Aalto University
School of Business

Data Journalism: An Outlook for the Future Processes

Information Systems Science

Master's thesis

Minna Rapeli

2013

Department of Information and Service Economy
Aalto University
School of Business

AALTO UNIVERSITY SCHOOL OF BUSINESS ABSTRACT
Department of Information and Service Economy 31.05.2013
Master's Thesis
Minna Rapeli

ABSTRACT

Data journalism is a rather new format of reporting news stories online, which utilizes open data as the basis of stories and illustrates findings visually. The data journalism creation process and the way the information is being presented are very different to that of traditional print journalism. The objectives of this research are to identify current data journalism creation processes, forecast the future process outlook and to discuss journalists' role in the process.

There has been very little academic research conducted in the field of data journalism previously. Therefore, the literature review within this thesis has a wide scope and is explorative in nature while the findings serve as the basis for creating a future model for data journalism creation processes. The viability of the created future model has been tested by interviewing three Finnish data journalism professionals.

Suggested streamlined process model for the future of data journalism proposes that in the future the two most work-intensive data journalism creation phases, data manipulation and story visualization, will be outsourced to organizations specialized in these phases. This model is generally accepted as a viable future scenario by the professionals interviewed. However, there is some concern about communication between journalists and the institution to which the data journalism creation process is being outsourced to. This notion of the importance of communication prompted the creation of two more models that take communication more into consideration: (1) data journalism consultancy model and (2) outsourced chain model for data journalism. Together these two models suggest four approaches for producing data journalism in the future, these are: (a) an in-house data journalism team, (b) partner up with an organization specialized in creating data journalism and outsource the whole data journalism creation process except story creation to it, (c) outsource each phase to a separate phase provider and (d) a hybrid model of the first and last mentioned approach. The role that journalists will have in the data journalism creation process of the future will alter according to which the model media organization decides to adopt.

Keywords

Open data, in-house data journalism team, programmer-journalist, data manipulation, information visualization

AALTO-YLIOPISTON KAUPPAKORKEAKOULU TIIVISTELMÄ

Tieto- ja palvelutalouden laitos 31.05.2013

Pro Gradu-tutkielma

Minna Rapeli

TIIVISTELMÄ

Datajournalismi on melko uusi verkkoympäristössä julkaistavien uutisten raportointiformaatti, joka käyttää avointa dataa uutisten perustana ja esittää lopullisia löydöksiä visuaalisessa muodossa. Datajournalismin tuotantoprosessi ja se miten tieto esitetään eroaa paljon perinteisestä printtijournalismista. Tämän tutkielman tavoitteena on kartoittaa tämän hetkisen datajournalismin tuotantoprosesseja, ennakoida prosessien tulevaisuudenkuvaa ja tutkia journalistin roolia koko prosessissa.

Datajournalismista on toistaiseksi tehty hyvin vähän akateemista tutkimusta. Tästä johtuen tämän tutkielman kirjallisuuskatsaus on laaja-alainen ja toimii tulevaisuuden prosessimallin perustana. Tutkielmassa esitetyn hypoteettisen mallin toteuttamiskelpoisuus testataan haastattelemalla kolmea suomalaista datajournalismin asiantuntijaa.

Esitetty modernisoitu tulevaisuuden datajournalismin prosessimalli ehdottaa, että tulevaisuudessa datajournalismin kaksi työläintä vaihetta, datamanipulaatio ja uutisen visualisointi, ulkoistetaan näihin vaiheisiin erikoistuneille yrityksille. Haastatellut datajournalismin asiantuntijat kokivat ehdotetun mallin mahdolliseksi tulevaisuudenkuvaksi.

Asiantuntijoiden mukaan malliin liittyy huoli journalistin ja ulkoisen yrityksen välisen kommunikaation onnistumisesta. Prosessimallin saama palaute kommunikaation tärkeydestä johti kahden uuden mallin luomiseen, joissa kommunikaatio on otettu paremmin huomioon. Mallit ovat (1) datajournalismin konsultointimalli ja (2) ulkoistettu datajournalismin ketjumalli. Nämä kaksi mallia ehdottavat neljää tapaa luoda datajournalismia tulevaisuudessa, (a) organisaation sisäinen datajournalismitiimi, (b) muiden vaiheiden paitsi uutisen kirjoittamisen ulkoistaminen yhteistyökumppanille, joka on erikoistunut datajournalismin luomiseen, (c) eri vaiheiden ulkoistaminen eri organisaatioille, jotka ovat kyseiseen vaiheeseen erikoistuneita (d) ja hybridimalli, joka yhdistää ensimmäisen ja viimeisen tavan. Tulevaisuudessa journalistin rooli datajournalismin luomisprosessissa vaihtelee sen mukaan kumman mallin mediaorganisaatio ottaa käyttöön.

Avainsanat

Avoin data, organisaation sisäinen datajournalismitiimi, ohjelmoija-journalisti, datamanipulaatio, informaation visualisointi

ACKNOWLEGEMENTS

First, I would like to thank Professor Matti Rossi from the Department of Information and Service Economy for being my thesis supervisor and for all the comments, feedback and support throughout my research and writing process. In addition, I would like to thank Maija Töyry, a professor of magazine journalism from the Media Department, School of Arts, Design and Architecture for providing consultative assistance. Next, I would like to give my sincere thanks to data journalism professionals interviewed for this research; Data journalism producer Esa Mäkinen from Helsingin Sanomat, data journalism producer Juho Salminen from Finnish Broadcasting Company (Yle) and programmer-journalist Teemo Tebest from Finnish Broadcasting Company (Yle) for providing insights in to the world of data journalism. I also wish to thank Professor Hannu Kivijärvi and the entire Master's seminar group for their valuable feedback. Finally, I would like to thank my brother and my friends for their support and inspiration.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

Data journalism as a conversation topic easily divides people into those who have a clear idea about what it is and those who vaguely know what it consists of. Similarly, data journalism divides news organizations into those who publish the format and those that do not. People with a clear idea of data journalism can easily pinpoint a couple of news organizations that are pioneers in the field, such as The New York Times, The Chicago Tribune and The Guardian. Usually these pioneers are large news corporations that have resources to use on creating something new and innovative for their readers. Indeed, data journalism is a rather new journalism phenomena that only started to appear in industry publications after 2006.

A steppingstone for data journalism has been the open data phenomena, where advocates require governmental data to be freely available online to everyone, so that it can be re-used and distributed for commercial or non-commercial purpose since taxpayers paid for the data in the first place. For news organizations that have been struggling with the changing news delivery environment and consumption habits, such free data presents as a free resource for stories. However, before any stories can be written, data needs to be processed into a more analyzable format. The aim of the entire data journalism creation process is to produce an online article that combines both written and visual part. Visualizations presented with a story are considered as a crucial part of data journalism and important information disseminators. An interesting thing about data journalism is the fact that there are not many news organizations publishing it, other than industry pioneers. For instance, the first newspaper in Finland to establish a team dedicated to data journalism was Helsingin Sanomat a year ago. At the beginning of this year another major news organization, the Finnish Broadcasting Company (Yle) established their in-house data journalism team as well. While news organizations have clearly started to acknowledge the potential of data journalism it has been adopted slowly as a reporting method. Perhaps there is something to be discovered in the data journalism creation process that has stopped it from becoming a reporting norm so far.

1.1 Background of the research and motivation

Data journalism is a format of journalism that challenges traditional ways of creating and presenting news. The process of creating data journalism has been rather undefined until recently. One of the first process descriptions was introduced last year when the Data Journalism Handbook (Gray, Bounegru & Chambers, 2012) was published. The term data journalism has been appearing in industry publications since 2006 but the concept has not been researched academically all that much. The open data phenomena initiated in 2003 through public sector information (PSI) directive (European Commission, 2003) was the catalyst for data journalism. By using data that governments provide openly and presenting findings in an understandable and interesting way, journalists have the ability to influence and educate people. In other words, journalists have the ability to move a data society into a knowledge society.

The process of creating data journalism however is currently rather technical, complex and time consuming. Journalists producing data journalism also need to have specific technical skills to create this format of news. The needed skills are heavily related to the use of code in order to transform data to a more analyzable format and in order to create interactive visualizations, thus journalists performing code required processes are commonly called programmer-journalists. Technical programming plays a major part in data journalism. Some professionals say that often 80% of the time goes to programming (Poikola, 2013; Tebest, 2013a). This represents an enormous amount of resources and effort used on something that easily goes unnoticed by a reader. Equally important or even more important is the visual that is a noticeable part of the final story as it has a crucial role as an information disseminator.

In addition, the complex nature of the creation process provides a great opportunity to explore whether there is any solution in near the future that could make it simpler for news organizations. If the creation process was less tedious, media organizations would be likely to publish more news in this format, resulting in a more informed society that makes smarter decisions.

1.2 Structure of the study

The following chapter describes the design of this thesis. It includes research questions, the research process, the philosophies that underpin the research, research strategy and data collection and validity sections. The research design is presented in the second chapter because it is important to understand how this research proceeds before introducing the literature review. Chapter 3 introduces the open data phenomena, which works in this research a lense through which data journalism is explored.

Chapter 4 defines data journalism, what it is and how it is currently created, as well as what skills are needed to create data journalism. General and specific examples are being presented, namely The Chicago Tribune. Finally, the relationships between open data business models discovered in Finland and data journalism phases are discussed.

Chapter 5 explores two of the most work-intensive phases of data journalism: data manipulation and data visualization. Google is explored as an example of a company that algorithmically aims to manipulate open data. This chapter also takes a deeper look into the world of visualization while it separates data visualization from visual information representations. Column Five, Visual.ly and Infogr.am present example organizations that have rather different approaches to creating information visualizations. Chapter 5 concludes the literature review in this research.

Chapter 6 presents the streamlined process model for the future of data journalism which was created as a result of knowledge accumulated from the literature review i.e., chapters 3-5. The feasibility of this model is empirically tested through interviewing three Finnish data journalism professionals. The interviewees and an analysis of the interviews are presented in Chapter 7. Chapter 8 discusses findings discovered through the interviews. Finally, chapter 9 presents concluding remarks about the research as well as limitations and suggestions for future research.

2. RESEARCH DESIGN

Research design refers to the way in which a research idea is transformed into a research project that can then be carried out in practice by a researcher (Shank, 2008). Shank (2008) elaborates the definition as, ”the term that encompasses decisions about how the research itself is conceptualized, the subsequent conduct of a specific research project, and ultimately the type of contribution the research is intended to make to the development of knowledge in a particular area” (p. 762).

This chapter first begins by establishing the research question after which the research process is described. Then the philosophies that underpin the research are defined, followed by the research strategy, data collection and research validity.

2.1 Research question

The research question for this thesis is **“**what will future open data journalism processes look like?” In order to answer to this question, current data journalism processes need to be explored. By studying the processes, the following sub question is being answered “what is going to be the role of a journalist in future data journalism processes?”

In other words, the current data journalism creation processes is defined first and then a possible outlook for the future is established. The purpose of the research is to understand data journalism a phenomenon and to describe it. Therefore, the chosen research method is qualitative. Open data serves as a framework through which data journalism is being explored.

2.2 Research process

Rudestam and Newton (2007) describe the research process as a research wheel (Figure 1). The research wheel approach is being used for this research. The wheel metaphor suggests that research is a recursive cycle of steps that are repeated over time, rather than a linear process.

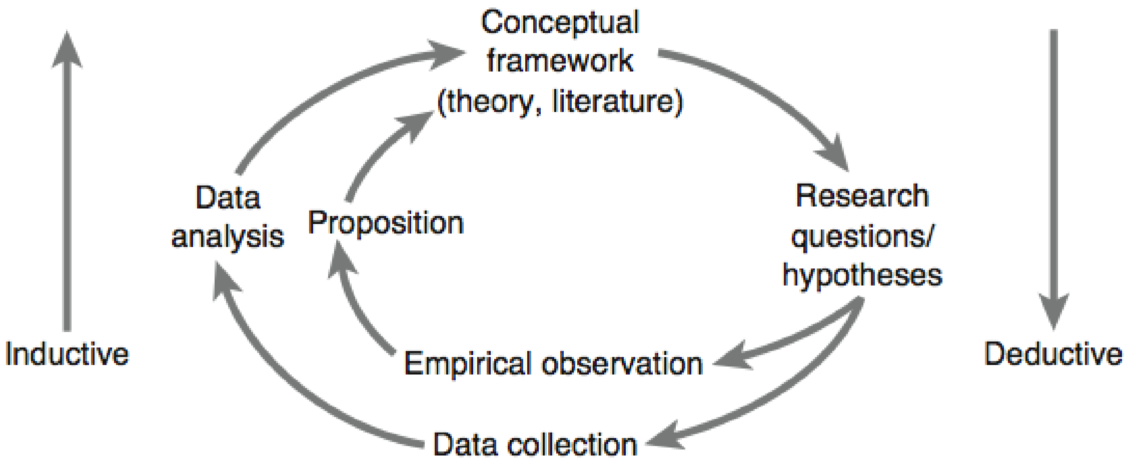


Figure 1 The research wheel (Rudestam & Newton, 2007, p.5)

The starting point for this thesis is a preliminary research question that is refined over the course of the study. The literature review serves as inductive logic in this study, where data is collected and analyzed to create a framework. The created framework or process model then establishes whether the likelihood that a certain claim about a phenomenon under study is most likely true. These claims serve as a premise from which implications are deduced. The implications are then constructed into propositions, which are empirically tested through interviews to see whether the suggested streamlined process model for the future of data journalism is feasible from the point of view of data journalism professionals. Deductive logic in this study is therefore used as a tool for determining whether the propositions presented by the inductively created framework are valid. (Given, 2008; Rudestam & Newton, 2007). This research aims to create knowledge about what the processes used to produce data journalism might look like in the future.

2.3 Philosophies that underpin the research

The philosophies that underpin research are also known as worldviews, paradigms, epistemology and ontology or broadly conceived research methodologies. They are perhaps explained at their simplest by Guba (1990) who depict them as "a basic set of beliefs that guide action" (p. 6) or as "a general orientation about the world and the nature of research that a researcher holds" (Creswell, 2009, p. 6). While philosophical ideas remain mainly hidden in research, they still influence the practice of research and therefore need to be addressed (Creswell 2009). The following table (Table 1) summarizes and explains the philosophies that underpin this research.

Table 1 The philosophies that underpin this research (adapted from Fitzgerald & Howcroft, 1998, p.160)

PARADIGM LEVEL	**Interpretivist** There is no universal truth. The researcher understands and interprets from their own frame of reference. Uncommitted neutrality is impossible. Realism of context is important.
ONTOLOGICAL LEVEL	**Relativist** Belief that multiple realities exist as subjective constructions of the mind. Socially-transmitted terms direct how reality is perceived and this will vary across different languages and cultures.
EPISTEMOLOGICAL LEVEL	**Subjectivist** Distinction between the researcher and research situation is collapsed. Research findings emerge from the interaction between researcher and research situation, and the values and beliefs of the researcher are central mediators.
METHODOLOGICAL LEVEL	**Qualitative** Researcher determines what things exist rather than how many there are. Thick description. Research is less structured, more responsive to needs & nature of research situation. **Exploratory** Research is concerned with discovering patterns in research data, and to explain/understand them. Research lays basic descriptive foundation. It may lead to generation of hypotheses. **Induction** Research begins with specific instances, which are used to arrive at overall generalizations that can be expected on the balance of probability. New evidence may cause conclusions to be revised. The approach is being criticized by many philosophers of science, but plays an important role in theory/hypothesis conception. **Deduction** The approach uses general results to ascribe properties to specific instances. An argument is valid if it is impossible for the conclusion to be false if the premises are true. The approach is associated with theory of verification/falsification & hypothesis testing.
AXIOLOGICAL LEVEL	**Relevance** External validity of actual research question & its relevance to practice is vital for the research, rather than constraining the focus to that researchable by ‘rigorous’ methods.

2.4 Research strategy

Wanderstoep and Johnston (2009) state that “qualitative research methods are inductive” (p.168). They continue that an inductive research strategy is a process of reasoning where observation precedes the theory, hypothesis, and interpretations (p.168). Qualitative researchers let the data “speak” to them, without the restrain imposed by structured methodologies and a preconceived idea of what they will find. (Wanderstoep & Johnston, 2009; Thomas, 2003)

This research uses “general inductive approach” (Thomas, 2003) to derive findings and conclusions. This approach has been reported in social science research by Bryman and Burgess (1994), Dey (1993), and Ezzy (2002, as cited in Thomas, 2003, p.2). The inductive approach is a systematic procedure used to analyze qualitative data where the analysis is guided by specific objectives i.e. research questions. The approach reflects frequently reported patters used in qualitative analysis. Usually inductive studies report a model that has between three and eight main categories in the findings. What makes the analysis strategy general is that no explicit label is given to the strategy. This gives more flexibility for interpretation and makes the analysis more straight forward as the analysis is not bound by certain rules of data analysis, such as open coding and axial coding which are commonly used in grounded theory. In addition, the results of an analysis may be indistinguishable from those that derive by using a grounded theory approach. (Thomas, 2003) Therefore, topics like data journalism that have hardly been studied as well as research that is exploratory in nature can benefit from the use of a general inductive approach rather than using other traditional qualitative approaches.

