels such as "interactive infographic" and "extended multimedia" (Jacobson 2012).

My study of news applications has several methodological weaknesses that partly arise from the lack of a tradition of analyzing this type of product. The selection and exclusion of material is hard. I collected a list of units I found to fit my criteria (being a journalistic product that contains custom written code to tell a story) and used the audience of the blog www.voxpublica.no to direct me to similar examples.[15] Thus, the selection strongly depended on both the performer and the audience that got to adjust the sample.

The initial goal for the sample was to provide some subcategories, genres, or archetypes of news applications similar to the study *Narrative visualization: Telling stories with data* (Segel and Heer 2010). With a set of 35+ variables distributed among the model, view, and controller components of a modern web application[16], I tried to capture some key features and groups of similar types. The working hypothesis was that through similar attributes some distinguishable traits would emerge, such as map-based, timeline-based, "top lists" or comparative applications. This approach did not result in any clear patterns or generalizable types of application, and this mode of analysis was abandoned for this subproject. The knowledge of the variation and overlapping features in the sample is still valuable.

[15] The blog post and list can be found at http://voxpublica.no/2010/10/nyhetsapplikasjoner-pa-web-hvem-hva-hvordan/

[16] Model-view-controller (MVC) is a software architecture that separates representations of information from the user's interaction with it. It has become a normal structure of many web development frameworks. MVC was invented by Trygve Reenskaug in the late 1970s (Reenskaug 2013).

The actual alignment among the journalistic functions was done physically with printed thumbnails of each application laid out on a large table. This template approach, or matrix analysis (Robson 2002, 458), rearranging according to theoretical concepts, was an alternative to the initial unsuccessful immersion approach.

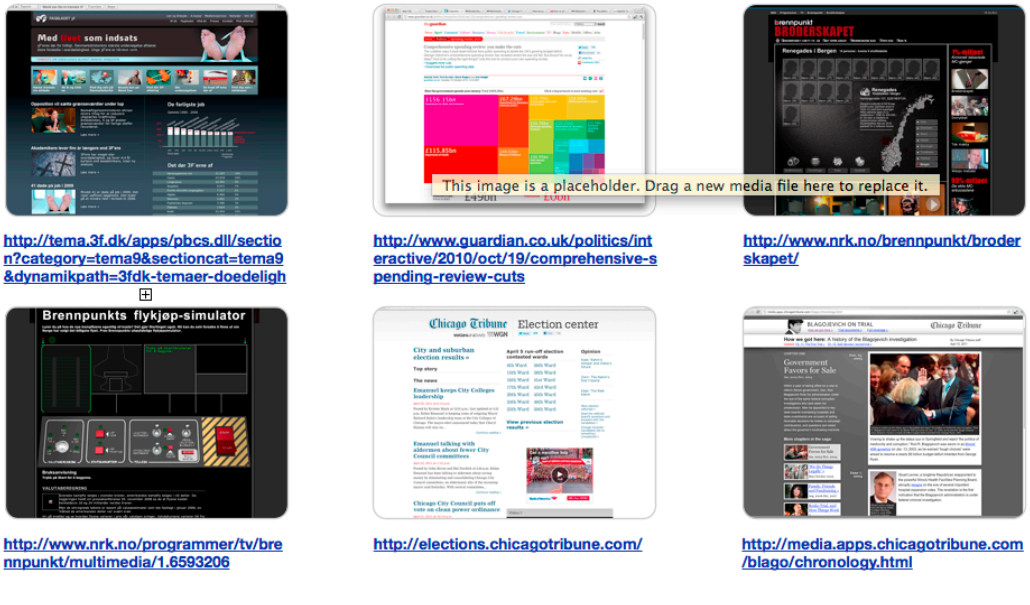


Figure 2: Example of miniature images that were printed on paper for each of the 79 applications and arranged on a table.

Further, the number of units was too low and too scattered across institutions to be considered a proper content analysis, and perhaps too big to be a web site analysis in a humanities tradition (e.g. Engholm and Klastrup 2004). The alignment with functions of the social contract also requires training, and no validation by inter-coder reliability was applied. As an exploration of a phenomenon, I consider these limitations relevant but bearable and the conclusion drawn sufficiently humble.

5.3 The work context

Paper II offers an interview study of Norwegian journalists who apply computing to their work.

The study of computational journalism as a social creative work practice requires different methodological means. C.W. Anderson suggests ethnographic studies, as he wants to expand from four frames in Shudson's classic typology of news to six angles "to give us some useful insight on the shaggy, emerging beast I have called computational journalism" (Anderson 2012, 18). This approach (observations, interviews) has proved useful in other studies of technological work in newsroom studies, including technological work that might fall under the term computational journalism, such as Royal's study, *The journalist as programmer* (Royal et al. 2012). For Anderson's frames to truly give useful insight on computational journalism (in relation to the six frames – political, economic, field, organizational, cultural, and technological), computational journalism must be a readily identifiable practice to observe and inquire about. This is an assumption we felt unsure whether or not to hold. We borrowed from the angles (economic, cultural, organizational) in Anderson's framework, but wanted to focus the study around computational journalism as a rhetorical craft in the Aristotelian/Heideggerian tradition. At least they are creating products: news applications.

We did attempt to formulate a questionnaire. We wanted to capture demographic variables (age, education, etc.) and preferential data (computer languages, tools, co-workers' fields, etc.). For this to be truly meaningful to us we needed something to compare this with – for instance, the general journalistic population or earlier studies of CAR journalists. We found this to potentially be a detour. No prior studies on computational journalism or CAR were – to our knowledge – carried out in Norwegian newsrooms, and we anticipated the results of a comparison with typical journalists being predictable. We also wanted the practitioners themselves to define what they do, so the questionnaire was rejected to keep the study as open as possible. Many of the topics from the questionnaire attempt were reformulated to be included

in the interview guide. The choice of interviews over more in-depth ethnographical case studies was mainly done to cover a broader spectrum of newsrooms than we could have afforded elsewise.

The choice to use interviews proved fruitful in providing us with rich data with descriptions of technologically advanced journalism production that indeed answered many of our questions. On the other hand, interviews alone did limit the analysis when it came to explaining using an actor-network perspective; an attempt at this was made in the first draft of the paper. Such analytical tools require more contextual and observational data than semi-structured interviews allow. In this regard, Anderson might be right about a more ethnographical approach.

The sampling was done using snowballing, with initial seeds being the by-lines from the Norwegian news applications from Paper I. This involved a bit of patience on the telephone as it turned out that the names accredited were not always the ones who actually did the technical work. We ended all the interviews by asking whom else the interviewee thought we should talk to. This turned out to be a good strategy. It took only a few interviews before the same names were repeated over and over. A full list of programming journalists in Norway does not exist, but our short list of interviewed journalists does define the top names from all the largest media institutions.

The post-data collection process followed a pipeline also used in the analysis of interview data in the last two papers: audio was transcribed into text, the text was read through as a whole, and then imported into *TAMS Analyzer* (Weinstein 2006) and annotated. The tags/categories that were used were made both based on the research questions and interview guide, but we also allowed for the creation of new tags. The new tags came from topics that emerged from the texts, e.g. elements that kept being discussed that we did not have a question directed towards. This process followed the step suggested by Creswell (2009) and others (e.g. Robson 2002).

5.4 Beyond the newsroom – design as a research method

Papers III and IV are studies where artifacts were designed and evaluated in the scope of computational journalism.

As a method of innovation in journalism, the product design approach has also made some marks in recent years. “Demos not memos” has followed as a slogan from within the community (Waite 2009). We typically only get to hear the success stories from research and development departments (R&D), and academic and business interests are not always aligned. Academic design efforts include alternate views on technology and journalism that we do not see in real-world newsrooms, such as alternative story structures for online news stories (Engebretsen 1999) or tools for finding sources through social media (N. Diakopoulos, De Choudhury, and Naaman 2012).

Quite a few interesting questions concerning technology and journalism cannot be answered by observing, interviewing, or analyzing products, simply because many particular applications of technology are not currently used in newsrooms – or even thought of as applicable in newsrooms. In such cases, designing and testing new ideas as technological artifacts lets us explore the potentials in constructed settings under a hypothetical light to figure out what such a combination can be. In that regard it is not a practice of the journalistic trade. Potentially, it could be so in the future, and we could learn why particular tools or methods do not align well with journalism. This could be valuable both in defining what journalism is as well as for the design of future tools and methods.

What R&D departments do is normally product design. They try to create new or better products and services. They are “designing interactive products to support people in their everyday and working lives”, to use a definition of interaction design from Sharp, Rogers, and Preece (2007). In R&D, design is, in contrast to design science research, operationalized as “professional design” with aims to create solutions as artifacts:

> *The key differentiator between professional design and design research is the clear identification of a contribution to the archival knowledge base of foundations and methodologies and the communication of the contribution to the stakeholder communities (Hevner and Chatterjee 2010, 15).*

It is as such a form of knowledge production more than the production of products as artifacts:

> *Design science research is a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence. The designed artifacts are both useful and fundamental in understanding that problem (ibid, 5).*

The focus of professional design is producing good products. In design science research, products are among the results. The understanding and knowledge this can create is dependent on the artifact that is designed. Other types of output are also produced, as summarized by Vaishnavi and Kuechler in Table 2:

Table 2: Recreation of table "Outputs of design science research" (Vaishnavi and Kuechler 2004)

	Output	Description
1	Constructs	The conceptual vocabulary of a domain
2	Models	A set of propositions or statements expressing relationships between constructs
3	Methods	A set of steps used to perform a task – how-to knowledge
4	Instantiations	The operationalization of constructs, models and methods.
5	Better theories	Artifact construction as analogous to experimental natural science, coupled with reflection and abstraction.

The process of producing these kinds of outputs has variations within the field, but a general overarching model over the process exists with guidelines for how it can be put into action (Hevner et al. 2004). The general model of design science research describes a circular, or iterative, process that breaks the process into smaller, more identifiable elements. Vaishnavi and Kuechler's model uses *awareness of problem, suggestion, development, evaluation,* and *conclusion* as the process steps.

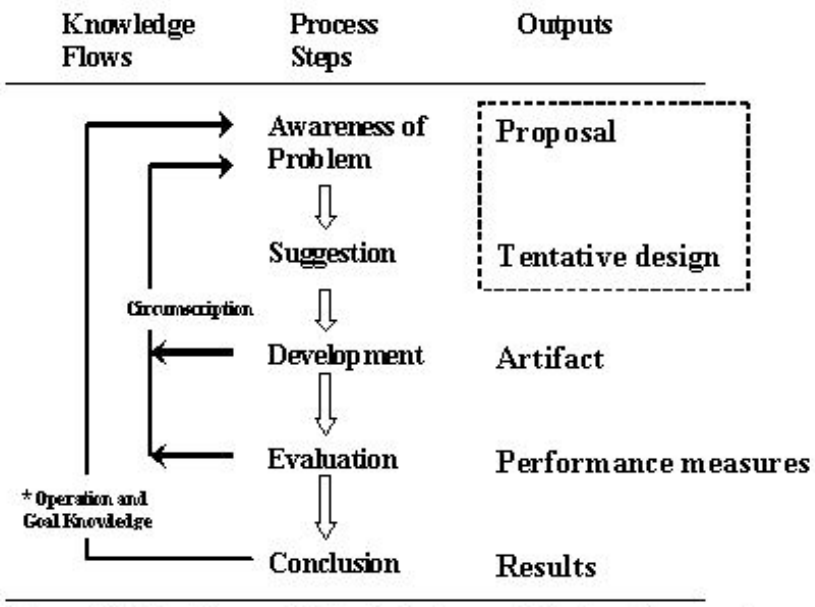


Figure 3: "The general methodology for all design science research" (Vaishnavi and Kuechler 2004)

Inside this model I find room to do the different steps in different ways. The differences in the role that design can play in design science is also noted by Mattelmäki and Matthews (2009). What development approach one chooses to use is one such area where I see different possibilities. On a spectrum of end-user involvement from ethnography to participatory design (Sharp, Rogers, and Preece 2007, 310), one could use any to decide on suggestions. One could develop low-fidelity prototypes on paper or implement full systems with programming teams using any given development strategy under development (scrum, extreme programming, waterfall model, etc.). Methods for evaluation also range the full available gamut of methods in academia. My choices in this regard fall under the "field" approach, in the lab/field/gallery distinction, with qualitative evaluation of use in an appropriately realistic context (Mattelmäki and Matthews 2009). The model is quite flexible. Indeed, in the Norwegian media research field, the prospects of using design to actively create new media texts are both discussed as a method for and carried out as media research; see Fagerjord (2012) for an overview.

5.4.1 How I used design science research

In order to account for how I have used design science in my research, I will align the processes I used to the research guidelines provided by Hevner et al. (2004):

1. Design as an artifact. Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
2. Problem relevance. The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
3. Design evaluation. The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
4. Research contributions. Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
5. Research rigor. Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
6. Design as a search process. The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
7. Communication of research. Design-science research must be presented effectively both to technology-oriented and management-oriented audiences.

This was used in both papers III and IV, but with some variations and deviations. The list contributes the *whats* for each step, but not the *hows*.

Paper III

1. The artifact consists of a model based on existing known and described algorithms that is tailored for Twitter messages in Norwegian and a graphical user interface that lets users interact with the algorithm's results.
2. The relevance for this project is well anchored, and was expressed by a team of journalists who analyzed Twitter messages manually after the 22/7 terror attack in Oslo.[17] The theoretic possibility of utilizing the wide variety of voices in social media to equip journalists with good overviews of public debates and sentiment is highly relevant in order to guide journalists towards relevant sources and arguments.

[17] The project in question is http://nrk.no/terrortwitter/ and the process behind this was described by Anders Hofsth at Nordiske Mediedager Spesial 2011 in Bergen 27.10.2011.

3. The evaluations of the artifact were done in two iterations. The first was done with two journalism students from the local university newspaper *Studvest*, and the second with professional journalists with a special interest or responsibility in social media. The graphical user interface (GUI) of the artifact was improved between the two iterations, as was the evaluation procedure. The procedure was a set of tasks (finding stories, trends, sources) under a think-aloud protocol, followed by a semi-structured interview.
4. The main contributions from this project are the experiences gained and verified by using the natural language processing and clustering techniques the model consists of, on localized data in a relevant context to journalism. Further, the theoretical assumption of the democratic aspects of social media needs to be balanced with the professional journalists' day-to-day management of sources, where the status and position of authors are key components in the assessment of the usability of a message.
5. The construction of the artifact is a recombination of known algorithms, and thus already well evaluated as separated parts. The evaluation of the design, on the other hand, is only evaluated by a small number of people. While experts in their field, their unfamiliarity with both the evaluation of information systems and treating a Twitter corpus as an object of analysis, limits the rigor of this research. The evidence provided is limited, and unfit for any quantitative analysis or attempts at explanation exceeding the sample of evaluators. They do, however, shed light on both the problem and suggestion addressed in this project.
6. The evaluation of this artifact was done in two iterations, both after the suggestion for problem solving was presented. The problems the evaluations identify are not implemented in a "final" or satisfying version that meets a real-world workday for journalists utilizing social media. As a search process the prototype required substantial efforts into identifying previous efforts into research on Twitter, on clustering, and natural language processing. The activity of consolidating findings and ideas from these works also functions as a negotiation and constant reevaluation of the artifact in relation to previous efforts. Development of software is in itself a search process, a process of finding the right solutions to make the software work. In my personal experience, trial and error is a key part of this (search) process.
7. The results have been reviewed and accepted for presentation and publication.

Paper IV

1. The artifact, the web application www.samstemmer.net, consists of a system for transferring data from the Norwegian parliament data API, a set of methods for treating this data, and various ways of displaying the results.
2. The problem relevance is collected from papers I and II, and consists of adapting news applications into continuous systems. This alteration to the news applications' format also asks how software can provide journalists with utility over time by automating methodological steps in analysis, in contrast to telling a story with static data.
3. The utility and quality of the design is partly evaluated. The goal for this project is to collect data on the suggestion level of the design science model: how can this data be treated to be interesting for journalists that cover national politics? The evaluation is therefore largely a data gathering process where "unstructured interviews are often used early on to elicit scenarios" (Sharp, Rogers, and Preece 2007, 211). The need for building a fully functional prototype might seem absent, as the goal is initial data gathering. If this was the sole purpose of the project it is, but I also wanted to explore how the API would behave in regards to stability over time as a data source, as this was an identified problem from the news applications paper. I also felt reluctant to bring over xml (what the API provides) or excel (what could have been a possible intermediary format) data to parliamentary reporters, as the profile for Norwegian parliamentary reporters (found in Allern (2001)) describes senior reporters as likely to be conservative in regards to how parliamentary reporting should be done. By providing a colorful, working draft containing different examples of how this data can be used, I hoped to gain better and more open-minded ideas, and corrections to my initial design. As such, this is not a full loop as in Figure 3, but only the inner loop (marked "circumscription" in the figure) that feeds back into the "awareness of problem" step of the process, and the "conclusion" provided is not a product, but better knowledge of the problem.
4. The contribution consists of mainly two things: A short list of unimplemented ideas the reporters suggested such a system should include, and a theoretical discussion concerning transparency when software is provided to journalists as opposed to created themselves (e.g. as by the interviewees in Paper II).
5. The rigor of the construction in this project is strengthened by the openness of the code the prototype consists of, which is open sourced and publicly available on

github.[18] The quality of the evaluation is sound in terms of qualitative semi-structured interviews, but not in terms of evaluating a product. The evaluation functioned as a space for discussion and brainstorming for how the parliament data can be transformed into a useful and usable tool for parliamentary reporters, while striving to maintain journalistic values.

6. This project lacks iterations as part of the search process. To identify "satisfying laws" of the problem domain, the criteria for when this type of suggestion can be said to be a success in parliamentary reporting, was the goal. In this sense the approach chosen was not necessarily the best or only approach, as a more user-centered design or contextual design approach from before the design suggestion was made would ensure the key stakeholders provide solutions within their own perspective.
   Underlying the prototype are some journalistic perspectives (found through the Retriever media database), visualization techniques (e.g. from the gallery in Bostock 2012), and statistical methods (e.g. from Poole 2005) that all required considerable search and research in finding and assessing as relevant and useful in order to arrange meaningful user-testing sessions where elements I wanted to discuss were exemplified. This guideline has not been followed as intended in design science research, with regards to an iterative process. The results must thusly be seen as the result of an initial workshop or data gathering for design requirements, and not a solution to the problem.
   Software development, as a search process, did involve iteration. Early versions were presented to fellow students and changed and expanded based on a hallway-testing-like approach.
7. Results for this work were presented to the academic community at the Future of Journalism 2013 conference, and are currently in review. The code from the project has been shared online and the prototype is publicly available at samstemmer.net.

In both of these design projects I deviate slightly from the guidelines, particularly by not thoroughly exploiting the core element of the design science process: iterations. Paper III contains revision between the pilot study with journalism students and

[18] Code can found at https://github.com/eiriks/samstemmer and freely used, adapted, shared, and studied by anyone.

professional journalists, but the results from the second round of interviews are not included in any further steps, other than a, perhaps, shortcut to project conclusions. Paper IV does not contain any design iteration at all. This was done, in both cases, as I felt I had qualitative data that was sufficient to create a research paper. These data are fundamentally connected to the artifacts, and the context they were presented in. In both cases the data describe aspects that matter in the design process for creating computational journalism. The aim in Paper III is to design a tool *for* journalists, as such fits the label design science as much as design science research, while Paper IV has a clearer aim of designing to gain knowledge of a domain rather than provide an artifact to solve a problem.

In the Scandinavian tradition of system design, I could have worked closer to the tradition of participatory design (Sharp, Rogers, and Preece 2007, 306). This would ensure proper anchoring in the user base, and the creation of solutions the target audience indeed requested. My goal was never to become a provider of tools for journalism, but I wanted to use design to test more theoretical concepts (democratic potential in social media and automation in watchdogging). A true participatory design project in this field would be an interesting endeavor, and presumably also an assumption underlying a fair share of the literature I described in Chapter 3.

5.5 Methodology appropriateness

The methodological choices I made in order to address the first part of my research question (how computational journalism is operationalized) were text analysis and interviews. Both of these approaches are well within the traditions of the social sciences, and their appropriateness is presumably quite clear. Both by analyzing journalistic output and interviewing practitioners I shed light on how this is practiced in Norwegian newsrooms. The appropriateness of design as a method for understanding perception, as an understanding of a phenomenon, is deserving of a detailed discussion.

The summary of outputs from design science research by Vaishnavi and Kuechler (Table 1) does not include outputs that seem likely to give insight into how a technology or its use is perceived. The framework, as summarized here, does not produce knowledge of how users of a technology feel or think about it. In the succeeding general methodology of design research (Figure 3), this kind of knowledge is included under "operations and goal knowledge" and "circumscription" that get fed into the model by evaluating and concluding. As one iteration of the model concludes, new knowledge on both the problem and suggested solution lays the foundation for the next iteration. Feedback from evaluations can both describe how the suggestion (prototype) solves the problem and how the problem in itself is understood and thus how it best should be addressed. While the design science methodology arguably can produce knowledge of the kind that can be used to answer question of perception, is it an appropriate choice?

Designing prototypes is resource-intensive work, and surveys, interviews, or experiments could more easily address the question of how computational methods are perceived in journalism. A survey could map a quantitative base where demographic variables could be used to explain how journalists express their perceptions of, relation to, and preference for different types of workers, methodologies, and technologies. Open-ended interviews can let such understandings unfold in richer, deeper analysis. A design approach still challenges the interviewees, or rather participants, as they partake in shaping an understanding, but in a different way. By providing tentative design solutions the participants are introduced to a process that is concerned with understanding in order to create solutions, a process the participant is invited into. An assumed or initial understanding of a problem can be adjusted as the participant sees how the designer has imagined a solution. Domain experts are likely to notice incomplete or skewed understandings of their own field, and as such are highly competent to adjust such problems. Tentative designs also bring theoretical understanding into practical experiences when problems can be addressed in visual and interactive prototypes. The kind of understanding a design approach requires demands both the designer and participants to think thoroughly through how a solution can be better or different, or how the problem best should be

understood to be adequately solved. It requires a different cognitive investment. The design approach does not necessarily provide better results than other methods; it creates a different form of knowledge, knowledge that is also formed by being experienced. Experience from evaluating a real-world, tentative design prototype is a form of "media experience" that does include a perceptional dimension (Gentikow 2005, 13–17). Design, as a method for understanding perception in this perspective, allows us to create experiences that later can be conveyed to a researcher and analyzed in more traditional social science research designs. This way, hypothetical fields and solution (e.g. imagined forms of media production and output) can be realized, explored and analyzed empirically.

6. Results

Paper I: *News applications – journalism meets programming*

In the analysis of news applications I find applications that deal with subject matter that fits nicely into a traditional content profile of media institutions. Politics, social issues, and economy were the biggest categories, and a typical scheme for a content analysis covers the subject matter. Further, the applications were found to map to the core of the media's social contract in regard to the information, watchdog, and arena functions. While the arena function did not manifest itself in the data, information and watchdogging were found to largely describe the aims for the applications, including the mix of these functions. As casual narrative information visualizations the news applications let the user re-explore and re-analyze the material as the journalists did, and offer a limited exploration of what the journalists found to be most important and/or informative. As data sources, open public data on social issues were found to be particularly important. The manifestation of news applications in traditional media institutions points to new skills and new personnel at play in the newsrooms – insight that informed the formulation of questions for the follow-up interview study of Norwegian programming journalists.

Paper II: *Computational journalism in Norwegian newsrooms*

Programming journalists in Norway are few, and certainly not always credited in by-lines of the stories they contribute to. In the newsroom they have adapted central values of journalism, such as a commitment to the social contract and a hunger to expose infringements committed by the powerful. They have clearly positioned themselves as journalists, not IT staff. Programming is underplayed as a tool like any in the journalists' toolbox and what is celebrated is the impact or significance of the stories they create, not the efficiency, elegance, or cleverness of the code it takes to make it. While the practitioners are aware of the technological possibilities they represent, focus is put on keeping output simple, and an aversion to bells and whistles is present in regard to both the technical and visual aspects of the work. This results

in more data journalism than computational journalism. While computational journalism has been proposed to free up time for journalism, the practitioners describe their abilities as important to keep up with the scale of digital data sets and tackle more day-to-day problems with software in order to fulfill the general requirements as journalists. Time and a boss that understands that this work is time-consuming are (still) the most important resources to a programming journalist in Norway.

Paper III: *The pursuit of newsworthiness on Twitter*

User-generated content, such as data from Twitter, represents a fantastic opportunity to tap into public opinion and public voices to keep journalism close to readers. The democratic element of allowing anyone a voice and a chance to contribute to a public debate could also be seen as a strength to journalistic integrity covering the areas in discussion. The design of a tool to evaluate data from Twitter included the adaptation of a set of known algorithms to cluster similar messages into larger groups in order to quickly get an overview of key entities. While boosting some linguistic elements of the messages before clustering helps, user-generated messages still contain so much noise that the signal can be hard to unambiguously detect. The kind of stories that expert evaluators found using the tool included mostly soft and human-interest news stories. The democratic promise of letting anyone express themselves was found to be intriguing but secondary when it came to finding solid stories that could be used in the media. Who the authors of the messages are is of such importance that I suggest future tools to position personas and networks as a first priority. The ability to find material should also include the function of hiding the known, the noisy, and the predictable.

Paper IV: *Watchdogging in code*

A goal mentioned by several programming journalists was the organization of news applications as continuous systems, or at least updated as new data are produced. One reason for the failure of this in Norwegian newsrooms is the lack of a mechanism for rewarding maintenance. By using APIs as data sources this problem is solved. The

samstemmer.net prototype explores this aspect of computing in journalism in relation to the Norwegian parliament's data API. Experienced parliamentary reporters evaluated the prototype in order to define preferred requirements for such a system and underlined an "expert mode" with as much and flexible data as possible as most relevant to them. Users from their audience for such a system were seen as "special interest", and the reporters' ability to get insight and test hypotheses was seen as the top priority. The equal access to data between the journalists and the audience was still seen as important. Large parts of the reporters' workflow are unfit for computational aid (old fashioned analog social networking), but some functions such as hypothesis testing and fact-checking were deemed promising for mixing with computational methods and meaningful in relation to running as a continuous system. The opacity software laying upon data in this system was compared to how reporters are frequently dependent on experts to provide answers, and was not seen as a major issue. The reporters saw the computing and numbers as dabbling in raw facts, while it is when these facts are filtered through them (and into stories) it becomes journalism. While software does enable exploration of new territory for parliamentary reporters, it also creates another frame (potentially) outside the journalists' control in regard to fully knowing how facts are produced – and thus could potentially weaken the journalists' accountability. Remedies for this are discussed in the following chapter.

7. Discussion

In this chapter I will discuss the findings in relation to the research questions. All data that are discussed were collected and analyzed in relation to the papers. This also includes some examples that are not quoted or mentioned in the papers.

7.1 Computational journalism output

News applications are the most visible form of joint journalism and programming projects from an audience perspective. These web applications are presented online, normally belonging to larger projects or "packages" of news items (articles, TV programs, etc.). In Norwegian newsrooms, teams consisting of people with different skills create this form of journalism. They still often depend on central programmer-journalists, people who can program but indicate their profession as being journalists or data journalists.

This form of journalism takes a lot of resources as in-house original reporting, often with investigative elements in the form of in-depth data analysis. In alignment with traditional news categories and the core journalistic functions of the social contract, news applications fall well inside the scope of online journalism, and represent a continuation of journalistic foci in their function and the subject matter covered. The visual display users see can be quite different from a typical text article in online newspapers, but while the technology places few restrictions on possible forms, the most common types of visualizations are maps, timelines, and charts – visualization types newspapers traditionally favor (Tufte 2001, 83).

When interviewing programmer-journalists (Paper II), they describe the work as a continuation of core journalistic practices, but also identify how this form of work can differ from traditional online journalism. One new aspect this form allows is personification, in the sense of making the story matter to the individual reader. As one interviewee put it, they disseminate "the unbroken line between the general and the particular". A typical example would be to present a story with a general impact

and a particular relevance to the reader (e.g. Norwegian school buildings are in a sorry state, and here you can inspect the report from your school). This is a result of how journalism as software deals with scale: “Proximity to you is important to obtain, and that is a luxury when working with computers and data-driven journalism. It’s merely a matter of fetching data for the whole country. Often you can get that, and then there is no reason to show moderation, as long as you present it well” (from an interview in Paper II). Not all data needs to be displayed, and what data are displayed can vary among the users. In a one-way communication medium the general story would have prominence, as time or space restricts the details in proximity to individual readers. As software, details in the general story can be served individually to different users, and proximity as a news criteria (Eide 1992, 66) can take precedence over the criteria for the general story. As such, news criteria can be juggled to better fit the readers’ position when presented on a reorganizable platform, and this is a journalistic reinvention.

Personification, or adaptability of news content to users, is one impact computing has had on online journalism. Other impacts include coverage of material that is too large to read through or analyze through manual labor, and newsroom-internal technological problem solving that enables newsrooms to connect their processes to external digital networks and data sources (e.g. ad-hoc encrypted communication for a whistle blower or the creation of a graphical user interface for other journalists to explore a database without knowing SQL). Newsroom-internal computational know-how enables newsrooms to maneuver well in the more technically advanced areas of digital media production and dissemination.

7.2 Creating a computational journalism culture

It is easy to point to the values of journalism. Sincerity, truthfulness, accuracy, and impartiality were values underlined by the American Society of Newspaper Editors in the 1920s (Schudson 2003). Balance, objectivity, fairness, freedom of speech, etc., have followed since. When describing the values of software (as a key part of computer science in newsrooms) it becomes less a matter of repeating acknowledged

values, but one could point to values such as efficiency, effectiveness, complexity, reusability, portability, readability, cost or time reduction, or elegant problem solving. A quick algorithm does not need to be objective in the perception of an audience, and a balanced story is not always an efficient way of getting a message through. The suggested values for each field can indeed contradict each other. Further, both journalism and computer science have different cultures that organize their values. These cultural characteristics can present major obstacles in merging into one practice or creating good collaborations, as noted by Cohen et al.:

> *Finally, it [computational journalism] faces cultural challenges, as computer scientists trained in the ways of information meet journalists immersed in the production of news. If it is able to overcome these hurdles, the field may sustain both public interest reporting and government accountability (Sarah Cohen, Hamilton, and Turner 2011, 66-67).*

The established ways, the status quo of Norwegian journalism, is a culture of partly tacit information of what journalism should be and how it should be made. This cultural tradition can certainly be an inhibitor for computational work, as computational solutions fall short of being natural or ordinary ways of solving problems. At the same time, this cultural ballast cannot be outright abandoned. Its slowness and protectiveness of the old ways of doing things contains brakes and checkpoints that also include an element of quality assurance and skepticism of miracle cures both inside and outside the newsroom walls. These cultural and organizational factors, such as an editorial chain of command that pinpoints responsibility and borders for what goes in terms of precision, fairness, etc., are concepts that non-news professional actors do not naturally provide. Outside efforts are in one sense free from these factors, but also lack them. Inside efforts, on the other hand, constitute an environment where technology is given very little significance.

The interviews for Paper IV provided an understanding of journalism as a filter that facts and data go through to become stories. Technology merely manages the facts, and it was seen as external to journalism. This was in line with the kind of impression

we anticipated when reviewing literature for Paper II – a world where a programmer would be a disturber of the peace, a world where technology would not be praised or embraced:

> *Research among reporters in various converging newsrooms in the US by Singer (2004) and Boczkowski (2004) shows similar experiences, citing turf wars and a general reluctance of journalists to innovate, share knowledge, embrace the new technology – even though those that do reportedly think they are better for it (Deuze 2005, 452).*

What we found was something completely different. They work in teams and "exploit each other's strengths" (from an interview for Paper II), programmers, designers, journalists, etc. We found that journalists that could program, and knew their way around solving problems through computers, still strongly underlined their position as journalists, not technologists. The technical expertise was described as nothing more than a means to meet journalistic ends. This is fortunate for the newsrooms these journalists belong to, but also worrying if they want to expand their work in this direction. The blend of technical skill and journalistic values is a very rare mix. A more traditional technologist would need to muffle his (or her) enthusiasm for technology and align to the cultural climate in the newsroom. "The newspaper industry has pushed away lots of skilled people", one of the programming journalists in Paper II noted. Journalism does not assimilate technology quickly and newsrooms can easily reject computer enthusiasts[19] and inhibit computational journalism. In order to strengthen software-oriented journalism production, technological work as a journalistic endeavor needs to be given some space where enthusiasm for technology can thrive.

The blurring of what a journalist does and thus what counts as "journalistic" is a process that develops over time. In *We are journalists,* Weber and Rall identify the inclusion of design work as journalistic work as one success factor for the New York Times' newsroom in regards to creating interactive information graphics:

[19] I use the term "computer enthusiast" synonymously with "computer geek" or "hacker", as it presumably comes through without the potential negative connotations of "geek" or "hacker". I could use the term "technologist", but "enthusiast", "geek", or "hacker" all contains aspects of pleasure and enjoyment, a love for the craft that "technologist" does not convey.

> *In this statement ['we are journalists' uttered by a NYT graphics editor], we recognize a paradigm shift that has occurred in the New York Times newsroom. What is new is that even programmers and designers belong to the journalistic team of the newsroom and define their task as a journalistic one. (2013, 163-164)*

Astrid Gynnild, in her theory on creative cycling, deals with this by favoring the term "news professionals" (2007). The theory also includes that news professionals manage multiple skills both individually and collectively: "At any given time, an unlimited number of skill level combinations are found among news professionals" (ibid, 88). In Weber and Rall's study and Gynnild's theory, convergent media production includes people with converging skillsets, too. As new skills are acquired and become normal, significant, and recognized by the higher ups in the newsroom's social hierarchy, they become "journalistic".

In a map of the Norwegian journalistic field, as presented by Hovden, this strong urge to underline work done by designers, programmers, etc., as journalistic represents an active choice of moving "left" in the field (see Hovden 2012, 69). From a (given or claimed) position as "agnostics" (more technically oriented, detached from investigative journalism and lower in impact and influence) they claim a role as "investigators" or even "educators" by underlining that they are journalists, that the social contract underlines their work, and that they aim for journalistic awards. The compass needle for computational journalists in Norway points in the same direction as the fields leading actors with weight to define values in the field. Both Papers I and II include evidence for this claim. The larger news categories in Paper I include politics and economy – categories that correlate with high journalistic capital. The extent that applications focus on powerful people and organizations, and as such orient toward the watchdogging function of journalism, also relates to high journalistic capital. The programming journalists in Paper II all worked in large newsrooms, primarily in national media institutions. They position themselves as journalists first, with the aims of uncovering and explaining, and stress the importance of informing the audience by keeping things simple. The best story in this universe is one that has impact and results in awards:

> *The dream-story is off course the one where I find something someone have tried to conceal. That goes without saying. That is the dream. Preferably the prime minister or someone like that. Something big. […] That is what any journalist dream of, and want to receive the SKUP-award[20] for (Programmer-journalist from Paper II).*

High-impact journalism and journalistic awards also suggest higher journalistic capital. The journalistic field as a whole has to a certain degree started to notice software-oriented news production as valuable. Awards are given to these kinds of endeavors (Bjørgan 2013), new jobs in this area are called for (e.g. www.aftenposten.no/digitalehoder), and work done in Norway in this field is lifted up as good journalism in international forums (Heftøy 2013), thus giving it credibility and status.

The expert parliamentary reporters from Paper IV did not see the information system as journalistic in serving data and computed results to them. It was external and somehow impartial. That the information system also functions as a frame that portrays the world according to latent methodical and visual preferences, and that these preferences should be in accordance with journalistic practice, needs to become newsroom-internal. This includes efforts to make them as transparent as possible, in order to be fair/balanced, by exposing the potential biases, assumptions, or considerations implemented in code. The gap that needs to be acknowledged in order to stay accountable in digital news is above all an understanding of technology as a companion (and antagonist) of agency in news production.

7.3 Journalistic values in software

In discussing journalistic values in computational journalism, I will concentrate on transparency. In computation and automation, transparency is only important when something goes wrong; in journalism, it is also important when things go right.

[20] The SKUP award is a Norwegian press award for excellence in investigative journalism, see http://www.skup.no/SKUP-prisen

Upholding transparency in computational journalism

Transparency as a goal is fundamentally a goal of exposing truth by revealing elements between facts and observers. While this can be seen as a positivistic way of thinking as an extension of "facts equal truth", it can also be seen as a critical way of thinking as removing intervening readings of facts in order to get a chance to create an independent reading within one's own perspective. What you then see is less of others' intermediary interpretations and more of the authorship of the creator of the fact itself. The transformation from facts to truth is in this perspective much like the transformation from data (collected symbols) to information (meaningful interpretations of data) – they both rely on a knowledgeable actor to do the transformation. That the end result is information that matches the originally recorded phenomenon that was described in the data is by no means a given.

Aligning computing with journalism is not merely a matter of picking up tools. As underlined by Diakopoulos, journalistic values need to be upheld (2010). Examples of the values mentioned are balance, accuracy, and objectivity. These concepts are hard (and disputed) but exist as guidelines for how journalism should be executed. Objectivity, for instance, has been suggested to be operationalized by balance (e.g. Lichtenberg 1991). If balance is to be considered a fair means, we need to know what we are presented with in relation to what is excluded or given lesser weight – in other words, transparency.

Transparency is an increasingly important goal for media institutions (Karlsson 2010), and while journalism has always depended on expert sources with greater knowledge in various fields, journalists are still expected to be able to ask questions in order to verify the reliability of the matter in question. In computational journalism, questions can be pointed in the direction of software: what is done to data and how, in order to produce the displayed results? In this regard it is clear that software development is not a neutral or objective craftsmanship, but yet another frame to behold the world through and it is not always easy to direct questions to computer programs. Software is also quite opaque by nature, and often does operate as a black box, as described by Latour: "[w]hen a machine runs efficiently, when a

matter of fact is settled, one need focus only on its inputs and outputs and not on its internal complexity. Thus, paradoxically, the more science and technology succeed, the more opaque and obscure they become" (Latour 1999, 304). This is a different problem in journalism than in many other professions, as journalists are expected to be both accountable to their audience and keep others accountable by demonstrating discrepancies between facts and expected or desired states of the world. This is how computational journalism can distort, obscure, or conceal the journalistic workflow and this is why automated journalism cannot operate the same way as other automated processes. The question of upholding journalistic values in computational journalism becomes above all a question of allowing and creating transparency.

Ideology can be written in software code, and this can reflect in the arguments or evidence a system produces. An example of this was given in the work for Paper II by a hacktivist software developer[21] who had made an application that displays results from school evaluations on a map:

> *... you are to map this linear data as a vector through a color space, and it turns out that almost all values cluster about here [point to the middle of a color space model drawn on a whiteboard], and if this is from red to green that will be very dull to look at because no points would be red or green, and they all would be a bit orange. What you need to do is to take these values and map them through an s-curve to spread them, and this is where the ideology lies, where the ideology becomes visible. If you have and agenda, a reason to do this, [even] as truthful as possible, the way you apply the curve to the dataset has a lot of impact in how the interpretation of the information will fall out. If you tune this hard in one direction, all schools in Holmlia[22] become totally red.*

The fact that the point of departure matters in the creation of software in journalism was also seen inside the newsrooms. In the words of a programming journalist from Paper II:

[21] As this interviewee did not have a newsroom affiliation, he was merely covered by the media for his efforts and insight, this interview was not included in the final sample of this study. The data were collected, transcribed, and evaluated before they were discarded to keep the data in the study within the same institutional category.

[22] Homlia is a densely populated, culturally diverse and socio-economically weaker part of Oslo, often used as an example of unequally in Norway. The schools in this area preformed below the national average in the national tests in question.

> *The journalists hands on the keyboard pushing keys into code is a defining factor for computational journalism, as the organizational rules and knowledge woven into the heart and soul of a journalist makes a journalist do things the journalist way. And that matters.*

The "journalist way" in this case points to obedience to the above-mentioned values, the social contract, and a need to stay accountable to an audience to maintain trust. In Paper IV the reporters held solid faith in numbers and data, but saw the technology as impartial. This can create a problem if journalists merely become heavier users of technology. Hamilton and Turner's report expresses a worrying "likely effect of computational journalism" in relation to tools:

> *The tools developed for reporters will likely need to be open-source or carry a very low cost of acquisition, since local papers and online news providers will be hard-pressed to make investments in accountability coverage. The tools will need to be easy to operate too, since journalists may not be given the time or training to use complex algorithms (Hamilton and Turner 2009, 12).*

Tools as described here are black boxes. Data go in, "facts" come out. Reporters without training will need simple systems, because they have no clue how the algorithm that produces the "fact" works. This is not "upholding values of journalism" as a journalist is expected to know how their facts are produced and on what basis they are drawing their conclusion. If we follow Diakopoulos's clause of upholding journalistic values and Flew's statement that computational journalism is a meaning-making enterprise, transparency risks being weakened by computation.

Proposal for transparency issue solutions

Using technology extends our capability, and if we need to be accountable for what is gained through technology, some measures can be taken. Some can be taken internally to raise the journalists/newsrooms' knowledge and thus give a chance to explain how facts are produced. Others can be applied to allow external forces to inspect, verify/falsify, and scrutinize computational methods.

1) Internal
    a. If journalists themselves write software they will be able to account for their software's results. Norwegian journalists typically choose education that does

not include engineering or computing (Hovden 2008), so increased knowledge of computing should be added. This can be done through courses given to journalists, but a likely better option is to include more technical training in journalism schools given the threshold this kind of work often represents. This knowledge can also be hired from outside journalism's traditional field of recruiting.

b. Apply algorithmists. As suggested by Mayer-Schönberger and Cukier, algorithmists are "experts in the areas of computer science, mathematics, and statistics; they would act as reviewers of big-data analysis and predictions" (2013, 180). They could also review computational journalism. This occupation is inspired from the media's concept of an ombudsman, and it reaches full circle if it is applied to computational journalism.

2) External

a. Publish code as open-source software. As the media's audience is often geographically bound, the institutions often do not directly compete and sharing code does not have to be sensitive in relation to competition. It can also result in getting improved code back to the newsroom from the open-source community (Groskopf 2011).

b. Publish open documentation on how the software works. A linked methodology page or automated documentation from well-formed docstrings, javadocs, or similar can allow recreation of key methodological steps.

c. Publish the raw data alongside the results.

All these suggestions follow the same pattern: allow external peer-review and ensure internal comprehension. There is also a lack of methods for evaluating journalism in code. In interface design, heuristic evaluations are used to identify usability problems in user interfaces. The various heuristics in use are based on principles of good design (e.g. Nielsen and Molich 1990). A set of journalistic design heuristics for both controllers and views could be merged from the sociology on news and HCI literature. This would presumably also make outsourcing of journalistic programming less problematic and less prone to misunderstandings.

Software as a beat

As watchdogs, the issue of creating transparency in software stays the same, but the software to see through is external. As journalists are also expected to perform journalistic investigations, this same problem of transparency gets flipped on its head: how can journalists investigate software that affects us, such as the systems used by tax authorities or Google?

This is an increasingly significant problem, and one that currently has no formal solutions in journalism. From the big-data context, the idea of "algorithmists" has been proposed in two flavors (internal and external), as the role that does this job. These algorithmists should, in cases of disputes, get access to "algorithms, statistical approaches and datasets that produces a given decision" (Mayer-Schönberger and Cukier 2013, 180), and that covers the same ingredients I have suggested to ensure transparency for external parts of computational journalism. One can see this as an investigative reporter covering software as his beat, or "algorithmic accountability reporting" in the words of Diakopoulos (2013). Exactly what we call them matters less – what matters is what they do. And what they in essence do is to reverse-engineer software to explain how it works. Reverse engineering is "the process of developing a set of specifications for a complex hardware system by an orderly examination of specimens of that system" (Chikofsky and Cross 1990, 13), a term also applied to software and other products. It is analyzing in order to move up the abstraction level from product to design model or specification, a matter of figuring out how a system works. Through software studies this has also become an important method in social sciences/humanities, under variations of black- and white-box testing (for a description see Bucher 2012).

My focus in this thesis is on computational journalism as something some journalists create and as a function third-party others can fulfill in aligning software development to society's need for information under the standards expected of journalists. The focus is internal to systems we have access to, as the computational journalism perspective is from the creators. Watchdogging software external to the newsroom is a field in need of more research, as part of computational journalism,

software studies, and engineering. As software becomes more and more ubiquitous and integrated in formal parts of society, the need to hold software accountable becomes inevitable.

7.3.2 Is automated watchdogging an oxymoron?

Implementing journalistic values in software is hard, but possible. Does this mean that autonomous machines will undertake journalism in the future, similar to how Narrative Science creates newspaper articles through computational means, without having journalists to "rattle off prose"?

Watchdog journalism can be summarized with three assumptions, that the media is: (1) autonomous, (2) acts in the public's interests, and (3) is able to influence dominant social groups to the benefit of the public (Franklin et al. 2005, 274). In Paper IV, the design part of the project consists of a prototype tool that aims to monitor the parliament's API. One way this was imagined by a journalist was as a tool that notifies a journalist if something unpredictable happens: the parliament votes down the government's plans, a vote result splits the parliament by gender, or some other predefined indicator of interestingness.

The system would thusly do its part (1) autonomously and (2) in the public's interest to monitor the parliament. If only the journalist is alerted, no dominant social groups would be influenced. Even if the journalist was to produce a massively influential news item, the system alone cannot be said to wield this power. But, the first two assumptions are also questionable. The system is not autonomous just because the data are untouched by human hands; the system is built upon and depends on the API. If the parliament found a reason to turn it off, or manipulate the truthfulness of the data it spews out, the information system would break or communicate misinformation. The permission of the data holder is a prerequisite for such a system. Further, the system should be in the public's interest, but what it is imagined as is in the journalists' interest. Alone the system is meaningless as it depends on intermediaries on both ends, and is still detached from an audience. What we have created here is not a watchdogging system, but a system that can make Norwegian

journalists better at monitoring the parliament, even in their sleep; a journalistic alarm system[23]. Automated watchdogging is, similar to other forms of efforts to computerize analytical work, a reminder that technology needs humans as much as humans need technology in computer-supported work.

7.3.3 Facilitating accountability journalism

To view journalism as a civic alarm system, that constantly poses a threat of exposing corruption or abuse of power, follows the idea of accountability journalism (Eide 2012, 391). Accountability journalism describes how media organizations and its journalists are accountable to the wider society in various ways (Franklin et al. 2005, 4–6), but also how an enlightened citizenry should be able to hold journalism accountable (Eide 2010; Franklyn et al. 2005). In computational journalism, as a result of the natural opacity of software and technologies, creating transparency is the central point is this regard. To enable citizens to hold computational journalism accountable is a matter of exposing how journalism is produced. Journalisms' agency in holding powerful actors in society accountable, as a part of what citizens should expect of journalism, in incorporated into Kovach and Rostenstiels' principles. In defining computational journalism as an overlap between computing and the purpose and goals of journalism, it is necessary to give an account on how these goal and purposes hold in practice. My studies can shed some light on this; I will quickly show how these principles stand in relation to my results.

Journalism's first obligation is to the truth. As discusses in Paper IV, computational journalism can hide how facts are produced, but as proposed in this discussion transparency can remedy this in various ways.
Its' first loyalty is to citizens. This is no different in computational journalism than other forms of journalism, interviews for Paper II confirms this traditional journalistic view also among programming journalists.

[23] A similar system, ChangeTracker, has been made and used to monitor the White House's webpage. See http://www.propublica.org/article/changetracker-howto for an introduction.

Its essence is a discipline of verification. As described by Flew et al. (2011) computational journalism is intended to be a meaning-making enterprise. In Paper I this shows in relation to presenting data as proof, and offering analysis based in this data as visual representations. The exposure of data opens up for inspection and validation of conclusions, if any are drawn. As discussed in Paper IV, the methods hidden in code can conceal how computed facts are established (cf. first principle). *Its practitioners must maintain an independence from those they cover/It must serve as an independent monitor of power*. Exactly who the journalist should be independent from, and in what ways, is not explicitly listed, but actors such as political parties, businesses and corporations are discussed later in the book. Journalism should not be carried out as a favor and independence should be understood as "nothing personal to gain", as an effort to operate as neutral as possible. The Freedom of Information Act (FOIA as an acronym, "Offentlighetlova" in Norway) requires public data to be exposed to the public, and is intended to make fair transitions of data independent of whom the requesting and requested parties are. Still, the exchange of data can be problematic, and can be (mis)understood as a favor, or be done with intentions of personal gain. Cohen (2011) point to other problems, such as a streetlamp effect, where some types of data gets a lot of exposure (e.g. crime maps), while other data gets no exposure. In Paper II programing journalists reported that they consider the access to data to be good, but that it is a larger problem that they have no way of knowing what data exists in governmental databases. In this regard they depend on finding particularly helpful clerks, or are left with filling out FOIA forms in cunning ways involving a lot of guesswork. Another problem noted by Cohen is that easy data can outweigh accurate data, for example by providing solid APIs for some datasets, and not for others. In some examples, she explains, these APIs are new independent systems that are not directly connected to the old systems, and data is manually moved from the internal to the public systems. As discussed, a system built to watchdog an API cannot be independent, and if such APIs also represent a "selected view" of the system it is supposed to expose, watchdogging becomes meaningless and potentially a highly efficient source of misinformation. All datasets that are interesting to journalists, if not complied by the

journalists themselves, are questionable in regard to who can gain or lose from its' exposure. All software-oriented forms of news production (cf. Table 1) need to address this in some way. Computational journalism is in this regard no different. Where it can be different, due to possibilities for including computational models, is that is can apply measures to detect interference (e.g. applying a models to detect if data has been tampered with, similar to how this is applied to image manipulation (Krawetz 2007), cheating in chess (Mcclain 2012) or plagiarism in academia (e.g. Gipp, Meuschke, and Beel 2011)). This can possibly strengthen journalisms' independence from data holders by providing new ways to scrutinize data, and through verification make it harder to use the media as tool of amplification of misinformation. Computational journalisms' independence from actors covered, and powerful actors in general, is similarly problematic to how journalism at large both exists in, and as part of society.

It must provide a forum for public criticism and compromise. As found in Paper I, the arena functions is not the aim for most news application. I deliberately omitted comment forms (many have these, but as part of the online site, not the applications) and web forums, in the selection. These elements are results of programming and journalistic goals, but as a part of the larger newsroom, not the news applications. Computational journalism is one of many ways newsrooms produce news, and should in regard to public scrutiny be considered as a part of a whole.

It must strive to make the significant interesting and relevant. News applications incorporate this, and choose to point to certain aspects of analyzed data. Story telling forms in data, such as the martini glass structure, interactive slideshows and drill-down stories (cf. Segel & Heer, 2010) are applied to achieve this.

It must keep the news comprehensive and proportional. This points to the journalistic layer between facts and data and the audience, which consists of explanations and interpretations. This is similar to the point on making the significant interesting and relevant. These explanations and interpretations, or frames, can be incorporated into software, as explained by the hactivist in Paper II: this is where the schools in Holmlia become "totally red" if the journalist so chooses. Even when such outcomes are taken into account, this can change with new data in continuous systems. If

dynamic data is used to back a statement (e.g. the schools in Holmlia are lagging behind) in the application or elsewhere, and this later changes (the new values for the schools in Holmlia are assigned less aggressively red colors) the statement is no longer proportional or comprehensive. If the dynamic data is embedded in a static online news story, inconsistencies can occur. This is also why the ability to "take a snapshot" of the application at a certain point in time was requested as a requirement for the watchdogging application in paper IV.
Its practitioners must be allowed to exercise their personal conscience. And this must be allowed regardless what kind of journalist, computational journalist included.

By going through the normative principles that I include in my definition, it is clear that computational journalism is compatible with these ideas. It also repeats findings found in the subprojects: software code can conceal truth (or at least the process that lead to it), graphical user interfaces frames the presentation of news (similar to how terminology, field/frame size and volume can frame a story in articles, video and audio) and the fact that these principles are vague (what is relevant or proportional, how independent can journalism in reality be, etc.) makes them something to strive for. Media Accountability Systems, efforts to regulate media as "a 'third force' of media regulation between the law and the market" (Brurås 2009, 120), are often based on transparency in creating "dialogue between journalism and society" (Eide 2012, 392) and such systems should provide ways to ensure what is concealed in software also can be cutinized by society.

7.4 Computational journalism as a process

As initially mentioned, journalism is often described as a process. This understanding makes journalism very pragmatic, and breaks it down to more manageable tasks that consecutively operationalize journalism. If computational journalism is to re-invent journalistic efforts, a technological answer could be to re-engineer journalism as a business process. This approach would analyze journalism as process, modeling it as a workflow from beginning to end and reorganize the involved steps in order to optimize the workflow. A typical key element is to change old, or include new,

software that supports this new workflow. The business at large (or section or portion) is modeled (e.g. though graphical representations in variation of workflow diagram) and reorganized to better fulfill some favorable outcome (e.g. spending less resources or producing better output). Both the need for new skills among employees, and new technologies are addressed in literature on how to orchestrate such change as business process reengineering (BPR) (Al-Mashari and Zairi 1999).

While the factors for success and failure for the changing of businesses as described (ibid) seem reasonable in some businesses, they do not seem appropriate in Norwegian newsrooms. Elements from Al-Mashari and Zairi are found in relation to computational journalism, such as the introduction of new job titles, the underlined importance of support and understanding from bosses, and the inadequacy of old reward systems. But the BPR concept at large aims to change businesses from the top down, with clear and quantifiably goals. This is not how Norwegian journalism is run. Norwegian journalists are to a large degree self-driven – to some degree in what content they choose to cover, but also how they choose to deal with the process between idea and end product. The process will vary among journalists and is, as other practices in knowledge production, hard for formalize as a complete business process. This independence is an important part of what journalism is; the flexibility or unpredictability of journalism underlines journalists as watchdogs that work on their own terms. The BPR perspective also assumes clear goals for businesses at large. While journalists are aware of their workplaces' need to make money, this it mainly the management's concern; journalists are concerned with their next story or next project. The goal of the next project, in relation to formalizing the process, is not unlikely to require a different set of assets (information, access to people, methods, etc.) in a different reconfiguration that the last. Journalism as a process is a good way of describing what journalists do (cf. Paper II), but it is self-contradictory and unfavorable when formalized as a whole in an information system. Journalism is flexible by default. Investigative journalism must stay flexible to be able to hold other accountable.

Some types of journalistic work can be changed in line this this reasoning, and

formalized with supporting information systems. Commentary of sports or legal trials and reviews of art and products are examples that, to a certain degree, already have this. Still, it likely has not made the job of assessing the quality of a piece of art or efforts on a sporting pitch any quicker or easier for the journalist.
An example observed at trade-shows is the implementation of "write to space" methods (systems that limits a journalists text to a word count through the whole process, to avoid too much time being spent on text that in the end will be cut by an editor). Knowing how much space a story will get is presumably helpful while managing time, but this also predetermines the significance of a yet-to-be researched story, and potentially restrains rigorous journalistic inspection in favor of more metered effort in accordance to the word count given.

What I do find, in relation to a more large-scale BPR perspective, is that computational journalism is not a business solution in order to change a larger organization. It is neither a particular technology nor infrastructure, but a way some journalists choose to change how they work, or how journalistic work is executed. It is a bottom-up change that emerges when the right circumstances allow it to. When information systems are to supply human labor in journalistic endeavors, they need to aid in smaller autonomous task and provide flexibility in reconfiguration to "yet to be invented" problems and scenarios.

7.5 Computational journalism in Norwegian newsrooms

As a software-oriented form of news production, computational journalism is operationalized lightly in Norwegian newsrooms. The skills that let journalists bridge the gap are scarce in the newsrooms, and scattered in the largest media institutions. A thorough understanding of technology is needed to switch from seeing it as tools to be used, to tools that can be created. When these skills are in place, problem solving in the newsrooms gains the benefits of computational thinking about news production, and allows for a platform-oriented journalism instead of a story-oriented journalism. Journalists who bridge this gap also see the relevance of the technical work as journalistic. News applications are one output of this journalism; while still

story-centric, this format gives the audience some control over some news criteria through interactive features and introduces a platform thinking that allows one application to tell multiple stories. Other outputs of computational journalism include business-internal systems for research and analysis, visualizations, and arrangement of computer-supported work for others, such as creating interfaces for databases of general value or providing encrypted communication for other journalists with this need.

Powers' previous research identified three main ways of viewing technological work in the newsroom: as continuity, as a threat to be subordinated, and as journalistic reinvention (2012). My studies support these angles of observation, but perhaps not as one might think.

Programming journalists underline computational journalism as a continuation of journalistic work in the digital realm. To non-programming journalists (e.g. participants in Papers III and IV) this is quite alien as programming and advanced use of software are simply not familiar problem-solving approaches. To them though, this represents possibilities for journalistic reinvention, an aspect the programming journalists are aware of, but underline to a lesser degree. Though computational journalism appears as alien to many journalists, they do not describe it as a threat, but it is avoided and segregated as technical – not journalistic – work.

In summary, computational journalism is emerging in Norwegian newsrooms. The required skills are sought after internally, and utilized in many of the steps in the journalistic process. They function both as digital handymen that can fix problems, but also, and preferably, as investigative reporters who use advanced computer software and programming as journalistic tools. The gap between those who can and cannot do this kind of work is negotiated in the newsrooms, where those with this kind of skill pull their work in the direction of the dominant values in journalism. The more broadly held unfamiliarity with this way of approaching journalism has put the computational journalists in a squeeze, but as this approach is given more positive feedback (awards, new job advertisements, positive press-internal coverage), it allows

computational journalism to appear as a natural direction for online and digital journalism.

Computational journalism follows its predecessors and competitors in terms of software-oriented news production, but exceeds the boundaries provided for these in some areas. The amount of overlap between the various forms is substantial. Newer news applications and other computed outputs of journalism, as well as discussions on mailing lists, blogs, and trade magazines (cf. NICAR-L, datadrivenjournalism.net or journalisten.no), do today contain tutorials, descriptions, and exemplifications of journalistic forms that fit the definition given in Chapter 4. The field is unsettled as it is still new and novel, but it is vibrant and existing beyond hypothetical academic articles.

7.6 Reservations and limitations

Whether or not generalizations can be made about qualitative research is a matter of discussion on what reliable, valid, and credible qualitative research is (cf. Golafshani 2003; Silverman 2001, 219–254). To avoid utilizing terminology inherited from positivistic branches of science, words such as “trustworthiness” and “rigor” are sometime used. The point is mainly the same: can this research be trusted to provide knowledge beyond anecdotal evidence for specific events. Qualitative research in general strives for understanding more than general truths, but the answer is often still the same, credible knowledge is possible through qualitative methods. Both qualitative and quantitative research depend on the same criteria for credibility, and demand proper use of the scholarly workflow (see Table 8.1 in Silverman 2001, 222).

This collection of articles gathers four quite different papers. They use different methodologies; they are written for different audiences and they are in different stages of the publication process. Paper I is written for an anthology in Norwegian, and written in a language intended to be as accessible as possible and it is loosely theoretically tethered. It is translated for this thesis. Paper II is written for the audience of the journal *Journalism Practice*, while Paper III is written for an

informatics conference, *Norsk informatikkonferanse*. This makes the collection of articles uneven and varied. This variability can be seen as uneven quality, but I prefer to view it as a result of exploratory work, where approaching a topic from different angles is a strength in creating initial accounts of objects with unclear boundaries.

As acknowledged in Paper I, computational journalism can have a direct line to an audience through products such as journalistic web applications and news applications. This points to the absent user perspectives in this thesis. How audiences experience news applications as a format, and what makes for good user experiences beyond good web design in this format, stand untouched. User perspectives could, and should, also fit into future heuristics for creating news applications. Succeeding user perspectives, wider organizational perspectives are lacking to better pinpoint computational journalism's position as an intended or allowed practice. Norwegian editors claim to see a great potential in data journalism (Øvrebø 2011), but we know little of how organization or editorial decisions are made for use or non-use of more software-oriented news production. The approaches I have used in this study only cover some aspects of the larger picture of journalism and news production.

How journalists in Norwegian newsrooms perceive computational methods is covered narrowly in my studies. I have interviewed journalists that work with social media (Paper III) and parliamentary reporters (Paper IV). These participants were chosen as experts in the domain the respective subproject dealt with, and the number of interviews per subproject was small. They are not a proxy for all journalists in Norwegian newsrooms. They represent some voices and some perceptions of computational methods. A wider selection (e.g. though a larger survey with a representative sample of the journalistic population) would give a better representation of a more general (or diverse) interpretation.

Computational journalism needs good theories that can help explain how both journalists and technology matter. My understanding of technology in society is influenced by newer socio-technological theories (e.g. Latour 1992; Orlikowski 2000; Kaptelinin and Nardi 2006), and these perspectives are used in my framing of

computational journalism. This tradition of scholarship often underlines contextual variables and observation as important methodological tools. My studies do not contain such in-depth contextual information of technological use, and this does the theories a disservice by not exploiting some key elements in allowing technological use and interaction with artifacts to become the center of attention. I still argue that journalistic theory, that to a large extent overlooks technological artifacts in its explanations, can gain from borrowing ideas from this tradition when dealing with technological aspects of journalism. Orlikowskis' adaptation of structuration theory also shows how this theory can function as a common denominator for technology studies and journalism studies (cf. Eide 1992; Eide 2012 and Orlikowski 2000). Future research can benefit form applying such a theoretical view, in order to contribute to an understanding that includes both social and technical aspects. As programming directly involves changing structural objects (software), studying programming might help identify how both actors and structures define journalism, as newsroom-internal software functions simultaneously as authoritative (or symbolic) and allocate (material) resources. The study of software design and the use of this software should shed some light on the complexity in which news is created. As an exploratory research project the search for good theories that capture computational journalism has proven challenging. Much, if not most, literature on journalism does not include a reasonable account of how technology functions in the production of news. My studies do not provide a theoretical framework to fill this gap, and more empirical research is needed to provide a descent account of technologies' position and impact on news production.

Is the model in Chapter 4 – my alignment of computational journalism – valid? Is it rigid? It is definitely not final, but I feel comfortable using it as an explanation for how computational journalism is different from earlier software-oriented forms of news production. It is based on my research, but also the existing literature I have been exposed to on the subject. My research is mainly based on trusting domain experts' (Papers II, III, & IV) explanations of how my understanding of computational journalism is shared or dismissed, partly or wholly, by them. The total

empirical data I have used are not big, but from highly specialized sources of knowledge. Still there are qualities in these studies that ensure a certain rigor.

News applications describes one activity in this field, it is easily observable internationally, and is not bound to the 79 applications I have analyzed. The patterns (and lack thereof) identified in Paper I represents a positioning of news application I feel confident in finding in news applications in general, at least from this approximate period of time. Paper II shows how central actors in Norway thinks about this field, and they share many key elements that are likely to be shared by other programmer-journalists elsewhere. In this paper we choose to focus on the elements of the craft that was shared across newsrooms, and not outliers in terms of things that was different. This is a weakness in this study, but an active choice we made together as co-authors to deliver an as clear as possible account where emphasis is given to the elements with strongest signal in the data. A strength in this study is the amount of discussion co-authorship requires as both the analysis and discussion is made by two authors and formulated in constant dialogue where all claims requires an agreed interpretation of the data. Findings from paper III identifies a discrepancy between what user-generated content is imagined to represent to journalism, and how journalists that could fulfill these whishes actually perceived this. User-generated content, to them, represented an interesting source of information, but in a workplace where the management of limited resources is pressing, the identified positives (democratic aspects) are secondary to daily needs, such as finding what "the usual suspects", already publicly known persons and organizations say. This was a finding that emerged from the analysis of the data, and not a prepared question. The summary of stories found show that the types of stories mostly are of human interest and soft-news. This is consistent with other studies involving UGC in journalism (e.g. Harrison 2009). As a case to show how computational methods are perceived in journalism, this approach of creating software *for* journalism was to a lesser extent successful. Particularly the lack of early user-involvement made the feedback on the softwares' operation as a journalistic function alien and less fruitful as a tool to create good interviews concerning software as journalistic. Paper IV puts software in the

position as a journalist, or as journalism, an approach that stimulated/provoked parliamentary reporters to not only identify what they want a parliamentary watchdogging information system to do, but also expressing how parliamentary reporting is done. As such this study works much better than Paper III to understand what journalism is, and how technology can aid in some cases and not in other.

I consider it likely that if my studies were to be repeated, or my data reanalyzed by other social scientists, that my conclusions would stand. As such I regard my model as a “stable for now” model of computational journalism, based on sound methods and humble conclusions.

References

Al-Mashari, Majed, and Mohamed Zairi. 1999. “BPR Implementation Process: An Analysis of Key Success and Failure Factors.” *Business Process Management Journal* 5 (1) (March 1): 87–112. doi:10.1108/14637159910249108.

Allern, Sigurd. 2001. *Flokkdyr På Løvebakken? - Sigurd Allern - Innbundet (9788253023168)*. http://www.bokkilden.no/SamboWeb/produkt.do?produktId=120881.

Amico, Laura, and Chris Amico. 2011. “Homicide Watch D.C.” *Homicide Watch D.C.* August. http://homicidewatch.org/about/.

An, J., M. Cha, K. Gummadi, and J. Crowcroft. 2011. “Media Landscape in Twitter: A World of New Conventions and Political Diversity.” *Proc. ICWSM* 11.

Andersen, Espen. 2013. *Datastøttet journalistikk*. IJ-forlaget. http://www.adlibris.com/no/product.aspx?isbn=8202404266.

Andersen, Michael. 2009. “Four Crowdsourcing Lessons from the Guardian’s (spectacular) Expenses-Scandal Experiment.” http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment/.

Anderson, C.W., Emely Bell, and Clay Shirky. 2012. “Post Industrial Journalism: Adapting to the Present”. Columbia Journalism School | Tow Center for Digital Journalism. http://towcenter.org/research/post-industrial-journalism/.

Anderson, Chris W. 2012. “Notes Towards an Analysis of Computational Journalism.” *HIIG Discussion Paper Series* 2012 (1). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2009292.

Baack, Stefan. 2011. “A New Style of News Reporting: Wikileaks and Data-Driven Journalism.” *Cyborg Subjects*. July. http://journal.cyborgsubjects.org/2011/07/style-news-reporting-wikileaks-data-driven-journalism/.

Becker, Hila, Mor Naaman, and Luis Gravano. 2010. “Learning Similarity Metrics for Event Identification in Social Media.” In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 291–300. WSDM ’10. New York, NY, USA: ACM. doi:10.1145/1718487.1718524. http://doi.acm.org/10.1145/1718487.1718524.

Berkowitz, Daniel A. 1997. *Social Meanings of News: A Text-Reader*. SAGE.

Berry, Dr David M. 2012. *Understanding Digital Humanities*. Palgrave Macmillan.

Bjørgan, Janne. 2013. “Digitale Vinnere.” *Journalisten.no*. April 26. http://www.journalisten.no/node/39788.

Boczkowski, Pablo. 2005. *Digitizing the News: Innovation in Online Newspapers*. The MIT Press.

Boczkowski, Pablo J. 2009. “Technology, Monitoring, and Imitation in Contemporary News Work.” *Communication, Culture & Critique* 2 (1): 39–59. doi:10.1111/j.1753-9137.2008.01028.x.

Bogost, Ian, Simon Ferrari, and Bobby Schweizer. 2010. *Newsgames: Journalism at Play*. The MIT Press.

Bostock, Michael. 2012. “D3.js - Data-Driven Documents.” *Data-Driven Documents*. http://d3js.org/.

Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things out*. MIT Press.

Breed, W. 1954. “Social Control in the Newsroom: A Functional Analysis.” *Social Forces* 33: 326.

Brurås, Svein. 2009. “Media Accountability Systems.” *Norsk Medietidsskrift* 16 (2): 120–139.

Bucher, Taina. 2012. “Programmed Sociality: A Software Studies Perspective on Social Networking Sites”. Dissertation, Oslo: UiO. http://tainabucher.com/wp-content/uploads/2009/08/Bucher_Ph.D.diss_.pdf.

Burn-Murdoch, John. 2012. “First Ever International Data Journalism Awards Launched.” *The Guardian*, January 19, sec. News. http://www.guardian.co.uk/news/datablog/2012/jan/19/global-data-journalism-awards-google.