The objectives for using general inductive approach include the following (Thomas, 2003):

1. “To condense extensive and varied raw text data into a brief summary format;

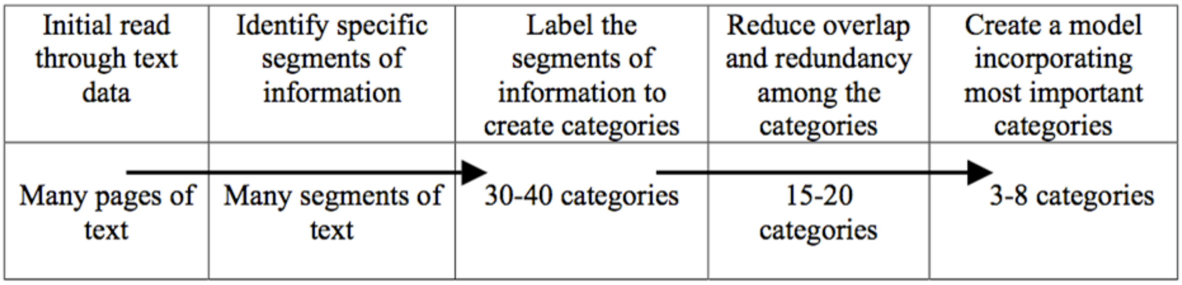
2. “To establish clear links between the research objectives and summary findings derived from the raw data and to ensure these links are both transparent (able to be demonstrated to others) and defensible (justifiable given the objectives the research);

3. “To develop of model or theory about underlying structure of experiences or processes which are evident in the text (raw data)” (p. 2).

According to Hahn (2008) “all major qualitative methods utilize coding techniques to help organize the overwhelming amount of data that is frequently collected during qualitative research” (p.7). Data coding plays a major role in this research and as a result a model that incorporates the most important categories found in the literature related to this research, has been created. The coding for this research follows the general coding process explained below. So far no researcher has claimed to be the final authority on the “best” way to code qualitative data (Saldana, 2009). Saldana defines code in qualitative research as:

> A code in qualitative inquiry is most often a word or short phrase that symbolically assigns a summative, salient, essence capturing, and/or evocative attribute for a portion of language-based or visual data. The data can consist of interview transcripts, participant observation field notes, journals, documents, literature, artifacts, photographs, video, websites, e-mail correspondence, and so on. (p.3)

The coding process begins with many pages of text data and ends with 3-8 categories that constitute a model or framework. Thomas (2003) summarizes the coding process used in inductive analysis as follow (Figure 2):



Initial read through text data	Identify specific segments of information	Label the segments of information to create categories	Reduce overlap and redundancy among the categories	Create a model incorporating most important categories
Many pages of text	Many segments of text	30-40 categories	15-20 categories	3-8 categories

Figure 2 The coding process used in inductive analysis (Thomas, 2003, p.6), adapted from Crewell, 2002, p. 266

Usually coding is a cyclical act as the first cycle of coding rarely addresses data perfectly. According to Saldana (2009) “ The second cycle and possibly the following ones of recoding further manages, filters, highlights, and focuses the salient features of the qualitative data record for generating categories, themes, and concepts, grasping meaning, and/or building theory” (p. 8).

Coding is heuristic and therefore an exploratory problem-solving technique without specific formulas to follow. Coding is only the preliminary step toward an even more rigorous and evocative analysis and interpretation (Saldana, 2009, p.8). Richards and Morse (2007) explain that coding is more than just labeling, it is linking, “it leads you from the data to the idea and from the idea to all the data pertaining to that idea” (p. 137, as cited in Saldana, 2009, p.8). Codifying happens when codes are applied and reapplied to qualitative data. This process allows data to be segregated, grouped, re-grouped and re-linked in order to combine meaning and induction. Coding enables the ability to organize and group similar coded data into categories because they share some characteristic. (Saldana, 2009)

In this research, the coding phase concludes with a framework or model that incorporates the most important categories. Based on the created model, propositions are created. Propositions are then either accepted or rejected based on interview data collected and analyzed.

2.5 Data collection and research validity

Data collection takes place in this research within two phases. First, literature is collected, analyzed and categorized as explained in the previous chapter. As the goal of this qualitative study is rather to present a range of perspectives and information on a topic than to make generalizations, the data is collected until redundancy is met. The redundancy criterion is met when the inclusion of additional data does not significantly add new information or understanding. Redundancy is accepted as a reasonable criterion as it is commonly accepted that qualitative research studies have much smaller samples than quantitative research studies. (Wanderstoep & Johnston, 2009)

Secondly, after sufficient knowledge is created based on phase one, the data is then collected in the form of interviews by using purposeful sampling technique. Purposeful samples are comprised of people based on a particular attribute (Wanderstoep & Johnston 2009). The people interviewed for this study comprised of professionals in the data journalism field who have experience in data journalism processes. They were asked to assess the feasibility of a created process model. This process increases the validity of the research and it is a common method in qualitative research called “Reflexive validity” (Wanderstoep & Johnston 2009, p.192) or “Stakeholder checks” (Thomas 2003, p.7). Thomas defines stakeholder checks as:

> Stakeholder checks enhance the credibility of findings by allowing research participants and other people who may have a specific interest in the research to comment on or assess the research findings, interpretations, and conclusions. Stakeholder checks may be carried out on the initial documents (e.g., interview transcriptions and summaries) and on the data interpretations and findings. (p. 7)

Data journalism professionals interviewed for this research were presented a series of questions (see Appendix 1) as well as, two models and asked to give their professional opinion about the streamlined process model for the future of data journalism.

3. OPEN DATA PHENOMENA

The spread of technology over recent decades has increased the level of comfort with data processing and the appetite for more of it. This cultural shift has been noted in dense newspaper displays and online. Governments around the world are rushing to open their data collections to the online public, hoping to get people involved in governmental decision making processes. This chapter discusses the open data phenomena and the value it provides.

3.1 Background of open data phenomena

The concept of open data derives from an epistemological background of at least three different aspects of openness: technological, non-proprietary and legal openness. Technological openness refers to concepts such as machine-readability, semantic web and linked open data. Here the emphasis is on connectivity of data and the possibilities to create smarter applications. Non-proprietary openness derives from the need for interoperability and inclusivity which are related to the idea of sharing and utilizing common resources. Finally, legal openness is linked to rights to use the data. Data must be licensed in such a way that gives users rights to exploit the data in a variety of ways, including commercially. (Halonen, 2012)

The public sector information (PSI) directive introduced in 2003 by the European Parliament has worked as a steppingstone for opening up government data to the public in Europe (European Commission, 2003). At the same time, the open data movement has become a global phenomenon with a simple demand: Government agencies should publish data they have gathered so that it can be freely used by everyone, re-used and distributed for commercial or non-commercial use since the data was paid for by taxpayers. This data constitutes of non-personal scientific, economic and geospatial data and reports accumulated over the years while its proliferation continues. According to Schellong and Stepanets (2011), opening up these data is expected to create great public value by "ensuring transparency and accountability, encouraging innovation and economic growth, educating and influencing people, and improving the efficiency of the government” (p. 2). Open data advocates, non-governmental organizations, individuals and researchers have been engaged in making data more available to the public as well as in shaping the open data discourse. (Schellong & Stepanets, 2011)

The Sunlight Foundation (2010), an advocator for greater transparency of the U.S. Government for instance, has identified ten principles for open data. Later on open data advocates, researchers and governments have compiled and adapted these requirements into an eight-point list (Halonen, 2012; Schellong & Stepanets, 2011; Tauberer, 2007):

1. “Data must be complete: all public data are made available. Public data are data that are not subject to valid privacy, security or privilege limitations;

2. “Data must be primary: data are collected at the source, with the finest possible level of granularity, not in aggregate or modified norms;

3. “Data must be timely: data are made available as quickly as possible to preserve the value of the data;

4. “Data must be accessible: data are available to the widest range of users for the widest range of purposes;

5. “Data must be machine-processable: data are reasonably structured to allow automated processing;

6. “Access must be non-discriminatory: data are available to anyone, with no requirement of registration;

7. “Data formats must be non-proprietary: data are available in a format over which no entity has exclusive control;

8. “Data must be license-free: data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed”.

The open data requirements, however, do not offer insight into what constitutes data but rather focus on the issue of openness and re-use. A commonly accepted definition of public data by the Sunlight Foundation (2010) and Graudenz, Krug, Hoffmann, Schulz, Warnecke and Klessman (2010, as cited in Schellong & Stepanets, 2011 p. 4) is, “data that is not subject to valid privacy, security or privilege limitations”. Providing data according to the eight-point requirements is the first step towards creating value out of open data. However, in order to create knowledge from data it needs to be transformed through stages into information and then knowledge. Schellong and Stepanets (2011) have captured this movement in Figure 3.

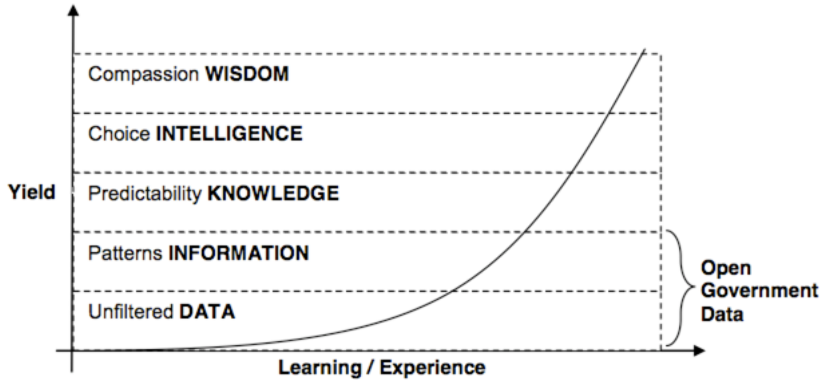


Figure 3 A conventional view of data and knowledge hierarchy (Schellong & Stepanets, 2011, p. 4)

Knowledge management literature provides a conventional view of hierarchical relationship of data, information and knowledge. Data here is assumed to be observations of the world or isolated facts. When data is put into a context and combined within a structure it results in meaning or purpose and therefore becomes information. By interpreting information and assigning it a meaning based on beliefs, perspectives, expectations or knowledge, information becomes knowledge. Facts exist within a mental structure that consciousness can process, such as to predict future consequences or to make inferences. When the mind uses knowledge to choose between alternatives, behavior becomes intelligent. Behavior can be said to be based on wisdom just as values and commitment guide intelligent behavior. During the process of creating knowledge the value of the various forms of data-information-knowledge increases through learning while data is increasingly refined. (Schellong & Stepanets, 2011; Tuomi, 1999) This means that open data has the power to move a data society to a knowledge society where citizens are better informed to make their decisions.

While open data advocates have demanded data become freely available, governments have come to acknowledge that providing unfiltered data as well as data that has been formed to information has benefits. One of the main governmental open data objectives is for citizens to use open data and create knowledge out of it that will benefit the society.

3.2 Governmental motivations for open data

Open data has been researched and it has been identified that certain areas such as political, economic and social aspects have potential for greater gains. Open data can be considered as the means for an end where the objectives are related to goals that are hoped to be achieved (Halonen, 2012). These objectives can be divided into internal and external value where internal objectives reflect increased internal understanding of an organization's work and their goals as well as the use of their resources which leads to better efficiency. External value arises from providing opportunities to businesses and empowering citizens. (Fioretti, 2010; Halonen, 2012)

Political motivations

The ideology of transparency comes from the fact that as citizens pay for public services they should know how and where their money is spent. In addition, as the data is recorded continuously it should be available for everyone to utilize. Therefore, by opening their data sets for public, governments invite citizens to think about how their money is used as well as to identify with wasteful behavior in public sector organizations. Thus transparency improves governmental efficiency. By doing so the open data movement which is, also known as “two-way online transparency encourages participative writing society where data can be used and reused again in order to foster democratic accountability” (Halonen, 2012, p. 18). Here citizens are assumed to take the role of armchair auditors that utilize data for creative activity and meaningfully engage in governmental activities. While it is hoped that this kind of behavior will lead to knowledge creation and better informed tax payers, it is also aimed at tapping the creativity of citizens and fostering service and product innovation. (Halonen, 2012)

Economical motivations

Open data is expected to stimulate economic growth, for instance, if citizens were to use open data to develop online applications that governments might have created themselves if they had the time and money or applications that government agencies had never contemplated. Therefore, open data not only save governments money but enables new commercial opportunities, entrepreneurship and creates new business models (Hammel, Perricos, Branch & Lewis, 2011). According to the UK’s Cabinet Office (2011), this “new market will attract talented entrepreneurs and skilled employees, creating high value-added services for citizens, communities, third sector organizations and public service providers, developing auxiliary jobs and driving demand for skills” (p.21). Through the use of creative data, disparate data and information can have a new dimension.

Social motivations

Citizens’ expectations toward governments as well as the aspirations and expectations people have and their sense of what they are entitled to are shaped by the culture they live in. The internet and social media are quickly altering those expectations: the way people expect to get information, make their voice heard, connect to others and receive services are all changing due to online connections and free information. Social media enables the conditions for the emergence of a mass of loosely connected, small-scale conversation campaigns and interest groups that might occasionally coalesce to create a mass movement. In an era of open data, governments everywhere will have to contend and work with this so called “civic long tail” (Leadbeater, 2011). Even if social media does not become a platform for overly political activity, it has already changed citizens' expectations about the way they communicate with the world, including governments (Leadbeater, 2011). While what the public require of governments is changing, governments themselves can also make use of data generated online by citizens. For instance, social media alongside online applications can offer a way to understand the shifting attitudes, interests and demands of the public. By analyzing and understanding this data cleverly and quickly, governments should be able to respond to emerging needs and even forestall them. Even if governments fail to react to publicly created information, there are people who re-use data to build new tools and services such as online applications to cater for those needs. (Leadbeater, 2011).

The true interest behind opening governmental data for public use is therefore to increase governmental transparency, get citizens involved in governmental decision making processes, to tap their into the public's creativity and to co-create knowledge with them. However, this kind of symbiosis is unlikely to happen automatically by placing data online. The internet provides access to information but it does not automatically give people the skills they need to create knowledge from what they find. This is where some researchers such as Fioretti (2010) say that **"**transparency is not enough without real interest and literacy in the masses" (p.25). By masses he means the combination of computer, digital media and traditional math skills necessary to correctly give context to sources, numbers and other information and to interpret everything as objectively as possible. In order to reach the public and to be effective, open data needs to be packed in a ways that most people care about and can quickly understand. (Fioretti, 2010)

Journalism has always been considered as a great trusted source that packs information in understandable formats and in interesting ways for the public. Deuze (2005) describes key characteristics of this profession as a number of discursively constructed ideal-typical values:

1. "Public service: journalists provide a public service (as watchdogs or 'newshounds', active collectors and disseminators of information);

2. "Objectivity: journalists are impartial, neutral, objective, fair and (thus) credible;

3. “Autonomy: journalists must be autonomous, free and independent in their work;

4. “Immediacy: journalists have a sense of immediacy, actuality and speed (inherent in the concept of ‘news’);

5. “Ethics: journalists have a sense of ethics, validity and legitimacy” (p.446).

A rather new form of journalism that has appeared alongside open data is called data journalism. It utilizes data as a basis for the stories and presents findings with interactive visualizations online. Therefore, data journalism has become an intermediary that delivers information for the public in a manner that is easy to understand.

This research concentrates on data journalism from the governmental open data perspective and therefore assumes that data is freely downloadable from governmental websites. Open data is different from so called “Big data” which is usually referred to as the data that businesses collect from their activities. However, private businesses are not obligated to publish their data and therefore are outside of this study. The following chapter introduces data journalism and its creation processes.

4. DATA JOURNALISM

Throughout its history, journalism has been shaped by various phases of technological innovation which has affected journalistic practices and forced the field to deal with novel work forms. New forms of work such as graphic design, computer-assisted reporting and photojournalism have all shaped the way journalism is practiced today (Powers, 2012). The latest phenomena, data journalism is the new form of work that news organizations have been struggling with. While data has an ability to create knowledge in society, it needs to be analyzed before any news organization can create stories based on it. This chapter explores data journalism processes and why this news reporting method has been adopted relatively slowly.

4.1 Data journalism and the role of programming

According to the inventor of the World Wide Web, Tim Berners-Lee, data driven journalism will have a bigger role in journalism in the near future (European Journalism Centre, 2012). Developing the techniques needed to use the available data more effectively, to understand the techniques and communicate and generate stories based on the data will present a huge opportunity for news organizations. Journalists will have a new role as sense makers and will make reporting more socially relevant. (European Journalism Centre, 2012)

Data journalism can be described at its simplest as a form of journalism that utilizes data as the basis for its stories. This definition, however, is very broad and mainly describes traditional journalism. Bradshaw **(**2012) provides a more insightful description of data journalism suggesting that it is based on “possibilities that open up when you combine the traditional “nose for news” and ability to tell compelling story with the sheer scale and range of digital information that now available” (para. 3). Accordingly, these possibilities can come at any stage of the journalism process. Programming, which is at the core of data journalism, can be used to: automate the process of gathering and combining information, finding connections between hundreds of thousands of documents, telling interactive stories through engaging infographics, distilling big numbers such as governmental spending and placing them into a context or explaining how a story relates to an individual. Data can therefore be the source of journalism, or it can be the tool with which the story is told or it can take both forms. (Bradshaw, 2012)

According to Holovaty, an award winning programmer-journalist and a pioneer of data journalism "doing journalism through computer programming is just a different way of accomplishing tasks that journalists do” (Niles, 2006, para. 4). According to Niles, generally these journalistic tasks include:

> 1. “Gathering information. This involves talking to sources, examining documents, taking photographs, etc. It is so called basic reporting;
>
> 2. “Distilling information. This involves applying editorial judgment to decide what parts of the gathered information are important and relevant;
>
> 3. “Presenting information. This involves shaping the distilled information into a format that is accessible to the readership. Some examples are writing style, photo color-correction, and newspaper page design” (para. 3).

The main role of programming here is in the use of automation whenever possible. Holovaty uses www.chicagocrime.org, an entirely automated website he created, as an example to explain the role of automation (Niles, 2006). Task one, the gathering of information, can be automated by creating a computer program that goes to the Chicago Police Department's website and gathers all crimes reported in Chicago and saves the data to a database every day (para. 5). Task two, distilling information, can also be automated in a similar way as an editor can apply editorial judgment to decide which facts in a news story are most important (para. 6). A programmer-journalist needs to decide which 'queries' should be made about the data. For chicagocrime.org, Holovaty decided that browsing by crime type, ZIP code and city ward would be in the interest of users and therefore used them as the basis for queries. Once the decision regarding what information to display is made, it is only a matter of writing the programming code that would automate it. Task three, presentation, can be automated, however it is particularly complex because when creating websites it is necessary to account for all possible permutations of data (para. 8). In the case of www.chicagocrime.org, Holovaty explains he needed to think about how the site should display crimes. In addition, other considerations such as what should happen if a crime's longitude/latitude coordinates are not available and what should happen when a crime's time is listed as "Not available" (para. 8). Although chicagocrime.org works as an example of a wholly automated website, Holovaty notes that often it is not possible to automate an entire project. In such cases he advices to use automation as often as possible.

Data journalism is a rather novel concept and has not been formally studied much although media organizations have gradually started to acknowledge its value as a vehicle to disseminate information. In July 2012, one of the first guidelines for data journalism was presented as Open Knowledge Foundation announced their Data Journalism Handbook (Gray, Bounegru & Chambers, 2012). The following chapter discusses the phases that go into creating data journalism.

4.2 The data journalism process

The following process model (Figure 4) was created as a result of combining written description of data journalism presented in the Data Journalism Handbook (Gray, Bounegru & Chambers, 2012). It visualizes eight phases of data journalism creation of which some overlap with each other.

Obtain data set

Although many institutions are working towards open data standards explained in chapter 3 acquiring data online until such standards are in place may be challenging and laborious for journalists unless it is in a machine readable format on governmental websites. Extracting data from PDFs, for instance, is very difficult as they do not retain much information on the structure of the data that is displayed within a document. Similarly, screen scraping among other methods is used to extract structured content from a normal web page with the help of a scraping utility or by writing a small piece of code that is able to turn data machine-readable format. (Grey et. al, 2012)

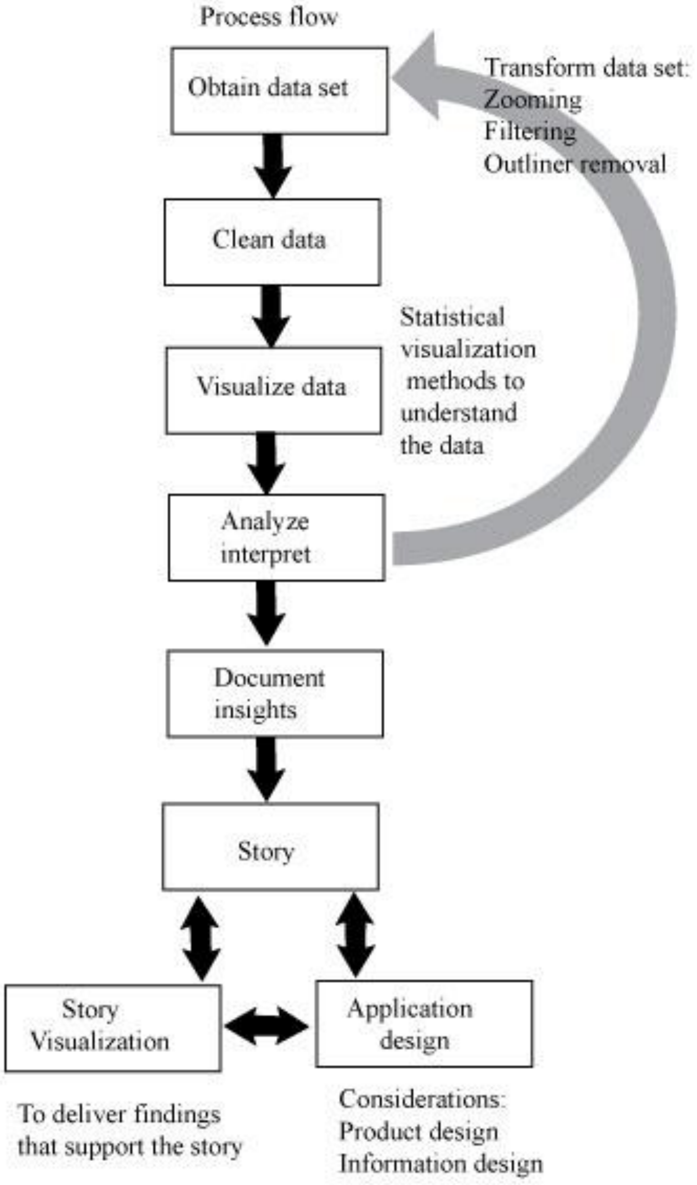


Figure 4 Data journalism creation process model (adapted from Gray, Bounegru & Chambers, 2012)

Clean up data

Once the data set is acquired, it needs to be examined for undocumented features and unneeded variables and other possible errors that may ruin a journalist's attempt to discover patterns in the collected data. One of the biggest problems with database work is that often data to be used for analysis has been gathered for bureaucratic reasons that vary from the needs of the public. Therefore, standard accuracy for those two is quite different and needs to be resolved. (Grey et. al, 2012)

Data visualization (statistical)

In order to be able to see and understand data, it needs to be visual. Statistical programs that display tables, charts and maps are generally used to find out insights from the data. New insights may emerge and become the beginning of a story, while other findings can be errors in the dataset, which are most likely to be found by making the data visual. (Grey et. al, 2012)

Analyze & interpret

Once data is in a visual format it can be analyzed and interpreted. Here the linkage between different variables is discovered. If a story idea has proven to be accurate, data should back-up the story or new ideas may be discovered. (Grey et. al, 2012)

Document insights

Details of link between stories and data are documented. Decisions whether more data is needed and how to modify data in order to inspect certain data patterns in more detail are made. Possible transformations are:

1. “Zooming: To look at a certain detail in the visualization by combining many data points into a single group;
2. “Filtering: To (temporarily) remove data points that are not in major focus;
3. “Outlier removal: To remove of single points that are not representative for 99% of the dataset” (Grey et. al, 2012).

Story

Patterns that were found from the dataset are now turned into insights and finally into a compelling story. Storyline is supported by raw data. (Grey et. al, 2012)

Story visualization (Information visualization)

The story and raw dataset is supported by informational visualization for the public. This form of visualization differs from data visualization used earlier for data analysis. Story visualization outcomes present information in an understandable and compelling format. (Grey et. al, 2012)

Application design

The story is accompanied by an application that allows readers to explore the data behind the stories through interacting with the visualization. By doing this they can find a story that is meaningful to them. (Grey et. al, 2012)

Since Holovaty was interviewed in 2006, more than 7 years ago, news companies have moved towards data journalism considerably slowly. With potential of open data journalism, it raises a question as to why the media has not embraced the field. Governments have been opening their data vaults gradually over the years, yet the biggest news organizations in the USA such as the New York Times & the Chicago Tribune as well as the the Guardian in the UK and the the Helsingin Sanomat as well as the Finnish Broadcasting Company in Finland, have established teams dedicated to data journalism. The following chapter looks into reasons why data journalism has been adopted slowly.

4.3 The current state of data journalism

Both inter-organizational and external factors have forced the struggle that media organizations have with integrating data journalism into their reporting. This chapter looks first into the internal and then external factors. Powers (2012) conducted a research study about how journalists perceive the role of technology and their attitudes towards changes that technology proposes. Powers discovered three technologically specific forms of work which describe the attitudes towards the changes.

4.3.1 Three modes of technologically specific forms of work in journalism

The future of journalism seems bright from the technology and open data perspective. However, there are considerations that technology and open data create for journalism as a profession. Changes in technologies used by news production teams do not only modify journalistic practices but they also introduce new roles that can be considered as "technologically specific forms of work" (Powers, 2012). Such forms of work have historically been a result of the introduction of photojournalism and graphic design to newsrooms as they have altered journalism processes. According to Powers (2012), technologically specific forms of work are “work practices rooted in the affordances of technical capacities that also make claims about the journalistic nature of such work” (p. 1).

Powers (2012) discovered three dominant frames of discussion or attitudes towards the development of technology which in practice tend to overlap with the forms of work: exemplars of continuity, threats to be subordinated and possibilities for journalistic reinvention.

Exemplars of continuity

New forms of work are seen as vehicles to ensure the continuity and strengthening of existing dominant occupational practices and values. Here the process of technical change is seen as an inevitable and the focus is on how to best manage this inevitability. Commercial considerations such as cost savings and greater efficiency in the production process are also stressed rather than occupational norms. In the spectrum of this form of work, journalists can both argue that nothing really has changed, e.g., working processes, as they still report, edit and write stories, while everything has changed as they use completely different tools to do all of their work. (Powers, 2012)

Threats to be subordinated

In the second attitude mode, new forms of journalistic work do not align with the core occupational norms and are therefore seen as negative and as a threat. In response to the perceived threats, journalists tend to subordinate and alienate the new forms of work by making it seem foreign, unnecessary and even dangerous to their core occupational practices. One such example is criticizing the utilization of computer-assisted reporting from a journalistic but not a technical perspective. Here journalists embrace the amount of information made available through computers but fault it for neglecting to provide context. Computer programmers and web developers are also often singled out due to their non-journalistic work form as it differs drastically from core journalistic norms. Generally these work forms are seen as necessary but insufficient forms of emerging work; necessary due to their technical expertise but insufficient because of their lack of journalistic know-how. Such criticisms often exist alongside the realization that the true technical potential of the Internet and new media can only be acquired with the technical expertise of programmers. (Powers, 2012)

Possibilities for journalistic reinvention

The third attitude mode considers new work forms positively as the basis for reinvention due to their capacities not only to contribute to existing practices and values but also to transform them into an uncertain future by creating new novel types of journalism. Change plays a part in this mode discussion, because unlike in the first mode change does not ensure continuity here. In addition, subordination contributes to this mode but is seen as something that can be mitigated and over come. The main idea is that the basis for reinvention is in combination with technical skills and innovative journalistic thinking. This kind of approach has led newsrooms to integrate print and online employees in a single newsroom. (Powers, 2012)

Reintegration into the newsroom environment has led to discussions about technologically novel specific forms of work, especially programmer-journalism to appear in industry publications during the past decade. For instance in 2006 Adrian Holovaty, talked about the value of understanding programming as “having the advantage of knowing what's possible, in terms of both data analysis and data presentation. It helps one think of journalism beyond the plain (and kind of boring) format of the news story” (Niles, 2006, para. 10). In addition, Brian Hamman who is part of The New York Times interactive news team notes “a journalist programmer is not the same person at the core as a programmer making tools for journalists. These [journalist programmers] are people who see journalistic problems and opportunities and just happen to have the technical skills to make things happen” (Glaser, 2007, para. 39).

Power's study sheds light on inter-organizational reasons why the adoption of data journalism has been slow. It can be assumed that during data journalism’s brief history, news organizations have generally treated data journalism as a threat to be subordinated and as such reluctantly invested in acquiring the needed resources. In fact, many news organizations have been struggling during the past decade to move from print only to online news publication and this transformation has possibly been a priority before considering moving to data journalism. Another issue that has stalled the use of data journalism has been a lack of resources, here referring to journalists with mathematical and coding skills. Although universities such as the Medill School at Northwestern University have established degrees where programmer-developers can acquire a Master's degree in Journalism, they have had difficulties attracting applicants. For instance, in 2008, a year after the establishment of the novel program, Medill announced that they were “still seeking coders interested in journalism” (Gordon, 2008), as they had six full scholarships available but no applicants for the program. By November 2011, four and half years later, only nine people, all of them awarded a full scholarship, had graduated from the program. The underlining aim of such programs is to give coders journalistic skills. (Gordon 2008; Foundacion Mebi, 2011; Glaser, 2007) This may leave some question as to why not train journalists to understand basic coding. According to Holovaty who studied journalism in the 90's, the issue was an attitude towards mathematics. He states: “I remember several times, the professor would say something like, “Well that would involve math, and that's why we all went to journalism school -- so we wouldn't have to learn math. Ha, ha, ha. And everyone would laugh,” (Glaser, 2007, para. 17). In the 90’s data was neither available nor used in a manner that is possible these days. Holovaty’s comment, however, reveals that generally those who decide to study journalism are not enthusiastic about mathematics.

While there has been a lack of resources, there has also been confusion between newsrooms and programmer journalists about what a programmer journalist can actually do. Minkoff (2011) represents one such case and contemplates it in her blog:

> The frustrating part is that so many of these jobs are asking for varied technical skills, but the "journalism" part is that you work in a newsroom. That is, you can be in the newsroom and use the technical skills without doing journalism, without being a journalist. (para. 4)

She then continues:

> So, when we're asked to use our programming skills to add a new commenting system, redesign an overall site, it's not what I'm looking for. No one makes you choose between being a reporter and a writer — you use the writing as a way to express your reporting. Why is that not more commonly acceptable for those of us who code? I don't spend a certain percentage of my day in journalism and a percentage in programming — I spend most of my day in programming in order to practice journalism. (para. 6)

Her final analogy really hits the core of the issue:

> Just like if you shoot your own stories, you're responsible for making sure your video camera works. Fine. But, if you are a video journalist, and you spend all day fixing video cameras, something's a little off. (para. 7)

Understanding the above mentioned internal and external factors gives some insight into why news organizations have been struggling to establish data journalism as a reporting norm so far. Some bigger news organizations, however, have managed to embrace data in their stories and therefore have become forerunners in the data journalism field. One such news organization that has embarked upon the early stages of journalism-programming is The Chicago Tribune which has established their own newsroom news application team. The team has a space in the paper dedicated to their work called Maps & Apps (Chicago Tribune, 2013) which showcases applications that were created as part of a story. In the following description, the app team's work process and how they interact with the rest of the newsroom is explored.

4.3.2 Data journalism at Chicago Tribune

The Chicago Tribune has a specialized data journalism team that has their own called Maps & Apps (Chicago Tribune, 2013). Unlike many teams in the field, the news app team at Chicago Tribune is a bit of a special group of hackers that was founded by computer technicians for whom journalism was a career change. Some of them have acquired masters' degree in journalism after several years coding for business purposes while others have previously worked for open government communities. The team works closely with editors and reporters to help: research and report stories, illustrate stories online and build updating web resources for those living in Chicago. They generally find work via face-to-face conversations with reporters who usually ask the team for help to write screen scrapers for a crummy government website, tear up a stack of PDFs' or otherwise turn non-data into something that is analyzable. As an outset of these actions, potential data projects for the team surface. The team's motivation for creating news applications is because they believe they have an impact on people and they hope people will find their own story from the data. (Boyer, 2012)

According to Boyer the team does not normally go about suggesting their ideas for applications. On the contrary, all app ideas come from the reporters and editors in the newsroom. Much of the work in the newsroom is reporter support but some of the work becomes a news application, a map, table or sometimes a larger-scale website. They have built strong personal and professional relationships with their peers in the newsroom so that when journalists have data that needs to be transformed they can go ask for the app team's help. (Boyer, 2012)

Based on Boyer's description of The Chicago Tribune's data journalism processes, their workflow can be mapped on the data journalism creation process model presented previously (p. 22). For further purposes of this study, The Chicago Tribune's nine processes have been compiled into four main workloads or phases and have been named accordingly to describe the phase best: finding story lead, data manipulation, story creation and story visualization. The created model (Figure 5) also shows who does each phase in the data journalism team. The Chicago Tribune's app team mostly completes the data manipulation and story visualization phases although they do have the ability to do the whole piece of news themselves.