Carr, David. 2012. “Innovation in Journalism Goes Begging for Support.” *The New York Times*, September 9, sec. Business Day / Media & Advertising. http://www.nytimes.com/2012/09/10/business/media/homicide-watch-web-site-venture-struggles-to-survive.html.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011. “Information Credibility on Twitter.” In *Proceedings of the 20th International Conference on World Wide Web*, 675–684. WWW ’11. New York, NY, USA: ACM. doi:10.1145/1963405.1963500. http://doi.acm.org/10.1145/1963405.1963500.

Chikofsky, E.J., and J.H. Cross. 1990. “Reverse Engineering and Design Recovery: A Taxonomy.” *IEEE Software* 7 (1) (January): 13–17. doi:10.1109/52.43044.

Chua, Reg. 2010. “What’s It All About?” *(Re)Structuring Journalism*. August 10. http://structureofnews.wordpress.com/2010/08/10/whats-it-all-about/.

Cohen, S., C. Li, J. Yang, and C. Yu. 2011. “Computational Journalism: A Call to Arms to Database Researchers.” In Asilomar, California, USA.

Cohen, Sarah. 2011. “Shared Values, Clashing Goals: Journalism and Open Government.” *XRDS* 18 (2) (December): 19–22. doi:10.1145/2043236.2043246.

Cohen, Sarah, James T. Hamilton, and Fred Turner. 2011. “Computational Journalism.” *Commun. ACM* 54 (10) (October): 66–71. doi:10.1145/2001269.2001288.

Coulter, Jones. 2012. “Investigative Reporters and Editors | Share, Interact with Data Easier with a PANDA in Your Newsroom.” *IRE*. January 23. http://www.ire.org/blog/on-the-road/2012/01/23/share-interact-with-data-easier-with-panda/.

Cox, Melisma. 2000. “The Development of Computer-Assisted Reporting.” *Informe Presentado En Association for Education in Jornalism End Mass Comunication). Chapel Hill, EEUU: Universidad de Carolina Del Norte*.

Creswell, John W. 2009. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Los Angeles: SAGE.

Curran, James, and Michael Gurevitch. 2005. *Mass Media and Society*. London; New York: Hodder Arnold ; Distributed in the U.S.A by Oxford University Press.

Cutting, Douglass R., David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections." In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 318–329. SIGIR '92. New York, NY, USA: ACM. doi:10.1145/133160.133214. http://doi.acm.org/10.1145/133160.133214.

Daniel, A., and T. Flew. 2010. "The Guardian Reportage of the UK MP Expenses Scandal: A Case Study of Computational Journalism." In *Record of the Communications Policy and Research Forum 2010*, 186–194. http://eprints.qut.edu.au/39358/.

DeFleur, Margaret H. 1997. *Computer-Assisted Investigative Reporting: Development and Methodology*. Erlbaum. http://www.getcited.org/pub/100150985.

Deuze, M. 2005. "What Is Journalism?: Professional Identity and Ideology of Journalists Reconsidered." *Journalism* 6 (4): 442.

Deuze, Mark, Axel Bruns, and Christoph Neuberger. 2007. "PREPARING FOR AN AGE OF PARTICIPATORY NEWS." *Journalism Practice* 1 (3): 322–338. doi:10.1080/17512780701504864.

Diakopoulos, N., M. De Choudhury, and M. Naaman. 2012. "Finding and Assessing Social Media Information Sources in the Context of Journalism." In *Proc. Conference on Human Factors in Computing Systems (CHI)*.

Diakopoulos, Nicholas. 2010. "A Functional Roadmap for Innovation in Computational Journalism." In .

———. 2012. "Cultivating the Landscape of Innovation in Computational Journalism", March. http://towknight.org/files/2012/04/diakopoulos_whitepaper_systematicinnovation.pdf.

———. 2013. "Rage Against the Algorithms." *The Atlantic*. October 3. http://www.theatlantic.com/technology/archive/2013/10/rage-against-the-algorithms/280255/.

Django Software Foundation. 2010. "Django | Writing Your First Django App, Part 1 | Django Documentation." 06. http://docs.djangoproject.com/en/dev/intro/tutorial01/#activating-models.

Eide, Martin. 1984. *Etter det vi forstår på politisk hold--: politikere og massemedia*. Universitetsforlaget.

———. 1992. *Nyhetens interesse: nyhetsjournalistikk mellom tekst og kontekst*. Oslo: Universitetsforlaget.

———. 2012. "Reconstructing Accountability: Essential Journalistic Reorientations." In *The International Encyclopedia of Media Studies*, by Angharad N. Valdivia. Oxford, UK: Blackwell Publishing Ltd. http://onlinelibrary.wiley.com/mrw/advanced/search/results.

EJC. 2013. "Homepage | Data Driven Journalism." Accessed August 7. http://datadrivenjournalism.net/.

Engebretsen, Martin. 1999. "Nyheter På Nettet: Fortellinger Eller Informasjonsnettverk?" *Norsk Medietidsskrift* 1999 (2). Institusjonar Og Økonomi. http://www.medieforskerlaget.no/archives/1186.

Engholm, Ida, and Lisbeth Klastrup. 2004. *Digitale Verdener: De Nye Mediers Æstetik Og Design*. [København]: Gyldendal.

Entman, Robert M. 1993. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication* 43 (4): 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x.

European Journalism Centre (EJC),. 2010. "Data-Driven Journalism: What Is There to Learn?" European Journalism Centre (ECJ). http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdf.

Fagerjord, Anders. 2012. "Design som medievitenskapelig metode." *Norsk Medietidsskrift*. https://www.duo.uio.no//handle/10852/34192.

Farr, Christina. 2013. "Narrative Science Goes beyond 'Robot Journalism' with CIA Investment." *VentureBeat*. June 5. http://venturebeat.com/2013/06/05/narrative-science-goes-beyond-robot-journalism-with-cia-investment/.

Fassler, Joe. 2012. "Can the Computers at Narrative Science Replace Paid Writers?" *The Atlantic*. April 12. http://www.theatlantic.com/entertainment/archive/2012/04/can-the-computers-at-narrative-science-replace-paid-writers/255631/.

Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift. 2011. "The Promise of Computational Journalism." *Journalism Practice* 6 (2): 157–171. doi:10.1080/17512786.2011.616655.

Franklin, Professor Bob, Martin Hamer, Mr Mark Hanna, Marie Kinsey, and Dr John E Richardson. 2005. *Key Concepts in Journalism Studies*. Sage Publications Ltd.

Futsæter, Knut-Arne. 2012. "MedieTrender 2011" February 12. www.tns-gallup.no/arch/_img/9100748.pdf.

Garrison, Bruce. 1998a. "Newspaper Size as a Factor in Use of Computer-Assisted Reporting." *Unpublished Paper Presented to the Communication Technology and Policy Division of the Association for Education in Journalism and Mass Communication, Baltimore, MD*. http://com.miami.edu/car/baltimore1.htm.

———. 1998b. *Computer-Assisted Reporting*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Gentikow, Barbara. 2005. *Hvordan utforsker man medieerfaringer?: kvalitativ metode*. IJ forlaget.

Gipp, Bela, Norman Meuschke, and Joeran Beel. 2011. "Comparative Evaluation of Text- and Citation-Based Plagiarism Detection Approaches Using Guttenplag." In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 255–258. JCDL '11. New York, NY, USA: ACM. doi:10.1145/1998076.1998124. http://doi.acm.org/10.1145/1998076.1998124.

Gitelman, Lisa. 2013. *Raw Data Is an Oxymoron*. MIT Press.

Golafshani, Nahid. 2003. "Understanding Reliability and Validity in Qualitative Research." *The Qualitative Report* 8 (4): 597–607.

González Veira, Xaquín. 2013. "Making of NYT's Mariano Rivera's Pitches | Xocas.com." Accessed September 19. http://www.xocas.com/blog/en/making-of-nyts-mariano-riveras-pitches/.

Gray, Jonathan, Lucy Chambers, and Liliana Bounegru. 2012. *The Data Journalism Handbook*. http://shop.oreilly.com/product/0636920025603.do.

Groskopf, Christopher. 2011. *PyCon 2011: Best Practices for Impossible Deadlines*. PyCon 2011. http://blip.tv/pycon-us-videos-2009-2010-2011/pycon-2011-best-practices-for-impossible-deadlines-4899490.

Gynnild, Astrid. 2007. "Creative Cycling of News Professionals." *The Grounded Theory Review* Volume 06 (2) (March). http://groundedtheoryreview.com/2007/03/30/1144/.

———. 2013. "Journalism Innovation Leads to Innovation Journalism: The Impact of Computational Exploration on Changing Mindsets." *Journalism* (May 22). doi:10.1177/1464884913486393. http://jou.sagepub.com/content/early/2013/05/19/1464884913486393.

Hacks/Hackers. 2010. "About Hacks/Hackers." *Hacks/Hackers*. http://hackshackers.com/about/.

Haik, Cory. 2013. "About | TruthTeller." Accessed August 8. http://truthteller.washingtonpost.com/about/.

Hallin, Daniel C., and Paolo Mancini. 2004. *Comparing Media Systems: Three Models of Media and Politics*. Cambridge University Press.

Hamilton, James T., and Fred Turner. 2009. "Accountability Through Algorithm: Developing the Field of Computational Journalism. Report from Developing the Field of Computational Journalism." Center for Advanced Study in the Behavioral Sciences Summer Workshop (Stanford, CA, July 27--31, 2009). http://dewitt.sanford.duke.edu/images/uploads/about_3_Research_B_cj_1_final report.pdf.

Harrison, Jackie. 2009. "USER-GENERATED CONTENT AND GATEKEEPING AT THE BBC HUB." *Journalism Studies* 11 (2): 243–256. doi:10.1080/14616700903290593.

Heftøy, Jens Egil. 2013. "SKUP - Stiftelsen for En Kritisk Og Undersøkende Presse." April 3. http://www.skup.no/177/9132.

Heidegger, M. 2001. *The Question Concerning Technology*. http://www.google.com/books?hl=en&lr=&id=BgYc9_ldWFYC&oi=fnd&pg=PA99&dq=Heidegger+Concerning+Technology&ots=wuMwN1SGv_&sig=iWJ-mS3l8vc3Py_QpCxW0feqUdE.

Hevner, A. R., S. T. March, J. Park, and S. Ram. 2004. "Design Science in Information Systems Research." *Mis Quarterly* 28 (1): 75–105.

Hevner, Alan, and Samir Chatterjee. 2010. *Design Research in Information Systems - Theory and Practice*. http://www.springer.com/business+%26+management/business+information+systems/book/978-1-4419-5652-1.

holderdeord. 2013. "Holder de Ord." July. http://www.holderdeord.no/home/faq#har-dere-en-faglig-inspirasjonskilde.

Holovaty, Adrian. 2006. "A Fundamental Way Newspaper Sites Need to Change | Holovaty.com". Article. http://www.holovaty.com/writing/fundamental-change/.

———. 2009. “The Definitive, Two-Part Answer to ‘Is Data Journalism?’ | Holovaty.com.” May 21. http://www.holovaty.com/writing/data-is-journalism/.

Houston, Brant. 1996. *Computer-Assisted Reporting: A Practical Guide*. New York: St. Martin’s Press.

Houston, Brant, Len Bruzzese, Steve Weinberg, and Inc Investigative Reporters and Editors. 2002. *The Investigative Reporter’s Handbook: A Guide to Documents, Databases, and Techniques*. Boston: Bedford/St. Martin’s.

Hovden, Jan Fredrik. 2008. “Profane and Sacred. A Study of the Norwegian Journalistic Field”. Doctoral thesis, The University of Bergen. http://bora.uib.no/handle/1956/2724.

———. 2012. “A Journalistic Cosmology.” *Nordicom Review* 33 (2) (December 1): 57–76.

Hullman, Jessica, and Nick Diakopoulos. 2011. “Visualization Rhetoric: Framing Effects in Narrative Visualization.” *IEEE Transactions on Visualization and Computer Graphics* 17 (12) (December): 2231–2240. doi:10.1109/TVCG.2011.255.

“Intelligent Information Laboratory @ Northwestern University - Projects - Stats Monkey.” 2013. *Projects > Stats Monkey*. Accessed September 19. http://infolab.northwestern.edu/projects/stats-monkey/.

Jacobson, Susan. 2012. “Transcoding the News: An Investigation into Multimedia Journalism Published on Nytimes.com 2000–2008.” *New Media & Society* (January 9). doi:10.1177/1461444811431864. http://nms.sagepub.com/content/early/2012/01/05/1461444811431864.

Kaptelinin, Victor, and Bonnie A. Nardi. 2006. *Acting with Technology Activity Theory and Interaction Design*. Acting with Technology. Cambridge, Mass.: MIT Press.

Karlsen, Joakim, and Eirik Stavelin. 2013. “Computational Journalism in Norwegian Newsrooms.” *Journalism Practice* (July 23): 1–15. doi:10.1080/17512786.2013.813190.

Karlsson, Michael. 2010. “RITUALS OF TRANSPARENCY.” *Journalism Studies* 11 (4) (August): 535–545. doi:10.1080/14616701003638400.

Keane, John. 2009. *The Life and Death of Democracy*. New York: W.W. Norton & Co.

Kiscuitwala, Kanak, Willem Bult, Mathias Lécuyer, T.J. Purtell, Madeline K.B. Ross, Augustin Chaintreau, Chris Haseman, Monica S. Lam, and Susan E. McGregor. 2013. “Weaving a Safe Web of News.” In *Proceedings of the 22nd International Conference on World Wide Web Companion*, 849–852. WWW ’13 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. http://dl.acm.org/citation.cfm?id=2487788.2488063.

Klinenberg, Eric. 2005. “Convergence: News Production in a Digital Age.” *Annals of the American Academy of Political and Social Science* 597 (January 1): 48–64.

Kovach, Bill, and Tom Rosenstiel. 2007. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect, Completely Updated and Revised*. Rev Upd. Three Rivers Press.

Krawetz, N. 2007. “A Picture’s Worth...: Digital Image Analysis and Forensics.” *Black Hat Briefings USA2007*.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. “What Is Twitter, a Social Network or a News Media?” In *Proceedings of the 19th International Conference on World Wide Web*, 591–600. WWW ’10. New York, NY, USA: ACM. doi:10.1145/1772690.1772751. http://doi.acm.org/10.1145/1772690.1772751.

Larsson, Anders. 2012. “Interactivity on Swedish Newspaper Web Sites – What Kind, How Much and Why?” *Convergence: The International Journal of Research into New Media Technologies*. http://uppsala.academia.edu/AndersOlofLarsson/Papers/1007934/Interactivity_on_Swedish_newspaper_web_sites_-_What_kind_how_much_and_why.

Larsson, Anders Olof. 2012. “Doing Things in Relation to Machines: Studies on Online Interactivity”. Uppsala University.

Larsson, Anders Olof, and Hallvard Moe. 2011. “Studying Political Microblogging: Twitter Users in the 2010 Swedish Election Campaign.” *New Media & Society* (November 21). doi:10.1177/1461444811422894. http://nms.sagepub.com/content/early/2011/11/21/1461444811422894.

Latour, Bruno. 1992. “Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts.” In , by Wiebe Bijker and John Law, 225–258.

———. 1999. *Pandora’s Hope: Essays on the Reality of Science Studies*. 1st ed. Harvard University Press.

Lessig, Lawrence. 2006. *Code Version 2.0*. New York.

Lewis, Seth C., and Nikki Usher. 2013. “Open Source and Journalism: Toward New Frameworks for Imagining News Innovation.” *Media Culture & Society* 35 (5) (July): 602–619. doi:10.1177/0163443713485494.

Lichtenberg, Judith. 1991. “In Defence of Objectivity.” In *Mass Media and Society*.

Loper, Edward, and Steven Bird. 2002. “NLTK: The Natural Language Toolkit.” *arXiv:cs/0205028* (May 17). http://arxiv.org/abs/cs/0205028.

Mattelmäki, Tuuli, and Ben Matthews. 2009. “PEELING APPLES: PROTOTYPING DESIGN EXPERIMENTS AS RESEARCH.” In *Engaging Artifacts*. http://ocs.sfu.ca/nordes/index.php/nordes/2009/paper/view/214.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

Mcclain, Dylan Loeb. 2012. “A Computer Program to Detect Possible Cheating in Chess.” *The New York Times*, March 19, sec. Science. http://www.nytimes.com/2012/03/20/science/a-computer-program-to-detect-possible-cheating-in-chess.html.

McGHEE, GEOFF. 2010. “Journalism in the Age of Data: A Video Report on Data Visualization by Geoff McGhee.” *Journalism in the Age of Data*. August 31. http://datajournalism.stanford.edu/.

Meyer, Philip. 1973. *Precision Journalism: A Reporter’s Introduction to Social Science Methods*. Indiana University Press.

Mitchelstein, Eugenia, and Pablo J. Boczkowski. 2009. “Between Tradition and Change A Review of Recent Research on Online News Production.” *Journalism* 10 (5) (October 1): 562–586. doi:10.1177/1464884909106533.

Morozov, Evgeny. 2012. “A Robot Stole My Pulitzer!” *Slate*, March 19. http://www.slate.com/articles/technology/future_tense/2012/03/narrative_science_robot_journalists_customized_news_and_the_danger_to_civil_discourse_.html.

Mulvad, Nils, and Flemming Tait Svith. 1998. *Vagthundens Nye Bolig*. [Århus]: Ajour.

Mulvad, Nils, Swithun helgen, and Flemming Tait Svith. 2002. *Vagthund I Vidensamfundet*. Århus: Ajour.

mySociety. 2013. “TheyWorkForYou: Hansard and Official Reports for the UK Parliament, Scottish Parliament, and Northern Ireland Assembly - Done Right.” Accessed June 24. http://www.theyworkforyou.com/.

Neuendorf, Kimberly A. 2002. *The Content Analysis Guidebook*. SAGE.

Nezda, Luke. 2012. “ClusteringInDepth. Methods and Theory behind the Clustering Functionality in Google Refine.” Accessed March 6. http://code.google.com/p/google-refine/wiki/ClusteringInDepth.

Nielsen, Jakob, and Rolf Molich. 1990. “Heuristic Evaluation of User Interfaces.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 249–256. CHI ’90. New York, NY, USA: ACM. doi:10.1145/97243.97281. http://doi.acm.org/10.1145/97243.97281.

NxtMedia. 2013. “Kamp Om Prisen for Datajournalistikk.” *NxtMedia*. Accessed August 6. http://nxtmedia.no/kamp-om-prisen-for-datajournalistikk/.

Nygren, Gunnar, Ester Appelgren, and Helge Hüttenrauch. 2012. “Datajournalistik - ett växande område.” *Nordicom Information* 34 (3-4): 81–88.

O’Brien III, John. 2010. *Your TwapperKeeper – Archive Your Own Tweets - Http:// Your.twapperkeeper.com*. http:// your.twapperkeeper.com.

OpenCongress. 2013. “About Open Congress - OpenCongress.” Accessed June 24. http://www.opencongress.org/about.

Orlikowski, Wanda J. 2000. “Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations.” *Organization Science* 11 (4) (July 1): 404–428.

Østbye, Helge. 2009. *Journalistiske nyorienteringer*. Edited by Martin Eide. Scandinavian Academic Press.

Øvrebø, Olav Anders. 2011. “Redaktører Hyller Datajournalistikk, Men Satser Ikke Selv « Vox Publica.” http://voxpublica.no/2011/04/redakt%c3%b8rer-hyller-datajournalistikk-men-satser-ikke-selv/.

Parasie, Sylvain, and Eric Dagiral. 2012. “Data-Driven Journalism and the Public Good: ‘Computer-Assisted-Reporters’ and ‘programmer-Journalists’ in Chicago.” *New Media & Society* (November 18). doi:10.1177/1461444812463345. http://nms.sagepub.com/content/early/2012/11/15/1461444812463345.

Park, S., M. Ko, Y. Liu, D. Y. Jin, and J. Song. 2011. “Improving Journalism through the Web: Framework for Media Bias Mitigation.”

Pilhofer, Aron. 2010. “Programmer-Journalist? Hacker-Journalist? Our Identity Crisis | Idea Lab | PBS.” April 22. http://www.pbs.org/idealab/2010/04/programmer-journalist-hacker-journalist-our-identity-crisis107.

Poole, Keith, Jeffrey Lewis, James Lo, and Royce Carroll. 2012. “CRAN - Package Oc.” *Oc: OC Roll Call Analysis Software*. January 24. http://cran.r-project.org/web/packages/oc/.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Cambridge University Press. http://www.google.com/books?hl=en&lr=&id=OmeQNHvcULoC&oi=fnd&pg=PR11&dq=spatial+models+of+parliamentary+voting&ots=1HfFZmjfQi&sig=J7ah1qtkIyi1LcYQdv2W5Z3O-jo.

Poole, Keith T., and Howard L. Rosenthal. 2011. *Ideology and Congress*. Transaction Publishers.

Pousman, Z., J. Stasko, and M. Mateas. 2007. “Casual Information Visualization: Depictions of Data in Everyday Life.” *IEEE Transactions on Visualization and Computer Graphics*: 1145–1152.

Powers, Matthew. 2012. “‘In Forms That Are Familiar and Yet-to-Be Invented’ American Journalism and the Discourse of Technologically Specific Work.” *Journal of Communication Inquiry* 36 (1) (January 1): 24–43. doi:10.1177/0196859911426009.

Poynter Institute. 1999. *When Nerds and Words Collide: Reflections on the Development of Computer Assisted Reporting*. Poynter Institute for Media Studies.

Pulimood, Sarah Monisha, Donna Shaw, and Emilie Lounsberry. 2011. “Gumshoe: A Model for Undergraduate Computational Journalism Education.” In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*, 529–534. SIGCSE ’11. New York, NY, USA: ACM. doi:10.1145/1953163.1953314. http://doi.acm.org/10.1145/1953163.1953314.

Reenskaug, Trygve. 2013. “MVC - XEROX PARC 1978-79.” Accessed October 10. http://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html.

Roberts, Graham, Shan Carter, and Joe Ward. 2013. “How Mariano Rivera Dominates Hitters.” Accessed September 19. http://www.nytimes.com/interactive/2010/06/29/magazine/rivera-pitches.html?_r=0.

Robson, Colin. 2002. *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*. Wiley.

Rogers, Simon. 2011. “The First Guardian Data Journalism: May 5, 1821.” *The Guardian*. September 26. http://www.theguardian.com/news/datablog/2011/sep/26/data-journalism-guardian.

———. 2013. *Facts Are Sacred: The Power of Data.* London: Faber and Faber.

Roppen, Johan, and Sigurd Allern. 2013. “Journalistikkens Samfunnsoppdrag - Akademisk | Cappelen Damm Undervisning.” Accessed October 23. https://www.cappelendammundervisning.no/undervisning/akademisk/tema/ijforlaget/product-detail.action?id=163071.

Royal, Cindy, Robert Bergland, Javier Diaz Noci, Maria Holubowicz, Marcus Messner, Ahmed El Gody, David Hon, Lisa Crawford, Sarah Noe, and David Domingo. 2012. *#ISOJ The Official Research Journal of the International Symposium on Online Journalism, Volume 2, Number 1*. BookBrewer.

Schudson, Michael. 2003. *The Sociology of News*. W W Norton & Company Incorporated.

Segel, E., and J. Heer. 2010. "Narrative Visualization: Telling Stories with Data." *Visualization and Computer Graphics, IEEE Transactions on* 16 (6): 1139–1148.

Sharp, Helen, Yvonne Rogers, and Jenny Preece. 2007. *Interaction Design: Beyond Human-Computer Interaction*. 2nd ed. Wiley.

Silverman, David. 2001. *Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction*. London [u.a.]: Sage.

Sirkkunen, Esa, Tanja Aitamurto, and Pauliina Lehtonen. 2011. "Trends In Data Journalism."

Sjøvaag, Helle. 2010. "The Reciprocity of Journalism's Social Contract." *Journalism Studies* 11 (6): 874–888. doi:10.1080/14616701003644044.

———. 2011. *Journalistic Ideology: Professional Strategy, Institutional Authority and Boundary Maintenance in the Digital News Market*. [Bergen]: University of Bergen.

Sjøvaag, Helle, Hallvard Moe, and Eirik Stavelin. 2012. "Public Service News on the Web." *Journalism Studies* 13 (1): 90–106. doi:10.1080/1461670X.2011.578940.

Sjøvaag, Helle, and Eirik Stavelin. 2012. "Web Media and the Quantitative Content Analysis: Methodological Challenges in Measuring Online News Content." *Convergence: The International Journal of Research into New Media Technologies* (February 7): 1354856511429641. doi:10.1177/1354856511429641.

Skardal, Thomas, and Thomas Jakobsen. 2007. "Readability Index." http://www.mortengoodwin.net/publicationfiles/webmining_2007_reportgroup6.pdf.

Star, Susan Leigh, and James R. Griesemer. 1989. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19 (3) (August 1): 387–420.

Stasko, J., C. Görg, and Z. Liu. 2008. "Jigsaw: Supporting Investigative Analysis through Interactive Visualization." *Information Visualization* 7 (2): 118–132.

Stavelin, Eirik. 2012. "Nyhetsapplikasjoner: Journalistikk Møter Programmering." In *Nytt På Nett Og Brett: Journalistikk I Forandring*, edited by Martin 1956- Eide, Leif Ove 1961- Larsen, and Helle 1977- Sjøvaag, 107–125. Oslo: Universitetsforlaget.

Steensen, Steen. 2010. "ONLINE JOURNALISM AND THE PROMISES OF NEW TECHNOLOGY -- A Critical Review and Look Ahead." *Journalism Studies*. http://www.informaworld.com/10.1080/1461670X.2010.501151.

Stodden, Victoria, Peixuan Guo, and Zhaokun Ma. 2013. "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy

Adoption by Journals." *PLoS ONE* 8 (6) (June 21): e67111. doi:10.1371/journal.pone.0067111.

Stray, Jonathan. 2011a. "Jonathan Stray » A Computational Journalism Reading List." January 31. http://jonathanstray.com/a-computational-journalism-reading-list.

———. 2011b. "The Overview Project - 3 Difficult Document-Mining Problems That Overview Wants to Solve." November 2. http://overview.ap.org/blog/2011/11/3-difficult-document-mining-problems-that-overview-wants-to-solve/.

Tauberer, Joshua. 2012. *Open Government Data: The Book*. http://opengovdata.io.

"The Oslo-Bergen Tagger." 2012. Accessed March 7. http://www.tekstlab.uio.no/obt-ny/english/index.html.

Tuchman, Gaye. 1972. "Objectivity as Strategic Ritual: An Examination of Newsmen's Notions of Objectivity." *American Journal of Sociology* 77 (4) (January 1): 660–679.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Graphics Press.

Vaishnavi, V., and W. Kuechler. 2004. "Design Science Research in Information Systems." January 20. http://desrist.org/design-research-in-information-systems/.

Viegas, Fernanda, Martin Wattenberg, and Sarah Cohen. 2010. "TimeFlow." *Timeline Visualization Application Http://flowingmedia.com/timeflow.html*. https://github.com/FlowingMedia/TimeFlow.

Waite, Matt. 2009. "The Key Lesson I Learned Building PolitiFact: Demos, Not Memos | Mattwaite.com." April 27. http://www.mattwaite.com/posts/2009/apr/27/key-lesson-i-learned-building-politifact-demos-not/.

Weber, Wibke, and Hannes Rall. 2013. "„We Are Journalists." Production Practices, Attitudes and a Case Study of the New York Times Newsroom." In *Interaktive Infografiken*, edited by Wibke Weber, Michael Burmester, and Ralph Tille, 161–172. X.media.press. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-15453-9_9.

Weinstein, Matthew. 2006. "TAMS Analyzer Anthropology as Cultural Critique in a Digital Age." *Social Science Computer Review* 24 (1) (February 1): 68–77. doi:10.1177/0894439305281496.

Willett, W., J. Heer, J. Hellerstein, and M. Agrawala. 2011. "CommentSpace: Structured Support for Collaborative Visual Analysis." In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, 3131–3140.

Wing, Jeannette M. 2006. "Computational Thinking." *Communications of the ACM* 49 (3) (March 1): 33. doi:10.1145/1118178.1118215.

Wolz, Ursula, Meredith Stone, Sarah M. Pulimood, and Kim Pearson. 2010. "Computational Thinking via Interactive Journalism in Middle School." In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education*, 239–243. SIGCSE '10. New York, NY, USA: ACM. doi:10.1145/1734263.1734345. http://doi.acm.org/10.1145/1734263.1734345.

Zelizer, Barbie. 2004. *Taking Journalism Seriously: News and the Academy*. SAGE Publications.

Part 2: The Articles

I. News applications – journalism meets programming

In the wake of the digitization of news production, a new, digital type of journalism is emerging: the news application. For example NRK's *Maktbasen [the power base]* [3] creates links between elected representatives and their duties in enterprises and organisations, the Guardian's *Comprehensive spending review: you make the cuts* [22] allows users to make cuts in Great Britain's state budget, while the New York Times' *Murder: New York City* [4] visualises where in New York City the police have investigated murders and allows users to manipulate variables such as the gender, age and ethnicity of victims and perpetrators to show when, where and with which weapon the murders were committed. These are only some examples of journalistic projects published by online newspapers, in which technology and journalism have been combined and program code has been published. Thus, producing a form of journalism written in software code.

This study is the result of a project, which mapped news applications in 2009/2010. The aim was to examine the purpose of this type of journalism, and what function news applications are intended to perform. News applications raise questions about what competencies and conditions are needed to create this type of journalism, and the fundamental aim of this study was to establish whether it is appropriate to categorise news applications as journalism.

The initial stage of this project was the creation of a list of all the news applications known to the author. The list was published in the online magazine voxpublica.no[24]. Readers of the magazine – among them journalists working for Norwegian online newspapers – added to the list with further examples, which were submitted through voxpublica.no or e-mailed. This set a trend in motion and the list was regularly updated.

[24] The blog post is available at http://voxpublica.no/2010/10/nyhetsapplikasjoner-pa-web-hvem-hva-hvordan/

There are several methodological issues that affect snowball sampling. The sample is quite small, and has some bias; in this case, bias in terms of which newsrooms would be represented (newsrooms with larger reach more likely to be included) and which countries included (Scandinavian countries and English speaking countries resulting from language apprehension). The advantage with gathering data in this way, however, is that it is possible to analyse more than just a few examples. By publishing the material during the process, external voices (voxpublica.no's readers) were able to contribute more data[25], and the result was a larger sample than would have been obtainable through traditional recruitment methods.

The result was a list of 102 news applications, both domestic and foreign. This was reduced to 79 by eliminating those not produced by established news organisations. The focus of interest was on how the established press, with its traditional position and function in society, uses these applications. The news applications collected were analysed according to thematic categories, their relationship to the core function of their social contract and the applications' (visual, technical and methodological) components.

The hope is to be able to determine if this phenomenon of news applications represents something genuinely new, or whether it is a development of traditional journalism? Nothing emerges from a vacuum, and the utilisation of computer programming by the press is not new. Among the predecessors of the news applications considered herein, are computer assisted reporting, data journalism and online journalism.

Computer assisted reporting

According to Melissa Cox (2000), 1952 was when the first computers were used for computer assisted reporting (CAR). CBS News' Washington correspondent, Walter

[25] Among the contributors to the data were employees from VG (national tabloid) and TV2 (national news producing television station).

Cronkite, used a computer to predict the election result in the race between Dwight Eisenhower and Adlai Stevensom based on counting early votes. When a gap between the experts in the studio (who expected a close race) and the computer analysis (which indicated a landslide victory in favour of Eisenhower) occurred, CBS hesitated to share the computed results. When they later published these computed results, they were criticised for not having trusted them, as Eisenhower did, indeed, win by a landslide. However, for the press in 1952 the computer was something unfamiliar and untrustworthy.

Today the use of information systems in newsrooms throughout the world is far more complex. Typewriters have been replaced with word processors, letters are now sent by email and large numbers of different information systems are utilised. Moreover, the majority of work is done in front of a computer screen, and "all" journalists use Twitter. Moreover, research, publishing and archiving – in other words, the whole production process – are today supported by information systems. In the example from 1952, a computer was used simply to edit data that was published as a news story. The information system was not used to support correspondent Cronkite with communication, extended memory or computing power (as an email reader or spreadsheet software does), but to assist him in predicting the election result. The computer support was not general, but specialised. The machine counted votes in Cronkite's place, and thereby extended his journalistic skills. He did not use a technical infrastructure to make his job easier, but created an infrastructure to do his job[26]. The difference between using a tool and creating a tool is a significant one.

Since 1952, computer supported analysis has been used to accurately predict the results of all American presidential elections. These methods have also been used in other categories of news and have developed in parallel with the entry of computers into society and newsrooms. In spite of this, computer programming has not been a

[26] Cronkite had a team of programmers that created the algorithm that gave the prognosis for the election results.

field of study taught in Norwegian journalism schools, and the number of stories produced using custom source code is limited.

Online journalism

"Most online newspapers have the same texts on screen as appears in the paper editions" writes Martin Engebretsen in *Norsk Medietidsskrift* in 1999. In the book *Digitizing the News,* Pablo Poczowski describes the process of reusing articles from a paper edition in an online edition as 'shovelware'. The texts are shovelled over to the online edition, with minimal changes, if any (Boczkowski, 2005). This is still largely the case for online newspaper stories worldwide. In parallel to this, original material is also produced for the web or uniquely adapted for it: for example, articles containing sound and film (Østbye, 2009), news games (Bogost, Ferrari & Schweizer, 2010), photo- and video-based stories and other formats often categorised under the somewhat unclear label, "multimedia". The enabling factors for these formats are found in the platform of the web itself.