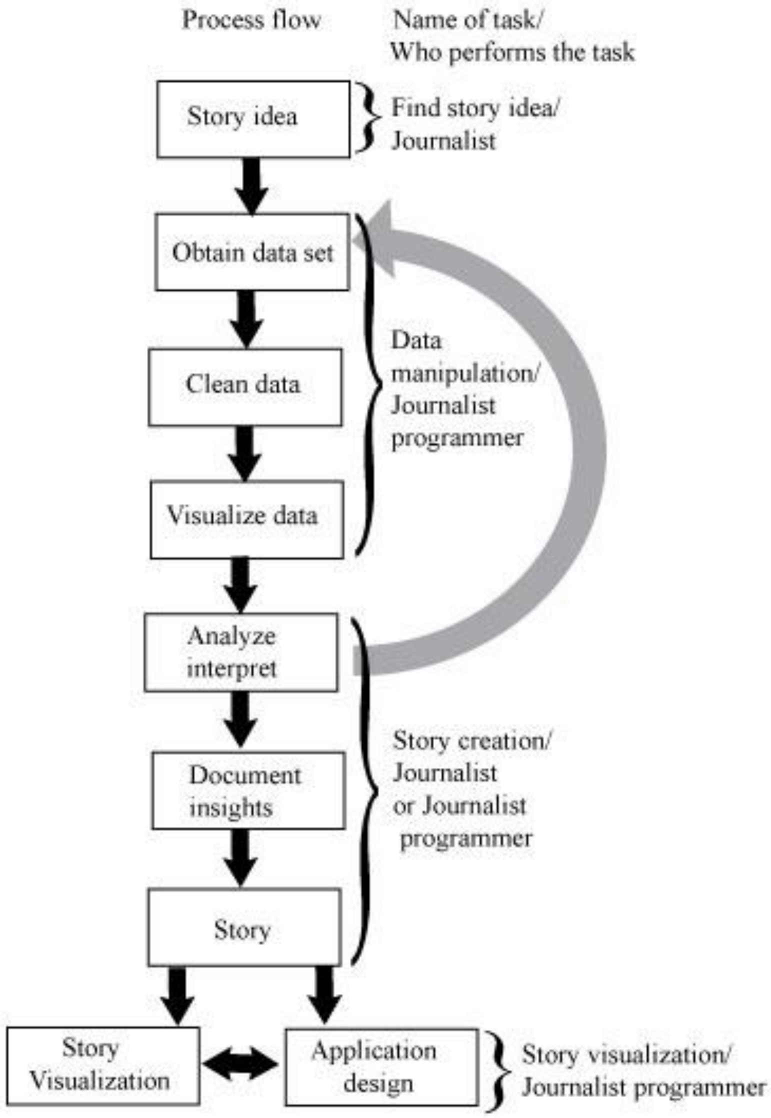
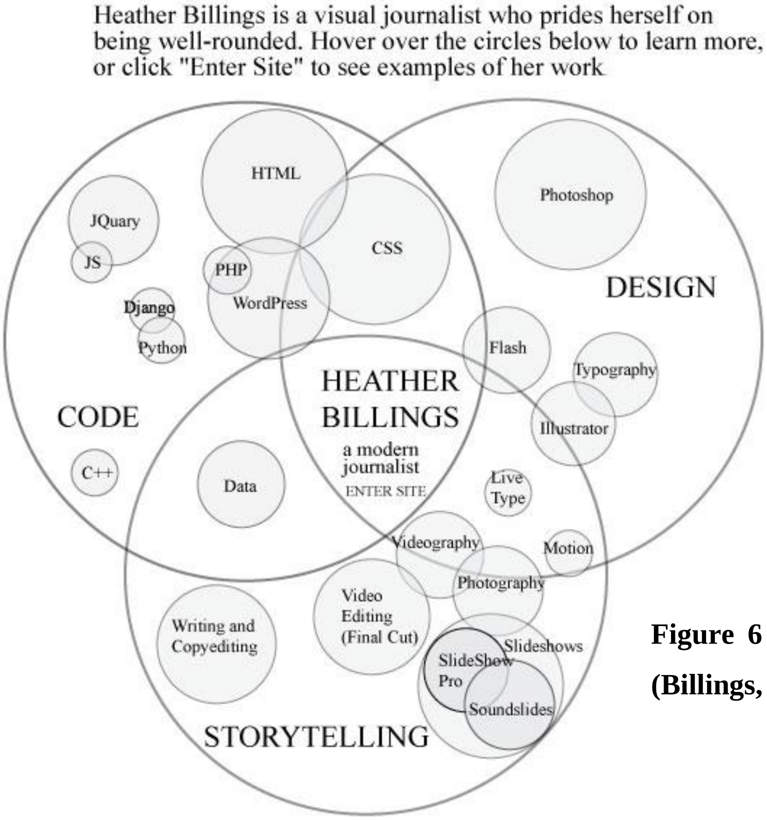


Figure 5 Chicago Tribune data journalism model 2012

Similar data journalism teams are slowly being established by other big news organizations as well. The scarcity of resources i.e., people with a combination of mathematical, coding, graphic and journalism skills has been one reason why data journalism is not used as much as it could be. In the following Heather Billings, a member of The Chicago Tribune's app team is used as an example of the skill set required by a journalist-programmer.

4.3.3 Skill set of a modern journalist

Heather Billings is a member of The Chicago Tribune news app team where she works as a news app developer (Heather Billings, 2011). On her website Billings describes herself as (Billings, 2012b):”I am, in a word, a nerd, and I think the future of journalism is in the hands of people like me” (para.1). The following figure displays Billings (2012a) skills. It gives an insight into the skill set that programmer-journalists consider to be beneficial.



Heather Billings is a visual journalist who prides herself on being well-rounded. Hover over the circles below to learn more, or click "Enter Site" to see examples of her work

Figure 6 A modern journalist’s skills (Billings, 2012a)

Billings divides her skills into three main categories: code, design and storytelling which overlap with each other. These three categories evidently represent the earlier mentioned journalistic tasks: gathering information (code), distilling information (storytelling), and presenting information (design). Billings considers code, design and storytelling as skills that are equally important for programmer journalists. She then divides the main categories into specific skill sets and stresses her knowledge of specific skills for readers by increasing the size of the bubble. Notable here is that as an application designer her data-bubble is not very big, as she probably does not deal much with data-analysis. However, she assumes this to change in the future (Billings, 2011): "From building apps to background stories, reporters work with numerical data in all kinds of ways. It's a practice that will no doubt increase in the future as more data becomes available all the time" (para. 1).

Billings explains that with stories it is difficult to convey the meaning of too many numbers to people without a visual. "Even a simple line chart can help in a city budget story, for instance, while more in-depth subjects like budget of USA (Washington Post, 2011) requires charts if they are to be understood" (para. 2).

While interactivity can increase understanding, it should be used according to Billings only if needed to avoid clutter. For instance, the previous example of budget of USA was "Created with a JavaScript library called Flot, which makes it easier to plot data on charts. If you're comfortable with CSS, HTML and a little jQuery, you should be able to create simple charts with Flot's defaults fairly easily" (para. 3).

All the programming languages that Billings uses and displays in her skills image may seem daunting for regular journalists. Similarly, the Data Journalists Handbook (Grey, Bounegru & Chambers, 2012) gives numerous examples of different open source programming tools that are used to process data journalism such as, Google Refine and Microsoft Excel to clean-up data; SPSS and R to do statistics; ArcGIS and QGIS to do geographic information systems; Google Spreadsheets for publishing and connecting with services such as Google Fusion Tables; Git for source code management; TextMate, Vim and Sublime Text for writing code; and a mix of MySQL, PostgreSQL and SQL Server for databases, to name but a few.

Indeed, it seems that data journalists are expected to be familiar with coding, design and storytelling. However, the Chicago Tribune data journalism model 2012 argues that data journalism’s workflow can be divided into four phases of which three seem to fit quite accurately with the skills previously mentioned, finding story lead, data manipulation (coding), story creation (storytelling) and story visualization (design). This begs the question as to whether some of the data journalism processes could be outsourced rather than doing them in-house, for instance, the most laborious ones such as data manipulation and story visualization. In the following, such outsourcing opportunities are explored through the perspective of open data business models in Finland.

4.3.4 Arising open data business models in Finland

Kinnari (2013) studied open data business models for the media industry in the Finnish sphere. While Finland is a small country it has a rather active open data community which makes it a good ground for open data business model analyses. Here it is assumed that similar business models as those found by Kinnari are also present outside of Finland.

Kinnari discovered businesses that operate in different areas of open data and designated to them profiles based on their business model. The profiles created were: commercial open data publisher, extract and transform, data analyzer, user experience provider and support services and consultancy. The following table (Table 2) is a partial extraction of Kinnari's findings. The open data business models called extract and transform, data analyzer, and user experience provider which are in the interest of this study.

Extract and transform business model

Data is extracted from its original source and transformed in to analyzable format. No data analysis is done at this stage. When there are several data sources, the transformation process plays a vital role as the data must be arrayed to the same format and into the same scale. Usually data requires some administration to ensure its integrity. This means correcting double records, missing information or otherwise incorrect information. Part of this work can be done by using clever algorithms, while part of the work requires hours of labor. (Kinnari, 2013)

Data analyzer business model

Businesses that fit this profile do data analysis to create new visualizations, while others do cross analysis of different data sources with advanced algorithms in order to provide valuable knowledge. Companies working in this field have either a solid business model or work towards a common good. (Kinnari, 2013)

Table 2 Finnish open data business models (Kinnari, 2013)

	Commercial open data publisher	Extract & transform	Data analyzer	User experience provider	Support services and consultancy
Companies	HSL Reittiopas HS Open	Louhos	Hahmota Cloud'n'Sci Asiakastieto	Duunitori Suomen turvaprojektit Gemilo ReittiGPS Reitit for iPhone Kansanmuisti Pikkuparlamentti	Flo Apps Logica
Offering	Data for others to analyze Facilitate events to encourage activity	Convert data in to an easier format for further analysis	Data visualizations Algorithm based data analysis	User experience created with help of open data sources	Consultation and subcontracting to help clients benefit from open data
Resources	Maintenance of the data	Data sources Open source community Developing the source code	Data sources Algorithms for visualization or numerical analysis	Data sources Development and maintenance of the user experience	In-house work Subcontracting

The boundary between data analyzers and data extractors/transformers is not always clear. In practice most of the data analyzers extract and transform needed data themselves as the raw data is rarely available in an analyzable format. They do use external data extractors and transformers such as Louhos, as one data source but the data available through these providers is limited. In addition, there are instances where data extraction and transformation must be done in-house before data can be analyzed such as when a customer brings in private proprietary data. (Kinnari, 2013)

User experience provider

The business idea is to utilize open data sources to create a valuable application directly for end users. In addition to the final product, they often need to extract the data from its original source as well as process and analyze the data. (Kinnari, 2013)

Based on Kinnari's findings it can be concluded that there are businesses created around specific areas of open data usage, at least in Finland. The following table (Table 3) summarizes discoveries presented here in chapter 4. The table explains how previously discussed examples of data journalism work phases, certain data journalism skills and business models are related to journalistic tasks in an open data environment. Reading from left to right, the table demonstrates that basic journalistic tasks such as gathering information for example, corresponds in data journalism manipulation of data, requires coding skills and could be done by companies operating in the extract and manipulation field.

Table 3 Summary of findings presented in chapter 4

Holovaty: “Basic journalistic tasks” (Niles, 2006)	Chicago Tribune's App team's current data journalism model: 4 Work phases	Heather Billings: Skill set of a programmer journalist (Billings, 2012)	Arising open data business models: Case Finland (Kinnari, 2013)
	Story idea		
Gathering information	Data manipulation	Code	Extract & manipulation
Distilling information	Story creation	Storytelling	Data analyzer
Presenting information	Story visualization	Design	Data analyzer/ User experience provider

Table 3 suggests that at least in Finland, there are existing businesses that could do certain data journalism work phases. These are likely to be data manipulation and story visualization. They are both often considered as the most time consuming phases in data journalism.

The following chapter focuses on exploring example organizations that are specialized in data manipulation and story visualization. Google is used as an example of a company that manipulates data through usage of algorithms and Column Five, Visual-ly and Infogr.am are examples of organizations specialized in creating information visualizations.

While data is the key to finding story ideas, it is the final story visualization that has the ability to create and transfer knowledge efficiently in today´s information loaded world. Written news stories are important but in order to reach the masses quickly, visualization has proven to be a helpful method. Visualization can support imagination, foster creative thinking, and help in organizing knowledge schemas (Pulak & Wieczorek, 2011). Therefore, visualization is an essential part of data journalism as the enormity of information needs to be systematically organized into a format that is easy to communicate. Visualization is explored in depth in the following chapter after looking into data manipulation.

5. DATA MANIPULATION & STORY VISUALIZATION IN DATA JOURNALISM

5.1 Data manipulation algorithmically

Hamilton & Turner (2009) discovered that "in recent years, ubiquitous computation has transformed the landscape of journalism. Computation has undermined business models, rebalanced the relative power of reporters and audiences, and accelerated the delivery of information worldwide" (p. 2). In the following, data manipulation is explored through practical example of Google. Google has taken interest in computational journalism. The company is known for its search engine technology however during the last decade it has expanded to many new avenues such as open data manipulation.

5.1.1 Google

Google is an empire of its own. Established in 1998 by Larry Page and Sergey Brin with a mission to "organize the world's information and make it universally accessible and useful", Google began as a web search engine provider. Over a decade, the company diversified into every corner of the internet from search engine to applications. Among Google's products are many tools that programmer journalists are talking about in the Data Journalism Handbook (Gray, Bounegru & Chambers, 2012).

Richard Gingras, the director of news and social products at Google emphasizes rethinking every aspect of journalism ecosystem including how the news is architected and how information gets produced today (Gingras 2012, April 12). According to Gingras, in journalism, such things include content architecture, narrative form of story, organizational workflow, reporter's toolkit and the usage of computational journalism, leveraging a social domain and rethinking site design. Gingras explains that Google's approach to news and information discovery has been and will continue to be "to connect the dots between a consumer's interest and informational needs and the most relevant available knowledge from the best possible sources" (August, 15).

Products such as Google Search, Google Plus and Google News use algorithms to find, cluster and present news in near real-time. In addition, they identify and harvest interesting and popular posts, mapping those posts to the interests of individual users and enable them to discover new people, new communities and new experiences. Gingras emphasize that Google's focus is on quality. Through algorithms it is possible to find the highest quality coverage from the best possible sources, the best article on any given subject or news story. (August, 15)

In his speech at Harvard University (Nieman Lab, 2012) Gingras talked about how the media industry has changed and why. While he gave small nuggets of information about what the future may be, he talked from multiple points of view and only those relative to this study are discussed here. As a seasoned veteran of new media he emphasized that while he currently works for Google, ideas he presents are rather personal observations than official Google announcements. He first talked briefly about the history of the media industry and explained that media organizations have been through transformations during times of technology change, with the biggest being a change in distribution. “A shift in distribution model frames your business model," says Gingras (Nieman Lab, 2012). He continues by suggesting that the biggest shift has been the evolution of marketplace niches that have been enabled by the internet. No longer is the vertical model of the newspaper organization possible, which caters everything and for everyone.

Computational journalism, according to Gingras provides significant opportunities. Computational journalism can be defined as (Hamilton & Turner, 2009):

> A 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. Computational journalism aims to enable reporters to explore increasingly large amounts of structured and unstructured information as they search for stories. (p. 2)

For instance, software can scan databases and social networks to identify and report patterns, which can then be reviewed by journalists for story ideas. Furthermore, these ideas would not have surfaced in any other way and therefore this unique opportunity increases the use of data in journalism. Tactics used in computational journalism generally are statistical analysis, regression analysis, correlation and matching, visualization, mashups and GIS (Geographic Information Systems), parsing, personalization and co-creation. (Daniel, Flew & Spurgeon, 2010) Many of them are used in data journalism as well.

In his speech at Harvard (Nieman Lab, 2012) Gingras threw out to the audience an interesting question to ponder over: "Can computational journalism not only help with stories but eventually become persistent, automated, investigative reports”? This question has a couple of aspects to consider. Firstly, persistency means having a persistent URL (Uniform Resource Locator) link under which all reports are collected. Gingras uses Wikipedia as an example and gives an interesting inside notion behind such an idea. According to him, 75% of readers these days go straight to the story page via search engines and only 25% of readers go through a home page to the story page. Furthermore, the number of readers going through a home page is diminishing all the time, hence it seems that the home page has lost its significance. Gingras suggests a “living article”, a real-time living resource that is owned by a reporter/editor as a potential solution. While Gingras does not mention this in his speech, Google actually has done an online experiment of such news layout in cooperation with The New York Times and The Washington Post during December 2009 and February 2010 called “Living Stories” (Google Living Stories, 2012).

Investigative reporting is typically considered as slow, high cost and possibly tedious, while it is valuable to the reputation of a news provider (Daniel et al., 2010). These setbacks can be offset by the use of algorithms to automate laborious data workloads such as what Google does with its search engine. A stream of automated open data analyzed using algorithms provides efficiency.

Data manipulation through the use of algorithms is only the first part of the data analysis process. Next, data needs to be made visually analyzable. Google clearly acknowledges this as they launched Google Public Data Explorer in 2011, a data visualization tool. On their website Google announced (Google Official Blog, 2012):

> Together with our data provider partners, we’ve curated 27 datasets including more than 300 data metrics. You can now use the Public Data Explorer to visualize everything from labor productivity (OECD) to Internet speed (Ookla) to gender balance in parliaments (UNECE) to government debt levels (IMF) to population density by municipality (Statistics Catalonia), with more data being added every week (February, 16).

The interface was developed to use Google's new data format Dataset Publishing Language (DSPL) which is openly available to anyone. “DSPL is an XML-based (Extensible Markup Language) format designed from the ground up to support rich, interactive visualizations”. (Google Official Blog, 2012) With Google’s powerful search engine combined with Google News, Living Stories and Google Public Data Explorer, the company seems to aim to establish a new ecosystem around data analysis, news presentation and distribution.

This study will now move from discussing data manipulation to another work-intensive data journalism phase, story visualization. Information visualization is a major part of moving data to information and knowledge and it has two roles in data journalism. Statistical data visualization is used to analyze data to find a story idea while the final story visualization requires an information design approach. Understanding the difference between these two visualization concepts is important as well as how they have originated. For instance Google's Public Data Explorer and other similar openly available data visualization interfaces are useful when analyzing data but their visual results do not translate well for the general public who vary in reading habits and statistical reading skills.

5.2 Communicating information visually

In the data journalism creation process model (p. 22), visualization is used both to analyze data to find patterns as well as to communicate findings. Visualization therefore has two functions that have developed from two different needs.

The university research field generally has two approaches to exploring the visualization of information delivery and display. One is through design schools where communication projects deal with visualization, packaging and delivering existing bodies of textual and visual information. The other approach is through computer science engineering schools, which generally focus on designing complex algorithms for gathering, synthesizing and organizing data into often impressive visual formats. The resulting structures in many cases are so complex that they raise issues related to clarity and function. (Woolman, 2002)

The following brief look into the history of information graphics explains how modern day data based visualizations have surfaced from two completely different interests and why they are aimed at two different audiences. One visualization branch developed visual graphics to match precise mathematical numbers because their audience was statistically minded. The other visualization branch was merely created to grab viewers' attention and to convey visually what was written in the story beside the graph.

5.2.1 Development of two distinctly different visualization formats

The origin of information visualization lies in prehistoric civilization times where rock engravings were used to depict stories (Rajamanickam, 2005). In the late 1700s time-series charts began to appear in scientific writings for the first time. At the time, people were interested in physical measurements such as time, distance and space and for astronomy, surveying, map making, navigation and territorial expansion. (Tufte, 2001; Friendly, 2006)

William Playfair [1759–1823] is widely considered to be the inventor of most of the graphical quantitative information formats widely used today. First, he invented the line graph and bar chart in 1786 and later the pie chart and circle graph in 1801 (Tufte, 2001). Perhaps the best statistical graphic ever drawn, according to Tufte is Minard's [1781–1870] illustration of the fate of Napoleon's army marching to Russia and back between 1812 and -1813 (see Appendix 2, Figure 1). This map has spatial dimension over time-series, which enhances the explanatory power of the graph when compared to, for example Playfair's quantitative information formats.

Six variables are found in the chart: the size of the army, it's location on a two-dimensional surface, direction of the army's movement, temperature on various dates during the retreat and declining size of the army during campaign. (Tufte 2001) According to Friendly (2006) Minard also created numerous other graphics and set the path towards incorporating pie charts when displaying information. In the early 1900's statistical graphics became main stream and graphical methods entered English textbooks. Numbers and, parameter estimates especially those with standard errors were established as precise while many statisticians considered visual graphics to be just pretty pictures incapable of stating a fact to three or more decimals. (Friendly, 2006)

In the late 1920's Otto Neurath introduced a visual communication system called Isotype (International System of Typographic Picture Education) in Vienna (see example Appendix 2, Figure 2). He aimed to convert profound research statistics into ideas and, ideas into a picture narrative (Rajamanickam, 2005). The Isotype was designed to be understood by people regardless of their language or cultural background. The idea was to use pictorial symbols, always of identical size which represented fixed amounts of information which could then be repeated to indicate larger quantities. Visually the strength was to give immediate graphic interest without the need for detailed captions or explanations. (Wildbur & Burke, 1998) The Isotype failed however, namely due to difficulties related to the enormity and complexity of iconic representation but it had profound impact on graphic design and iconography. The influence can still be seen today in road and transportation signs as well as in software user interfaces. (Rajamanickam, 2005)

In the 1930's Henry Beck mapped out London's underground (see Appendix 2, Figure 3). This map is considered to be the most successful infograph by far as it continues to accommodate the ever expanding rail network. According to Rajamanickam (2005) two design strategies were used for this map:

> The map places importance on function over precise geography while it minimizes topography of above ground. All commuters need to know is which line to take, where to change lines, and what are the preceding stations. The map meets this need by utilizing

> simple lines (uncluttered layout), color (differentiates the lines), clear typography (makes text easy to read), and symbols (differentiate stations from interchanges) (p. 8).

Towards end of the decade, simple infographics started to appear regularly in newspapers such as USA Today, with their 'Snapshots' graphics appearing in a sidebar below the fold on the front page of every issue. These daily snapshots usually displayed the results generated from national surveys in a simple way and they were meant to visually evoke the topic in question. (Bogost, Ferrari & Schweizer, 2010)

In 1957 computer processing of statistical data began with the creation of FORT R AN which was the first high-level language for computing. Developments in computational power and display devices allowed the advance of interactive statistical applications and true high-resolution graphics were developed. Many of the statistical graphic advances up until the mid-1980’s were concerned with multidimensional qualitative data as it allowed analysts to see relationships between progressively higher dimensions. (Friendly, 2006)

During this time, major newspaper publishers such as USA Today and The New York Times pursued infographics as they had both the resources and the technology that smaller publishers lacked. The prosperity which existed in the 50's and 60's lead to the use of the Isotype style of infographics as well as cartoonish style, where graphical data was displayed alongside extraneous detail, in the 1960’s and 1970’s. Both of these styles were popularized by Nigel Holmes (see example Appendix 2, Figure 4) and were claimed to be a reaction against the overly functionalist data graphics of the mid-century. These two styles aimed to make graphics visually more appealing. The styles worked as an introduction to the high-gloss but low-synthesis graphics of the USA Today newspaper. In 1970’s more professional visual artists entered into the fields of print publishing and advertising which placed emphasis on visualization rather than data. (Bogost et al., 2010)

The New York Times published sophisticated infographics frequently between 1965 and 1980 and therefore established themselves as the main proponent of the format in the newspaper field for decades. The infographics that USA Today published at the time were different from those of the The New York Times as they needed daily data which lead them to focus on soft and inconsequential questions. Despite their soft focus, USA Today's daily graphics raised the bar for infographics by changing readers' expectations. (Bogost et al., 2010)

A major revolution in the field of information visualization display happened in 1982 when George Rodrick's drew the first colored weather map, which appeared in the USA Today (see Appendix 2, Figure 5). It was an instant success as it became one of the newspaper's most popular features. By spreading the map over a full page and using a combination of colors, maps, tables and annotation, he explained visually why things happened and therefore changed dull and often hard to understand information into something very interesting and accessible. (Pompilio, 2004; Rajamanickam, 2005)

In the 1990's, the speed of technology development enabled accessibility to digital computer software leading to the digitalization of infographics. Towards the end of the decade, the Internet started to boom and more information was available than ever before. Post the millennium year, website publishing became somewhat popular among newspaper organizations and soon thereafter some newspapers, such as The New York Times, The Chicago Tribune and The Guardian noticed the potential of digital infographics, and formed graphic departments to create such graphics to complement stories for the print and online news. (Bogost et al., 2010)

The introduction of the Web 2.0 phenomena has changed the way information is expected to be displayed to consumers who now want to have their own input into the content. The increased use of information technology introduces higher expectations for from media houses because the lifecycle of information has become so much shorter. The focus is no longer on images printed on paper, but rather on the system that processes and displays the data. From a graphic design point of view, the challenge is now to develop the structure of data and the space it inhabits. (Woolman, 2002)

As history shows, statistical graphs are well established and mature field. The importance of graphical displays hit its highest point with Tukey’s groundbreaking Exploratory Data Analysis in 1977 (as cited in Tufte, 2001, p. 8) and Tufte’s books in the 1980's. Statisticians consider data as “the king” and are interested in effective and precise ways of visually representing numerical data. According to Gelman and Unwind (2002) “the right comparison is a key as numbers on their own make little sense and therefore graphics should enable readers to make up their own minds on any conclusion and possibly elaborate their findings” (p. 2). Statisticians tend to use standard graphic forms such as scatterplots and time series as they enable the experienced reader to quickly absorb a large amount of information while the inexperienced viewer might find them challenging and unimaginative to read. This view leads to problematic situations as statistical graphs do not communicate well outside of the academic world (p.2).

Programs that data journalists use to analyze data are generally based on statistics and therefore data visualizations created during this phase have not proven to be efficient for communicating information to the public. In a world where the main idea of the story needs to be visually compelling so that it can transfer information and knowledge, a different approach is needed and it is called information visualization.

5.2.2 Presenting information visually

Visual information communication is a vast research field. Numerous terms such as; information visualization (InfoVis), information graphics, visual representation, information illustration, data visualization, scientific visualization, infographics, statistical data visualization, visual analytics, data-base visualization and possibly some other ones populate the field. Not only are there numerous terms but many of them are vaguely defined and tend to overlap with each other. This creates a problematic situation as today’s graphical information communication area draws from different fields, such as computer science, design, statistics, communication design, data mining, psychology, and the visual arts and communication.

Information visualization or Infovis developed in the 1930's from visual 'snapshot' graphics based on national surveys to evoke viewer emotions about the news topic in question. They were later popularized in the 1960's and 1970's by Nigel Holmes' cartoonish style. As USA Today started to publish infographics daily their focus moved to more soft and inconsequential questions due to a lack of time for in-depth numerical analysis. Those past events have affected the style they use and this is why infographics are created in the first place. According to Gelman and Unwind (2002) those who create non-statistical information visualizations generally are more interested in grabbing the readers' attention and telling them a story by providing more contextual information. Once the reader is committed to a graph, innovative and sometimes even statistically inefficient graphical displays can be effective as they make the reader think about the data. The more time a reader spends trying to understand a graphic, the stronger is their unconscious emotional commitment to that graphic. This result in the reader being satisfied that they found a solution even if it was something that they should have already know. (Gelman and Unwind, 2002) In addition, Gelman and Unwind note that graphic designers themselves think of novelty as an end to itself, with the goal of evoking a reaction, “Hey, I’ve never seen this before,” followed by, “Of course—that’s a really good way to display this data” (p. 6).

Helfand (1997) quotes Aristole when explaining the philosophical antecedents of communication design “spoken words are the symbols of mental experience and written words are the symbols of spoken words.” (p.16) According to this, visualization for designers is an aim at compressing mental experience into visual symbols or illustrations. Within lies the notion that “A picture is worth a thousand words” referring to the fact that a complex idea or mental experience can be conveyed with just a single image. Mental experience is hardly ever linear, meaning that it contains so called “texture” according to Helfand. Texture is what statisticians generally consider undesirable “chartjunk” whereas visual designers often consider information design without it as an “aesthetic fatland”: “Information dense and visually unimaginative” i.e., information without experience (Helfand, 1997, p.16).

Continuing developments within the interactive media field and the growth of the internet has changed the nature of searching for information. The term “information highway” has been replaced with “information landscape, an interactive and dynamic landscape where online applications allow people to experience and explore by interacting or playing with information” (Hefland, 1997, p.22). While infovis applications have embraced this new landscape, truly data intensive applications seem to be stuck in the “Information fatland as they are over-mapped, under-designed and anonymous” (p. 22). This result comes from an effort to simplify navigation through vast amounts of data, while texture itself is often eliminated altogether to render complex content through inappropriately simple means. In screen-based media, such graphics often result dull and disappointing outcomes. (Hefland, 1997)

5.2.3 Visualization in human cognitive processing

Ever since the 1970's education researchers have been doing systematic empirical research about the role of visualization in human cognitive processing (Schnotz, 2002). Theories such as dual coding theory (Clark and Paivio, 1991; Paivio, 1986) conjoint processing theory (Kulhavy, Stock & Kealy, 1993) and multimedia learning theory (Mayer, 1997) have aimed at trying to prove how and whether spatial text adjuncts can help when learning information. Consequently, they all generally agree that images have supportive effects on communication, thinking and learning under certain circumstances. Accordingly, visualizations do not need to have a professional appearance, but rather they stress the relationship between the task demands and the learner's prior knowledge and cognitive abilities. Therefore, a sufficient understanding of how the human cognitive system interacts with visualizations is required to design effective visual displays. (Schnotz, 2002)

According to Peterson (1996) "representations are objects or events that stand for something else" (as cited in Schnotz, 2002, p. 102). Text and visual displays are categorized as external representations. They are "understood when a viewer constructs internal mental representations of the content described in the text or visualization" (Schnotz, 2002, p. 102) Therefore, understanding text and visuals is a task-oriented construction of mental representation. External representations are based on different sign systems, namely symbols and icons introduced by Peirce (1906, as cited in Schnotz, 2002, p. 102). According to Schnotz "words and sentences are examples of symbols as they have arbitrary structure and are associated with the designated object by a convention" (p. 102). Graphs have more common with icons than with symbols although their definition deviates slightly from that of icons. "Graphs are characterized by an abstract kind of structural commonality with the designated object" (p. 103). As texts and visuals displays are based on different sign systems they belong to different classes of representations.

These two classes are: descriptive (texts) and depictive (visuals). (Schnotz, 2002) According to Schnotz a general negation such as ‘No pets allowed!’ or a general disjunction such as ‘Seat reserved for infirm people and mothers with babies’ are straight forward as long as it is assumed that the language of the text is understood. Whereas visualizations of such negations and disjunctions need to be presented through a series of pictures and viewer needs to have previous knowledge of the signs to be able to interpret them. Therefore, visual or “depictive representations are especially useful to gain new information from already known information” (p. 104). When using a familiar context and the right texture to present data, visualizations become a powerful way to move information to knowledge.

5.2.4 Design of visual information: Tell the story and transfer knowledge

Visualizations aim to make it easier to make sense of the context. They also aim to communicate impressions and tell stories. Masud, Valsecchi and Ciuccarelli (2010) created visualizations as a process within the DIK (Data-Information-Knowledge) continuum framework to illustrate the visualization continuum (Figure 7). According to Bellinger, Castro and Mills (2004, as cited in Masud et al. 2010, p. 446) visualizations represent the process from a designer's perspective as data moves to knowledge, where each visualization is seen as a transformation artifact in the continuum. The continuum begins with raw data that simply exists but does not have meaning beyond its existence. It can take any form and be readily usable or not. Only after data has been given meaning by way of relational connections can it become information. This meaning can be useful but does not have to be. Finally, the appropriate collection of information with the intent of being useful is classified as knowledge. Within this Data-Information-Knowledge continuum, visualizations are not the final outcome. Instead, they are seen as part of the transformation process within the DIK continuum. (Masud, et al. 2010)

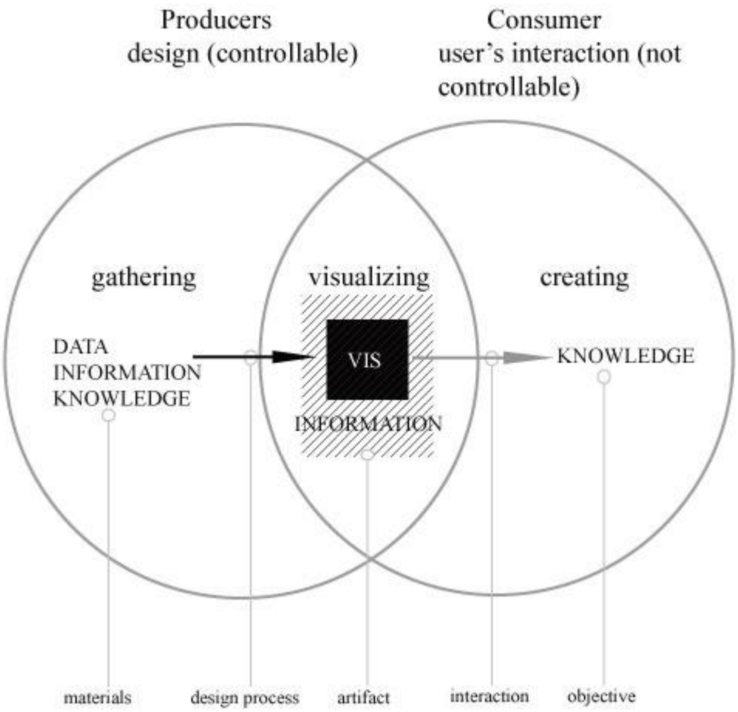


Figure 7 Visualizations as a process within the DIK continuum (Masud, et al. 2010, p. 446)

Accordingly, visualizations on their own are able to accumulate data, information or knowledge (materials) whereas visualizing them in an artifact is information which eventually creates new knowledge (objective) for the viewer (Masud, et al., 2010). Therefore, different kinds of visualizations create different kinds of knowledge.

The DIK process consists of two main parts: producers and users. The process begins with designing visualizations. "Just as information is selecting, ordering and relating data, visualization is always selecting and ordering" (Masud, et al., 2010, p. 446). The visualization process is about deciding what and how to show the given data set or information. Due to this, according to Zins (2007) "visualizations are generally considered as information in the "universal domain" (as cited in Masud, et al, 2010, p.446). The second part consists of users' interacting with the visualization and therefore the results of this part are not completely controllable by the designer.

Different researchers have established different categories of knowledge for their research purposes. For instance, Nonaka (1994) discovered declarative, procedural, causal, local and relational knowledge while Jong & Ferguson-Hessler (1996) found situational, conceptual, procedural and strategic knowledge. Masud et al. (2010, p. 447), however, have compressed these categories into three when it comes to visualizations: declarative, procedural and conditional. Declarative knowledge includes "visualizations that take data, abstract or not, and convert it into information allow the user to know something and to help them make assumptions about the data" (Masud et al., 2010, p. 447). This process explores "know-what or know about information" as it gathers data and visualizes it. Procedural knowledge refers to "visualizations that tell a story and communicate pieces of crystallized information as opposed to just presenting raw data" in a visual format (p. 447). The recipient can then use this information to understand something or know how to do something. Finally, the aim of conditional visualization is not to take data as a starting point, but rather to "transfer knowledge in a collaborative context" (p. 447). Here visualizations do not only communicate how to do things but they also transfer knowledge about when and why the recipients should use their knowledge. The fundamental difference here is the viewer's ability to take action in the context of knowledge transfer. (Masud et al., 2010)

In data journalism process model visualization happens in two phases, first during data analysis where declarative knowledge is created through statistical data visualizations by the analyst. As explained in the previous paragraph, this type of knowledge is only the first layer of knowledge and convey viewers about know-what or about information. Depending on the aim of the data representation, data visualizations need to be then transformed through design steps in order to create procedural or conditional knowledge. However, the final knowledge that a reader creates depends on the reader's interaction with the design and is therefore not controllable by designer as noted previously.

The final phase of data journalism creation is at least as important as data analysis itself because this is where knowledge is conveyed to readers. As previously discussed, statistical visualizations do not provide utility value as an end visualization, therefore outcomes of such programs as Google's Public Data Explorer are effective for data analysis but not for data representations for end viewers. Consequently, this means that analyzed data needs to be recreated into a visual format that has texture and context, which requires a certain kind of visualization knowledge. In another words, visualizations of data analysis need to be visually represented. In the following examples, organizations demonstrate the different ways to achieve the final visualization.

5.2.5 Column Five, Visual.ly & Infogr.am

This chapter discusses different approaches to achieving a final visualization. Column Five is a design based agency that represents traditional high-end local information visualization solutions as each visualization is tailor made. Visual.ly is an online community with a pool of design professionals and finally Infogr.am is an online service based on pre-designed infographic templates.