In the paper *Nyheter på Nettet: Fortelling eller informasjonsnettverk?* (News on the web: stories or information networks?) Engebretsen describes the basic prerequisites of the web as a platform for journalism:

> *The paper edition works […] within the framework of a culture of written letters developed on the basis for all graphical technology: marking of symbols on a two-dimensional plain. When this technology is used to produce a lasting representation of human verbal language, a linear sequence of words and sentences, is a natural consequence. One produces "frozen speech" (cf. Ricoeur, 1993). The online newspaper […] is produced and distributed by aid of a different technology than the graphical. The online newspaper is produced by aid of particular digital editors and design tools, and is coded in accordance to the protocol HTML (Hyper Text Markup Language), which enables distribution over the global computer network Internet. The technological frame for the online newspapers' news mediation is thus not defined by what is possible to mark on a paper surface, but what is possible to code in HTML (Engebretsen 1999).*

HTML places very few limitations on journalism. As Engebretsen further explains, it cannot be assumed that a news article on the web will be identical to the paper version. HTML is a format for digital information and this opens up further possibilities for online journalism: content can be copied, computed, moved and manipulated in innumerable ways, even after the material has been published. Adding different forms of interactivity to the individual news stories for the user to interact with requires the re-use of existing code (e.g. exploiting default functionality provided by browsers) or by writing custom code. When interaction are added to individual stories, this is called "medium-interactivity" (Larsson, 2012).

Interaction with news applications can produce changes in the graphical user interface that the reader is presented with, and this is possible due to the prefabricated functions that execute these changes and which are published as part of the material. While verbal language in graphical technologies produces "frozen speech", software code produces "frozen labour".

> *Marx referred to technology as "frozen labour" - work and its values embedded and inscribed in transportable form. Modern information technologies similarly embed and inscribe work in ways that are important to policymakers, but […] are difficult to see. […] [A]rguments, decisions and uncertainties […] are hidden away inside a piece of technology or in a complex representation. Thus, values, opinions and rhetoric are frozen into codes, electronic thresholds and computer applications. Extending Marx, then, we can say that in many ways, software is frozen organizational discourse. (Bowker & Star, 2000)*

While the online journalist has liberated herself from the typographer and layout artists' idea of frozen speech, she has also inherited the technologist's notion of freezing down and commodifying work. When software code is written to execute a function, this function can be tirelessly run with the same results each time. The work can then be separated from the worker and published online. The online journalist's platform is a fundamentally different starting point for journalistic publication. It is a platform where not only content ("frozen speech") but also methods ("frozen labour") can be published. This is a condition for the news applications: changeable content, methods for manipulating the content, and an additional layer of presentation for the

audience. The news application differentiates itself from other news material by the publication of code as a part of the individual news story. This is also the basis for interactivity in online newspapers; i.e. the software code enables change.

The precondition for computer supported journalism is constantly improving as more and more of society's data becomes available online or in a digital format. What and how newsrooms publish on the web is increasingly varied. Furthermore, it is becoming easier to publish online.

Data and data journalism

The display of data is a journalistic tradition. The Guardian newspaper's first edition in 1821 contained an example of this; a table of schools in Manchester and Salford with accounts showing number of pupils and the schools' annual expenses, revealing how many children received a free education, providing a measure of how many children were living in poverty in these cities.

Data centred journalism, or data journalism, is still a regular feature in most forms of news media. Graphs and tables are used particularly frequently during elections, but in other subject areas such as sports, economy and society; the practice of publishing this data is also routine. To differentiate the nuances in terminology, this study will use 'data journalism' to describe journalism based on one or more data sets published as part of a news story, and 'computer-supported journalism' for journalism based on software-enabled methods. The term computer-assisted reporting has been used for so many tasks in journalism (e.g. email interviews, online searches, spread sheet usage, etc.), it is arguably about to lose its function, and just "journalism" remains. Computational journalism, a term originating in the technological fields, emphasises computing, although this is so far exemplified in the community by tools and methods rarely used in newsrooms. Data refers to structured symbols, which can be transformed into information by interpretation.

As a society we generate enormous amounts of digital data in an increasing number of areas in our lives. Public bureaucracies, private enterprises and organisations all

collect data, and each one of us collects, generates and organises data about ourselves and others in ever increasing quantities, via social media, personal databases (e.g. contact lists on mobile devices), customer registers, and membership databases of various types. The many who have the opportunity to gain from digitizing seem to have done so. A proportion of this material ends up in the hands of journalists who use it for news applications.

The reason to create applications is data that needs treatment. Therefore, it follows that news applications are a modern form of data journalism which offer additional methods for treating and presenting the data online. In the collection of news applications examined it is evident that public data (data which the media can openly claim or demand access to according to the freedom of information act) is very important. Over half of the applications are designed for this type of data, including education, elections, public spending and other public concerns. All data collected by the government (and not covered by the personal data protection laws or classified) is, in principle, accessible to the Norwegian public.

A few applications (7pcs) are based on heterogeneous material collected via traditional journalistic research. The material can be of various sorts (documents, notes, interview notes, photographs, video, etc.), and then the application becomes an online hub for information concerning the story. Examples include TV2's *Følg piratjakten her [Follow the pirate hunt] [1]* where information on the ships, events, ships logs and weather conditions in the Gulf of Aden are channelled through a tailored portal for the story. Another example is vg.nos *Hvorfor forsvant Jarle [Why did Jarle disappear] [2]*, an interactive story that takes us through the final observations and documentations concerning the disappearance of a man named Jarle. It provides a timeline and an interactive map that the reader can use to navigate through the known facts of the story.

Some applications can be described as crowdsourcing applications: these collect data, knowledge or insights from their audience. Examples of these will be discussed below.

The remaining applications handle data derived from various sources. This data is either purchased from data warehouses, given to or purchased from private enterprises or organisations, or collected from the web (e.g. Twitter and Wikileaks). Typical of this are *mashups*, which provide combinations of different datasets that produce a richer context, comparison or functionality. A relevant goal here is the information cocktail, which is an explosive mix of datasets that independently offer poor levels of information. Certain (12pcs) applications are clearly combinations of datasets[27]. A good example of the effects this can have is nrk.nos *Maktbasen [the power base] [3]*, where data concerning elected representatives is combined with data about commercial enterprises: incompetent/disqualified representatives are then exposed.

As long as the working material is in a digital format, the possibilities of making applications are good. The diversity of data sources and matter show a significant amount of creativity and versatility behind each piece of work. It is evident that public data is important for data journalism and therefore important to the development of news applications.

Journalistic methods – published online

Online journalists were liberated from "frozen speech" and static surfaces, and inherited the idea of "frozen labour" from the digital platform. In the 1952 example from NBC, code was written to calculate the estimated election result; it could be run multiple times, but only for this one task. Code is generally very specific, and while reusability is an ideal in programming, adaptation for each project is often necessary. As long as the structure and data types (and other meta data) of two data sets is not 100% identical, two different adaptations of software code will be needed to treat them. Software is rigid, and so is sometimes described as frozen organisational discourse. This means that an organisation's problem-solving methods can be

[27] Usage of map services such as Google Maps, Yahoo! Maps or Open Street Maps are not included in this number.

transformed into formal methods using a computer; then the way in which a problem is solved often affects the result. A news application produced by journalists will potentially differ from an information system produced by bureaucrats, external consultants or activists. The journalist's method of problem solving is one component of the product.

Software consists of methodological steps that make it possible to work on digital data. This code is published in news applications as a central part of the product. This leads to the questions: What methods are used? What kind of work is the programmer enabling us to do?

After examining collected material, it is first and foremost evident that variety is stronger than a clear pattern. Secondly, it is clear that the work is already done. Users are not presented with a table – cf. the first edition of the Guardian – and then a set of functions to apply to this table. Nor are they presented with applications to which they can upload their own data, as with a traditional computer program. Users are presented with a tool already populated with data, and already resembling an analysis. Graphs and figures, tables for comparison, and timelines are all provided. Users are then able to manipulate with these prefabricated graphical representations of the data. In this regard the main function of the application is not to perform tasks, but to disseminate information. The applications allow users to see what the journalist saw, understand what the journalist understands, and simulate what the journalist did, to understand the data presented. The audience is not exposed to dismissed hypotheses, nor to areas in which the results are unclear and give little insight. Failures are weeded out. It is not the raw data that is published. Preliminary work, is therefore at least as important as the limited functionality of the published application.

Some aspects of the preliminary work are visible in the results, however. Comparisons and rankings, and steps between levels of aggregations are the main functions used to understand and disseminate the data; in particular, the approach of breaking down the data into geographical areas is frequently used. There are several possible reasons why this method of understanding datasets is used so frequently. The

difficulties with the selection of material favours such applications because they are relatively easy to spot. Furthermore, the results of sub-sets of this kind of data for different countries, cities, municipalities and the like are another potential reason. Pride is often geographically bound, and to be able to compare one's own town with other towns is often of interest beyond the factual subject matter of comparison. A third factor is the universal interest of location bonded data, as all people have relationships to places (e.g. place of birth, home, holidays) and aggregation over a national/regional/local area covers everyone, to a greater or lesser extent, given some interest in the subject matter, or by pure patriotism.

Aggregation in time is also common. Timelines, as known from the paper editions, are easy to understand and provide chronology describing a course of events. News applications can use the same principle and may in these cases be seen as a modernisation of timelines, typically adding some features for interaction. To keep track of different levels of aggregation (in time, geography, or other unit types), applications often provide an automated summary of the current levels based on relevant variables, and these may be summarised using colour coding or with comparisons. For example, see how the variables in *Murder: New York City [4]* change by adjusting the timeline or the left-hand menu.

Among the less frequently used methods are network analysis, linguistic (quantitative) analyses and demographic analyses. It is common for one application to contain several different methods of manipulating and visualising the data, and to offer simple interactions such as filtering, ranking and navigation.

Graphical presentation

The principle for all graphic technology is imprinting symbols on a two-dimensional surface. However, HTML is not bound to this. That does not mean that HTML does not have a graphical presentation; rather, that it is not restricted to just one. A website is a graphical and aesthetic piece of work, but it does not have to be a static two-dimensional surface. It is changeable and can potentially be altered with interactions

based on prefabricated code. Interaction design, as a profession, has supplemented graphical design with knowledge of the design of interchangeable surfaces. When these surfaces are utilised to present data, multiple fields overlap.

Journalism often presents data using different forms of illustrations, which display information in a way that is more sophisticated than in pure data tables. The press often utilises information graphics, which offer a graphical representation of information, data or knowledge conveyed quickly and clearly. Signs and pictograms belong to this tradition. In the intersection with computer-enabled visualisations is the research field and practice of information visualisation:

“The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system” (Hearst, 2003). Where information graphics have understanding and simplicity as their goal, information visualisation has precision and facilitation of analysis.

The news application, as a product of software code, can find itself torn between the clarity of the information graphics on the one hand, and the precision of information visualisation on the other. The simplifying elements in information graphics can come in conflict with information visualisation’s demand for “graphical integrity” (Tufte, 2001) and can be said to be an artistic aid to understanding. The danger with pure information visualisation is that the result requires media literacy that quickly surpasses the typical demands of the general newspaper reading audience. News media tend to avoid using graphs and illustrations beyond a certain level of complexity (Tufte, 2001). To present raw material for analysis in a newspaper can be judged as half-finished work, and information visualisations are among the methods traditionally used for expert analysis. Programming as a profession is indeed closer to producing information visualisations than information graphics. So we can ask: What is presented in news applications? The answer: a potpourri of computer-generated visualisations of various complexities. Common graphs such as bar graphs, line graphs and pie charts are frequently used, but rarely alone. Interactive maps, heat

maps, timelines, scatter plots and lists, data tables and tools to interact with the data such as custom menus and more traditional web elements such as forms, buttons and links are used in varying combinations.

In spite of the fact that the graphs are computer-generated, the visualisations presented in the sample of applications focus on quick and clear dissemination, and very few examples contain the complexity often found in the field of information visualisation. A journalist is much more skilled in communication than in advanced analysis. What is communicated is via analysis, but in a simple form that does not demand expert skills from the journalist or the audience. Usability, understanding and simple dissemination are prioritised over complex visualisations. Consequently, this can be described as "casual information visualisations" (Pousman, Stasko & Mateas, 2007). News applications present data and have much in common with the information visualisation field, but do not necessarily fulfil the formal requirements of the field. Another characteristic is the designing of elements for narrative visualisations, such as the martini glass structure, interactive slideshows and drill-down stories (Segel & Heer, 2010). These are methods of compiling stories through visualisations, and as such offer parallels to journalistic styles such as the inverted pyramid.

Are news applications journalism?

News applications represent a new direction for journalism, but: What are they orienting towards? Are they an attempt to move towards subjects traditionally beyond the newsroom's scope? Do they deal with 'hard' or 'soft' journalistic matter? Are they a form of entertainment?

In order to answer these questions, the applications were categorised according to online news categories (Sjøvaag, Moe & Stavelin, 2012). A low number of units indicates that the selection is unsuited to providing definite answers, but the distribution between the categories indicates that the categories are appropriate. The highest-scoring categories are politics, social issues, economy and crime; all subject

areas that are typically central to a newsroom's production, indicating consistency with traditional editorial preferences and choices.

Economy	11
Crime	9
Social issues	23
Politics	25
Accidents	5
Culture and entertainment	3
Sports	1
Weather	1
Technology and science	0
Other	1

Table 1: Applications according to category, n=79

It is clear that the material does not represent any shift towards entertainment, gaming or other areas of focus, but is related to current media categories. One core requirement for news media is the social contract. Do these applications try to fulfill this contract?

The information function, the watchdog function and the arena function constitute the core of the news application's social contract (Østbye, 2009). These functions define a field of work that separates journalism from the remainder of the media, and raises the expectations for (good) journalism. The journalistic stories should sufficiently overlap with these functions; news applications are no exception.

There is no obvious method for checking journalistic accuracy within these categories. The categories are fuzzy and not mutually exclusive. To shed some light on news applications in relation to their social contract, the applications used in this study were placed in the triangle between the three core functions. The result was as follows:

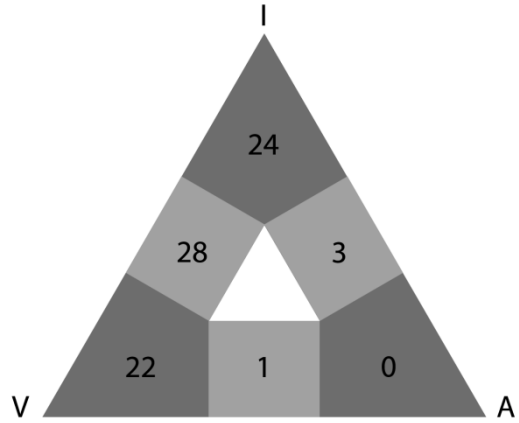


Figure 1: Each corner represents each of the functions; V – watchdog, I – information, A – arena. The areas between the corners represent combinations of the closest functions. One application did not fit this format.

Documenting and monitoring of powerful institutions and individuals are located in the corner for watchdogging. *Maktbasen [the power base]* [3], *Buskerudbenken [The Parliamentary bench row from Buskerud county]* [5] and *How much does it cost to be a regent?* [6] are examples of this. These are applications that monitor or track data which can be traced directly to responsible persons or organisations with power. The applications point *in* towards central power.

The corner for information (on important events and happenings) denotes applications, which aim to make some kind of statement on society through datasets and visualisations. Examples are unemployment (*The jobless rate for people like you* [7]), what constitutes national welfare (*The world's best countries* [8]), and more local issues e.g. where you can find the closest recycle station in the city (*Her kaster du bosset [Discard trash here]* [9]). These are applications that point *out* towards society at large and seek to explain and inform. Within the material collected are examples that underline an area outside this triangle too; applications showing which countries voted for each other in the Eurovision song contest belong in the culture and entertainment category, and it is difficult to argue that they provide important information for citizens making everyday decisions. They do not fulfil the criteria for hard news of resulting in material impact on people's lives (Franklin et al., 2005). These applications can therefore be said to lie outside the triangle. Nevertheless, they

show some gravitation towards the information function; knowledge of solidarity between neighbouring and sister nations in multinational contests is not categorically unimportant.

The third corner of the triangle, the arena function, is empty in this analysis. To create and support the exchange of words is, strangely, an area that fits the news application format to a much lesser extent. Both discussion forums, comment fields and the possibility of contacting the newsroom are normal functions, which online newspapers provide and utilise. Many of the applications also have comment fields, as with many other online news stories, but none of the applications in this collection are examples of a newsroom publishing an application for discussion as a news story. Some tendencies in this direction do, however, exist: *Minnesota slowdown* [10] lets the audience annotate graphs so that the community's observations can be commented upon. In the information visualisation community this type of collaborative analysis is an ongoing area of research; see, for instance, Willett et al. (2011). *Minnesota slowdown* was indeed a collaboration between the Minnesota Public Radio and the visualisation lab at UC Berkeley.

Extending the definition of a news application to include a comment field would give this corner of the triangle some content. Comment fields are the work of programmers, as are discussion boards, chat and the inclusion of comments from Twitter and Facebook. But they stretch the definition so that general computer support is included, and this study's focus is where software code is used in concrete ways to disseminate news. Therefore, the corner remains empty.

Even when using broad categories, some categories emerge which lie between them. Similarly, even when the three main corners are accounted for, the larger part of the material remains uncategorised. But firstly there are two small categories:

Between the arena and watchdog functions is a singleton (a category/set with only one element) entitled *Investigate your MP's expenses* [11], a crowdsourcing project where the audience contributed by sifting through thousands of scanned documents which constituted a central part of the evidence in Britain's 2009 scandal about MPs'

expenses. With a simple user interface readers can find documents related to representatives of their own constituencies (or other more prominent politicians) and click to mark the document in one of four content categories (claim, proof, blank, other) and one of four interest categories ("not interesting", "interesting", "interesting but known", "investigate this!"). In this way the Guardian channelled its resources in the direction most likely to yield good material, and the audience had the opportunity to influence the process as active participants. In a silent collaboration orchestrated by the newspaper, the coalition of Guardian/audience held their elected representatives to account. A point worth mentioning concerning this crowdsourcing project, is the fact that the audience is used as "processing power", not as a data source. It is more common that people contribute with information/facts and judgment to such projects; as is the case in the next category.

Between the arena and information functions is a small category defining other crowdsourcing projects. In 2009, VG produced *Vaksineguide.no [the vaccine guide]* [12], a portal providing information about where and how the mass vaccination program for the swine flu was organised in Norway's municipalities. The public health service was struggling with the task of informing the people, but VG let their readers collect information from all 429 municipalities and sent it to a national portal. The work was done wiki-style, so that everyone could edit and update the information. A simpler example from this category is the Bergens Tidene *Sett fingeren på trafikkproblemene [point out the traffic problems]* [13], a plain map application on which people can indicate with a marker and an explanation, what they perceive to be traffic flow issues. Common in this category is that the readers collate and share socially relevant information organised by the newsroom.

The last and biggest category that emerged when organising the applications according to their function in the social contract was found between the information and watchdog functions. Typical examples are applications handling data concerning people in positions of power, but that also provide information to the audience and lay down a premise for or against those in power. Multiple newsrooms have applications, which deal with war. One example is *Faces of the fallen* [14], which

communicates the toll on "our side" of wars with year-book style photographs, simple graphs and diagrams. Crime also falls into this category; e.g. *Mapping UK's Teen Murder Toll* [15] and *Tödesopfer rechter Gewalt 1990-2010 [Fatalities of right-wing violence 1990-2010]* [16]. Among these are crime maps. These are visualisations built on maps that show where and what kinds of crimes have been registered on police records. *Murder: New York City* [4] serves as an example. Other apps in this category handle traffic data (*Døden på veiene [Death on the roads]* [17], *Crash: Death on Britain's Roads* [18]), schools (*Fixing DC's Schools* [19]), and the environment (*Her kan oppdrettsanleggene bruke gift [here fish farmer are allowed to use poison]* [20]*, De hemmeligholdte GSM-basene [the secret GSM-base stations]* [21]).

The fact that this category is large can possibly be explained by the type of data it processes. Crime, wars, the environment and education are large topics strongly anchored in politics. These are topics that rely on results and factual information in order to influence the underlying policies. The journalistic task is not to expose the secret, but to disseminate and clarify known facts. In this way the newsroom puts itself in the position of supplier of material for debate, presenter of evidence, and potential critic of those in power. While the collated material indicates that the arena function is prioritised lower, information on important events and the monitoring of powerful figures – and the middle way between these – is in large supply. The functional area of news applications overlaps well with the idea of the social contract.

Technology, competency and coalitions

News applications are utilised for tasks that merge with the core function of journalism. They constitute small stand-alone information systems published via online newspapers (and, potentially, via other digital platforms). They treat data related to the newsroom's agenda or which is of general interest or use.

The technology used suggests that non-traditional personnel are at work in the newsrooms: news applications are web applications, and web applications are

software, which runs in a browser. They are differentiated from traditional websites by their capacity to process tasks on the web, in contrast to the traditional method of using locally installed software for each operating client. They disseminate news stories which the newsrooms would traditionally cover using articles or broadcast.

The technologies utilised are interpreted by a browser, and this guides the technology choices. Online newspapers sites are often high traffic sites, which means it is sensible to limit processing on the server side, cache anything that can be cached or simply publish content as plain text files. Consequently, the applications are simple in functionality and typically do all the processing in the browser.

JavaScript and flash are the technologies most often used, and both run on the client side. JavaScript is a scripting language that is implemented in all modern browsers, and adds the capability to allow richer interaction with web documents. Flash is a multimedia platform from Adobe. Flash files produced on this platform are published online (as binary .sfw-files) and are typically embedded in HTML documents. Flash supports animated vector graphics and audio/video and has its own programming language (ActionScript) that enables rich interaction. Both ActionScript and JavaScript can be extended with a third party code through (web) APIs. Common examples of this are integration with Google Maps, Twitter, Flickr or Youtube. It is hard to say, from studying websites, which technologies are chosen on the server side, but it is clear that there is a need for programming in the newsrooms[28].

Few journalists have this competency. Programming is not included in the curriculum in Norwegian journalism schools. It is more likely that this competency will be found in the newsrooms' IT-departments than in the main newsroom. Still, these applications are being produced, and they deal with subjects which are core journalism. Someone is collaborating. Two of the applications examined in this study are signed by an external consultancy; in other words, hired competency. The rest

[28] Languages such as Java, Ruby, Python and PHP or frameworks such as .NET are examples of server-side technologies used in the collected material.

have no explicit markings of having been produced outside the newsroom; indeed, many show clear signs that they were created internally.

The work and future of the hybrid journalist

This study has offered an introduction to a field which marks the intersection between journalism and programming. The applications examined show that strong thematic links to the traditional functions of journalism exist within data journalism, as well as a reflection of the social contract. These "hybrid journalists'" methodologies fall inside the newsroom's traditional toolbox, albeit with one exception: computer programming. Consequently, quantitative methods and casual information visualisations are frequently used. It is not unreasonable to state that the methodological spectrum is particularly utilised in the field of journalism today.

Furthermore, very little is known: who are these programming journalists? How do they adapt to the newsroom context? Where do they come from? Are they hired consultants? How much of a strain on resources is this type of journalism – to the newsrooms and to the programmers? Where does the competency for casual information visualisations come from? It is known that Norwegian editors regard data journalism positively, but they also view the competency for this type of journalism as lacking in their own newsrooms[29]. What do the editors think of results in the form of news applications? This is critical because without the editors' say-so, nothing happens.

And what does the audience think? Is this entertaining journalism? Do they understand the program specific functions offered in these applications? To user-test news applications from this kind of perspective would provide researchers and newsrooms with more knowledge of what is considered important to news applications. On the production side there are also many unknown factors; for

[29] See comment to the 2011 editors' survey at http://voxpublica.no/2011/04/redakt%C3%B8rer-hyller-datajournalistikk-men-satser-ikke-selv/

instance, are we lacking some solid heuristics for good news applications to ensure journalistic integrity and quality?

Applications mentioned

Below is a list of the applications used as examples in this study, with urls and a short explanation of each application. For a better understanding of each application, please visit the URL.

1. Følg piratjakten her [Follow the pirate hunt here], tv2.no, http://www.tv2.no/nyheter/spesial/piratjakten/, [visited 13.10.2011] A web portal for information concerning Norways' military efforts in the Gulf of Aden in 2009.
2. Hvorfor forsvant Jarle? [Why did Jarle disappear?], vg.no, http://www.vg.no/nyheter/innenriks/hvor-er-jarle-ramberg/, [visited 02.12.2010] Interactive feature that consolidates information concerning a disappearance case.
3. Maktbasen [The power base], nrk.no, http://www.nrk.no/maktbasen/, [visited 13.10.2011]
   Database over Norwegian politicians with relations to duties and commitments to Norwegian businesses and associations.
4. Murder: New York City, nytimes.com, http://projects.nytimes.com/crime/homicides/map, [visited 13.10.2011]
   Crime map that shows locations for murders, with demographic variables for victim and perpetrator.
5. Buskerudbenken [the Parliamentary bench row of Buskerud county], bt.no, http://buskerudbenken.dt.no/, [visited: 02.12.2010]
   Portal that tracks activities from the members of Parliament from Buskerud county.
6. How much does it cost to be a regent?, texastribune.org, http://www.texastribune.org/library/data/rick-perry-donors-appointed-regents/, [visited: 13.10.2011]
   Gives an overview over Texan teaching institutions board members at monetary contributions to the governors political party.
7. The jobless rate for people like you, nytimes.com, http://www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html, [visited 13.10.2011]

Stacked line graph that shows unemployment rates for different demographic groups in the USA.

8. The world's best countries, thedailybeast.com (Newsweek), http://www.thedailybeast.com/newsweek/2010/08/15/interactive-infographic-of-the-worlds-best-countries.html, [visited 13.10.2011]
   A ordering of the worlds countries with features to compare and examine what parameters that counts in favor or against the countries positions.
9. Her kaster du bosset [Discard trash here], bt.no, http://www.bt.no/bolig/Her-kaster-du-bosset-1769186.html, [visited 28.01.2011]
   Interactive map over locations of recycling stations for plastic and special/toxic waste.
10. Minnesota slowdown, minnesota.publicradio.org, http://minnesota.publicradio.org/projects/2008/07/16_minnesota_slowdown/, [visited 28.01.2011]
    Visualization of employment rates from different trades in Minnesota. Contains possibilities to annotate parts of the graphs with comments and discuss these annotations.
11. Investigate your MP's expenses, guardian.co.uk, http://mps-expenses.guardian.co.uk/, [visited: 13.10.2011]
    Application to spread the workload of investigating documentation for Members of Parliaments' spending in Great Britain.
12. Vaksineguiden.no, vg.no, http://www.vg.no/spesial/svineinfluensa/, [visited 13.10.2011]
    National portal for information concerning the mass vaccination for swine flu in 2009/2010.
13. Sett fingeren på trafikkproblemene [Identify the traffic problems], bt.no, http://www.bt.no/nyheter/lokalt/Sett-fingeren-p-trafikkproblemene-1913327.html, [visited: 28.01.2010]
    Interactive map where readers can mark and comment on locations they find troublesome or dangerous in relation to traffic.
14. Faces of the fallen, washingtonpost.com, http://apps.washingtonpost.com/national/fallen/, [visited 13.10.2011]
    Applications that tracks deaths of American soldiers in the Iraq and Afghanistan

wars. The content in organized by demographic parameters and every solider has a unique URL with individual information.

15. Mapping the UK's teen murder toll, bbc.co.uk, http://news.bbc.co.uk/2/hi/uk_news/7777963.stm, [visited 13.10.2011]
    Shows data on murders of teenagers in the UK. Includes map, graphs and links to relevant news stories for each of the cases.
16. Todesopfer rechter Gewalt 1990-2010 [Fatalities of right-wing violence], zeit.de, http://www.zeit.de/gesellschaft/zeitgeschehen/todesopfer-rechter-gewalt, [visited 13.10.2011]
    Shows murder cases in Germany where right-wing extremists are convicted or suspected as perpetrator. The application lest users compare parameters such as motive, weapon and demographic parameters to look for patterns.
17. Døden på veiene, de unge trafikkofrene [Death on the roads, the young traffic victims], bt.no, http://www.bt.no/nyheter/lokalt/dodenpaaveiene/ungdommene/, [visited: 08.07.2011]
    Shows traffic victims in Hordaland and Sogn og Fjordane from 2000-2010. Portrait photos in year book style. Clickable parameters to make selections. Links to relevant news stories from each victim.
18. Crash: Death on Britain's Roads, bbc.co.uk, http://news.bbc.co.uk/2/hi/8401344.stm, [visited 13.10.2011]
    Shows 10 years for traffic accidents in Great Britain, visualized in a map, as graphs and as a time line. Uses a separate tab to break down demographic parameters, and variables such as time of day, road conditions, etc. The latter part includes commentary from experts, professors in traffic safety.
19. Fixing DC's Schools, washingtonpost.com, http://www.washingtonpost.com/wp-srv/metro/interactives/dcschools/scorecard.html, [visited 13.10.2011]
    Data on schools visualized on a map that can be interacted with to find schools to do particularly well or poorly in different areas. Contains detail pages for each school.
20. Her kan oppdrettsanleggene bruke gift [Here the fish farmers are allowed to use poison], tv2.no, http://www.tv2.no/nyheter/innenriks/her-kan-oppdrettsanleggene-bruke-gift-3195698.html, [visited 13.10.2011]
    Shows fish farms that have obtained exemptions from the general law against usage of toxins, in use against salmon louse (Lepeophtheirus salmonis). This is a simple

Google Maps based solution, but uses clustering to aggregate at different zoom levels.

21. De hemmeligholdte GSM-basene [The secret GSM-base stations], nrk.no, http://www.nrk.no/programmer/tv/brennpunkt/multimedia/1.6696825, [visited 13.10.2011]

    Overview over 9708 GSM base stations in Norway, displayed on a map. Lets users choose a point on the map to be presented with base stations according to distance to this point. The unclear/disputed health hazards these installations poses was made into a big news story. The data this app is based on is obtained through a court case initiated by the NRK journalist team.

22. Comprehensive spending review: you make the cuts, guardian.co.uk, http://www.guardian.co.uk/politics/interactive/2010/oct/19/comprehensive-spending-review-cuts, [visited 13.10.2011]

    Tree graph of Great Britain's state budget, where users can cut in the different sections. Minister of finance (George Osborne) budget cut suggestion from 2010 serves as a comparative element to users own cuts.