Column Five

Column Five is a creative agency located in Newport Beach, California co-founded in 2007 by Ross Crooks, Jason Lankow and Josh Richie. They started by creating and distributing infographics and other visual content. Soon after, they were hired by others to help develop content strategies. In addition to design, they have backgrounds in writing as well. These days the company specializes in infographic design, data visualization and digital PR. Crooks, one of the co-founders, believes that visualization is a key these days as: “visual content is not only more appealing, but it is also more shareable, which can lead to the organic spread of the message”. Column Five also works with a lot of private companies’ proprietary data to uncover interesting stories. According to Crooks they have also been able to establish relationships among the journalism community (Oetting, 2012):

> We have been able to build really strong relationships with journalists at large and small publications by offering them interesting, well-designed visual content that they can

> feature. These graphics can be featured alone, but they are often a great complement to a topic on which they are already writing. The key to building these relationships is that we are always coming to them with something of value, not “pitching” them on a branded message or asking for a favor. There is no pressure from our end. We see them as partners that should see a mutual benefit to the relationship. (para. 7)

Crooks explains that his design team can help media organizations to understand the data as well as the story and publishing opportunities available (Oetting, 2012):

> Although we typically work directly with our clients, we also work with a number of advertising and PR agencies to help create infographics for their clients. Our team specializes in parsing data sets, researching and crafting a strong narrative for infographics. While this is something that can be learned, there are many mistakes that are easy for someone inexperienced to make. These include data visualization errors, incorrect volume or format of content and inappropriate tone. As a team, we have worked on thousands of these projects over the years, so we are very familiar with what it takes to make a great infographic. (para. 10)

Visual.ly

Visually is an online design community established in 2011. It claims to be a one-stop shop for the creation of data visualizations and infographics tas it brings together marketing gurus, data nerds and design junkies based on shared interests (see example Appendix 2, Figure 6). The community consists of more than 35,000 designers and partners, including some of the world's leading publications and brands. The community enables designers to promote their work. They are also building a data visualization tool that will allow everyone to quickly and easily create professional quality designs with their own data as well as the possibility to partner up with publication organizations. Visually has a “Marketplace” that is a platform for ecommerce and project management as it enables buyers and sellers of infographics to communicate with each other. (Visual.ly, 2012)

Infogr.am

Infogram is an online service startup created by Uldis Leiterts in 2011. It allows anyone to create infographics quickly from pre-designed, templates including charts, videos and maps (see Appendix 2, Figure 7). Data can be imported from MS Exel and CSV-files (Comma-Separated Values) or edited with their online Exel compatible spreadsheet program. (Infogr.am, 2012) While designs showcased online are rather similar to those that can be created with MS Exel and therefore are seen as data visualizations rather than information visualizations, the idea is very interesting.

These organizations are just a couple of examples of businesses that currently inhabit the field of visualization. While they all have a very different approach to achieving final visualization, they all demonstrate that there are institutions out there that can do end visualizations for news organizations. These businesses represent different degrees of outsourcing, from a fully outsourced option to an option where a news organization can acquire a pre-designed template that they can utilize according to their needs.

In the following chapter, all the knowledge presented so far in this research is combined into a model which works as the starting point for the empirical part of this research.

6. A STREAMLINED PROCESS MODEL FOR THE FUTURE OF DATA JOURNALISM

As a result of studying literature related data journalism processes, existing open data business models and example organizations, the following streamlined process model for the future of data journalism (Figure 8) was created.

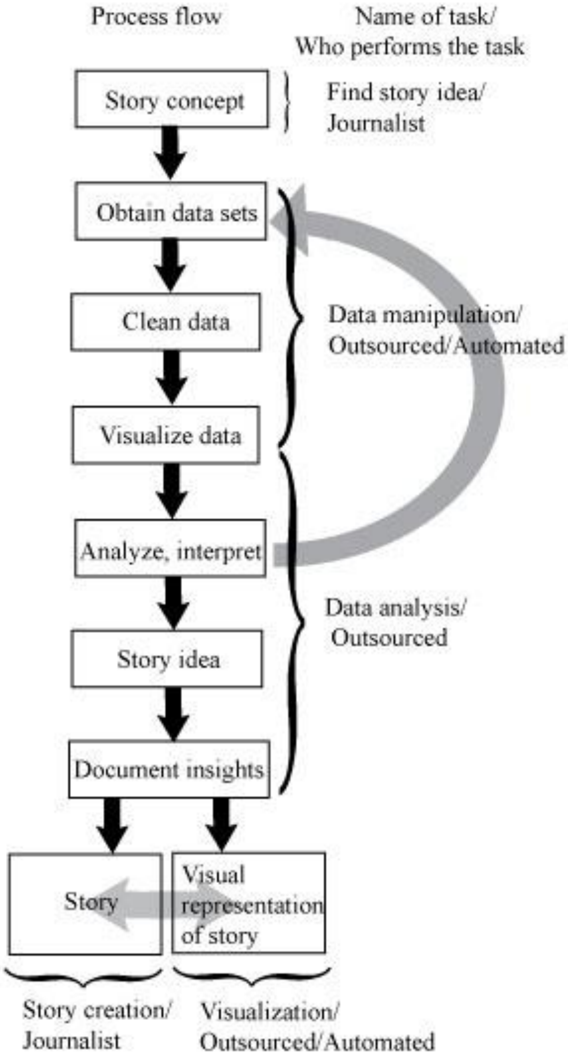


Figure 8 Streamlined process model for the future of data journalism

The model begins with a story concept that is understood as part of a broader story idea. The role of data is to tell the story as the initial story concept is being narrowed down through manipulating and cross analyzing large datasets of a newsworthy item. Comparing this starting point to that of the Chicago Tribune’s 2012 data journalism model (p. 30) where specific story ideas begin the process and the role of collected data is to support the initial story idea, the future model suggests that the role of data changes from being a supporting actor to being the true source of the story idea. Therefore, discovering the final story idea will happen only after data has been manipulated, statistically visualized and analyzed.

According to the future model, the data manipulation phase should be outsourced to businesses specialized in such work or automated through algorithms such as what Google does. During this phase numerous different datasets are compared and cross analyzed to narrow down the initial story concept to an interesting story idea. The cost and speed of the phase should be the deciding factors when deciding whether to outsource or automate in the future. The outcome of the phase is a final data set to be used to create a statistical visualization.

The next phase is data analysis. Here the dataset is explored through statistical visualization. Visualization of data is essential as the data set is large and visualization is the only way to find frequencies and/or anomalies that can be used as a story idea. Data analysis is performed by someone who has training and knowledge in statistical analysis. The final story idea is decided on after the analysis and insights from the news item are documented. Outsourcing of this phase is a viable option.

Story creation in this future model is the first phase where journalists contribute after bringing in the story concept. At the same time, visual representations of the story are created based on insights documented in the previous phase. Visualization is outsourced to entities like Five Column and Visuall.ly. Visual representations of the story are the visuals that are intended for the final user. Currently, visual representations range from still infographics to moving information films and downloadable interactive applications to name but a few. Infogr.am, for instance, has created pre-designed infographic templates which customers can attach to their own data set in order to display information graphically. In the future it can be assumed that such templates might become a rather standardized part of the visualization process and possibly become interactive and living in a sense that they are constantly automatically updating government data and displaying it in real-time. In that sense, visual representations of a story become living visualizations that can be used repeatedly as sources of news.

One notable difference between the streamlined process model for the future of data journalism and the Chicago Tribune data journalism model 2012 is that the role of a programmer journalist is diminished, as most phases where they are needed currently are either to be automated or outsourced in the future. According to the future model, as businesses specialized in the tasks that programmer journalists currently do become more established, the need for programmer journalist will diminish.

6.1 Propositions for the empirical research

Based on the streamlined process model for the future of data journalism, the following propositions are presented:

P1: All data journalism phases except story creation phase will become automated or outsourced in the future.

P2: Phases in the data journalism process will not become automated or outsourced in the future.

These propositions serve as a vehicle to move the research into the empirical phase. The constructed streamlined process model for the future of data journalism describes what future processes might look like and through interviews the feasibility of the model was tested.

7. INTERVIEWS

Three professionals were interviewed for this research. All the interviewees have been working in the data journalism field two or more years, and work at one or another of the major Finnish news corporations. The interviewees have also lectured about data journalism at universities, seminars and workshops. Thus, they represent the best source in Finland for acquiring information concerning data journalism and therefore they were chosen to be interviewed.

7.1 Interviewees

Esa Mäkinen

Mäkinen works as a data journalism producer at Helsingin Sanomat (HS), Finland's leading daily newspaper. He is a one man powerhouse that can do story writing, data manipulation, data analysis and data visualization. Mäkinen started to work with data journalism around 2009/2010 while working as a journalist at HS. At university, he studied political sciences. The current data journalism team that he leads was established in May 2012 after the running of a successful trial year before. The team includes Mäkinen who works first as a journalist and secondly as a programmer as well as another team member who works as a graphic designer who also programs. Mäkinen is a lecturer of data journalism at universities in Finland. The HS data journalism team frequently gets story ideas from other journalists within the company. Mäkinen also organizes HS Open workshops which gather together journalists, graphic designers and programmers to brainstorm how they can innovatively utilize open data for data journalism purposes. He defines data journalism as having: “two parts; a data manipulation phase and a visualization phase”.

In The HS newspaper, 1/3 of the stories that come from data journalism use governmental open data as a resource, other times data is either collected directly from online sources by using coding programs or data comes from HS archives or data is collected by asking readers questions. HS has a separate website Datajournalismi (Helsingin Sanomat, 2013) where stories featuring visual data are published. A couple of examples of Mäkinen's work are: realtime map of rush hour traffic in Helsinki area (Mäkinen, 2012) and an index that measures relationship between United Nations development aid and actual development of a country receiving the aid in Africa (Mäkinen & Ämmälä, 2012).

Juho Salminen

Salminen works as a data journalism producer at the Finnish Broadcasting Company (Yle), Finland's national public service broadcasting company. At university of applied sciences, he studied journalism. He has experience as an associate editor and developer of online journalism. He was in a team that published the first iPad-version of a magazine in Finland. Salminen is also a data journalism lecturer at universities in Finland. He started to work in the data journalism field in 2011while he was still working as an associate editor at Suomen Kuvalehti. Currently, Salminen leads a data journalism team called PlusDesk established in January 2013 which consists of two producers, two graphic designers and a programmer-journalist. His job is to decide what projects to take on based on resources and skills of his team. Salminen defines data journalism as “a process of acquiring large sets of data, discovering something interesting and illustrating it visually”.

PlusDesk mainly utilizes governmental open data for their stories, however, occasionally they collect data by asking readers questions. While PlusDesk has only been in existence 10 weeks (as of March 11), the team has been involved in over 40 stories that utilize visual data. The team does data journalism for both television and online. They do not have an official website that would contain their projects but there is an online spreadsheet (PlusDesk, 2013) that lists stories created by the team. A couple of stories that Salminen has been involved in are: How many single people in different age categories and areas in Finland have mortgage (Salminen & Tebest, 2013, February 17) and The best winter holiday weather in 10 cities in Finland during 1961-2012 (Salminen & Tebest, 2013, February 22).

Teemo Tebest

Tebest works as a programmer-journalist for the Finnish Broadcasting Company (Yle). He works within Salminen's PlusDesk team. Tebest has been working in the data journalism field for over three years and in particularly with data visualization for around six years. At university he studied engineering. He is primary a programmer but also carries out journalistic work. He shares a passion for data visualization and storytelling with the rest of the team. Tebest writes his own blog called Datajournalismi (Tebest, 2013b) and frequently lectures about data journalism at universities in Finland. Previously he has worked as a researcher at the Tampere University of Technology and has published numerous articles relating to social and hyper-media.

He participated in the Apps4Finland application creation competition in 2012, putting in two entries. One of those entries was Datavaalit which won the first price. Tebest defines data journalism as “telling a story through data and interaction”.

On his personal website (Tebest, 2013c) he lists data journalism projects that he has been involved in. Some of Tebest’s infographic examples are Foreign company ownership of Finnish companies (Tebest & Jaakkola, 2013) and Mining in Finland (Tebest, 2012, August 24).

7.2 Analysis of interviews

When the interviewees were asked about the role of programming and its usage in data journalism, the interviewees said that programming is used on a daily basis as it is a tool that enables data journalism to exist. Approximately 80% of a time is used for programming in data journalism. Reasons stated as to why data journalism has yet to become a reporting norm were because news organizations in Finland have not pushed online journalism enough. Online journalism has not been a priority as the field is still in a transition phase from print to online journalism. News organizations have been struggling to increase their revenue and they do not see business opportunities in data journalism. Therefore, they have not been willing to spend on it. Also, media organizations do not really know what data journalists do as there has been a lack of information. An additional problem is that those who have the required skills do not know that they could be employed by news organizations as they are generally engineers or computer scientists.

The biggest challenges in the data journalism field mentioned were a general lack of support, tradition and best practice. On the positive side, for those who are doing data journalism, it gives a great sense of doing something great as they are creating something entirely new. Other challenges mentioned were, the problems associated with finding a story idea, a lack of understanding how data could be utilized. On a technical level, one big challenge was mentioned, the problem of how to go about helping journalists understand that MS Word is not good to use when trying to reuse numerical data.

The interviewees were presented with a copy of the Chicago Tribune data journalism model 2012 and asked to identify similarities and differences between this model and their own processes. Everyone identified similar processes to those used by The Chicago Tribune. However, all interviewees also said that their data journalism teams create stories both on their own and based on story ideas that come from other journalists within the organization. Therefore, both teams' function equally to create stories, where as The Chicago Tribune team focuses on supporting story ideas that come from journalists only. According to interviewees, the length of a phase during the data journalism creation process differs project to project as sometimes data is easily acquired and in the right format, whilst other times data is in the wrong format and additionally, the data manipulation process takes a long time.

The interviewees had the same opinion about in regards to the most work-intensive phases, in terms of time and resources used. All interviewees agreed that both the data manipulation and the visualization phases take the most resources. Obtaining governmental open data nearly always requires phone calls to a bureau or, a waiting period to receive the data and once the data is received it is often in the wrong format. Open data is hardly ever available online as the suggested streamlined process model for the future of data journalism assumes. Coding takes a lot of time as code needs to be created from scratch as there are no templates. The phases seen as the most challenging were data cleaning, analysis and interpretation and yet again coding for visualizations.

When questioned about the use of automation, the interviewees said that it is not used much. Mäkinen said that he occasionally codes programs that can automatically collect data from a source in a timely manner. Salminen, on the other hand, said that the team he works with does not do any automation, but said they were aware that it is possible and they might use automation in the future. Salminen’s PlusDesk team does all of the phases themselves as their main function is to be a data journalism team within the news organization. They do however, co-operate with the internet development team within the organization when needed. Mäkinen, on the other hand, outsources phases such as data cleaning and visualization sometimes to gain efficiency in terms of time and cost. He said he occasionally may outsource a whole process to a trusted source but in such situation he must know the story he is planning to publish well. He emphasized that he also needed to be certain about the visualization he wants to go with the story before handling the process over to someone else.

All the interviewees agreed that the visualization phase was the most likely to be outsourced in the future. They all seemed to be aware that there are companies who do pre-designed visualization templates. Phases that cannot be outsourced or automated were obvious to the interviewees. They related to story creation, namely brainstorming and story writing. Tebest also noted that the amount of communication that occurs within the team during the creation process would be a problem if visualization was automated. He said that any visualization he has ever done needed some brainstorming and customizing to fit a news story. All the interviewees hoped that the data manipulation phase would at some point be outsourced, especially the obtaining data sets and coding parts.

There was a substantial difference between the news organizations when it came to sourcing data. Mäkinen said that he uses open data in only 33% of stories and would use it a lot more if it was available online. He also uses data from the HS archives as well as data collected online or scrape data from websites. Salminen said that 80% of his team's projects so far have been based on open governmental data. He said they have also used data from archive or collected it themselves. Tebest explained that governmental bureaus have their own legacy systems that produce documents in unknown formats which are not compatible with universal formats such as MS Excel. Sometimes agencies have been able to format these documents to Excel, but often they do not know how to do it.

The streamlined process model for the future of data journalism presented to the interviewees was not rejected by them. The data journalism producers said that it looks feasible, while the programmer-journalist was more skeptical about it. Tebest's skepticism about the process model related to a communication issue. There is constant communication between those who do different phases in data journalism. This communication aims to avoid misinterpretation of data. In his opinion, rather than outsourcing all phases to different organizations, it would be better to outsource the whole process to one trusted organization that is specialized in data journalism. Mäkinen, on the other hand, questioned whether integrating the different phase providers would be more efficient compared to in-house team in terms of time and money. In his opinion, in order for the model to be successful news organizations would need to have close relationships with the companies that they outsource the work to in order to ensure flexibility. He emphasized that it is important to achieve consistency with visualizations. Therefore, rather than having many organizations involved in the visualizations, he would outsource to one business that has a visualization style that complements the news organization's style and provides quality that they can trust.

All the interviewed agreed that automation would speed up if software programs used to complete the work became more integrated with each other. Tebest mentioned there is little to no integration between the programs he currently uses, even those provided by Google. Total automation of certain phases received general skepticism but all interviewees agreed it would be appreciated should it become possible. The following table (Table 4) briefly summarizes the key findings from the interviews.