Illustrations of selected news application

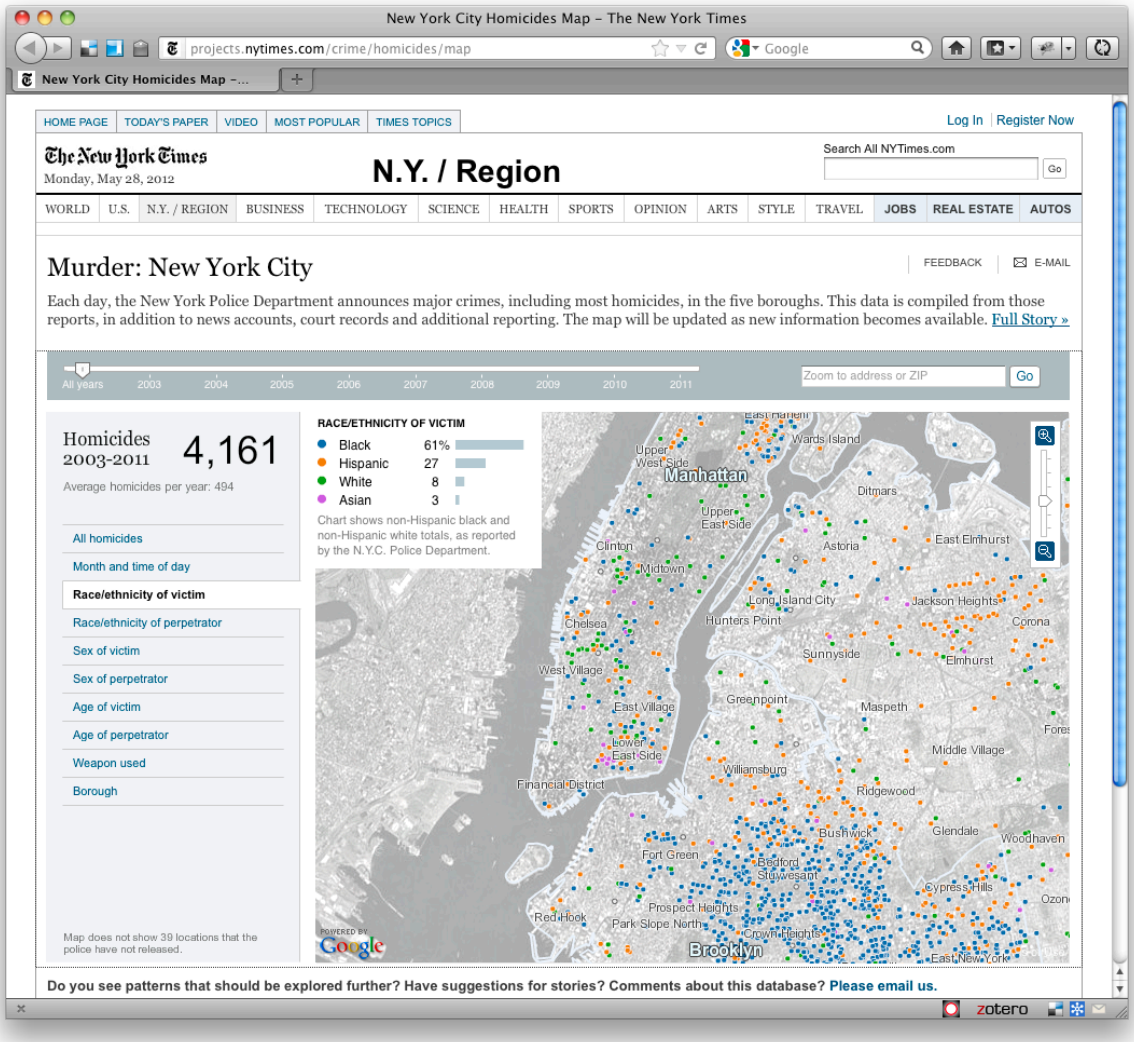


Figure 4: New York Times, Murder: New York City. Crime map with time line. [5]

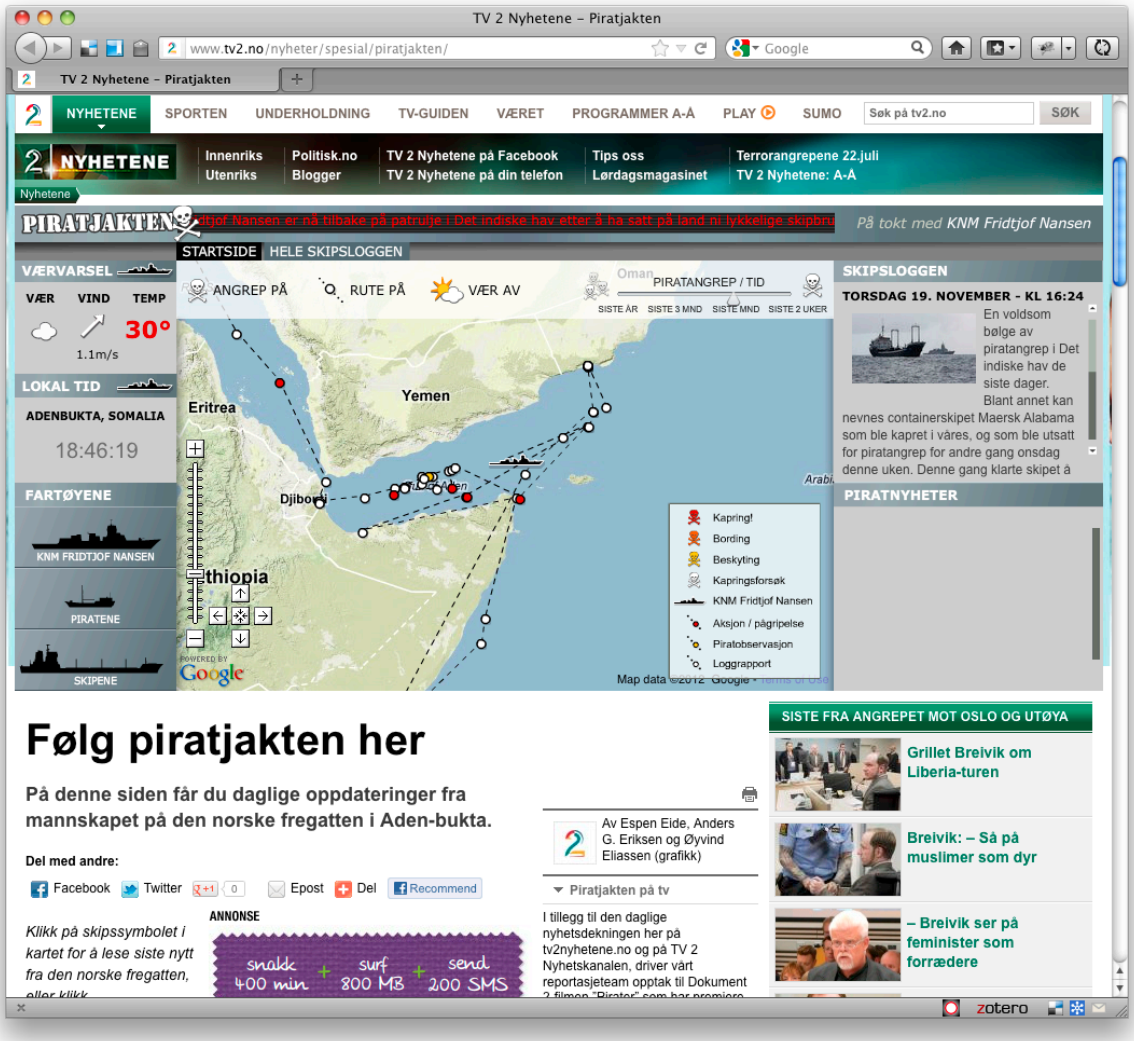


Figure 5: TV2 Følg piratjakten her. Portal for data on Norwegian sea craft safety operations on the Gulf of Aden. [1]

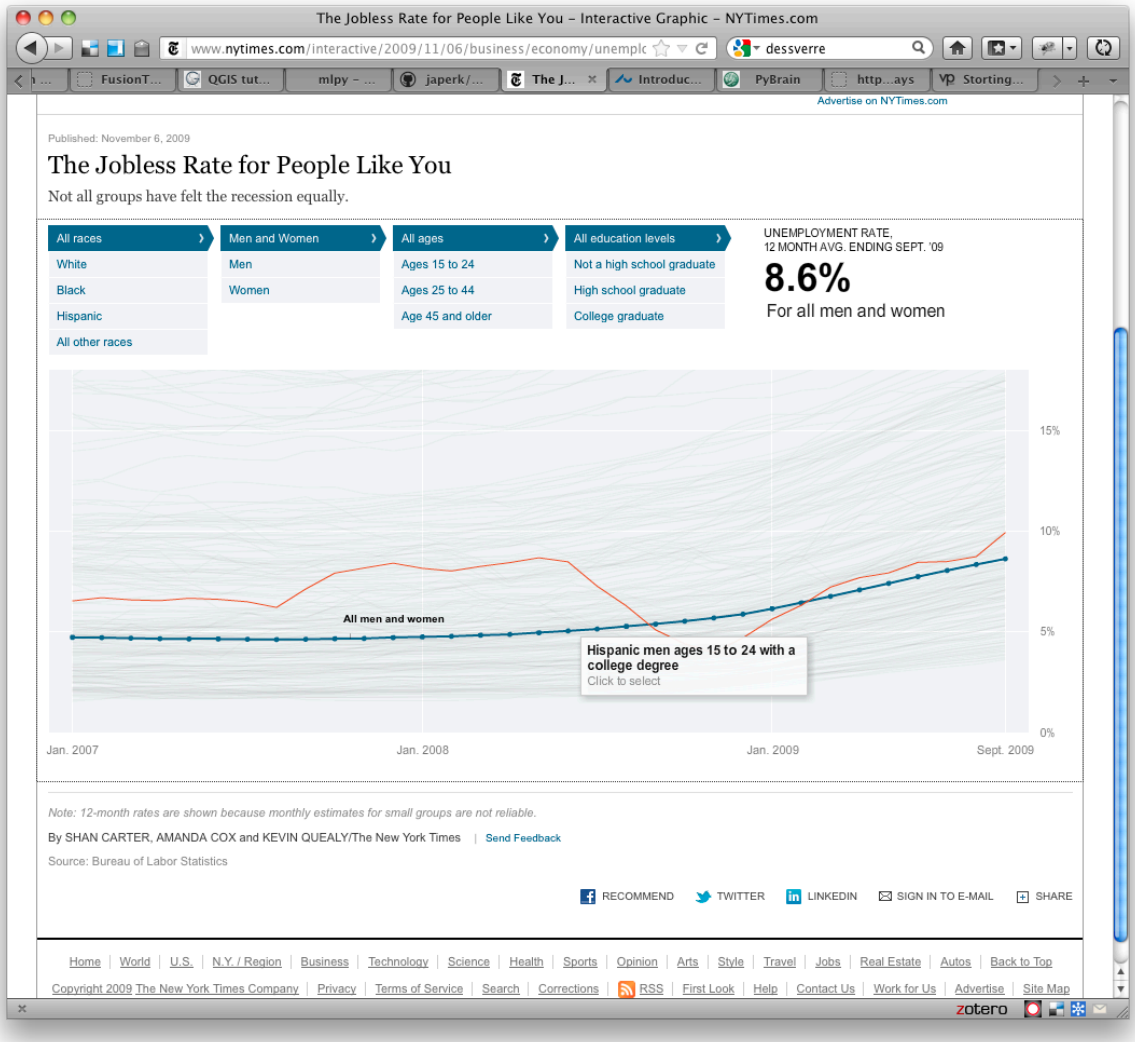


Figure 6: New York Times. The jobless rate for people like you. Interactive visualisation of the unemployment rates according to demographic variables. [8]

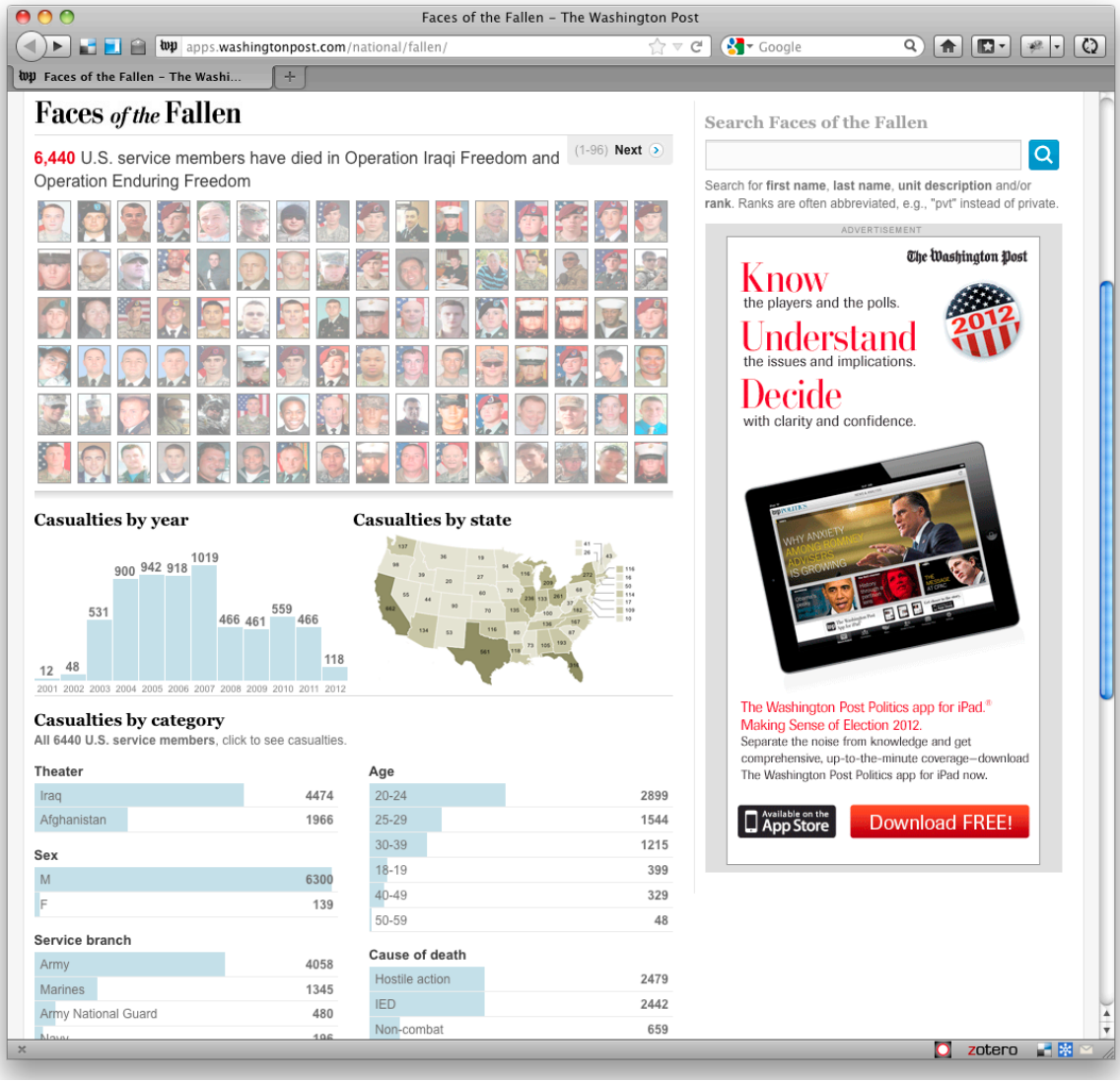
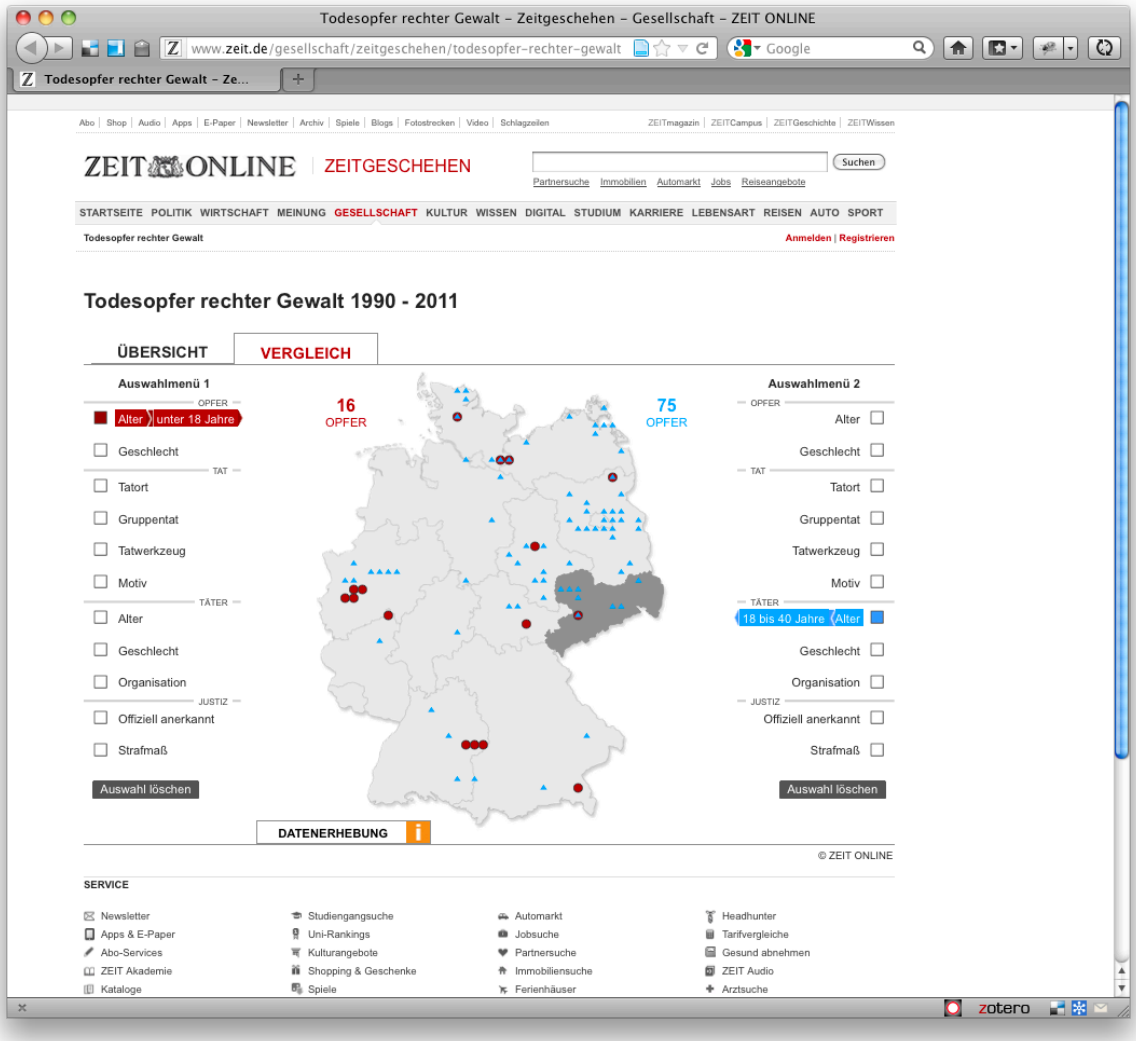


Figure 7: Washington Post. Faces of the fallen. Portal for victims of war in the USA led operations in Iraq and Afghanistan. [15]



References

Boczkowski, Pablo, 2005. *Digitizing the News: Innovation in Online Newspapers*. The MIT Press.

Bogost, Ian, Simon Ferrari, and Bobby Schweizer, 2010. *Newsgames: Journalism at Play*. The MIT Press.

Bowker, Geoffrey C., and Susan Leigh Star, 2000. *Sorting Things Out*. MIT Press.

Cox, Melisma, 2000. "The Development of Computer-assisted Reporting." *Informe Presentado En Association for Education in Jornalism End Mass Comunication). Chapel Hill, EEUU: Universidad de Carolina Del Norte*.

Engebretsen, Martin, 1999. "Nyheter På Nettet: Fortellinger Eller Informasjonsnettverk?" *Norsk Medietidsskrift* 1999 (2). Institusjonar Og Økonomi. http://www.medieforskerlaget.no/archives/1186.

Franklin, Professor Bob, Martin Hamer, Mr Mark Hanna, Marie Kinsey, and Dr John E Richardson, 2005. *Key Concepts in Journalism Studies*. Sage Publications Ltd.

Larsson, Anders, 2012. “Interactivity on Swedish Newspaper Web Sites – What Kind, How Much and Why?” *Convergence: The International Journal of Research into New Media Technologies*. http://uppsala.academia.edu/AndersOlofLarsson/Papers/1007934/Interactivity_on_Swedish_newspaper_web_sites_-_What_kind_how_much_and_why.

Østbye, Helge, 2009. *Journalistiske nyorienteringer*. Edited by Martin Eide. Scandinavian Academic Press.

Pousman, Z., J. Stasko, and M. Mateas, 2007. “Casual Information Visualization: Depictions of Data in Everyday Life.” *IEEE Transactions on Visualization and Computer Graphics*: 1145–1152.

Segel, E., and J. Heer, 2010. “Narrative Visualization: Telling Stories with Data.” *Visualization and Computer Graphics, IEEE Transactions On* 16 (6): 1139–1148.

Sjøvaag, Helle, Hallvard Moe, and Eirik Stavelin, 2012. “Public Service News on the Web.” *Journalism Studies* 13 (1): 90–106. doi:10.1080/1461670X.2011.578940.

Tufte, Edward R., 2001. *The Visual Display of Quantitative Information*. Graphics Press.

Willett, W., J. Heer, J. Hellerstein, and M. Agrawala, 2011. “CommentSpace: Structured Support for Collaborative Visual Analysis.” In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, 3131–3140.

II. Computational journalism in Norwegian newsrooms

Joakim Karlsen, Faculty of Computer Sciences, Østfold University College, Norway. E-mail: joakim.karlsen@hiof.no

Eirik Stavelin, Department of Information Science and Media Studies, The University of Bergen, Norway. E-mail: Eirik.Stavelin@infomedia.uib.no

This article examines computational journalism as a craft practised in Norwegian newsrooms.Based on in-depth interviews with expert practitioners in six of the largest newsrooms in Norway, we find that computational journalism represents a continuation of traditional (investigative) journalism. While the skills and tools necessary to do this kind of journalism diverge from the typical journalist's, the values and aims align well with tradition. Even though computation enables journalists to cope with the size and scale of journalistically appealing datasets, we find little evidence for computational journalism to increase the efficiency of doing journalism or in any other way rid journalists from low-level technical work.

KEYWORDS communication technology; computer-assisted reporting; computational journalism; journalism; online journalism; rhetorical craft

Introduction

Computational journalism is a contemporary term for journalistic work done in the intersection between journalism and computing. This type of work seems to be more common now, and according to Terry Flew et al. (2012) this has to do with the increasing availability of data and tools and the dynamics often associated with Web 2.0. In this article we are not claiming that computational journalism is a growing trend or that this practice is going to improve or save journalism. The aim of this study is to identify, describe and analyse the work practices of "computational" journalists in Norwegian newsrooms. We are going to find out who they are, what they think is a good computational news story, why and how they create news stories, and how they see their place in their respective newsrooms. To frame and analyse

these findings we are going to develop an understanding of computational journalism as a rhetorical craft, using the Aristotelian concept of techné, building on theory from design, writing and science studies (Buchanan 2001; Johnson 2010; Wickman 2012). This paper offers some insights necessary to avoid the pursuit of research on this phenomenon propelled only by hopes, promises and hype.

In the following we will review the literature on computational journalism before framing computational journalism as a rhetorical craft. After a presentation of the findings from interviewing practitioners of this craft in Norway, we will analyse and discuss these results to reach a better understanding of computational journalism as a performative and productive activity in Norwegian newsrooms.

Journalism and Computing

According to Melissma Cox (2000), computers have been used by journalists to produce news stories since 1952. This practice has often been labelled "computer-assisted reporting" (CAR), and has been supported by an active international community.

The organization for Investigative Reporters and Editors (IRE) holds an annual conference on CAR, and the community offers a wealth of books on the topic in bookstores. From the early 1970s, Philip Meyers' book, Precision Journalism: A Reporter's Introduction to Social Science Methods (2002), included advice on how to use computers in reporting. The book describes a work process that is very similar to what other textbooks on journalism recommend (collect, store, retrieve, analyse, reduce and communicate), and offers advice on how computers and methods from the social sciences can support each activity. Depending on what part of the work process has been emphasized, other labels have been proposed for similar practices, such as data journalism (Gray, Bounegru, and Chambers 2012), data-driven journalism (Lorenz 2010) and database journalism (Loosen 2002).

As noted by Christopher Anderson (2011), the empirical literature on computational journalism is sparse. We have found only one study that investigates how computational journalism is actually performed in the context of the newsroom.

Cindy Royal (2012) has undertaken an ethnographic study of the work practices of the New York Times (NYT) Interactive News Technology Department. Her findings are relevant for this study. The first is that the people working in the department see themselves as journalists first. They value technology insofar as it supports their journalistic work. The second finding is that the NYT group understands the process of producing news applications as lightweight, fast and flexible. The third finding is that the group identifies with the hacker culture emphasizing creativity, innovation and collaboration. The fourth finding is that the important skills needed besides journalism are general problem solving and the mastering of Web technology.

Matthew Powers (2012) identifies three ways of understanding new technologies in news production, as exemplars of continuation, as threats to be subordinated and as the basis for journalistic reinvention. The first perspective is common and researchers find that traditional values in journalism are resilient (Weinhold 2010), and that technophobia is rare amongst journalists who easily see new tools as useful if they support existing practices (O'Sullivan and Heinonen 2008). The second perspective is observed by researchers who focus on conflicts and friction occurring when technology and new forms of work are introduced in newsrooms (Mitchelstein and Boczkowski 2009; Robinson 2011). Turf wars are common and journalists are reluctant to accept innovative technologies in their newsrooms (Singer 2004; Boczkowski 2005; Deuze 2005). The practice of data-driven or computational journalism is seen by some as posing core epistemological challenges to journalism (Parasie and Dagiral 2012). A majority of research conducted specifically on "computational journalism" can be placed within Powers (2012) third perspective of "technology as the basis for journalistic reinvention". This research can again be separated into two groups. The first group offers a hypothetical perspective, exploring ideas on what computational journalism might be. The second group offers a practical design science perspective, creating and testing prototypes of tools for doing computational journalism.

Starting with the hypothetical, Flew et al. (2012) argue "that computational journalism techniques may provide new foundations for original investigative

journalism and increase the scope for new forms of interaction with readers". They explain that the utility value of computational journalism comes when it frees journalists from the low-level work of discovering and obtaining facts to allow greater focus on the verification, explanation and communication of news. This utility value was emphasized in a recent study of journalists' reaction to robot journalism as performed by the Statsheet network (van Dalen 2012). Hamilton and Turner (2009) report from a workshop on computational journalism and explain that:

> *computational journalism aims to enable reporters to explore increasingly large amounts of structured and unstructured information as they search for stories. At the same time though, computational journalism offers a new way to help sustain the watchdog reporting on which democratic citizenship so clearly depends.*

Sarah Cohen et al. (2011) turn the question of how computer scientists can participate in shaping computational journalism to the database community and outlines ideas for a system which is "a cloud for the crowd" to support collaborative investigative journalism. Nicholas Diakopoulos (2012) suggests a model for systematic innovation in this field by construction of a framework for combining journalistic and computer science concepts into new ideas for computational journalism.

More practical approaches build on methods from computer science (how to achieve various goals using computation) and theory from journalism (what is important to journalists and journalism), and are typically evaluated with journalists as participants to reveal journalistic utility. Examples of this include a framework for media bias mitigation (Park et al. 2011), tools for finding sources and eyewitnesses (Diakopoulos, De Choudhury, and Naaman 2012) and tools for computer-supported analysis for user-generated content (Diakopoulos et al. 2011).

Computational Journalism as a Rhetorical Craft

Scholars have used an Aristotelian view of craftsmanship to explain productive activities like scientific inquiry (Wickman 2012), writing (Johnson 2010) and design (Buchanan 2001). In line with this we promote an understanding of computational journalism as an activity that depends on a certain kind of techné or a deep understanding of the underlying causality of the practice. In "The Question

Concerning Technology" (Heidegger and Stassen 2003), Martin Heidegger explains Aristotle's fourfold causality of artefacts by using a silver chalice as example. The material cause is what the artefact is made of (silver). The formal cause is the form or shape given to the artefact (a cup). The final cause is what the artefact is used for, in Heidegger's example, a sacrificial rite. The moving cause is the craftsman or silversmith. Heidegger underlines the co-dependence of the four causes and that the craftsman's duty is to bring forth the artefact, in this example the chalice, letting the end determine the choice of material and form. We will continue by explaining computational journalism by its fourfold causality. After this we will relate this concept of techné to the concepts of episteme and rhetoric.

Data are the material cause of computational journalism. In this context data generally refer to quantitative variables, structured in tabular, tree or graph structures. Tax records, school examination results, financial reports, membership lists, election results and stock prices are frequently used examples. Unstructured data, such as documents, images and audio-visual material can also be subject to computation and is used. For an overview of what data have been used in previous journalistic projects, see The Data Journalism Handbook (Gray, Bounegru, and Chambers 2012) or the data journalism awards gallery (Global Editors Network 2012).

The formal cause of computational journalism is most often information visualizations or info graphics. To relax the differences between the subject area of information visualization in computer science and the more pedagogical and artistic info graphics style (both in graphics and subject matter), an alternative term, casual infovis, have been proposed for this kind of graphics (Pousman, Stasko, and Mateas 2007). Narrative theory has been applied to understand online interactive graphics as storytelling (Segel and Heer 2010). Bogost, Ferrari, and Schweizer (2010) give a good overview of the different types of interactive news storytelling being developed today. Susan Jacobson (2012) has analysed multimedia news packages produced by the NYT, emphasizing the use of hypertextual, interactive and social media elements. Steen Steensen (2010) claims that terms like the ones used by Jacobson are too

narrow and technology oriented to be useful when understanding new forms of online journalism.

The final cause of computational journalism, or what function it is meant to have in society, is not that different and can easily be aligned with how traditional journalism has been perceived to play a role in society. Kovach and Rosenstiel's (2007, 4) nine principles to help journalist to fulfil their task to "provide people with the information they need to be free and self-governing" applies to computational journalism too. They report that most journalists think that pursuing the truth is important and that journalism must serve as an independent monitor of power.

The journalist or programmer or the journalist-programmer is the moving cause of computational journalism. She combines journalistic skills and value systems with programming skills to bring forth the finished story based on the data, the form and the purpose of the case she is working on. The skills involved in transforming data into a useful form can be classified according to Bloom's taxonomy of learning objectives (Bloom et al. 1956; Kratwohl, Bloom, and Masia 1964; Harrow 1972). The taxonomy consists of a cognitive, psychomotor and affective domain. We propose that in computational journalism computational thinking (Wing 2008; Hu 2011) or programming in an abstract sense are important cognitive skills together with proficiency in investigative reporting. The mastery of applied programming techniques is important psychomotor skills. The affective skills associated with computational journalism are to value the traditional journalistic principles as revealed or described by Kovach and Rosenstiel (2007).

There are two more aspects of the ancient understanding of techné that are useful when framing computational journalism. The first is the close relation between techné (craftsmanship) and episteme (science). The second is how techné relates to, or can be understood as, rhetoric. The work practice of computational journalists resembles that of researchers, where the commitment to truth is paramount. Philip Meyer (2002, viii) relates science to journalism: "Scientific method is still the one good way invented by humankind to cope with its prejudices, wishful thinking, and perceptual

blinders. And it is definitely needed in journalism". Chad Wickman (2012) conceptualizes scientific inquiry as a productive technical art using the Aristotelian concept of techné and the four causes. He locates rhetoric in the production of artifacts needed by the scientist in the process of inquiry. He writes that: "Knowledge production for these scientists involves an on-going negotiation between instruments, technical procedures, material artefacts, visual repre- sentations, and the physical reality that they construct through their inquiry". When the scientist uses these artefacts to communicate and support his knowledge claims, the technical and the rhetorical merge. Framed this way it is possible to see how technology, knowledge and rhetoric play together in scientific inquiry, and we think this is the case also for "journalistic inquiry" performed by computational journalists.

Research Design - Finding, Selecting and Interviewing the Journalists

To find these journalists we started with Eirik Stavelin's (2012) lists of news applications and the names of journalists found in each of the by-lines. One of several criteria for inclusion in this list was that the application "convey news, where the journalist has written code, himself or in collaboration with a developer" (ibid, 107). There are only a few Norwegian newsrooms that have contributed to the list so we were able to create a shortlist of potential candidates quickly. After this we called the informants to assess whether they fit our criteria: producing journalism by computational means. These conversations led to the inclusion of additional candidates. The reason for this is that people who work with journalistic programming often are omitted from the by-lines.

During the preliminary telephone interviews we realized that having programming skills was too narrow a criterion for inclusion in the study. Journalists belonging to the CAR tradition, without knowledge of programming, can still have good knowledge of traditional information retrieval methods, practical database management and quantitative methods, and can be said to perform computational journalism. We included two journalists within this tradition in the sample. We ended

each pre-interview by asking who else we should talk to, thus discovering relevant sources unknown to us beforehand.

We conducted full interviews with 11 people, but have included only nine in our analysis because two were not working in a newsroom context. In our sample we have representatives for three types of newsrooms; television broadcasting, newspapers (still in print) and online newspapers. Six large (national/regional) news organizations are represented: the two largest news-producing television channels; the two largest daily national tabloids; and the two largest regional daily papers. These are the biggest media organizations in Norway, with the largest reach in the population. All organizations also produce online newspapers, and the data journalists interviewed produced content for all platforms. The organizations have their base in Oslo or Bergen (or both), the two biggest cities in Norway. The lack of local news producers and small town affiliations is likely a product of the snowball sampling methodology and the possibility that very few local newsrooms practise computational journalism on a regular basis.

We chose a qualitative approach with semi-structured interviews. The interview guide consisted of five relatively open questions. The first question was definitional; what is computational journalism to you? We deliberately left this question ambiguous to better capture the interviewees' perspective on the phenomenon. Journalism can refer to both their work practices and the finished news stories. The next question was about good methods and techniques when doing computational journalism. To be certain that we covered the topic we prepared four follow-up questions. How do you get access to data? How do you prepare the data for analysis? How do you analyse the data? How do you tell stories about the data to the readers? The third question focused on the skills needed to do computational journalism, both general and specific. The fourth question was about what kind of support the journalists doing computational journalism need? The follow-up questions focused on composition of teams, collaborative work processes, technical infrastructure and tools. The last question was about how the production, publication and form of the "computational" news stories themselves are different from other types of journalism.

The interviews lasted between 50 and 110 minutes. All the interviews were taped, transcribed and imported into TAMS Analyzer, a tool for computer-assisted qualitative data analysis (Weinstein 2006). Here we followed the suggestions of John Creswell (2009) for steps to take when analysing qualitative data. After transcoding the interviews we read the transcripts to get a general sense of the data. The next step consisted of detailed coding. We used both contextual codes linking answers to questions and a hierarchy of thematic codes. Some of the topics were coded based on the goal of the study and some of the codes emerged by reading the transcripts. The last two steps consisted of selecting and grouping statements based on codes and preparing the summary of the results.

Findings

We have explored computational journalism as a craft, focusing on the work practices of the interviewees. This means that we learn about this phenomenon from the single perspective of the craftsperson. We can relate what they say about the newsrooms they belong to, without being able to verify this by other accounts. We will now summarize how the interviewees perceive themselves and their work practice in the newsrooms before going on to relate our findings from the interviews according to the fourfold causality of computational journalism: material, formative, final and moving causes.

In the Newsroom

When asking about the role played by the computational journalists in the newsrooms, we anticipated stories of conflict. When no conflicts were related, we asked more directly and got answers describing this as a problem of the past.

> *Now we have developers working in the newsroom, but it wasn't like that before. It felt like I was stepping on everybody's toes back then, as you were partly designer, partly developer, partly journalist, and partly operations person, right. I remember that as being very tiresome.*

Now the dependency between journalists and developers seems stronger than the differences in work cultures. One respondent reported that they most often work in

teams to “exploit each other’s strengths”. While some worked in formal groups (multimedia journalism, newsroom IT, etc.), all worked in loosely coupled non-formal cooperation with other journalists and newsroom staff.

All our respondents had positions inside the newsrooms and not in the ICT departments. These departments are often separated, and cross-departmental cooperation is rare. One respondent told us that: “the relationship to IT is increasingly institutionalized and formalized and alienated. It is a sad tendency, but that’s that”. An effort to bypass organizational ICT infrastructure was found in all the newsrooms. Extra software that needed installing, databases that needed to be created, and the set-up of servers to host Web applications, are all examples of tasks that were done without the help of the ICT department. This said, the respondents also reported that in extraordinary cases, e.g. large WikiLeaks dumps, this non-cooperation state could be overcome.

The respondents reported that resources in the form of soft- or hardware are not a problem. The technology needed to undertake computational journalism is relatively cheap and available and for the most part already exists in newsrooms. The limiting factors are not the technical infrastructure but according to one of the interviewees, “time and goodwill” granted from the editors. He continued and said that it is difficult to get the other newsroom staff to understand that, “computational journalism takes time. Visualization takes time. Analysing takes time. Fetching [data] takes time”. Time and goodwill from editors were repeatedly mentioned as key resources when doing computational journalism; “Time is always a limited resource. And of course a boss that trusts you a hundred per cent, because time is ticking”.

The Material Cause: Data

The main findings when it comes to data concern access. In Norway access to public data is regulated by freedom of information legislation that favours transparency and public inspection. Although all interviewees had stories of troublesome bureaucrats complicating matters, there was a clear consensus among them indicating that the access to data is good. One informant told us that: “The available [data sources] are

rather good, and people have been very helpful in those cases I've been out nagging. I have almost been surprised by how well it has gone". One interviewee, a senior reporter with a long track record in computer-assisted reporting revealed: "I've only used a right of access application once".

While data access offers few problems, some issues of the commercialization of public data were recurring in interviews. Some governmental bodies are allowed to charge money for data, and put barriers around the data in ways the current legislation does not (arguably) address. The lack of quality geographical data is one example of this, which was mentioned by several of the interviewees. One of them put it like this: "What annoys me the most is map data in Norway, it is a pain in the ass, that county borders and municipality borders, that I shall not be able to fetch them somewhere in shape-file format".