Table 4 Key findings of the interviews

Question presented	**Summary of answers**
Who was interviewed?	- Data journalism producer Esa Mäkinen (HS) - Data journalism producer Juho Salminen (Yle) - Programmer-journalist Teemo Tebest (Yle)
Why is data journalism not used as much as it could be?	- Online journalism is not a priority in Finland - Not seen as important, no business opportunity there - Lack of knowledge how to do it - Not enough people who know how to do it - Skilled people do not realize news organizations could be potential employer
Biggest challenge when creating of data journalism?	- Lack of support, tradition and best practice - Finding story idea - To understand possibilities it offers & how to utilize data
The most work-intensive process phase in terms of time or resources used?	- Data manipulation & coding for visualization - Acquiring data from governmental agencies
The most challenging phase?	- Data cleaning & visualizations - Analysis, interpretation and understanding data
Use of automation?	- Mäkinen's team uses self-created code programs that collect data from sources timely - Not used in Salminen's team
Are phases done in-house or outsourced?	- Mäkinen's team sometimes outsources data cleaning and sometimes even the whole

	project (except story creation) - Salminen's team does everything in-house
Which phases could be outsourced/automated?	- Coding & visualization
Which phases cannot be outsourced/automated?	- Story creation - Visualization is unlikely to be automated
How much do you use open governmental data?	- Mäkinen's team uses it in 33% of the stories - Salminen's team uses it in 80% of the stories
What is the biggest problem with using open governmental data?	- Not available enough online/ wrong format - Data acquisition takes a lot of time - Governmental bureaus' legacy systems produce non-compatible formats with MS Excel or similar
Is streamlined future model feasible?	- Generally accepted as feasible - Received some skepticism
What do you think about outsourcing other phases, except for the story creation phase?	- It is possible, there are already signs that visualization can be done with predesigned template - Requires good integration of processes - Constant communication is a must between those who execute different processes
Is there anything that would prohibit outsourcing other phases, except story creation phase?	- Lack of process integration would be a major problem - Lack of communication during the process

8. DISCUSSION

Data journalism is a reporting style that is slowly taking off in news organizations. In essence, visualizing stories based on data is nothing new. It is the story creation process and the way that information is being displayed that have changed. Through this research, certain building blocks of data journalism have become clear:

1. Data journalism publishing platform is online environment
2. Visualization is an inseparable part of the story
3. Visualizations are interactive

While data journalism is getting more and more attention, media organizations, both globally and in Finland, are divided into those who publish data journalism and those who do not. Those that understand the building blocks mentioned above will find it easier to deliver news in this way.

The professionals interviewed did not reject the streamlined process model for the future of data journalism presented earlier. Therefore, proposition one (P1) which stated “all data journalism phases except story creation phase will become automated or outsourced in the future” has been accepted. However, through the interviews it was discovered that there are considerations related to the model that need to be addressed.

Firstly, the streamlined process model for the future of data journalism was created from an open data perspective and assumed that by looking for patterns and cross analyzing huge amounts of open governmental data downloadable from online resources, an interesting story idea would eventually emerge through data manipulation and analysis. In reality, such investigative data journalism does not happen all that often as obtaining and cleaning data sets takes a lot of time. However, it can be assumed that this will change in the future when governments move to publishing their data in universally compatible formats such as MS Exel. Until then, data manipulation remains as a work intensive part of data journalism that can be either done by an in-house team or outsourced to a business specialized in this field.

Secondly, the importance of communication in the streamlined process model was not fully addressed. Therefore, two new models for the future of data journalism were created with a focus on communication (Figure 9 & Figure 10).

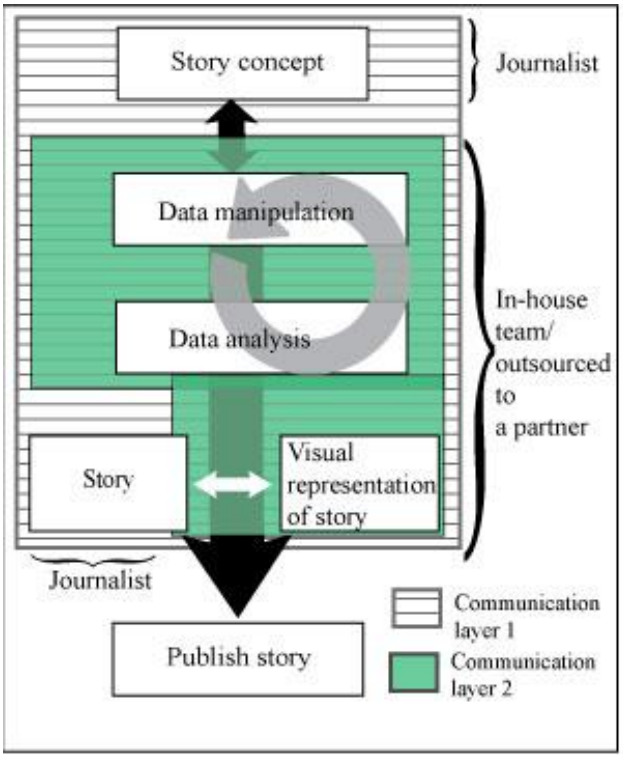


Figure 9 Data journalism consultancy model

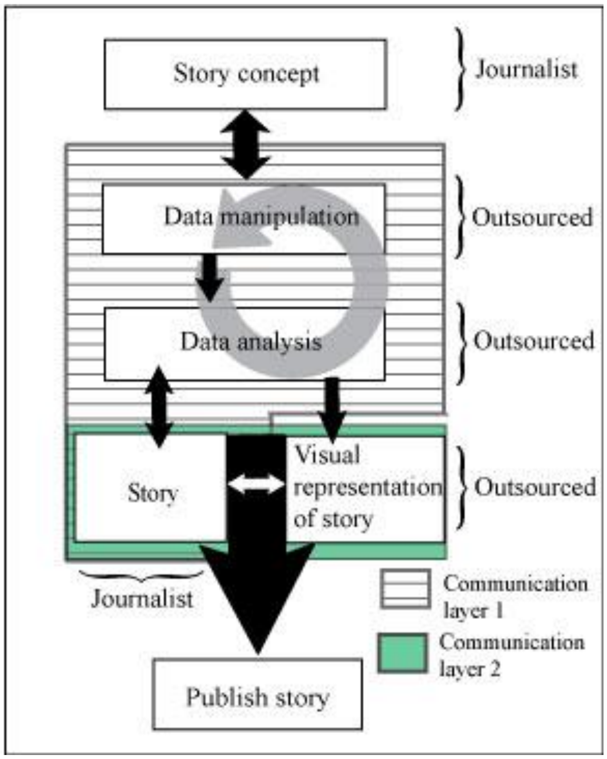


Figure 10 Outsourced chain model for data journalism

Data journalism consultancy

This model (Figure 9) suggests two options: (1) to have an in-house data journalism consultancy team or (2) outsource the entire process to a partner organization specialized in data journalism.

First option, having an in-house team, is the currently the most common process model. This model correspond the Chicago Tribune’s data journalism model 2012, but the approach in this model is from the point of view of communication. Here the data journalism team is part of the newsroom, they may look for and publish stories independently or stories may be brought to them. The process begins with a story idea or concept, the data journalism team then goes through data manipulation, data analysis and visualization processes. Communication happens within two overlapping layers, between the team and journalists (layer 1) and among members of the data journalism team (layer 2). Here journalists are part of the process but in the background. The role of a journalist is to communicate a vision for the story. In this model, the task of the journalist is to come up with a story concept, write the story based on information discovered from data analysis and to provide a vision in terms of the style and look for the final visualization. The longer a team produces data journalism together, the easier it will be for them to automate visualization and coding processes based on previous works.

The second option that this model suggests is to outsource the whole data journalism creation process, except the story creation phase, through collaboration with a company that is specialized in creating data journalism. Currently such companies are few and far between, however this is likely to change in the future. One example of an organization that specializes in data journalism is Journalism++ which was established 2011. Journalism++ is one of a chain of companies who along with individuals are part of data journalism network that shares the same brand. Partnerships with specialized organizations in the data journalism field can be best described as consultancy based. As with many business areas, consultants specialized in a field's processes are used to deal with the processes that are difficult or resource wise not efficient to handle in-house. Partnering up with a company specialized in data journalism and building a long term trusted relationship with them, is the best option to achieve close communication. Although, outsourcing without creating partnership can be considered but then communication between the journalist and the company that does the data journalism processes may become a problem. Outsourcing processes to data journalism consultants may be the easiest option for media organizations that have not published data journalism previously as there is no need to create a team within the organization. The data journalism consultancy model might be an option for long investigative story concepts that take a lot of time and resources, whether the news organization has an in-house data journalism team or not. Those media organizations that have an in-house team probably publish data journalism more on a daily basis and concentrate on topics that are more informative rather than investigative.

Outsourced chain model for data journalism

In this model (Figure 10), the data journalism creation phases, except the story creation phase, are outsourced to different phase providers. Communication within this model happens in two layers. After initiating a story concept, the journalist hands on the project to a company specialized in data manipulation having communicated the data need (layer 1) and steps temporarily out of the process. After the data has been manipulated, it is handed on to a company that specializes in data analysis. The journalist rejoins the process here communicating with data analyzers about the information needed for the story. Data may need a new round of manipulation, in which case data is returned to the company that does this phase, and then sent to the analyzers again. In order to find the right information, this cycle may be repeated a couple of times. Once the data analysis provides the needed information, it is passed on to a company responsible for the visual representation of story. Here the second layer of communication (layer 2) becomes effective as the journalist communicates the story's visualization needs.

The role of journalist in this model is more like a team manager who communicates the information he/she needs for the story and supervises the following phases within the chain. The journalist's tasks in this model are creating story concept, writing the story and managing communications, especially making sure the vision has been properly understood in each phase. This model may be an option for those media organizations that have experience with data journalism as it provides the flexibility to outsource some phases but not necessarily the whole process. A hybrid model from the data journalism consultancy model and the outsourced chain model for data journalism is also possible. In the hybrid model an in-house team would outsource, for instance, data manipulation and visualization phases but would do the analysis phase in-house.

In the end, the question remains to be whether to have a data journalism team do it all or to outsource the whole process. Moreover, when it comes to outsourcing an additional question is whether to partner up with a company that does the whole process or outsource to different phase providers.

9. CONCLUSION

This research studies data journalism, current data journalism processes and forecast the future process outlook by using the open data phenomena as a lens through which data journalism is explored. The aim of this research was to answer the primary research question, “what will future open data journalism processes look like?”, and the secondary research question, “what is going to be the role of a journalist in future data journalism processes?”. In order to answer these two questions, the research reconsiders current data journalism creation processes and provides an alternative process model. Furthermore, as data journalism creation processes may change in the future, so will the role of a journalist.

The open data phenomena and data journalism are important research topics because together they have the ability to make society smarter as people will become better informed about the issues that matter to them. The availability of information fosters creativity, which will lead to service and product innovations as well as new jobs. Journalism has always been considered as a great trusted source that packs information into an understandable format and in an interesting way for the public. Open data works as a resource for news stories while data journalism plays a key role in disseminating information.

This research begins by looking into the open data phenomena and governmental motivations behind providing data freely for citizens to utilize. It was discovered that just placing data online does not mean people will utilize it. In fact for data to be meaningful it needs to be easily understood and packed in a way that catches people’s attention. One such format is data journalism. The study then focuses on understanding the essence of data journalism by defining current processes and skills needed to create it. The skills and process phases found were then mapped out as basic journalistic tasks. Finally arising open data business models found in Finland were discussed. A summary of the findings is presented in Table 3. The resulting conclusion found that there are businesses at least in Finland that are specialized in areas that correspond to certain phases of data journalism. This finding suggests that certain data journalism phases, namely data manipulation and story visualization which are the two work intensive and time consuming could be outsourced in the future. These two phases were then studied more closely. Google is used as an example company that achieves data manipulation through automation using algorithms. In fact, it was discovered that Google not only creates the tools that journalists use daily to create data journalism but they also aim to create a new ecosystem around data analysis, news presentation and distribution through the use of technology. Examples of this are Google News, Living Stories and Google Public Data Explorer.

Visualization plays a major role in data journalism as it is used to analyze data as well as in the final presentation alongside a news story. In fact, this research argues that data journalism visualizations require design to be efficient information disseminators. Visual information design is a complex design process that requires knowledge. Therefore, it should not be confused with data visualization generated by statistical computer programs. While data visualizations serve data journalists well in the analysis phase, they do have a tendency not to translate well for the general public who vary in their reading habits and statistical reading skills. Therefore, data visualizations need to be transformed into visual representations of the data to disseminate knowledge in a collaborative context. Example companies used in this research were Column five, Visual.ly and Inforgr.am who all represent organizations that have different approaches to achieving information visualizations. Consequently, they all represent the outsourcing option described in the streamlined process model for the future of data journalism.

Based on the literature review and the above findings a new streamlined process model for the future of data journalism (Figure 8) was created. This model provides an answer to the main research questions of this study and suggests that other phases except for the story creation phase could be outsourced or automated. The model's feasibility was then tested through interviewing three Finnish data journalism professionals. Furthermore, the empirical part of this research revealed that the created model was generally seen as feasible from a professional data journalist’s point of view, although certain considerations concerning communication management between different data journalism creation phases needed more though.

This notion that communication is of importance is then taken in consideration resulting in two new models being created, the data journalism consultancy model and the outsourced chain model for data journalism. These two models describe four approaches to creating data journalism in the future. They also raise the question as to whether to use an in-house data journalism team, or weather to outsource part or all of the different phases or perhaps use a hybrid from both versions. For news organizations that have not previously published data journalism, outsourcing the entire data journalism creation process to a partner company represents an easy way to achieve data journalism. However, news organizations that already have in-house data journalism teams may be interested in outsourcing certain work intensive phases.

A sub question for this study related to the role of a journalist in future data journalism processes. This was answered through the data journalism consultancy model and the outsourced chain model for data journalism. These models suggest that the role of a journalist in future data journalism processes will differ depending on which above mentioned approach the newsroom decides to adopt. In other words, these models suggest that communication defines the role of a journalist when it comes to data journalism creation processes in the future. The models propose that a journalist's role is going to be either to write a story and communicate vision of the story to a team that does the rest of the processes or to write a story and act like a team manager who communicates the information needed for the story while he/she supervises phases done in the process chain.

The findings from this research suggest that in the future, data journalism will face changes when it comes to the data journalism creation process and the final news piece. Outsourcing of data journalism’s entire creation process or some phases of it will become a feasible option once there are established organizations that are specialized in doing the processes. The final product i.e., the news article accompanying the visualization will likely turn into Google’s suggested way which is a real-time interactive online application that automatically updates itself and is located under a persistent URL.

Finally, this research provides a completely new insight into creating data journalism by combining data journalism creation processes and arising open data business models found in Finland. By combining these aspects, outsourcing data journalism creation processes will become a real, viable option for media organizations in the future.

9.1 Practical implications

This research offers insights into how the data journalism creation process is likely to change in the future. Due to this change, the role of journalists' in the process will change as well. These changes will have practical implications on the field of journalism. This will affect how news organizations, education systems and journalists perceive data journalism.

Firstly, this research has found that the role of in-house data journalism teams will diminish in news organizations of the future. This implies structural changes and considerations for newsrooms. In addition, media organizations will no longer be strongly divided in terms of those who can publish and cannot publish data journalism as the data journalism creation process becomes a service that any news organization can purchase.

Secondly, the research illustrates that as the need for in-house data journalism teams decline, the need for programmer journalists will diminish as well. This implies that universities should reconsider whether there is need for establishing programs that educate programmers about journalism on a larger scale. The proposed future models suggest that story writing remains in the hands of journalists while data manipulation and story visualization can be outsourced either to a partner organization specialized into data journalism creation or to different organizations specialized into these phases. Therefore, there is no longer a need for journalists who do programming in newsrooms in order to create data journalism stories.

Thirdly, the findings suggest that in the future the role of a journalist will still be to write the publishable article, however they will also need to manage communications with the organization that they decide to outsource the data journalism creation to. The outsourcing approach, whether the whole creation process or just certain phases, will determine the degree of communication management and the journalist's role.

Additionally, in the future, as standardized data structures are introduced and algorithmic data manipulation becomes an everyday practice the way that Google suggests, the data manipulation phase may be completely automated. Therefore, the most work-intensive and time-consuming phase will become something that is handled mostly by computers.

Furthermore, the findings imply that in the future the focus will turn to story creation and final visualizations. Information design plays a major role in creating visualizations for data journalism. Technically these visualizations are likely to be real-time data updating interactive applications with persistent URLs. These applications will be used to repeat news story instead of archiving it after it has been published.

Finally, open data journalism presents a business opportunity for those interested in either data manipulation or story visualization. Currently there are very few organizations specialized specifically in either data journalism creation phase. The sooner people establish businesses around these processes, the faster the future scenario suggested in this research will became a reality.

9.2 Limitations

The aim of this research is not to be conclusive about any topic presented but rather to explore the phenomena called data journalism from various perspectives. The scope of the research was wide and the aim was to look into the factors related to data journalism. Due to the scope of the topic it was only possible to present limited information about each topic.

The concept of data journalism has been researched very little which affected the use of academically proven literature in this research. One of the first descriptions of data journalism processes was presented only a year ago. In addition, while data journalism is a concept that has only been conceptualized recently, the field is driven by technological developments which are constantly transforming.

The research design presented in chapter two proposes some limitations as well. Firstly, the main research question has a future oriented component. The future represents the big unknown and therefore one can only claim that in the future, anything is possible both theoretically and practically. In addition, the main research question leans toward subjective research results. In other words, this research represents subjective ideas about creating data journalism. The results are based on interpretations of literature on the subject and the empirical data. These results may be difficult or nearly impossible to replicate even if another researcher was to take in the same worldviews (see chapter 2.3) that the researcher has. Secondly, the sample size limits the findings of this research. The suggested future model was tested by interviewing three data journalism professionals from two major Finnish news organizations. The sample size of three is rather small to create a comprehensive understanding of data journalism. Decisions about the sample size were based on the fact that only two Finnish news organizations have dedicated data journalism teams and the professionals interviewed are key people in either of those teams. Thirdly, the empirical study was carried out by interviewing people. Had another kind of method been used such as observation of data journalism teams and their processes in action, the results might have varied.

Finally, the final results of this research are propositions about the future not definitive descriptions. Therefore, the results do not claim to be generalizable. The future models of data journalism creation processes presented in this research might become real or they might not.