Getting an overview of what data exist offers a challenge to journalists investigating public records. One interviewee said: "the job of finding out what data actually exist, that is perhaps the largest job, I think". Further, the public servants who serve the data access requests from journalists are often not trained to export and transfer digital records. One of the interviewees put it like this: "there is often a lack of knowledge at the other end. They do not know how to retrieve the data. They have data and they have databases, they have all sorts of stuff, but they have no clue how to export it". A common obstacle is the PDF file format, intended to store documents with a stable visual form across platforms, but which is often used to store tabulated data. This creates an extra headache for a computational journalist, who has to scrape the PDFs to get the data into a table format.

The Formal Cause: Info Graphics and Storytelling

The informants described the ideal of "drillable" interfaces to data where both the overall and the detailed view are represented. This presentation should frame and empower the reader to become an investigator/journalist. The global and the local, the journalist and the user are connected by data aggregated to generalities from singular facts. One journalist said that he wanted to convey the "the unbroken line between the

general and the particular". But at the same time they emphasized the difficulty in finding the right balance between the drillable dataset and storytelling. To make data available on the Web with search and filtering tools generates little interest. One of the informants puts it like this: "when we have a large dataset we often give the audience the whole package with lots of buttons and analytic tools, etc. Then people aren't really interested".
Computational journalists need to filter, select and tell the major trends in the data to attract readers. Several of the informants reported that after experimenting a lot, they have ended up treating datasets more as "internal sources" than artefacts to be published online. One said that it is a "human being, a journalist, an editor, that chooses the facts that should be mapped".

Most of the interviewees were wary of their own technological interest and passion, and seek to resist the temptation to add all the possibilities afforded by Web technology. They try to overcome "the 'see what I can do' phase". One interviewee with a technical background said it is "very easy to become a technician" and he had to remind himself not to "forget that there is a reader in the other end". The most important thing is to convey a clear message to the reader, and most readers are not experienced data wranglers. The end products need to be "'for dummies' to get the message out".

The Moving Cause: Journalism by Computation

The backgrounds of our respondents included both journalism school and IT-related degrees. All but one had higher education of some form. While some had no formal technical education, and had worked their way into more technical tasks, those with journalism degrees mentioned training courses in various technical fields. Of the technically educated interviewees none had studied journalism. When asked what kind of backgrounds would be relevant in a computational journalism team, the respondents mentioned programming, design, typography, info graphics, usability, databases, Web and journalism. They all emphasized that to be a computational journalist you need the double vision of technology and journalism, regardless of how teams are put together or the development of editorial support functions. The

journalist in our sample with the highest formal education in computer science put it like this: “we have had developers here, but when it comes to the journalistic bit it all falls to pieces. We have also had journalists here, but when it comes to the development bit it all falls to pieces. That combination is terribly hard. There are very few that have that competence”.

When it comes to cognitive skills the logic of investigative reporting has primacy. Almost all of the respondents claimed that curiosity is the one trait you cannot do without, clearly formulated by one experienced journalist with an IT background as: “Curiosity. Curiosity. Curiosity. That is...the essence”. This curiosity has to go hand in hand with problem-solving capabilities and a sense of logic. As one journalist put it, “I guess they [computational journalists] have a...pronounced sense of logic..., problem solving is very important”. This curiosity and problem solving is not equivalent to what a programmer or developer does, but is scripted by the long tradition of doing investigative reporting. One of the respondents went so far as to claim that investigative journalism is programming. He says: “Systematically going through material is investigative journalism and investigative journalism is really a manual form of programming”.

When characterizing the hands-on or psychomotorical skills involved in computational journalism work, several of the informants indicated end-user programming as most relevant. They use common applications marketed and sold by technology giants such as Microsoft (Excel, Access) and Google (docs, maps, refine, fusion tables). The spreadsheet is a central tool, also when cooperating with non-technical journalists. One of the informants put it bluntly “the essential skill to do data journalism is one: learn Excel”. According to many of the interviewees L/M/W+AMP (linux/mac/windows, apache, mysql and php) is important. This stack of technologies includes a database, a Web server and a scripting language for presentation. Statistical methods and packages are rarely used, but some informants reported using simple methods for cluster, network and regression analysis. The informants were explicitly cautious about using statistical methods - considering this outside their field of competence. Overall they had a pragmatic relationship to technology and

emphasized that it is important to be "able to easily obtain new skills" and "choose the right tool for the job".

The Final Cause: Accountability

The respondents' view on what affective skills a computational journalist should have are in line with the journalistic tradition. One of the interviews said that a computational journalist should "feel committed to the social contract of the press". All in all most of the respondents were committed to hard news and fulfilling the accountability function of journalism.

When asked to define computational journalism the respondents did not offer narrow or distinct definitions. They all gave wide and open definitions that included many different activities and forms. One of the respondents said that he "uses computational journalism for everything from research and fetching data to visualizations" and that computational journalism is to "find new ways to both find stories, and to tell stories". Another respondent offered a similar wide definition, "computational journalism is everything from the simplest use of Excel to heavy tools that enable journalism that is impossible without these tools". They emphasized that computational journalism empowers them to do more. Analysis of data is central to this. They were able to "analyse large datasets and find support for existing theories, or find new truths, new trends".

Discussion

We will now interpret and discuss our findings according to the fourfold causality of computational journalism as craft and its newsroom context. To begin with data, as the material to obtain analyse and convey, data are at the core of computational journalism. Access to data is perceived as good by the Norwegian journalists interviewed in this study. This is crucial to them, but must represent a risk to the data owners. The transmission of, or denied access to data is an important event in the constant negotiation for control and power between these parties. Several strategies can be effective when you own data that you do not want to share with the public.

One is to “flood” the journalists with data, and hope that they will never be able to analyse it. Another strategy is to make a bet that the journalists have little time on their hands and package the data strategically. The journalist’s main tool is a computer, and available sources in the right format will be preferred. Non-controversial data can be made available as well-structured Excel sheets and controversial data can be hidden or given out in rasterized PDFs. This said, our results indicate that this is not a big problem with the journalists in our study. Several of them report that they get funding to do “deeper” investigative reporting regularly, resulting in “fresh” stories based on previously unused data.

When it comes to info graphics and storytelling, our results show that most of the journalists interviewed for this study have ended up preferring traditional linear storytelling, emphasizing computation as useful for research rather than presentation. Consequently, the journalists focus less on giving the readers access to “raw” data. They choose design elements that support linear narrative rather than free exploration. Examples of this are the use of timelines, maps, writing, sound and video. These forms are simple, relatively quick to make and limited in functionality. Graphics, lists, tables, grids, searching and filtering are elements that demand more skills and time. One explanation for this trend can be that the advanced journalistic artefacts, as for instance the drillable dataset, do not get enough positive feedback from the readers to make it worth the (considerably lengthy) time it takes to create it. A positive take on this is that the audience needs time to understand and appreciate these new forms of online journalism; that new genres will develop over time, which have a useful balance between data and story. Another explanation can be that the established tradition of journalistic storytelling gives primacy to linearity. The function of journalistic stories is seldom to let the audience explore, but to explain and convey ideas already thought out. Form follows function or rephrased by journalist and information graphics expert Alberto Cairo as “function constrains the form” (Cairo 2012).

When it comes to skills, the training necessary to undertake computational journalism is different from what has traditionally been on offer at journalism schools. A

computational journalist needs to master both the inverted pyramid structure of journalistic storytelling and basic iteration statements found in any programming language such as the “while”, “for each” and “for” statements. To a certain extent programming is manual and repetitive work, especially when preparing data for analysis. There are general rules to follow when putting together a database and making sure that the quality of the data is good. The fundamental activity is to normalize the database using SQL (Codd 1970). When doing this, errors in the data are often discovered. When this “drawing by numbers” job has been completed the journalistic inquiry can begin. How data are joined, analysed and presented is the direct result of journalistic sense-making combined with proficiency in programming. This is often an inductive process of trial and error, not guided by reasoning alone. The programming becomes inseparable from journalism and vice versa. That said, this “programming-as-journalism” should not be mystified unnecessarily. Programming can be directly compared to other journalistic modes of expression like writing and photography, but as you do not want to hire any photographer as photojournalist - you do not want to hire any programmer to do computational journalism. The development of computational journalism has led to the hiring of new hands and the acquisition of new skills by older journalists. But our respondents, of both these types, gave primacy to journalism skills when explaining what they do. “The journalist’s way” of doing things is a prerequisite for doing computational journalism within a news organization. The technical skills are subordinated to the unbroken tradition of journalism. Why do the computational journalist themselves subordinate their technical skills to journalism? One possible explanation is that the cognitive and affective skills necessary to do computational journalism are easily aligned with the journalistic tradition, while the psychomotor skills are not. This represents the difference in how mental and manual labour is valued in the society and in the newsroom. To put it bluntly: programming is viewed as only a technique and therefore something not worth talking about.

According to Terry Flew et al. (2012), the aim of applying computation to journalism is to free “journalists from the low level work of discovering and obtaining facts, thereby enabling a focus on verification, explanation and communication of news”.

More journalistic work can be done in less time with fewer errors. In tasks where work is, and also previously was computable, this is a reasonable result to expect. In our study, on the other hand, aspects of speed, consistency or accuracy of computers, aspects that could be said to support the accountability function of journalism, are rarely mentioned. On the contrary, time is not something you use less of when doing computational journalism, but more. Programming is a tedious process and added complexity demands constant checking and rechecking of the facts. This is in line with how our respondents chose to define computational journalism. They said they use computers where their colleagues use telephones, microphones and shoe-leather. Sometimes they write a computer program in C#, sometimes they find, install and use a new software tool to get a job done, and sometimes they use the more advanced features in Excel. Computing is just another tool in the toolbox when aggregating the knowledge needed to tell revealing news stories, alongside notebooks and physical archives. Computational journalism, as framed by programming journalists in Norwegian newsrooms, is a therefore mainly a continuation of journalistic work practices. One exception is worth mentioning though, and that is the use of crowdsourcing, a method interviewees spoke highly of. The potential of using human judgment on the Web to gather and verify information is substantial, and it is possible to see how that can transform investigative journalism in some cases.

In the Newsroom: Fading Conflicts and a Bright Future?

To become a programming journalist you must accept that journalism goes before programming. You need to be a journalist "by conviction" to avoid conflict and to be able to thrive in the newsroom. It is important to distance yourselves from the technologists working in the ICT department. Your fellow journalists should not be in doubt whether you belong to the newsroom or ICT. You need to bypass ICT by choosing lightweight technical approaches and find solutions that do not require direct assistance from ICT. This finding is comparable to what Cindy Royal (2012) found in her study of the NYT Interactive Department. Programming journalists at NYT see themselves as journalists first and technologists second.

The NYT Interactive Department has established a lightweight "rapid prototyping"-based work process, while we find that stable routines are not quite established in Norwegian newsrooms yet. It seems that it is unclear to many what the limitations and possibilities of computing are. There is no "Interactive News Technology Department" in any of the news organizations we visited. The computing journalists in our study are working alone and in small teams, figuring out how to best collaborate as they go. Royal (2012) describes a hacker culture emphasizing creativity, innovation and collaboration. Our informants subscribe to the same values, but are at the same time careful to emphasize that innovation should happen within the boundaries of the journalistic tradition. These differences in findings suggest a cross-national comparison of practices as future research. What differs among different media systems, and what properties of these systems facilitate high-quality computational journalism?

The skills needed to do computational journalism are valued by the newsrooms, but are often "black boxed" by editors, non-programming newsroom staff and the computational journalists themselves. In the future, it is important that the professionals interviewed in this study are allowed to share the knowledge with other journalists, including the more technical aspects of their practice. Computational journalism is a craft where journalism and computing merge into one process where both skill sets are used simultaneously. To be able to create innovative and journalistically sound products, the performer of this craft needs a whole understanding of material, form, technique and purpose. We think that it is impossible to outsource the programming and create an ultimate journalistic machine. Technology in itself cannot solve the challenges that are important to journalists and democracy. Craftsmen with knowledge to build, wield and aim the technology are needed. The interviewees in this study are pioneers in a valuable craft, a craft that needs to be nurtured and given priority if it is to fulfil the potential so clearly identified by editors, researchers and the computational journalists themselves.

Conclusion

In this paper we have examined the work practices of computational journalism in Norwegian newsrooms. The traditional journalistic process and values are followed, but supplemented with software both as pre-fabricated and project-specific programs. Typically the job starts with a dataset, either collected on the journalist's own initiative or by collaborators in the newsroom. The access to data in Norway is perceived as good, while the process also includes obstacles of both legal and practical matters. When the data are analysed and facts or trends are found, finding a suitable form to present the results can be a challenge. Advanced info graphics are considered too complex. Aversion for bells and whistles, and a preference for simplicity and clarity, results in the use of linear narratives with timelines, maps, text, sound and video. While the cognitive and affective skills needed to do computational journalism align smoothly with traditional journalistic values, the psychomotor skills (the use of computational techniques) represents a very different practice. When undertaking computational journalism, the creative process is not separable from coding, comparable to how writing cannot be separated from authorship. In the current literature we find great hopes for computational journalism, but we suggest modest expectations in this regard, at least if we anticipate these changes coming from within newsrooms. According to the expert practitioners in Norway, the utility of computational journalism is not to free journalists from the low-level work of discovering facts or freeing up time. The utility of computational journalism is rather the development of new forms of data-driven and user-driven journalism that have the potential to fulfil the traditional hopes for and promises of journalism *per se*.

References

Anderson, Cristopher W. 2011. ‘‘Notes Towards an Analysis of Computational Journalism.’’ HIIG Discussion Paper Series 2012 (1). http://papers.ssrn.com/sol3/papers.cfm?abstract_id 2009292.

Bloom, Benjamin S., Max D. Engelhart, Walker H. Hill, Edward J. Furst, and David R. Krathwhol. 1956. Taxonomy of Educational Objectives. The Classification of Educational Goals, Handbook I: Cognitive Domain. New York: David McKay.

Boczkowski, Pablo J. 2005. Digitizing the News: Innovation in Online Newspapers. Cambridge, MA: MIT Press.

Bogost, Ian, Simon Ferrari, and Bobby Schweizer. 2010. Newsgames: Journalism at Play. Cambridge, MA: MIT Press.

Buchanan, Richard. 2001. ‘‘Design and the New Rhetoric: Productive Arts in the Philosophy of Culture.’’ Philosophy and Rhetoric 34 (3): 183-206. doi:10.1353/par.2001.0012.

Cairo, Alberto. 2012. The Functional Art: An Introduction to Information Graphics and Visualization. Berkeley, CA: New Riders.

Codd, Edgar F. 1970. ‘‘A Relational Model of Data for Large Shared Data Banks.’’ Communications of the ACM 13 (6): 377-387. doi:10.1145/362384.362685.

Cohen, Sarah, Chengkai Li, Jun Yang, and Cong Yu. 2011. ‘‘Computational Journalism: A Call to Arms to Database Researchers.’’ In Proceedings of the 5th Biennial Conference on Innovative Data Systems Research, January 9□12, 148□151. Asilomar, California, USA: ACM.

Cox, Melissma. 2000. ‘‘The Development of Computer-assisted Reporting.’’ Paper presented to the Newspaper Division, Association for Education in Journalism and Mass Communica- tion, Southeast Colloquium, March 17-18, University of North Carolina, Chapel Hill.

Creswell, John W. 2009. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. 3rd ed. Los Angeles: Sage.

Deuze, Mark. 2005. ‘‘What Is Journalism? Professional Identity and Ideology of Journalists Reconsidered.’’ Journalism 6 (4): 442-464. doi:10.1177/1464884905056815.

Diakopoulos, Nicholas. 2012. ‘‘Cultivating the Landscape of Innovation in Computational Journalism.’’ Tow-Knight Center for Entrepreneurial Journalism. http://www.nickdiakopou- los.com/wp-content/uploads/2012/04/diakopoulos_whitepaper_systematicinnovation. pdf.

Diakopoulos, Nicholas, Munmun De De Choudhury, and Mor Naaman. 2012. ‘‘Finding and Assessing Social Media Information Sources in the Context of Journalism.’’ In Proceedings of the Conference on Human Factors in Computing Systems (CHI). http://research.microsoft. com/en-us/um/people/munmund/pubs/chi_2012.pdf.

Diakopoulos, Nicholas, Mor Naaman, Tayebeh Yazdani, and Funda Kivran-Swaine. 2011. ‘‘Social Media Visual Analytics for Events.’’ In Social Media Modeling and Computing, edited by Steven C. H. Hoi, Jiebo Luo, Susanne Boll, Dong Xu, Jin Rong, and Irwin King, 189-209. London: Springer.

Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift. 2012. ‘‘The Promise of Computational Journalism.’’ Journalism Practice 6 (2): 157-171. doi:10.1080/17512786. 2011.616655.

Global Editors Network. 2012. ‘‘DJA 2012: The Gallery of All the 2012 Data Journalism Awards Shortlisted Projects.’’ http://www.wizehive.com/voting/dja2012.

Gray, Jonathan, Liliana Bounegru, and Lucy Chambers. 2012. The Data Journalism Handbook. Sebastopol, CA: O’Reilly Media.

Hamilton, James T., and Fred Turner. 2009. ‘‘Accountability through Algorithm: Developing the Field of Computational Journalism. Report from Developing the Field of Computational Journalism.’’ Center for Advanced Study in the Behavioral Sciences Summer Workshop, Stanford, CA. http://dewitt.sanford.duke.edu/images/uploads/about_3_Research_B_cj_1_finalreport.pdf.

Harrow, Anita J. 1972. A Taxonomy of the Psychomotor Domain: A Guide for Developing Behavioral Objectives. New York: David McKay.

Heidegger, Martin, and Manfred Stassen. 2003. ‘‘The Question Concerning Technology.’’ In Martin Heidegger: Philosophical and Political Writings. New York: Continuum.

Hu, Chenglie. 2011. ‘‘Computational Thinking: What It Might Mean and What We Might Do About It.’’ In Proceedings of the 16th Annual Joint Conference on Innovation and Technology in Computer Science Education, 223-227. New York: ACM.

Jacobson, Susan. 2012. ‘‘Transcoding the News: An Investigation into Multimedia Journalism Published on Nytimes.com 2000-2008.’’ New Media & Society 14 (5). http://nms.sagepub. com/content/early/2012/01/05/1461444811431864.

Johnson, Robert R. 2010. ‘‘Craft Knowledge: Of Disciplinarity in Writing Studies.’’ College Composition and Communication 61 (4): 673-690.

Kovach, Bill, and Tom Rosenstiel. 2007. The Elements of Journalism: What Newspeople Should Know and the Public Should Expect. 1st rev. ed. New York: Three Rivers Press.

Kratwohl, David R., Benjamin S. Bloom, and Bertram B. Masia. 1964. Taxonomy of Educational Objectives, the Classification of Educational Goals. Handbook II: Affective Domain. New York: David McKay.

Loosen, Wiebke. 2002. ‘‘The Second-level Digital Divide of the Web and Its Impact on Journalism.’’ First Monday 7 (8). http://firstmonday.org/ojs/index.php/fm/article/view/977. doi:10.5210/fm.v7i8.977.

Lorenz, Mirko. 2010. ‘‘Data-driven Journalism: What Is There to Learn? (Stanford, June 201 . . .’’ Stanford. http://www.slideshare.net/mirkolorenz/datadriven-journalism-what-is-there-to- learn.

Meyer, Philip. 2002. Precision Journalism: A Reporter’s Introduction to Social Science Methods. 4th ed. Oxford: Rowman & Littlefield.

Mitchelstein, Eugenia, and Pablo J. Boczkowski. 2009. ‘‘Between Tradition and Change: A Review of Recent Research on Online News Production.’’ Journalism 10 (5): 562-586. doi:10.1177/ 1464884909106533.

O’Sullivan, John, and Ari Heinonen. 2008. ‘‘Old Values, New Media.’’ Journalism Practice 2 (3): 357-371. doi:10.1080/17512780802281081.

Parasie, Sylvain, and Eric Dagiral. 2012. ‘‘Data-driven Journalism and the Public Good: ‘Computer-assisted-reporters’ and ‘Programmer-journalists’ in Chicago.’’ New Media & Society. doi:10.1177/1461444812463345.

Park, Souneil, Minsam Ko, Ying Liu, Dal Yong Jin, and Junehwa Song. 2011. ‘‘Improving Journalism through the Web: Framework for Media Bias Mitigation.’’ In Proceedings of the 3rd International Conference on Web

Science. Koblenz, Germany. http://www.websci11.
org/fileadmin/websci/Posters/88_paper.pdf.
Pousman, Zachary, John Stasko, and Michael Mateas. 2007. ‘‘Casual Information Visualization: Depictions of Data in Everyday Life.’’ IEEE Transactions on Visualization and Computer Graphics 13 (6): 1145-1152. doi:10.1109/TVCG.2007.70541.
Powers, Matthew. 2012. ‘‘‘In Forms That Are Familiar and Yet-to-be Invented’: American Journalism and the Discourse of Technologically Specific Work.’’ Journal of Communica- tion Inquiry 36 (1): 24-43. doi:10.1177/0196859911426009.
Robinson, Sue. 2011. ‘‘Convergence Crises: News Work and News Space in the Digitally Transforming Newsroom.’’ Journal of Communication 61 (6): 1122-1141. doi:10.1111/ j.1460-2466.2011.01603.x.
Royal, Cindy. 2012. ‘‘The Journalist As Programmer: A Case Study of The New York Times Interactive News Technology Department.’’ #ISOJ The Official Research Journal of the International Symposium on Online Journalism 2 (1).
Segel, Edward, and Jeffrey Heer. 2010. ‘‘Narrative Visualization: Telling Stories with Data.’’ IEEE Transactions on Visualization and Computer Graphics 16 (6): 1139-1148. doi:10.1109/ TVCG.2010.179.
Singer, Jane B. 2004. ‘‘Strange Bedfellows? The Diffusion of Convergence in Four News Organizations.’’ Journalism Studies 5 (1): 3-18. doi:10.1080/1461670032000174701.
Stavelin, Eirik. 2012. ‘‘Nyhetsapplikasjoner: Journalistikk Møter Programmering.’’ In Nytt Pa˚Nett Og Brett: Journalistikk i Forandring, edited by Martin Eide, Leif Ove Larsen, and Helle Sjøvaag, 107-125. Oslo: Universitetsforlaget.
Steensen, Steen. 2010. ‘‘Online Journalism and the Promises of New Technology.’’ Journalism Studies 12 (3): 311-327. doi:10.1080/1461670X.2010.501151.
Van Dalen, Arjen. 2012. ‘‘The Algorithms behind the Headlines.’’ Journalism Practice 6 (5-6): 648-658. doi:10.1080/17512786.2012.667268.
Weinhold, Wendy. 2010. ‘‘Letters from the Editors.’’ Journalism Practice 4 (3): 394-404. doi:10.1080/17512781003643228.
Weinstein, Matthew. 2006. ‘‘TAMS Analyzer Anthropology as Cultural Critique in a Digital Age.’’ Social Science Computer Review 24 (1): 68-77. doi:10.1177/0894439305281496.
Wickman, Chad. 2012. ‘‘Rhetoric, Techné, and the Art of Scientific Inquiry.’’ Rhetoric Review 31 (1): 21-40. doi:10.1080/07350198.2012.630953.
Wing, Jeannette M. 2008. ‘‘Computational Thinking and Thinking about Computing.’’
Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 366 (1881): 3717-3725. doi:10.1098/rsta.2008.0118.

III. Newsworthiness on Twitter

Although journalists follow social media and social media do contain newsworthy insight and knowledge, this type of data is rarely analyzed for journalistic purposes. This paper presents a cluster analysis tool that supports a journalistic inquiry into Twitter messages. By tailoring cluster analysis methods this tool produces subsets of a text collection with similar texts to speed up the quest for interesting user generated content. Results show promise for this kind of computer supported analysis, but also outlines problems both with the methodology and the assumptions that good arguments or novel opinions are enough to be considered newsworthy. Evaluation by professional journalists show that in spite of an acceptable quality of algorithmic sorting, the focus in future applications should put more emphasis personas, and identifying "the usual suspects".

Introduction

Social media has grown to become an arena for anyone to participate in debates and share opinions on any topic. For news outlets, this offers new opportunities to discover stories and issues of debate and public interest. Journalists are among the elite of Twitter users (Kwak et al. 2010; A. O. Larsson and Moe 2011) and thus do get insights and stories from this use trough his or hers circle of Twitter contacts. But mostly the data are an ephemeral glimpse of individual tweets. In cases where data from social media can offer new or additional insights, stories or sources for journalist, methods for analyzing data and finding texts of interests are needed. This paper presents experiences with a custom built tool to cluster Twitter texts (tweets) into groups of similar tweets by tailoring clustering techniques for Twitter.

The first obstacle with social media data is the size of the datasets. While Norway is small, and Twitter is not widely used (8% of Norwegian internet user accessed Twitter weekly in the 4th quarter of 2011, (Futsæter 2012)) the datasets exceed what a journalist can be expected to read through in her normal workflow. The second is the unfamiliar mix of participants of debaters and commentators. While a journalist is

trained in questioning experts and authorities and finding ordinary people for cases, the cacophony of voices and opinions from everyman on Twitter makes it hard to decide: is it the messages from traditional sources – authorities in various sectors – that are to be examined or the most prominent, popular or clever sayings of 'the public'? Further, methods should be independent of the topics of the analyzed data, and reduce the time needed to get a fair overview over the data.

In order to explore the possibilities to utilize Twitter-data, I built a software tool that puts tweets into chunks of manageable sizes and offers users a navigation interface with simple manipulation functions. The approach is based on the bag of words model in natural language processing, and is at the core an effort to extract a structure in the data that isn't explicitly found as meta-data. The applications take large amounts of tweets and groups those who use similar words into clusters that gets presented to a journalist, so that the journalist can review groups of similar tweets to quicker gain an overview over the material. Functions for further exploration and keeping track of single texts and clusters of texts allows for a rapid analysis of user generated content on Twitter.

Related Work

Prior research on Twitter has revealed that actors who are important in the old media landscape also hold key position on Twitter, and Twitter usage follow media events (An et al. 2011; De Longueville, Smith, and Luraschi 2009; Kwak et al. 2010; D. A Shamma, Kennedy, and Churchill 2010). Through the aim of understanding the microblogging phenomenon, different studies have found different user-types (De Choudhury, Diakopoulos, and Naaman 2012; De Longueville, Smith, and Luraschi 2009; Java et al. 2007; Larsson and Moe 2011) and also, that in spite of being a noisy media, a potential for news media to "analyze, interpret and conceptualize a system of collective intelligence, rather than in the established practice of selection and editing of content" (Hermida 2010). In this context tools for journalists has been constructed, both to find sources (Nicholas Diakopoulos, De Choudhury, and Naaman 2012) and to analyze user feedback on media events (N. Diakopoulos et al.

2011). My approach explores the data with intentional blind eye to the authors' status to emphasize the democratic possibility of letting anyone through with their perspective. Ways of detecting events from larger streams of social media data is also developed (H. Becker, Naaman, and Gravano 2011; Weng and Lee 2011) including ways of identifying relevant and useful messages from Twitter (H. Becker, Naaman, and Gravano 2011; Luo, Osborne, and Wang 2012). The studies of politics on micro blogs also reveal that social media usage "shadow" real world events to the degree where election results (Tumasjan et al. 2010) and the change of topics in TV-broadcasted debates (David A. Shamma, Kennedy, and Churchill 2009) can be algorithmically detected. Topics that do not spike, but linger, constantly low-volume ongoing arguments, falls outside of these methods, my approach includes such discourses.

Outside academia relevant projects such as the Knight foundation funded associated press project (The Overview Project (Stray 2012)) also apply clustering techniques to aid analysis of textual data for journalistic enquiry. While the overview project is aimed at optical character recognition (OCR) type corpora, this project aim is events in social media texts.

Much focus has been put on understanding what Twitter is, and who the users of Twitter are and how they do discuss topics that are of interest to the news media (politics, disasters, media events, etc.), but less has been done to apply this to ways of analyzing this to find news stories. I want to explore the possibility to analyze accumulated data for a topic in a way that supports a journalists' need to quickly get an overview over what is discussed.

Design Process

The scenario is simple: a journalist collects tweets for a topic related to her work and end up with too much data and too little time. To read through all texts is unrealistic, so how can she get an overview over what the Twitter-data contains?

The initiation for this project was a presentation of a journalistic project where a team of journalists had analyzed Twitter material concerning the 2011 terrorist attacks in Oslo[30]. The team had spent considerable resources on reading though every single tweet. In such a sensitive case this might be necessary when the data is to be published, but in smaller cases the manual labor of this team can be exchanged with computational means to ensure that this kind of data gets some analytical attention instead of no attention.

This paper relates to a real world problem and the initial criteria for the prototype was collected in dialogue - by telephone interview - with the team leader from the NRK project and followed up though emails to work out details and adjustments.

The tool should be flexible enough to handle data from all sorts of topics; it should take into account the use of multiple languages used in debates in Norway and offer users a way to assess and store material of particular interest to the journalist. The clusters should represent grouping of the material where similar things are discussed, and reduce the work of looking through one large dataset to look at fewer clusters of similar texts.

While meta-data such as time and location can aid event detection (Becker, Naaman, and Gravano 2010) very few of the tweets that I have collected from Norway contains location information (typically 2-3%). I discarded the time dimension as some topics are continuous, and multiple clusters with the same topic are undesirable for this experiment. As a result of this the clustering must be based on the texts themselves. This differentiates this study from others. While time is very important to make sense of the world, the exclusion of time in the clustering of the tweets allows for topics of low quantity per time unit but with continuous discussion to be gathered.

[30] The NRK project can be seen at http://nrk.no/terrortwitter

Clustering Tweets

Clustering as a method is closely related to information retrieval and search. It can also be used for other activities such as browsing (Cutting et al. 1992) and identification of redundant pieces of information (Nezda 2012). My intention was to utilize clustering to expose themes of topics of discussion in a larger dataset.

Hierarchical clustering has the advantage that the number of clusters does not need to be known a priori, and this makes sense in a corpus with more or less unknown content. A disadvantage is scaling, as a matrix containing distance measure between all documents needs to be created, and this is computationally expensive. Other methods such as k-means takes the number of desired output clusters as an input, but do not require a distance matrix. To overcome this problem I used the Buckshot algorithm (Cutting et al. 1992) to initialize k-means with the results from a hierarchically clustered random subset. While speed still is an issue, the memory needed is limited to what can be found in a typical desk- or laptop computer. For the sake of this experiment processing time is not decisive. While in a real world scenario speed is key and a quicker method of clustering is needed.

As distance measure the Euclidian distance was used for the hierarchical clustering and the cosine angle distance was used for the k-means clustering.

The operationalization of the clustering algorithm was done in Python, drawing on the natural language tool kit (NLTK) (Loper and Bird 2002) and the Oslo-Bergen tagger (a grammatical tagger for Norwegian) for word categorization and lemmatization (“The Oslo-Bergen Tagger” 2012). By removing unwanted word categories (assumed less rich in information such as determinatives, conjunctions, pronouns, etc.) revealed by the tagger, vectors were built with the tf-idf value of each word. The tf-idf value is a numerical description of how important a word is to a document in a collection of documents.

Twitter contains a lot of data that’s hard to categorize (noise). The messages are short, often with unconventional abbreviations and slang is heavily used. Spelling

mistakes are not uncommon. The bag of words model gets weaker as a result of this. Initial results pointed out some further alteration in the algorithm. Tweets shorter than seven words were excluded from the clustering and presented to the user as a single cluster. I also added a short list of stop words. The reduction of the set is done in effort to condense the concentration of significant and meaningful words. Further an effort was put into boosting particularly meaningful words:

Hashtags is a way for authors on Twitter to label a tweet as in a context or topic. This is done by using the number sign (#) as a postfix to any word. Examples can be #Oslo, #politics, #obama, etc. These words I consider as more valuable than others as the give topical information, and I give them a boost by adding a fixed value to the tf-idf previously calculated for this word. The same is done with nouns, verbs and user mentions (identified by the postfix "@" to a single word) with decreasing boost values.

By using the Buckshot algorithm, the number of clusters (k) is decided by the outcome of a smaller initial hierarchical clustering procedure. In the following k-means algorithm a troublesome problem occurs. If a tweet has no or little overlap with one of the existing clusters, it still needs to be put in one. This results in some very large, bloated and inconsistent clusters that offer little aid to understand what is discussed. As a remedy to this I added a breakout mechanism to the k-means; if the distance between a tweet and the nearest centroid of any current cluster is too great, the new tweet is added as a new cluster. The algorithm stops when there are no (or fewer than n) alterations between this and the last iteration. To explore the algorithms ability to cluster material - retweets (that are identical or largely overlapping texts) were removed. Retweets can be fetched back in at a later point, and by doing so the clustering without retweets makes it is clearer to see what texts become clustered together. While retweets are often used as a key factor in understanding Twitter, the exclusion of retweets allows for a clearer view of what different texts that gets grouped together, and also limits the effects of having the same message repeated over and over

A further alteration was done in the presentation of the clusters. To quickly get a sense of what is typical in a cluster I ordered the clusters by the sum of the tf-idf values for each word in a tweet. The clusters were also labeled with the highest-ranking tf-idf word for the individual cluster against all the clusters. The clusters were presented to the participators as rectangles in a one-leveled treemap where the size of the rectangle corresponds to the number of tweets in the cluster (see figure 1). The clusters were labeled with the cluster size, with top key words revealed by hovering over each rectangle. By clicking on a rectangle, the clusters' content is displayed in a side-by viewer.

Flow of the Algorithm

The clustering algorithm starts with input of how many tweets to fetch from the database. The Oslo-Bergen tagger preforms lemmatization and determines word categories, and words from unwanted categories are removed. Text with very few words (n<7) are removed from further processing but kept as a separate cluster (so these tweets sill can be found in searches and though browsing meta data in the application GUI).

Further vectors are created for each tweet, with tf-idf values representing each word per tweet. A boost is added to @mentions, verbs, nouns and hashtags. Through extensive experimentation I ended up using {+ 0.15 for @mentions and verbs, +0.3 for nouns and +0.5 for hashtags}, a set of values that worked well in practice. A distance matrix is then created for a random sample of the vectors and a hierarchical clustering is performed. The centroid of each cluster from the hierarchical clustering is calculated and used as seeds for the k-means clustering. The k-means is performed on all vectors and is ended when zero (or fewer than n if number of iterations is greater than nn). If a tweets has a distance to a current centroid that is greater than 0.009 (where 1 is identical and 0 is no overlap) a new cluster is created from the text.

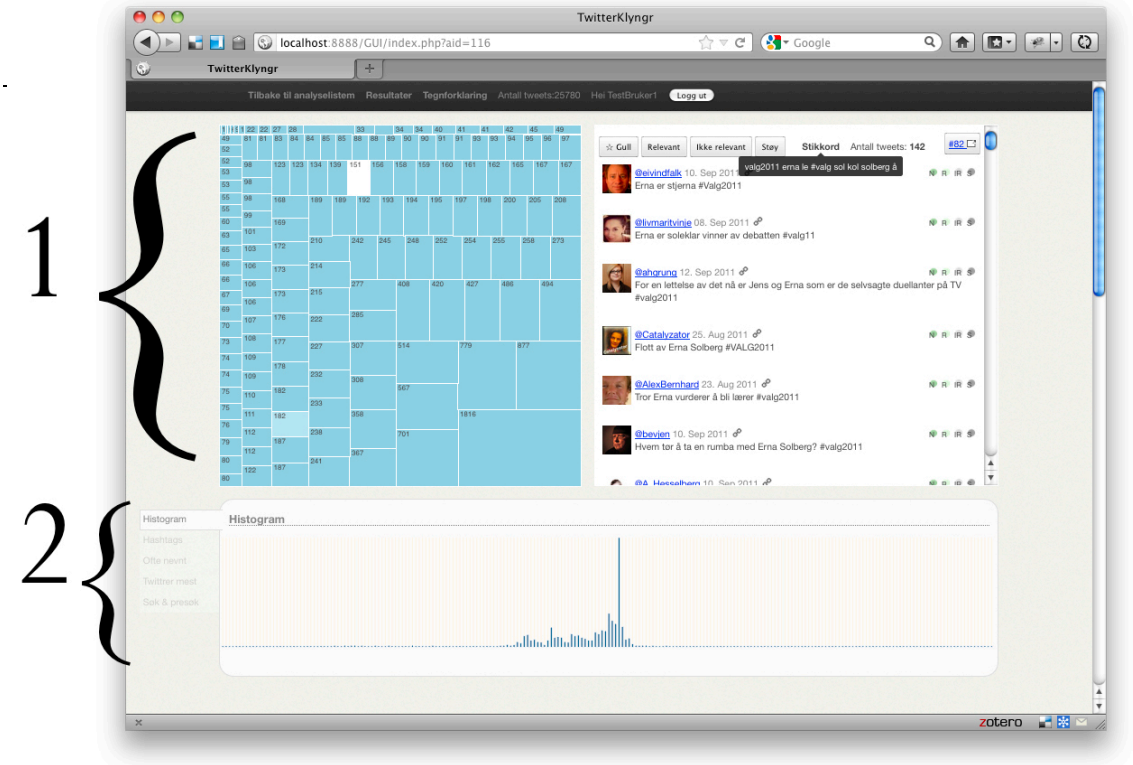


Figure 1: Screenshot of the graphical user interface. 1 points to a single layer tree graph where each cluster is a rectangle (the cluster display). Hovering each cluster shows the top key words, hashtags and users for this cluster as a tool tip. By on clicking a cluster the tweet in this cluster is displayed in the viewer to the right. On top of the viewer the user can prioritize the cluster (pure gold, interesting, not interesting, noise). Single tweets can be annotated in the same matter (to the right of each tweet). In the end of a session the user can view her annotated clusters and tweets in prioritized order. 2 points to Meta data (histogram, hashtags, users and @mentions by frequency, and search. By clicking on items from the Meta data (a day, a hashtag, a user, etc.) the clusters that contain these tweets are highlighted in the cluster display and the tweets are displayed in the viewer

Data

Data for the study was collected using YourTwapperKeeper (O'Brien III 2010). This is an application for fetching and storing tweets collected from the Twitter API. My application shears database tables with YourTwapperKeeper. The data was collected over 8 months, from May to December 2011 and are search results based on hashtags used in

1) A broad debate: the 2011 Norwegian local election (n=25780, hashtags: #valg, #valg11, #valg2011)
2) A narrow debate: the EU Data Retention Directive (n=6279, hashtag: #DLD).

The narrow debate is a smaller set and was used as a warm-up exercise to let the participants get to know the system before the larger set with election data was examined.

This data was chosen over the potential use of standardized data collections from the text mining community to offer professional Norwegian journalists data that likely is relevant to their day-to-day activities, in their own language.

Study

To gain understanding of how the clustering application and method appeals to journalists an exploratory study was conducted. After a pilot study with students from the local student newspaper a total of seven test-sessions was done with professional journalist. These were selected from national, regional and local media institutions (TV, online and paper). All participants had experience with social media as a source for news production, and all used Twitter in their workplaces. The participants were relatively young (from 26 to 45 years old) and digitally literate, but also relatively experienced in the news business (from 4 to 18 years). The session with each participant lasted about an hour (M = 64 min), including a five-minute introduction and a twenty-minute interview after the participant had used the system. The participants were unpaid.

The test sessions were conducted in the participants local work environment when possible. Five sessions were done in silent rooms connected to the main newsrooms, while two were done in a seminar room at the university campus.

The experimental procedure was an introduction to the system followed by a small dataset for the participants to get to know the interface and functionality. A larger dataset was then presented with the task of 1) evaluating the clustering ability and 2) find material that they thought could be worth reporting or looking further into.

The testers were encouraged to talk allowed during the test session. The audio from both the test sessions and the interviews was recorded and fully transcribed. The researcher also recorded observations.

Criteria for evaluation was the perceived utility of the approach, in qualitative terms, as described by the testers in regard to what they considered to be important when searching for stories, sources or trends in the data.

Findings

Evaluating the clustering ability is an exercise in finding overlap in the algorithmically generated groups, and the participants' expectations to what such groups of sub-theme texts should contain. The search for material worth reporting is a search for "interestingness" and building on the literacy gained from the first task, an exercise in applying journalistic interest to a given dataset. The latter exercise can tell a lot of how the clustering approach should be executed.

Some overarching results include the unfamiliarity with working with Twitter data as an object of analysis; another is the experience that overwhelmingly large datasets are discouraging in spite of being broken down into smaller units (156 and 183 clusters) and; that while the participants had different fields of interest and different ideas of what could be worth looking further into, much of the same focus was put on who the authors of the tweets are and what functions they have in the context they are in. Finding sources was equally interesting to finding good stories or noteworthy arguments.

Deciding where to look – gaining literacy

The bag of words model for natural language processing with a clustering strategy that incorporates a simple distance measure creates clusters that are statistically similar, but occasionally semantically and pragmatically inconsistent. The sentences "I don't hate you, I love you" and "I don't love you, I hate you" are considered equal. While both sentences are related to strong feelings of love and hate, this kind of grouping did lead to some realization of methodological shortcomings.

> *Here is one cluster that has clustered together both school and school election, and election and win. It has clustered some tweets about school elections, and also politics about schools (P6).*

This characteristic becomes a problem when two sides of the same debate utilized the same words to express very different ideas:

> *Here the key words are 'to stop', so there is a lot of 'stop DLD', but also a lot of 'DLD stops'. 'Could DLD have stopped 22/7?' vs 'your donation to stop DLD (P2).*

As much as this is a shortcoming in the chosen methodology it becomes a matter of media literacy in practice. "So it is as much about stopping the Data Retention Directive as if the Data Retention Directive may be to stop anything. So it is both sides of the debate" (P2).

This lead to the question of what kind of texts that would be "hidden" by the noise in clusters with poor key words and unclear topics, particularly in large clusters that the participants considered too big to properly look though:

> *Some of the categorizations - the clusters - are totally uninteresting while some are interesting. So what this actually does is to cluster together some conversations, some discourses that are found on Twitter, so that you do not have to do it yourself. It also categorizes the stuff that isn't interesting, but you can't quite know whether to trust it fully, because it only says that all these 1816 tweets doesn't fit into any other topics. But there might be interesting stuff here; only it hasn't found any common denominator that they fit (P6).*

The flipside of this feature is that topics that are known can be scattered across multiple clusters and thus diluted. "This story should have had a lot more tweets. The one with the teacher in Kongsberg or wherever, I can't understand this properly. Why are there only 12 tweets here?" (P5) The expectation of what a cluster should be and what the clusters from my algorithmic clustering are - is slightly different. The participants expected subsets of a debate to be thematically divided or "threads". Not linguistically similar texts. In spite of this, evaluation of the algorithm shows promise for further tailoring and adaptation of the algorithm and possibly even more

important: help identify and highlight clusters that contain texts that are expected to be interesting (e.g. containing known named entities, etc.).

The clusters that did yield most positive feedback were typically small in size and considered clear (absence of noise). These were perceived as more specific and more interesting.

> *It's the same here, here's one that has 'moe' and 'borten' and 'ola', 'bort' and '2011', so here you have tweets about Ola Borten Moe, that suggests that by skimming though this cluster –with 63 tweets – we can say something about what people think of him in this data. Generally, by looking at smaller clusters, you get more specific key words (P6).*

Although the datasets were big and the interest fields of the participants are different, some clusters were identified and commented by multiple testers.

> *Here it has come across something that has to do with first times. So if you look for first-time voters, this cluster would be very interesting (P2).*

This cluster did not gravitate towards a named entity and functions as an example of how this methodology can construct clusters concerning concepts simply through the similarity of wording.

> *This is something I could have checked out to see if I wanted to write a story about; someone that has voted for the first time. 'Looking forward to vote for the first time tomorrow'. Perfect, we give him a call to ask if we can join him. It is like this: you can show up at the polling station, but then you risk going to the wrong station, or maybe you can't find any first-time voters when you're there, or they say no (P4).*

The overall experience with evaluating clusters to identify where interesting chunks of a debate are to de found in this tool, all results point towards purity. The lack of noise and immediate consistence (strong signal) is good, and this was found in smaller clusters that tend to have more specific key words.

Divide and Conquer

The prototype offers simple methods to extract and highlight content based on various meta-data (dates, hashtags, mentions) and free text search. The ability to save

clusters or tweets to a user-specific list also strengthens the focus on encompassing and including subsets the user finds interesting or pertinent. The participants had no complaints concerning this, but through use the opposite function – exclusion – were found lacking. Just what to exclude varied with the participants (dis)interest.

Some found particular authors noisy: “To opt out of threads that @nicecap participates in would have help a lot" (P1). Who to pay attention to and who to exclude is important.

> *If you can remove content, create a Twitter group of media people and politicians, and then search for immigration... […] don’t get me wrong, the ‘important’ people must be abstracted (P7).*

While others would like to remove geographically bound clusters: "Rana, election in Rana, I'm not interested in that. Trondheim, I'm not interested in that" (P3). To be able to remove subsets based on meta-data was also demanded.

> *This is clearly something Swedish I am not inn on. #Acta here seems like something Swedish. Høyre [the conservative party] is Norwegian .. Yes. #FRP [the progress party], #AP [labour party], #EU [the european union]. #2pl - this is some football stuff? Is it possible to remove tags from the dataset? Or mark this #2pl as irrelevant for instance? (P3).*

During testing it became clear that while the testers did have special fields of interest and ideas on how to find data in their field, the massive amounts of data invites to explore and investigate outside of their daily topical spheres. When clusters and subsets (e.g. by hash tag) were identified as related to something they recognized (e.g. media stories; events; persons; or locations) they had the need to exclude it to see what is left when this is removed. The focus on finding the interesting is also a matter of removing what is known and uninteresting, and this should be included in the requirements list for a future tool.

Finding Stories

The initial idea for the tool was to aid journalists in finding stories; follow-up stories and sources in social media, or to confirm that the journalist has a fair overview of a debate. When asked to identify trends, tweets or other noteworthy findings that could

be used for stories in their workplace, the participants all found something to show. This does not mean that these findings would actually end up in print; broadcast or web media, the unaccustomed setting of a user test session, with a researcher present, does skew these story-findings, but the results might illuminate what sort of stories this kind of tool does inspire.

The already mentioned first-time voter angle is a story that is regularly told in relation to elections, other findings the participants found are similar to this in regard that they are stories that often is covered: property taxes (P4), voters reactions after voting and election vigils (P3). Another category are Twitter-related stories such as small political parties that has few votes but that generate much interest and debate on Twitter. Among other Twitter-related angles was the distribution between the national party leaders and parties in mentions and activity, and while this was commented upon frequently, the seemingly predictable results (the distribution looks a lot like the election result) was not considered worth reporting upon.

> *It's just the usual suspects here. It's definitely the political left that is most active in social media. That is almost a story. Which [parties that] are mentioned most. It is red-green. Perhaps it's an effect of the current government; it's hard to tell. But it is interesting never the less (p3).*

The stories that were identified while browsing are mostly curiosities that are immediately surprising or extraordinary. "That looks like a story, if there is a 94 year old man that is going to the municipal council, that is a story" (P3). The fact that journalists are looking for stories in the margins of normality were underlined repeatedly.

> *This is an odd cluster –with the key word "go"/"come on". It indicates… it is almost like someone sits in a sporting arena and cheers at the debate. It spans from 'go, go, go' and nothing more to someone cheering for the TV hosts to someone cheering for a party. This could be a curiosities-story (P6).*

A small entertaining surprise does not necessary suffice on an election night though:

> *This is a story. It's witty, if Ferjelista [the ferry list] .. it's dead funny. 'Ferjelista is likely to have two representatives in Volda municipial council'. On an election night there are a lot angles, but on general grounds that is one witty piece of information, but it depends if we would have time to make it (P3).*

The same participant identified a situation where a politicians' (Oddny Miljeteigs') gesticulation on TV was freely interpreted by the Twitter audience as sign language and entertainingly spread as simultaneous interpretation.

As an exception to light-hearted and soft-news-type findings one untold politics story of regime critic was identified: "Sad that deceased from the 22/7 are not removed from the ballots for AP (labor party) in Oslo. This is actually a big story" (P1).

User generated content (UGC) is found to add to the soft- and human-interest news in other studies where user can contribute directly to the media organizations (Mark Deuze, Bruns, and Neuberger 2007; Harrison 2009). My findings indicate that journalists find the same kind of news categories also when the data isn't handed in to them, but when they are looking for stories in UGC.

Finding Sources

The facts that debates on Twitter allows for anyone to share their ideas and comments is received with a certain ambivalence by the informants. On the one hand the democratic aspects and openness for new potential voices are welcomed in warm terms. On the other hand, the need to identify "the usual suspects" is complicated by the share number of voices. Who the authors of tweets are, is of paramount importance.

> *You gain insight into what people care for, you do. And what parties they talk about. It is interesting that so many talk about Venstre [liberal party]. But if you are looking for sources - if you are to get hold of sources - I would not have picked some random person (P7).*

Consequently familiar logos, organizations and persons were appreciated. "It is very pleasant when they show what the stand for [points to an avatar with a political logo]" (P7). While authorities were identified with quotable insights, some positions can be excluding;

> *Thor Bjarne Bore is an editor in a newspaper, so there is no point using him as a source (P1).*
>
> *There is a lot of official stuff here, that is boring. [Interviewer asks: Official?] Yes, official accounts [points to a profile of an NGO] (P3).*

Profiles of politicians and political parties were considered quotable, while the average Twitter user were approached with a very different caution, and whether they could be quoted was taken into a much more thorough consideration.

> *It is also a discussion of press ethics, to what extent we can… -it is something we discuss continuously – whether we can fetch material from social media and incorporate it into stories. I think that what people say on Twitter is public, I do (P5).*

The possibility to waste time and efforts on someone that didn't deliver a strong enough case was clearly expressed.

> *It is relevant to find who this is; I wouldn't use just any person sitting at home in his bedroom being angry at something. But this one there is a systems developer and information flow geek, liberalist, Dag B, a blogger. That means that he works with relevant things in this matter (P7).*

One participant had come up with a strategy to cope with this issue, and kept track of no-authoritative "potentials" by following them for a while before potentially making contact for a story. "It has happened a few times, so it shows that it is possible to find people that are sources, and not authoritative sources" (P6).

As sources, the non-professional commentator needs to have more than an interesting comment to fit the journalists' criteria: "the most important thing to separate is those who speak about something that happens in the media, and those who speak about something truly unique" (P1). Identifying who authors of Twitter texts are is important to a journalist for many reasons, independent of how clever or interesting the content of the tweet is. A key reason for evaluating tweets so sternly in context of their authors is the journalistic practice of accountably, related to public sources in checking for spin and with civilian sources in checking for competency, relations, credibility, etc.

Conclusion and Further Work

The approach presented in this paper has some limitations with regards to noise and ways to reveal a clusters' content quickly, but results also indicate that cluster analysis can aid the analysis of Twitter messages. The usefulness is related to the quality of the clusters, and the quality of the clusters is a matter of meeting the journalists' expectations.

Contribution: Improving clustering for tweets

Clusters of texts are considered good if they are immediately recognized as related to something the journalist already knows and there is a strong coherence within the cluster. Known entities should have a high priority both on order to examine these in particular, but also in order to exclude them. Ways to improve the clarity (signal) in the clusters are needed.

One idea for a better result is to crowd-source or manually construct the initial clusters before running k-means; another is to base the k-means only on data that contains named entities or other textual favorable features. Who people are is very important to the journalists in this study, and it is reasonable to assume this as an occupational trait. As such, the idea of finding an exceptionally clever insight or argument from anyone regardless of their position in society must come second after the journalistic need to anchor voices to positions and groups of people. This can be potentially be operationalized by basing the initial centroids on texts that contain such entities.

Other more crude approaches that exclude material could also be applied (to exclude texts based on the lack on machine recognizable entities, frequency of spelling mistakes, text length, etc.). This might increase the experienced utility, but adds a bias towards particular groups of authors and dims the democratic aspects of using social media as sources.

Elements that worked well in this study is the general idea of reducing the amount of units a journalist needs to examine to get an overview of twitter data, browsing by

metadata and using tf-idf to label and order sub-sets of data. The deliberate omission of the time aspect in the analysis worked without any problems in this study. The histogram that was offered for exploration obtained the chronological order and amount of texts in time. Improvements in the application include the ability to exclude data (from persons, hashtags, etc.), navigate by user types (politicians, media-people, organizations, etc.) and producing an even clearer signal.

In spite of the democratic promise of social media where anyone can participate, who the authors of messages are is very important to journalists. The "usual suspects" for journalists are people of power (politicians, organizations, celebrities and experts of various kinds), and in the category for "John citizen" they are looking for "cases"; persons that can exemplify a bigger phenomenon. Who the authors of tweets are were examined when tweets of interest were found and is too important for journalists too not include in the design of an analytics tool. Prior research projects have identified roles of Twitter authors though directories (An et al. 2011) and such lists could be used for filtering in an analytics tool for journalists. A similar approach was used to identify tweets that link to established media institutions (N. Diakopoulos, De Choudhury, and Naaman 2012) an approach that can be implemented to support the need to filter out opinions that is likely to be third hand information.

Acknowledgments

I will direct a special thanks to the NRK- team for valuable feedback throughout this project, and Matthew Weinstein for modifying TAMS Analyzer to support Scandinavian languages – an valuable improvement for doing the analysis of the data in this study.

References

An, J., M. Cha, K. Gummadi, and J. Crowcroft. 2011. "Media Landscape in Twitter: A World of New Conventions and Political Diversity." Proc. ICWSM 11.

Becker, H., M. Naaman, and L. Gravano. 2011. “Selecting Quality Twitter Content for Events.” In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM’11).

Becker, Hila, Mor Naaman, and Luis Gravano. 2010. “Learning Similarity Metrics for Event Identification in Social Media.” In Proceedings of the Third ACM International Conference on Web Search and Data Mining, 291–300. WSDM ’10. New York, NY, USA: ACM. doi:10.1145/1718487.1718524. http://doi.acm.org/10.1145/1718487.1718524.

De Choudhury, M., N. Diakopoulos, and M. Naaman. 2012. “Unfolding the Event Landscape on Twitter: Classification and Exploration of User Categories.” In Proc. CSCW.

Cutting, Douglass R., David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. “Scatter/Gather: a Cluster-based Approach to Browsing Large Document Collections.” In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 318–329. SIGIR ’92. New York, NY, USA: ACM. doi:10.1145/133160.133214. http://doi.acm.org/10.1145/133160.133214.

Deuze, Mark, Axel Bruns, and Christoph Neuberger. 2007. “PREPARING FOR AN AGE OF PARTICIPATORY NEWS.” Journalism Practice 1 (3): 322–338. doi:10.1080/17512780701504864.

Diakopoulos, N., M. De Choudhury, and M. Naaman. 2012. “Finding and Assessing Social Media Information Sources in the Context of Journalism.” In Proc. Conference on Human Factors in Computing Systems (CHI).

Diakopoulos, N., M. Naaman, T. Yazdani, and F. Kivran-Swaine. 2011. “Social Media Visual Analytics for Events.” Social Media Modeling and Computing: 189–209.

Diakopoulos, Nicholas, Munmun De Choudhury, and Mor Naaman. 2012. “Finding and Assessing Social Media Information Sources in the Context of Journalism.” In Proc. Conference on Human Factors in Computing Systems (CHI). http://research.microsoft.com/en-us/um/people/munmund/pubs/chi_2012.pdf.

Futsæter, Knut-Arne. 2012. “MedieTrender 2011” February 12. www.tns-gallup.no/arch/_img/9100748.pdf.

Harrison, Jackie. 2009. “USER-GENERATED CONTENT AND GATEKEEPING AT THE BBC HUB.” Journalism Studies 11 (2): 243–256. doi:10.1080/14616700903290593.

Hermida, Alfred. 2010. “TWITTERING THE NEWS.” Journalism Practice 4 (3): 297–308. doi:10.1080/17512781003640703.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. “Why We Twitter: Understanding Microblogging Usage and Communities.” In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, 56–65. WebKDD/SNA-KDD ’07. New York, NY, USA: ACM. doi:10.1145/1348549.1348556. http://doi.acm.org/10.1145/1348549.1348556.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. “What Is Twitter, a Social Network or a News Media?” In Proceedings of the 19th

International Conference on World Wide Web, 591–600. WWW '10. New York, NY, USA: ACM. doi:10.1145/1772690.1772751. http://doi.acm.org/10.1145/1772690.1772751.

Larsson, Anders Olof, and Hallvard Moe. 2011. "Studying Political Microblogging: Twitter Users in the 2010 Swedish Election Campaign." New Media & Society (November 21). doi:10.1177/1461444811422894. http://nms.sagepub.com/content/early/2011/11/21/1461444811422894.

De Longueville, Bertrand, Robin S. Smith, and Gianluca Luraschi. 2009. "'OMG, from Here, I Can See the Flames!': a Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires." In Proceedings of the 2009 International Workshop on Location Based Social Networks, 73–80. LBSN '09. New York, NY, USA: ACM. doi:10.1145/1629890.1629907. http://doi.acm.org/10.1145/1629890.1629907.

Loper, Edward, and Steven Bird. 2002. "NLTK: The Natural Language Toolkit." arXiv:cs/0205028 (May 17). http://arxiv.org/abs/cs/0205028.

Luo, Zhunchen, Miles Osborne, and Ting Wang. 2012. "Opinion Retrieval in Twitter." In Sixth International AAAI Conference on Weblogs and Social Media. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewPDFInterstitial/4592/5044.

Nezda, Luke. 2012. "ClusteringInDepth. Methods and Theory Behind the Clustering Functionality in Google Refine." Accessed March 6. http://code.google.com/p/google-refine/wiki/ClusteringInDepth.

O'Brien III, John. 2010. Your TwapperKeeper – Archive Your Own Tweets - Http://Your.twapperkeeper.com. http:// your.twapperkeeper.com.

Shamma, D. A, L. Kennedy, and E. F Churchill. 2010. "Conversational Shadows: Describing Live Media Events Using Short Messages." Proceedings of ICWSM.

Shamma, David A., Lyndon Kennedy, and Elizabeth F. Churchill. 2009. "Tweet the Debates: Understanding Community Annotation of Uncollected Sources." In Proceedings of the First SIGMM Workshop on Social Media, 3–10. WSM '09. New York, NY, USA: ACM. doi:10.1145/1631144.1631148. http://doi.acm.org/10.1145/1631144.1631148.

Stray, Jonathan. 2012. "The Overview Project." Accessed April 30. http://overview.ap.org/.

"The Oslo-Bergen Tagger." 2012. Accessed March 7. http://www.tekstlab.uio.no/obt-ny/english/index.html.

Tumasjan, A., T. O Sprenger, P. G Sandner, and I. M Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment." In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 178–185.

Weng, Jianshu, and Bu-Sung Lee. 2011. "Event Detection in Twitter." Proc. of ICWSM. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2767/3299.

IV. Watchdogging in code

Open government initiatives have led to the exposure of data on elected representatives' actions in parliament, allowing journalists to automate some watchdogging tasks - to track votes, questions and issues at the heart of our democracy. This study explores the scope of automating watchdogging and discusses the topic in reference to in-depth interviews with expert practitioners in parliamentary watchdogging, alongside the practical development of prototype software for watchdogging the Norwegian parliament. Prototype software has been developed and evaluated by parliamentary reporters, and ideas for future systems have been collected accordingly. Results from interviews suggest that given the distinctive workflow associated with political reporting, computational journalism represents an area that is expected to fulfill a useful role rather than be perceived as a threat.

Keywords: computational journalism, political reporting, computing, watchdogging

Introduction

The political reporter is described when the idea of the 'fourth estate' was introduced (Franklin et al. 2005), underlining that journalists should watchdog the powerful. First in line in a democracy stand the parliament and its representatives. Today digitization has caught up with the Norwegian parliament, as was demonstrated in the opening of an application programming interface (API) for some of its core data on elected representatives in 2012[31]. This continuously updated data includes questions asked, votes cast and relationships to working groups, counties and parties. In contrast to the parliament website, this API offers structured data in a machine-readable format creating a solid starting point for public inspection by means of computers. To political journalists, the API offers an opportunity to monitor data automatically, as well as to add computational steps for analyzing progress. "Automation of process and content is the most under-explored territory for reducing

[31] http://data.stortinget.no/

costs of journalism and improving editorial output", writes Anderson, Bell, and Shirky (2012). This paper presents a study in which an application to track parliamentary data was built and evaluated by expert political reporters, to explore this possibility in regard to political reporting. Results highlight great possibilities for computational journalism in parliamentary reporting, but also show how this practice can be problematic in terms of transparency.

Open data[32] initiatives spread quickly and terms such as 'gov 2.0' and 'open government' are used to underline the imperative for governments to be as transparent as possible. While open data can introduce new challenges to journalism (Sarah Cohen 2011), it mainly represents raw data that can offer insight into public institutions' recordings. Leading media institutions also stress the importance of transparency on new platforms (Karlsson 2010). The programming journalists' job includes efforts to uphold this transparency, from data source to story through code.

The important questions concern what and how to compute in order to maximize the usefulness of data to political reporters: What can we learn from the data? What analytical steps should be added? How can it be constructed so as to be useful over time? How transparent should it be? And, what do parliamentary reporters feel about introducing computing to their field?

The application

The application built for this project is a web application that can be inspected at www.samstemmer.net[33]. It routinely gathers data from the parliament API, calculates various scores and values for representatives, parties, etc. and displays the results online.

[32] Public open data are free from copyright, unpatented data produced and shared by public institutions.

[33] The application was awarded 1st place in the apps4norge programming competition. www.apps4norge.no

In designing the application some key factors were emphasized: The tool should be continually updated to offer utility over time, it should obtain facts stories can be based upon and add an analytical layer to the data to offer support for analysis.

Continually updated

News applications as a form of journalism have to date had limited success when creating continually updated systems (Karlsen and Stavelin 2013). While this is an idea put forth as an ideal or goal for such projects, in practice there are divers reasons why it rarely occurs. One is the short shelf life of online news; another is the lack of an organizational culture for maintaining projects over time. News applications as we know them today are designed to tell stories (Stavelin 2012), but if they were to be updated (preferably automatically) they would become tools of monitoring and research. APIs as data sources, in contrast to data dumps or web scraping, offer a stability and predictability that should make the extra effort worthwhile.

Fact-obtaining

A continually updated tool would to a certain extent take the story out of the journalists' hands. As the data changes, so would the stories. Thus, a fair aim would be a system that tracks and stores data providing the basis for stories without integrating a storytelling angle. This is an attempt to operationalize one of Terry Flews' hopes for computational journalism:

> *Ultimately the utility value of computational journalism comes when it frees journalists from the low-level work of discovering and obtaining facts, thereby enabling greater focus on the verification, explanation and communication of news. (Flew et al. 2011)*

In relation to parliaments, a number of such systems already exist, and features from projects such as the civic hacking project govtrack.us (Tauberer 2012), opencongress.org (OpenCongress 2013) and theyworkforyou.com (mySociety 2013) serve as references and sources for ideas. These projects are organized by non-profit NGOs. Some of the basic operations applicable to parliamentary data reveal similarities between journalists and NGOs, who try to share knowledge of the inner

workings of our democracy; however, a critical media organization is also likely to have requirements that NGOs do not to provide one-size-fits-all solutions to.

Facts on social issues are often a matter of debate. An example concerning the post-party owned media and parliament politics is the matter of whether or not promises are kept by political parties. It is not always obvious when reading parliamentary bills whether they contradict any promises. One Norwegian non-political organization, *holderdeord.no* ("do they keep their word") aims to do just this: fact-check promises against parliament votes. The organization's source of inspiration is political scientist John Keane and his idea of a *monitory democracy* (holderdeord 2013). Keane notes the change from a post-1945 representative democracy to a contemporary monitory democracy in regard to journalism:

> *The change has been shaped by*
> *a variety of forces, including the decline of journalism proud of its commitment to fact-based 'objectivity' (an ideal born of the age of representative democracy) and the rise of adversarial and 'gotcha' styles of commercial journalism*
> *driven by ratings, sales and hits (Keane 2009).*

One does not have to agree with Keanes' view of journalism to argue that media organizations should monitor parliament. This leads to the question of to what extent current political journalists are ready to outsource parts of their fact-obtaining activities to NGOs/non-profit organizations or software engineers or indeed be in competition with them.

Analytical

An analysis can arguably not be automated since it includes interpretation and understanding, and the extent to which that can be computed is a matter of philosophical debate in artificial intelligence. However, methodological steps developed and utilized in academia can be (and frequently now are) computer-supported, so that the time required from inputting raw data to results for interpretation is significantly reduced. A system monitoring parliamentary data for a media organization could and should benefit from this fact and include relevant methods from the field. In political science spatial maps are one such method (Poole

2005) which, following rigorous testing with different data sources and researches (Poole and Rosenthal 2011), should be ready to deploy in Norwegian newsrooms. To aid interpretation of data, visualization if often applied as we process images quicker than words, and visualization can aid in analytical task such as pattern detection, comparisons, anomalies, etc. Visualizations are not analysis; it is a method for aiding the interoperation of data. An analyst would choose what forms of visualization he thinks best would help answering his questions, parallel to how he would choose metrics to analyze a phenomenon. To formulate methods and visualizations in code as an automated process, and present this to an analyst spares the analyst from choosing data, methods and visualization type, but inscribes the programmers ideas of appropriateness for these choices.

Method

This study combines two aims. One is to collect knowledge concerning what an automated parliament watchdogging information system should be, according to experts in parliament reporting. The second is to obtain insight into how they experience the emerging technological possibilities in their field. Both ends are perused by the same methodological procedure: a design science approach (A. R. Hevner et al. 2004) with semi-structured interviews as tool for evaluation and explanation.

Design and implementation

The normal chronology in software engineering is to collect requirements for a system in advance of writing any code or producing any prototypes or solutions. This study reverses that order. An experienced parliamentary reporter in Norway, the professional we want to support is typically an older man holding a degree in subjects such as history or political science (Allern, 2001); confronting him with a printout of the central parts of the data from the API would presumably result in a blank face. It is much easier to critique and discuss possibilities when something tangible, clickable or graphical is presented. I therefore created ways to display the data from the API as graphs, lists and tables, made the system automatically update itself when new data is

uncovered and developed a few models for journalistic angles to explore and discuss. Much inspiration for the design was found in projects already mentioned: some ideas emerged by exploring the data, some journalistic hypotheses were borrowed from news media, found through searching in the Retriever media database[34], and some ideas were found in collections of information visualizations.

The prototype is written in Python, based on Django[35] (Django Software Foundation 2010), utilizes d3js (Bostock 2012) for visualization, and reuses code from academia for some computational methods, such as the optimal classification implementation in R (Poole et al. 2012), the Norwegian version of the readability tool (Skardal and Jakobsen 2007) and the term frequency inverse document frequency implementation in the Natural Language Toolkit (Loper and Bird 2002).

The key difference between the prototype constructed in this project and most news applications in today's online news scape is the continuity of the data in the system. While most journalistic (as well as academic) projects are based on data that is collected once and then analyzed, the data in this project is updated and recalculated every time new material emerges. When a question is registered or a vote is over, the system includes the new data and adjusts accordingly. The system checks for new data every hour and collects any new items. The feasibility of this challenge is not in question; it works well as this is precisely what APIs are designed to do. The application has been running live without problems since March 2013.

To connect to an API the conceptual models in my system must match the models offered through the API. This part of the work took some time and required a good understanding of the domain and solid documentation. When the models match and data flows correctly, the true question follows: what can we learn from the data, what would a parliamentary reporter look for, what analysis can or should be derived from the data? These were among the questions that constituted the interview guide.

---

[34] More info at http://www.retriever-info.com/en/om-oss/

[35] Django "the web framework for perfectionists with deadlines" - a web framework with roots in the news industry

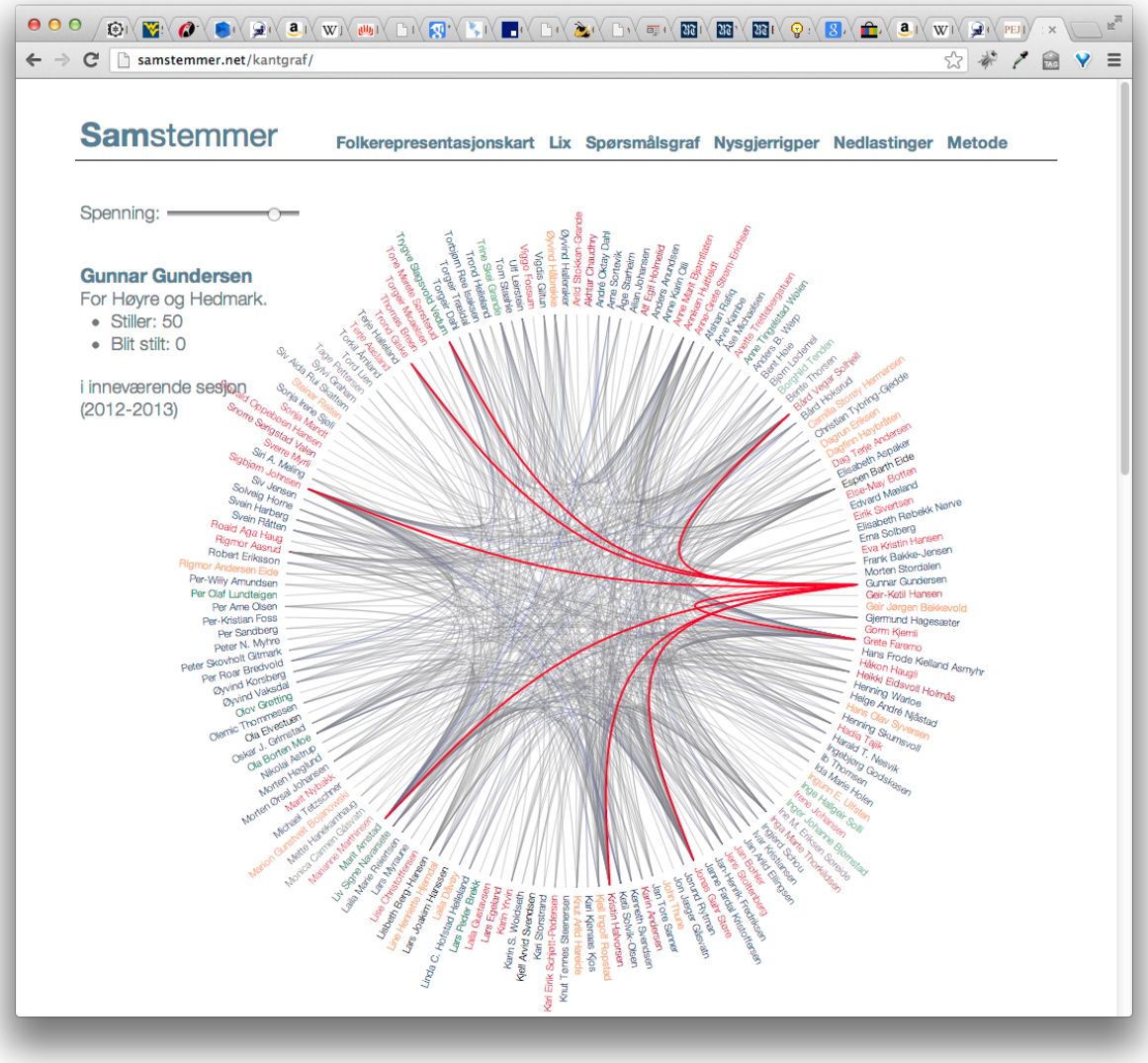


Figure 1. Screenshot of one view of data in the system. Interactive edge graph showing recipients and raisers of questions in parliament this term (2012-2013).

Interviews

Five in-depth semi-structured interviews were conducted between 16th and 18th April 2013. Two interviews were held in the parliament building, where the journalists had access to office space, and three took place in quiet offices adjoining the main newsrooms to which the journalists belonged. The participants were unpaid. The interviews lasted approximately 45 minutes with a median length of 46 minutes. The session consisted of a short introduction to the system with the participant in front of a laptop to ensure they had a reasonably good understanding of the system, and a set of mostly open-ended questions. The URL to the application was given to the

participants in advance of the interviews for them to know some of the system in advance of the interviews. The interviews were recorded, transcribed and analyzed using TAMS Analyzer, a tool for computer-assisted qualitative data analysis (Weinstein 2006). The transcribed interviews were coded in accordance with the interview guide, as well as with themes that emerged from reading the transcribed texts.

The participants were chosen for experience (median work experience with parliamentary reporting was 15 years) and connection to central newsrooms. All are members of the Parliament press gallery (Stortingets Presselosje), from where their names were picked as key reporters by a fellow researcher who had previously worked as an advisor at the parliament. The journalists belonged to 1) the public national broadcaster *NRK*, 2) the commercial public broadcaster *TV2*, 3) the number one online and number two printed newspaper *Verdens Gang*, 4) the number one printed newspaper *Aftenposten*, and 5) a *freelancer* with prior affiliation to several major newsrooms. The journalists produced content for all platforms (web, TV, radio and newspaper).

Results & analysis

The results from this study consist of the initial prototype (samstemmer.net), lessons learned from the design work and the analyzed data from the interviews.

The journalist is in the (coding) details

One thing that became clear while programming samstemmer.net is that choices that largely steer both the truthfulness and perception of the results are in the coding itself. Assumptions, emphases and omissions are embedded in the code. This constitutes an exact parallel to how tone and vocabulary in written text is telling of positive or negative attitudes to aspects, arguments or perceptions in any journalistic story. The crafting of code is the crafting of a frame that sets boundary for the data that is processed and displayed and the prevalence and priority of this premade arrangement for an analysis. It is too a matter of "selection and salience" (Entman 1993). Such

frames can also distort the interpretations of the data, for instance by bending the graphical integrity of a visualization. Edward Tufte refer to this as the "Lie Factor" of a visualization (Tufte 2001). The framing of the data can also be done without altering the data values, but by attributing the values with various emphases. Mapping color to graphical elements can be done by separating the color space in different ways, discriminating or favoring some variables, by creating larger contrast, weight, visual space etc. The fact that the code is run without the analyst (the journalist in this case) reviews the code in advance, also makes it possible to directly tamper with the analysis (outside the frame), e.g. by adding exceptions that removes or alters particular entities variables.

Results from interviews

As the interviews were conducted after a prototype was constructed, central parts of the results are simply ideas, improvements and feedback on the current system. These features and improvements are included as a list of unimplemented ideas together with the source code for this project and can be found on github[36].

The second level of results include the journalists' views on automation and computation in relation to their work, the possible competition from other interested parties on open data and insights on whether a tool such as this should offer different views to reporters and the public.

The majority of the visualizations and analysis offered in the prototype, confirmed insights or impressions the reporters already had although they could not always pinpoint the precise source. They could explain and relate to most of the features, typically responding that the data did not include many surprises, and for instance quickly assessed who asked most questions of whom in Parliament:

> *There are no surprises here, it is typical that he is a member of the topical committees against the ministers. - P2*

[36] https://github.com/eiriks/samstemmer#ideer-og-innspill-fra-erfarne-norske-stortingsreportere

In spite of many such realizations that the open data from the API confirmed knowledge these reporters already had, four of the interviewees identified something in the data they considered to be a story. As a technological approach to journalism, none of the reporters had produced products such as this. Except for the freelancer, they all knew whom in their newsroom they could talk to in order to produce systems of this kind.

Attitude towards computation in political reporting

The inaccessible nature of software code makes it hard to see exactly what methods and conditions that have been applied to produce the data on display. This did not worry the reporters to any significant degree. They could all list examples where this is standard practice and experts, organizations and special interest groups presented facts that the reporter had little chance (or methodological interest) of double-checking for validity or methodological implications worth mentioning.

> *I see no problems with it. I do, however, see that you can get very interesting stuff from it. – P2*

> *My gut reaction is of course that this is terrible. Right? - Because everything that is automated is too simple. But I think it can be useful, but it is as with all tools, they need to be used right. – P1*

The perspective that an algorithmic approach is useful to political journalists and that it opens up new possibilities was shared among all the participants. Finding positive sides was considerably easier than identifying negative ones. The second quote above was the only truly skeptical voice. Computing is not seen as "threats to be subordinated" but almost purely "as possibilities for journalistic reinvention" (Powers 2012). Few nuances emerged in relation to utility, but the notion of computing and journalism merging and computing potentially being a part for the journalism was met with little support.

> *This is very useful; it is an extremely useful tool that can never replace journalistic craftsmanship. So I'm not afraid journalism - classical journalism – will disappear because one uses computational methods. I think it opens up new possibilities - new fields to explore – but you can't get around that you at one point need to go in and make the hard journalistic assessments in relation to 'what is a story' – P3*

Algorithms are not journalism and cannot replace journalism, that was the theme. The reason why the journalists drew the line here was further explained and is presented in the following section.

Journalism is more than systematizing data

> *As soon as you talk assessment, then all comparison of data is much more complicated than numbers and facts. - P2*

> *No, as long as you use it for what it is, systematized raw data as we have here is not in itself journalism. It becomes journalism only when you put it into context. - P3*

Facts alone do not constitute journalism. The journalist reads, interprets and finds a suitable context for the facts so that they become journalism. The dividing line between raw data as facts and numbers and what the journalists considered their role was clear: systematizing raw data is not journalism.

This has perhaps more to do with the specialized workflow of political reporters than any theoretical description of a generalized journalistic process. It is easy to argue that systematizing data is indeed a key part of any fact-producing process and thus a cornerstone of journalism.

Further clues as to why computing falls outside of the journalists' idea of political reporting follow in the descriptions of their workflow.

Key sources and methods

To figure out how the participants work from idea to story, I asked about key source types and methods they frequently utilize. Some elements of the workflow can perhaps be supported by aid of computers?

> *In all future political journalism there will be a great deal of chemistry, person-based, oral sources, the management of trust concerning this, leaks, interviews – it cannot be replaced. –P4*

The role of the interview and other oral sources as key methods for political reporters was echoed in the material and falls in line neatly with prior descriptions of Norwegian parliamentary reporting (Allern 2001; Eide 1984). The reporters do not act as autonomous "investigators" in this regard, they also represents a channel for various interest groups to get their word out. The gallery of sources incudes a wider arsenal that what the data API contains.

> *It is personal relations with politicians or others – other types of people, not only politicians. From organizations, from the bureaucracy, from very different types of jobs too, they come with suggestions and tips. Nine out of ten stories start this way. – P5*

Numbers were suggested several times. Eighty percent of the time oral sources are the key sources; nine out of ten stories are initiated from outside, etc. Computing will presumably never substitute the interpersonal relationships so many of the stories depend on, and this is perhaps why the feedback on the lack of threat computing poses was so strong and united. The political reporters saw their methods as not computable. A third key method also became apparent:

> *What we often do is to develop hypotheses – based on our knowledge of the central actors – that is how we often work. – P3*

The confirmation or invalidation of a hypothesis, with comments from both political wings, is often used as a story. The prototype includes examples of such projects where data is arranged to answer specific assertions and allegations. The non-interview sources the journalists claimed to use, included text documents and websites, particularly the parliament's own website where the minutes and verbatim transcribed accounts of parliamentary meetings are found. It is at this end that computing and automation can best be applied. A good hypothesis that can be answered or invalidated by data, and automated through a continual process of updating will not only aid in the production of one story but potentially multiple stories, as the data and potentially the answer to the question change. An example can

be a party's stand in a political issue, witch actors who are prominent as letter-writers or posters of questions, or the distribution among political fields covered in a time period.

Computing for or by journalists?

When the deadline is due, it is so that the audience will receive new information served at the next dispatch. The audience of media organizations can be seen as the end-users of a system such as samstemmer.net. Alternatively, news organization staff, the reporters, can be considered as constituting the end-user. The potential difference in requirements will identify what aspects the journalists hold important to themselves in contrast to their audience.

The potential different systems would likely differ similarly to how information system often have advanced features for advanced users, features that can be turned on and off. In a journalistic context one could also imagine the possibility to "freeze" the system (, or a part of the system), to include in a news story, a status quo at a certain point in time.

> *It is a hard question. In theory there should not be one [difference between internal/external tool], because the parliamentary reporters role is to tell the audience of the important events that occur at the parliament, and so the important things he can see on such a dashboard. But at the same time it is clear that it is very useful information to us when we do journalism, we constantly navigate according to old knowledge – what we know of different politicians, the different parties, where to expect conflicts. Who we can expect to oppose an issue and who will support it. Who the allied parts are. And all this is information the audience necessarily does not need to know, but we do to know where to look for stories. So, I would say that a dashboard would be useful to us, but also have different significance and use for the audience. –P1*

> *Perhaps it is. But I mean that in principle it's worrying, if I may [say that]. The information should be equal to all at any given time. – P3*

> *Ideally I would like more options. Ideally I would – for internal use – have as much data as possible, with all imaginable possibilities to combine data. But I think for external use, the more data you offer and more choices you offer – the more you can confuse the majority of users. –P2*

The principle that the journalists and audiences should have access to the same information was one concern, while the interest for as much information as possible as specialist in the field often is too much for their audience was another. The acknowledgement that what they know and care for of politics regularly exceeds what is actually transmitted was also noted. If the system were to be made publically available, it should be offered as a special interest service because, again, it is the filtering of information through journalists that creates journalism. The service itself might not be considered journalism, in spite of containing elements of a journalistic workflow.

In summary

No initiatives to make use of the data were done in newsrooms. While some had discussed them internally, none had made any efforts to connect to the API and analyze the data. This kind of approach was seen as very useful and promising and does not represent a threat to the parliamentary reporter, but represents new fields and opportunities to explore. The limited transparency in software was not perceived as any more problematic than many other sources lacking opacity, and while journalists and their audience in theory should have access to the same information, the journalists interviewed believed their information requirements exceeded general audience expectations. Computing, or the systematizing of raw data, was seen as clearly separate from journalism that filters such facts and places them in a context.

Discussion

A tool such as samstemmer.net is not neutral. Although it treats all the representatives with the same methods and displays the data to all visitors in exactly the same way, the programming decides the outcome in much the same way as choice of

terminology and angle does in traditional articles. Determining which methods are explained and revealed or the ways in which values are computed can shift the interpretation of the results. Choices that decide in which direction a story will go are written in code. These are decisions that are often described as “journalistic”, as they define the frame in which information is relayed to the audience. Presented in code, these choices become less accessible and visible to journalists if the journalists themselves do not write it.

The “bias” of the system can be underlined with aid from framing theory. The theory describes how a frame is the product of selecting some aspects of perceived reality and make the more salient in a text (Entman 1993). This is an inescapable framework as something always is first, biggest, clearest or last, smallest, most unclear, also in computer systems. In online news “above the fold” is described as the pars of the webpage that is visible without scrolling down, a digital equivalent to a folded newspaper. Software also has these default or initial views. This is one frame, the inclusion/exclusion from this space. Further the salience of information on the screen, type size, contrasts, zoom level or numerical range in a graph or map, all are example of indication of importance in the visual space. This are the result of many different types of choice, they can be journalistic assessments, but are likely also influenced by programmers, graphical designer, etc.

The communicating text (in my case the prototype) contains these choices of inclusion and salience. The resulting focus the various elements of the prototype gets can easily be read in the server logs that records what elements gets clicked on the screen. A redesign of the “home” screen will affect the distribution of clicks among the clickable elements. In a computer system the visual display on screen is one part of the system where this is observable and also usable as rhetorical effects (Hullman and Diakopoulos 2011), but also the code can be analyzed to identify how and why some data gets displayed and some do not. While the data from the API is ‘raw’ in the sense ‘not processed by the prototype’ when it is collected, it indeed is processed when scores are computed, rankings made and other variables attributed through the

analytical layer the system creates. Loads of choices that possibly influence the interpretation of the data are made even before it is viewable on screen.

The communicators, the authors of the system, see the world though their own frame onto the world. In good software this frame is well aligned with the receivers frame, in this case the journalists analyzing data from the parliament. This is one reason why the users of a system should help design it. The receivers view is accessible through the frames in the texts and the frames of its authors. Further, the text is both created and interpreted in a social cultural context. A journalistic conflict narrative can inspire a gender battle in what other see as a simple minor vote in parliament.

Journalism has rules and guidelines for accountability, such as the inclusion of opposing voices in a text, editorial approval before publication, and the acknowledgement of sources. Conforming to these rules is incorporated in the journalistic process, yet programming as journalism does not feature such systematic checklists, and the transparency of the work delivered is limited. Editors will have to approve a text before publication, indirectly also approving the frame the text provides. Editors typically do not read source code and the audience is unlikely to have the opportunity to do so. This creates uncertainty of exactly what the data we see is, and how it has been treated. To stay accountable someone should be able to reproduce what the system displays, if not the editor or audience by themselves, someone else should aid them. This is in direct parallel to how the scientific community strives to uphold reproducible results in computational research (Stodden, Guo, and Ma 2013).

To illustrate:

In this project the distance between all members of parliament was computed based on votes. This makes it possible to find people guaranteed to oppose each other. The array of possible distance measures includes at least a dozen more or less suitable measures. I opted for percentage as the simplest solution. I compared the votes between two arbitrary representatives and counted every time they voted the same over all the votes.

Alternative 1) As expected, this gives a percent distance where 100% is identical voting. As the parliament has a substitution system, not all representatives have to vote on all the votes (as long as those present represent the distribution of the parties and the number of votes exceeds 50% attendance). Two representatives not being present at the same time, but rarely agreeing, would end up very close. Not being present together is not agreeing.

Alternative 2) is to only use the votes where the first of the two MPs are present. This also creates odd results when some MPs hardly ever vote. And the distance from MP1 to MP2 is not necessarily the same as the distance from MP2 to MP1.

Alternative 3) is created by choosing only the votes where both MPs were present and cast their vote. This might seem as an obvious methodological choice, but as three produce quite different results (for comparisons with MPs with high records of absence) it matters because the methodology is written in a language not easily accessible to most journalists and editors and inaccessible for the audience.

The three results would constitute a fact obtained by the system, but the results would vary depending on the version chosen. These kinds of pitfalls were not on the interviewed reporters' radar, similar to how building such systems were not. This suggests that fact-checking is a potential problem in computational journalism; as for providing tools, "they need to be used the right way", as one participant put it. They also need to be constructed the right way, and journalists should ideally know the inner quirks of the system, if not build it themselves. Editors will at times need someone to explain how computed results are made, is order to truly argue that accountability is upheld.

Another question relevant for building journalistic system on APIs as data sources is whether or not the design of the API influences the kind of watchdogging that can reasonably be done. To start with I assume the intent of the API is to honestly expose data, as opposed to a presenting a selective rosy picture of the institution, and that this functions as a more effective way of dealing with data that likely would be demanded through freedom of information queries. The API creates another frame in witch

some selected parts of the activity is described. The next hurdle is the one-liner '*"raw data" is an oxymoron'* offered in a anthology with the same name (Gitelman 2013). Data exposed through the API of the Norwegian Parliament might be seen as raw data, but it is created in a particular context, for purposes that is likely very different from what a journalist intends to use it for. It is not given that what the journalist believes that data describes in fact is correct in the perspective the journalist creates. Such issues should naturally be possible to unravel through the documentation of the API, but the 'rawness' of the data is a point worth noticing as it raises the question of the reasonable limits for data-reuse.

In contrast to many APIs from social media platforms that exposes data as streams of data ordered by time where the newest data comes first (e.g. Twitter, Flickr, etc.), the parliament API offers a quite normalized view of entities that reflects a relational database. While the API does include the foreign keys to other tables (or API endpoints in this case), it also returns redundant basic information about the key. This makes it easy to display who votes what in a parliament vote, but harder to assess what length of service, position/opposition status, or other features the MPs have, since these richer descriptions are exposed in other parts of the API (other endpoints). The kind of question the interviewed reporters wanted to ask (e.g. is the opposition voting more or less in tune that the position?, how many days left do the MPs have to achieve the parliaments' (famously generous) pension?, does gender divide the parliament hall different that the parties?, etc.) are possible to answer, but not through one endpoint. Thusly the journalist must collet all necessary data, normalize it, and then start the inquiry. In this light a fully normalized mirror of the database would be better to journalists. Very few of the efforts, both implemented in the prototype and requested by the reporters, was possible to answer simply by data from a single endpoint. To conclude, the variation and unpredictability of the questions journalist ask suggest that the API is modeled to be as true to the database it represents, so that the journalist can query their databases as if it was a true copy of the original.

To approach political journalism as a design problem aligns poorly with the idea of a journalist as a generalist since the skills involved are highly specialized. There is also

a mismatch with the traditional skillset of political journalists in Norway, who typically studied history or political science (Allern 2001) with little or no focus on technology. Moreover, the interpretation of data and data visualization requires literacy outside of traditional journalistic skillsets.

> *Anyone can construct a spatial map using computer programs I discuss in subsequent chapters. But the maps are worthless unless the user understands both the spatial theory that the computer program embodies and the politics of the legislature that produced the roll calls. A practitioner must be able to stand before an audience of her peers and explain the meaning of the spatial map. (Poole 2005)*

Specialists with the types of skills mentioned by Poole, such as quantitative theory and politics, are required to read the type of maps his method produces. It is not far removed from the education of political reporters, but does not align well with the work process he was exposed to while working, where personal contact with informants and inside tips constituted most of the sources for his stories.

While the "old school" of parliamentary journalists constitute the inner circles of expertise in their field, they are unlikely to represent any dramatic change in work practices. They should be consulted to inform future efforts, but it is more likely that younger, more technologically advanced successors will take this kind of journalism to our media screens. This papers' main contribution is the discussion of accountability in software as journalism, as the gains of computation also can represents a journalistic setback in transparency.

References

Allern, Sigurd. 2001. *Flokkdyr På Løvebakken? - Sigurd Allern - Innbundet (9788253023168)*. http://www.bokkilden.no/SamboWeb/produkt.do?produktId=120881.

Anderson, C.W., Emely Bell, and Clay Shirky. 2012. "Post Industrial Journalism: Adapting to the Present". Columbia Journalism School | Tow Center for Digital Journalism. http://towcenter.org/research/post-industrial-journalism/.

Bostock, Michael. 2012. "D3.js - Data-Driven Documents." *Data-Driven Documents*. http://d3js.org/.

Cohen, Sarah. 2011. “Shared Values, Clashing Goals: Journalism and Open Government.” *XRDS* 18 (2) (December): 19–22. doi:10.1145/2043236.2043246.

Django Software Foundation. 2010. “Django | Writing Your First Django App, Part 1 | Django Documentation.” 06. http://docs.djangoproject.com/en/dev/intro/tutorial01/#activating-models.

Eide, Martin. 1984. *Etter det vi forstår på politisk hold--: politikere og massemedia*. Universitetsforlaget.

Entman, Robert M. 1993. “Framing: Toward Clarification of a Fractured Paradigm.” *Journal of Communication* 43 (4): 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x.

Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift. 2011. “The Promise of Computational Journalism.” *Journalism Practice* 6 (2): 157–171. doi:10.1080/17512786.2011.616655.

Franklin, Professor Bob, Martin Hamer, Mr Mark Hanna, Marie Kinsey, and Dr John E Richardson. 2005. *Key Concepts in Journalism Studies*. Sage Publications Ltd.

Gitelman, Lisa. 2013. *Raw Data Is an Oxymoron*. MIT Press.

Hevner, A. R., S. T. March, J. Park, and S. Ram. 2004. “Design Science in Information Systems Research.” *Mis Quarterly* 28 (1): 75–105.

holderdeord. 2013. “Holder de Ord.” July. http://www.holderdeord.no/home/faq#har-dere-en-faglig-inspirasjonskilde.

Hullman, Jessica, and Nick Diakopoulos. 2011. “Visualization Rhetoric: Framing Effects in Narrative Visualization.” *IEEE Transactions on Visualization and Computer Graphics* 17 (12) (December): 2231–2240. doi:10.1109/TVCG.2011.255.

Karlsen, Joakim, and Eirik Stavelin. 2013. “Computational Journalism in Norwegian Newsrooms.” *Journalism Practice* (July 23): 1–15. doi:10.1080/17512786.2013.813190.

Karlsson, Michael. 2010. “RITUALS OF TRANSPARENCY.” *Journalism Studies* 11 (4) (August): 535–545. doi:10.1080/14616701003638400.

Keane, John. 2009. *The Life and Death of Democracy*. New York: W.W. Norton & Co.

Loper, Edward, and Steven Bird. 2002. “NLTK: The Natural Language Toolkit.” *arXiv:cs/0205028* (May 17). http://arxiv.org/abs/cs/0205028.

mySociety. 2013. “TheyWorkForYou: Hansard and Official Reports for the UK Parliament, Scottish Parliament, and Northern Ireland Assembly - Done Right.” Accessed June 24. http://www.theyworkforyou.com/.

OpenCongress. 2013. “About Open Congress - OpenCongress.” Accessed June 24. http://www.opencongress.org/about.

Poole, Keith, Jeffrey Lewis, James Lo, and Royce Carroll. 2012. “CRAN - Package Oc.” *Oc: OC Roll Call Analysis Software*. January 24. http://cran.r-project.org/web/packages/oc/.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Cambridge University Press. http://www.google.com/books?hl=en&lr=&id=OmeQNHvcULoC&oi=fnd&p

g=PR11&dq=spatial+models+of+parliamentary+voting&ots=1HfFZmjfQi&sig=J7ah1qtkIyi1LcYQdv2W5Z3O-jo.

Poole, Keith T., and Howard L. Rosenthal. 2011. *Ideology and Congress*. Transaction Publishers.

Powers, Matthew. 2012. "'In Forms That Are Familiar and Yet-to-Be Invented' American Journalism and the Discourse of Technologically Specific Work." *Journal of Communication Inquiry* 36 (1) (January 1): 24–43. doi:10.1177/0196859911426009.

Skardal, Thomas, and Thomas Jakobsen. 2007. "Readability Index." http://www.mortengoodwin.net/publicationfiles/webmining_2007_reportgroup6.pdf.

Stavelin, Eirik. 2012. "Nyhetsapplikasjoner: Journalistikk Møter Programmering." In *Nytt På Nett Og Brett: Journalistikk i Forandring*, edited by Martin 1956- Eide, Leif Ove 1961- Larsen, and Helle 1977- Sjøvaag, 107–125. Oslo: Universitetsforlaget.

Stodden, Victoria, Peixuan Guo, and Zhaokun Ma. 2013. "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals." *PLoS ONE* 8 (6) (June 21): e67111. doi:10.1371/journal.pone.0067111.

Tauberer, Joshua. 2012. *Open Government Data: The Book*. http://opengovdata.io.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Graphics Press.

Weinstein, Matthew. 2006. "TAMS Analyzer Anthropology as Cultural Critique in a Digital Age." *Social Science Computer Review* 24 (1) (February 1): 68–77. doi:10.1177/0894439305281496.

Screenshots

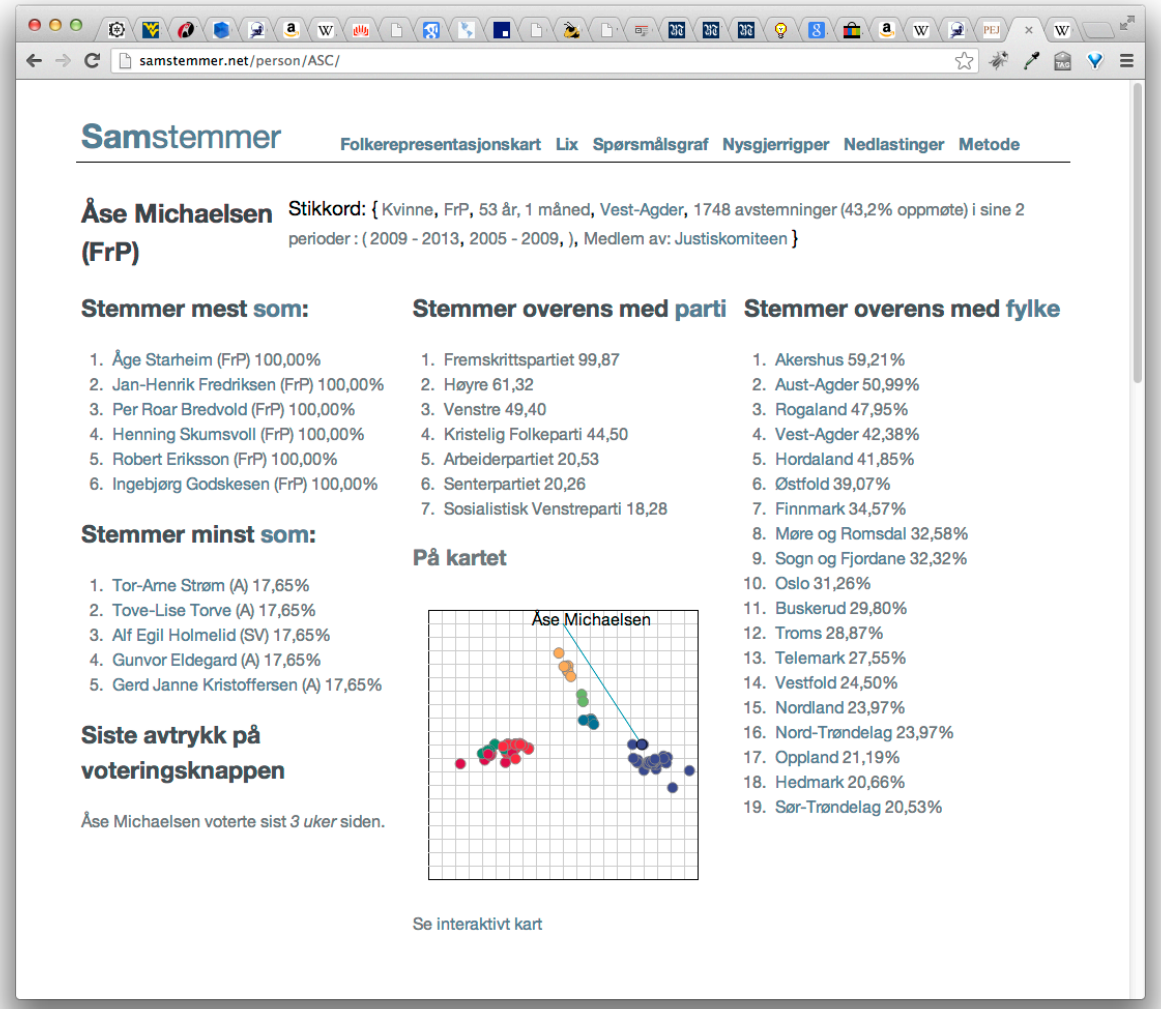


Figure 2: Detailed page of individual Member of Parliament. The left column shows persons voting the most and least similar, the center column shows voting similarities for the median vote of each of the seven parties in parliament (top) and the position on the optimal classification map (bottom).

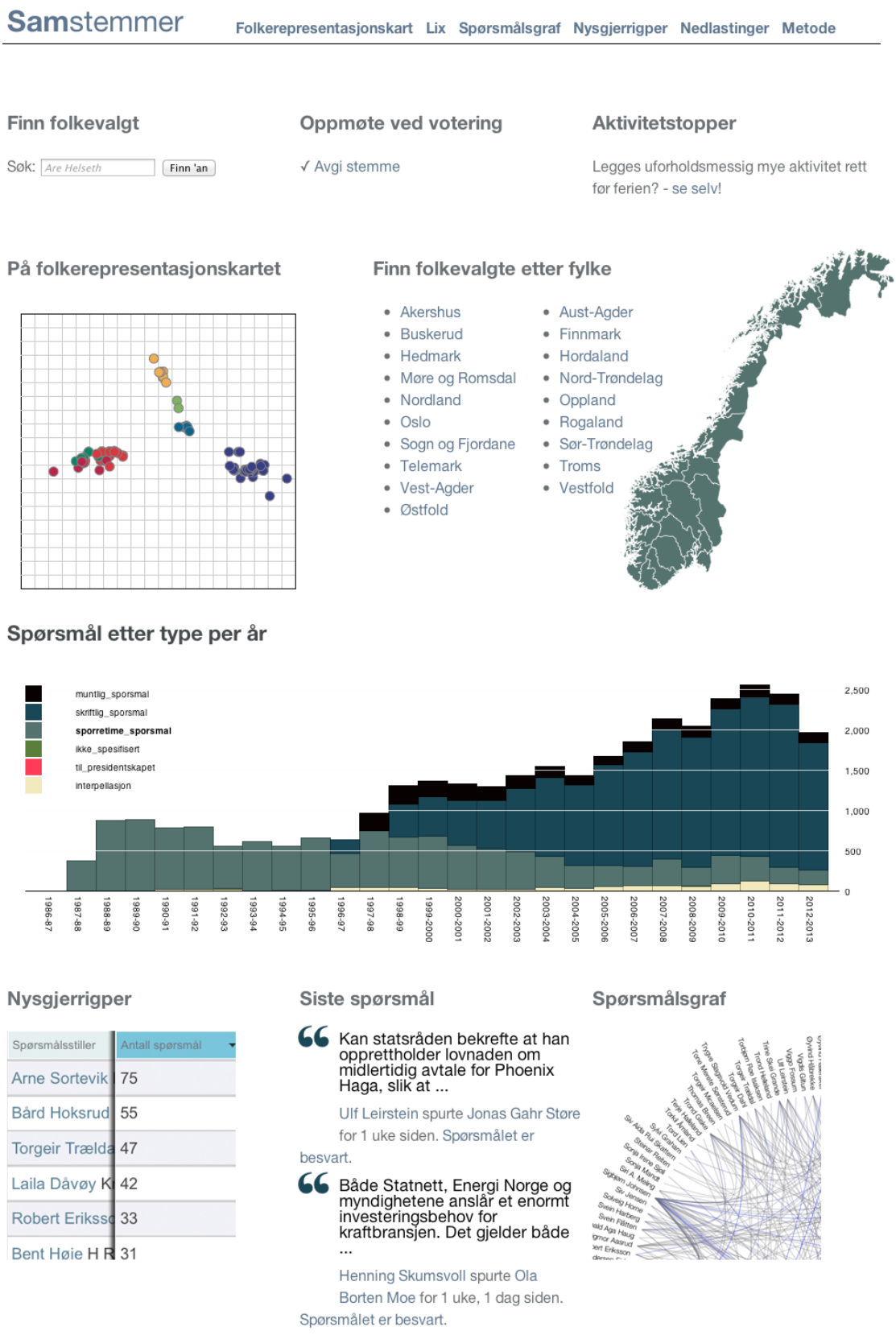


Figure 3: Front page of samstemmer.net, allowing multiple entries into data from the Norwegian parliament.