9.3 Suggestions for future research

A natural way to broaden this research topic is perhaps to discover if there are companies abroad that are specialized in specific data journalism process phases. Are there, for instance, other companies outside of Google in the field of computational journalism, what do they currently do, what is their business model and how do they perceive the future of data journalism. What is technologically needed to automate data manipulation completely? Are there any companies specialized in visualization that would design and create real-time open data applications?

The communication process in both models should be investigated. What are the possible benefits and problems of both models? Should the communication process be reconsidered? Moreover, which model presented would ultimately be better to achieve data journalism?

Finally, how would media companies see the data journalism creation process as a purchasable service? Would they be interested in buying such service and begin to publish data journalism? If there was a great demand for such service, this would likely increase new business in this field.

10. REFERENCES

Bellinger, G., Castro, D., & Mills, A. (2004). Data, information, knowledge, and wisdom. Retrieved March 30, 2013 from http://www.systems-thinking.org/dikw/dikw.htm

Billings, H. (2011, December 15). How Journalists can use Flot to turn numbers into visual stories. Retrieved September 23, 2012, from http://www.poynter.org/how-tos/newsgathering-storytelling/156079/how-journalists-can-use-flot-to-turn-numbers-into-visual-stories/

Billings, H. (2012a). A modern journalist. Retrieved September 23, 2012, from http://www.heatherjaybillings.com/

Billings, H. (2012b). A modern journalist. Retrieved September 23, 2012, from http://www.heatherjaybillings.com/about.htm

Bogost, I., Ferrari, S., & Schweizer, B. (2010). *Newsgames: Journalism in play*. Cambridge, MA: The MIT Press.

Boyer, B. (2012). How the news apps team at Chicago Tribune works. In Gray, J., Bounegru, L., & Chambers, L (Eds.), *The data journalism handbook*. Sebastopol, CA: O'Reilly Media. Retrieved August 21, 2012, from http://www.datajournalismhandbook.org/1.0/en/in_the_newsroom_2.html

Bradshaw, P. (2012). What is data journalism? In Gray, J., Bounegru, L., & Chambers, L (Eds.), *The data journalism handbook*. Sebastopol, CA: O'Reilly Media. Retrieved August 21, 2012, from http://www.datajournalismhandbook.org/1.0/en/introduction_0.html

Bryman, A., & Burgess, R.G. (Eds). (1994) *Analyzing qualitative data*. London, England: Routledge.

Cabinet Office. (2011). *Making open data real: A public consultation*. UK: HM Government. Retrieved August 20, 2012 from http://www.cabinetoffice.gov.uk/sites/default/files/resources/Open-Data-Consultation.pdf

Chicago Tribune. (2013). Maps&Apps. Retrieved October 2, 2012 from http://www.chicagotribune.com/news/data/

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*, 149–210.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Pearson Education.

Creswell, J. W. (2009). *Research design:* Qualitative, quantitative and mixed methods Approaches. Thousand Oaks, CA: Sage.

Daniel, A., Flew, T. & Spurgeon, C. (2010). *The promise of computational journalism.* In McCallum, K (Ed.) *Media, democracy and change*: *Refereed proceedings of the Australian and New Zealand Communications Association Annual Conference (pp. 1-19).* Canberra, Australia: Australia and New Zealand Communication Association. Retrieved October 10, 2012, from http://eprints.qut.edu.au/39649/1/39649_flew_2011008025.pdf

Deuze, M. (2005). What is journalism? Professional identity and ideology of journalists reconsidered. *Journalism, 6*, 442-464.

Dey, I. (1993). *Qualitative data analysis: A user-friendly guide for social scientist*. London, England: Routledge.

European Commission. (2003). *Directive 2003/98/EC of the European Parliament and of the Council: On the re-use of public sector information.* Retrieved August 16, 2012, from http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf

European Journalism Centre. (2012). Data driven journalism. Retrieved September 27, 2012, from http://datadrivenjournalism.net/about/faq

Ezzy, D. (2002). *Qualitative analysis: Practice and innovation*. Crow's Nest, NSW: Allen & Unwin.

Fioretti, M. (2010). *Open data society*: *A research project about openness of public data in EU local administration.* EU:DIME. Retrieved August 11, 2012 from http://www.dime-eu.org/files/active/0/ODOS_report_1.pdf

Fitzgerald, B., & Howcroft, D. (1998). *Competing dichotomies in IS research and possible strategies for resolution (pp. 1-17).* ICIS'98 Proceedings of the International Conference on Information Systems. Atlanta, GA: Association for Information Systems Atlanta.

Friendly, M. (2006). *A brief history of data visualization*. Toronto, ON: York University.

Fundacion Mepi. (2011, November 10). Knight Foundation extends Medill Journalism Scholarships for programmers. Retrieved October 5, 2012, from http://www.fundacionmepi.org/index.phpoption=com_content&view=article&id=280:knightfoundation-extends-medill-journalism-scholarships-for-programmers&catid=62:mediashift

Gelman, A., & Unwin, A. (2011). *Infovis and statistical graphics: Different goals, different Look.* Unpublished manuscript. Department of Statistics and Department of Political Science, Columbia University, New York, NY/ Department of Mathematics, University of Augsburg, Germany.

Gingras, R. (2012, April 12). Google's Richard Gingras: 8 questions that will help define the future of journalism. Retrieved September 26, 2012, from http://www.niemanlab.org/2012/04/googles-richard-gingras-8-themes-that-will-help-define-the-future-of-journalism/

Gingras, R. (2012, August 15). Google's Gingras: 'The future of journalism can and will be better than its past'. Retrieved September 26, 2012, from http://www.poynter.org/latest-news/top-stories/185089/googles-gingras-the-future-of-journalism-can-and-will-be-better-than-its-past/

Glaser, M. (2007, March 7). Web focus leads newspapers to hire programmers for editorial staff. Retrieved October 4, 2012, from http://www.pbs.org/mediashift/2007/03/web-focus-leads-newspapers-to-hire-programmers-for-editorial-staff066

Google Living Stories. (2012). Retrieved November 6, 2012, from http://livingstories.googlelabs.com/

Google Official Blog. (2011, February 16). Retrieved November 6, 2012, from http://googleblog.blogspot.fi/2011/02/visualize-your-own-data-in-google.html

Gordon, R. (2008, May 26). Still Seeking Coders Interested in Journalism. Retrieved October 16, 2012, from http://www.pbs.org/idealab/2008/05/still-seeking-coders-interested-in-journalism005.html

Graudenz, D., Krug, B., Hoffmann, C., Schulz, S. E., Warnecke, T., & Klessman, J. (2010). Vom open government zur digitalen agora. ISPRAT Whitepaper, ISPRAT e.V., Hamburg, Deutschland.

Gray, J., Bounegru, L., & Chambers, L (Eds.). (2012). *The data journalism handbook*. Sebastopol, CA: O'Reilly Media. Retrieved July 28, 2012, from http://www.datajournalism handbook.org/1.0/en/index.html

Guba, E. G. (1990). *The paradigm dialogue*. Thousand Oaks, CA: Sage.

Hahn, C. (2008). *Doing qualitative research using your computer: Introduction, coding, and the big picture*. Thousand Oaks, CA: Sage

Halonen, A. (2012). *Being Open About Data: Analysis of UK Open Data Policies and Applicability of Open Data*. The Finnish Institute in London. Retrieved July 18, 2012 from http://www.finnish-institute.org.uk/images/stories/pdf2012/being%20open%20about%20data.pdf

Hamilton, J. T., & Turner, F. (2009). *Accountability through algorithm: Developing the field of computational journalism* Unpublished manuscript. A Center For Advanced Studying the Behavioral Sciences, Stanford University, CA. Retrieved October 14, 2012 from http://dewitt.sanford.duke.edu/wp-content/uploads/2011/12/About-3-Research-B-cj-1-finalreport.pdf

Hammell, R., Perricos, C., Branch, D., & Lewis, H. (2011). *Unlocking growth: How open data creates new opportunities for the UK*. UK: A Deloitte Analytics Institute Paper. Retrieved August 20, 2012, from http://www.deloitte.com/assets/Dcom-

UnitedKingdom/Local%20Assets/Documents/Market%20insights/Deloitte%20Analytics/uk-mi-da-unlocking-growth.pdf

Helfand, J. (1997). *Six (+2) essays on design and new media*. New York, NY: William Drenttel.

Helsingin Sanomat. (2013). Datajournalismi. Retrieved March 11, 2013, from http://www.hs.fi/kotimaa/aihe/datajournalismi/

Infogr.am. (2012). Retrieved November 20, 2012 from http://infogr.am/#tour

Jong, T., & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31(2)*, 105-113.

Kinnari, T. (2013). *Open Data Business Models for Media Industry: Finnish Case Study*. Unpublished manuscript. Aalto University School of Business, Helsinki, Finland.

Kulhavy, R. W., Stock, W. A., & Kealy, W. A. (1993). How geographic maps increase recall of instructional text. *Educational Technology Research Development, 41*, 47–62.

Leadbeater, C. (2011). *The civic long tail*. London, UK: Demos. Retrieved August 23, 2012 from http://www.demos.co.uk/files/Civic_long_tail_-_web.pdf?1315915449

Masud, L., Valsecchi, F., Ciuccarelli, P., Ricci, D., & Caviglia, G. (2010). *From data to knowledge: Visualizations as transformation processes within the Data-Information-Knowledge continuum (pp. 445-449)*. London, UK: 14th International Conference on Information Visualization. Retrieved August 17, 2012, from www.researchgate.net/publication/221360612_From_Data_to_Knowledge_-_Visualizations_as_Transformation_Processes_within_the_Data-Information-Knowledge_Continuum/file/9fcfd5006b40145ed6.pdf

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychology*, *32*, 1–19.

Minkoff. M. (2011, July 30). What's it like looking for a programmer-journalist job? [Web log comment]. Retrieved September 23, 2012, from http://michelleminkoff.com/2011/07/30/what-is-it-like-looking-for-a-programmer-journalist-job/

Mäkinen, E. (2012, October 26). Realtime map of rush hour traffic in Helsinki area. Retrieved March 20, 2013, from http://www.hs.fi/kotimaa/HSfin+ruuhkakartta+n%C3%A4ytt%C3%A4%C3%A4+liikenteen+pullonkaulat/a1305610085608

Mäkinen, E., & Ämmälä, A. (2012, October 26). Relationship between United Nations development aid and actual development of a country receiving the aid in Africa. Retrieved March 20, 2013, from http://www.hs.fi/ulkomaat/HS-vertailu+Kehitys+hidasta+Afrikan+suurimmissa+avunsaajamaissa+/a1305609772891

Nieman Lab. (2012, May 16). Google's Richard Gingras: We are at the beginning of a journalism renaissance. Retrieved September 26, 2012, from http://www.niemanlab.org/2012/05/googles-richard-gingras-we-are-at-the-beginning-of-a-journalism-renaissance/

Niles, R. (2006, June 5). The programmer as journalist: A Q&A with Adrian Holovaty. Retrieved September 20, 2012, from http://www.ojr.org/the-programmer-as-journalist-a-qa-with-adrian-holovaty/

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organizational Science, 5(1)*, 14-37.

Oetting, J. (2012, August 23). Tech Profile: Column Five. Retrieved November 20, 2012, from http://www.agencypost.com/tech-profile-column-five/

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press.

Peirce, C. S. (1906). Prolegomena to an apology for pragmaticism. The Monist.

Peterson, D. (1996). Forms of Representation. Exeter, UK: Intellect Books.

PlusDesk._(2013). Spreadsheet. Retrieved March 15, 2013 from https://docs.google.com/spreadsheet/pub?key=0AtSvRKpOIo3OdDZWOHVQQkF1M3Z5b0RFb1JRSm42bkE&single=true&gid=1&output=html

Poikola, A. (2013, February 18). Johdattelua datajournalismiin: Mitä uutta on datajournalismissa, datajournalismin työprosessi ja muutamia esimerkkejä. Retrieved from http://www.slideshare.net/apoikola/20120123-johdattelua-datajournalismiin

Pompilio, N. (2004). Graphic Evolution: By creating USA Today's weather map, a news art revolutionary set the tone for modern design. *American Journalism Review, 13(2)*. Retrieved July 22, 2012, from http://www.ajr.org/Article.asp?id=3642

Powers, M. (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)*, 24-36. Retrieved September 24, 2012, from http://jci.sagepub.com/content/36/1/24.full.pdf+html

Poynter. (2011, December 15). Heather Billings. Retrieved September 23, 2012, from http://www.poynter.org/author/heatherbillings/

Pulak, I., & Wieczorek, M. (2011). *Inforgraphics- the carrier of educational content.* Unpublished manuscript. Institute of Educational Sciences Jesuit University of Philosophy and Education/ Department of Educational Technology and Media Pedagogical University, Krakow, Poland. Retrieved July 27, 2012, from http://www.weinoe.us.edu.pl/files/23-Pulak_Wieczorek-Tomaszewska_1.pdf

Rajamanickam, V. (2005). *Infographics Seminar Handout* [Lecture notes]. National Institute of Design, Ahmedabad/ Industrial Design Centre, Indian Institute of Technology, Bombay, India. Retrieved July 20, 2012, from http://www.informationdesign.org/downloads/Infographic_Handout.pdf

Richards, L., & Morse, J. M. (2007). *Readme first for a user's guide to qualitative methods*. Thousand Oaks, CA: Sage.

Rudestam, K. E., & Newton, R. R. (2007). *Surviving your dissertation: A comprehensive guide to content and process*. Thousand Oaks, CA: Sage.

Saldana, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.

Salminen, J., & Tebest, T. (2013, February 17). How much single people in different age categories and areas in Finland have mortage. Retrieved March 22, 2013, from http://yle.fi/uutiset/katso_paljonko_ikaisillasi_sinkuilla_on_asuntolainaa/6498190

Salminen, J., & Tebest, T. (2013, February 22). The best winter holiday weather in 10 cities in Finland from 1961 to 2012. Retrieved March 22, 2013, from http://yle.fi/uutiset/loyda_itsellesi_suomen_paras_talvilomasaa/6505878

Schellong, A., & Stepanets, E. (2011). EU *Unchartered waters – the State of Open Data in Europe*. CSC Public Sector Study Series 1. Retrieved August 26, 2012, from http://www.epractice.eu/en/library/5269740

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review, 14(1)*, 101-120.

Shank, G. M. (2008). Deduction. In L. Given (Ed.), *The SAGE encyclopedia of qualitative research methods (pp. 208-209)*. Thousand Oaks, CA: Sage.

Sunlight Foundation (2010, August 11). Ten open data principles. Retrieved August 22, 2012, from http://www.sunlightfoundation.com/policy/documents/ten-open-data-principles/

Tauberer, J. (2007, December 8). 8 principles of open government data. Retrieved August 22, 2012, from http://www.opengovdata.org/home/8principles

Thomas, D. (2003). *A General Inductive Approach for Qualitative Data Analysis*. Unpublished manuscript. School of Population Health, University of Auckland, New Zealand. Retrieved August 28, 2012 from http://www.fmhs.auckland.ac.nz/soph/centres/hrmas/_docs/Inductive2003.pdf

Tebest, T. (2013a). Datajournalismi. Retrieved March 15, 2012, from http://teelmo.info/teelmo/app/seminar-2013-datajournalismi_otavan_opisto/datajournalismi.html#/step-39

Tebest, T. (2013b). Datajournalismi blog. Retrieved March 15, 2013, from http://datajournalismi.blogspot.fi/

Tebest, T. (2013c). Personal website. Retrieved March 15, 2013, from http://teelmo.info/#portfolio

Tebest, T. (2012, August 24). Mining in Finland. Retrieved March 25, 2013, from http://ohjelmat.yle.fi/mot/arkisto/mot_hullut_paivat_kaivoksilla

Tebest, T. & Jaakkola, J. (2013, February 28). Foreign companies' ownership of Finnish companies. Retrieved March 20, 2013, from http://yle.fi/uutiset/nama_ulkomaalaiset_omistavat_kotimaisia_porssiyrityksia__selvitys_avaa_salatut_omistukset/6505130

Tufte, E. R. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tuomi, I. (1999). Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *Journal of Management Information Systems, 16(3)*, 107-121.

Visual.ly. (2012). Retrieved November 20, 2012, from http://visual.ly/about

Wanderstoep, S. W., & Johnston, D. D. (2009). *Research methods for everyday life: Blending qualitative and quantitative approaches*. San Francisco, CA: John Wiley & Sons.

Washington Post. (2011, May 20). In search of a debt deal. Retrieved October 10, 2012, from http://www.washingtonpost.com/wp-srv/special/politics/sorting-out-the-budget-proposals/

Wildbur, P., & Burke, M. (1998). *Information graphics: Innovative solutions in contemporary design*. London, UK: Thames & Hudson Ltd.

Woolman, M. (2002). *Digital information graphics*. London, UK: Thames & Hudson Ltd.

Zins, C. (2007). Conceptual approaches for defining data, information and knowledge. *Journal of the American Society for Information Science and Technology, 58(4)*, 479-493.

11. APPENDIXES

Appendix 1 Interview questions

I. Background

1. Name

2. Job title

3. Job profile

4. How long you have been working with data journalism?

5. What roles you have had (journalist, graphic designer, programmer etc)?

6. What is your definition for data journalism?

II. The data journalism team

1. How did your data journalism team come about?

2. How many members you have in your team?

3. Describe members’ roles and their background.

4. In your opinion, what percentage (%) of data journalism generally requires programming?

    4.1. How about graphic design?

5. Can you give me viewpoints why the data journalism is not generally used as much as it could?

6. What is the biggest challenge when it comes to creating the data journalism?

III. The data journalism creation process

1. Chicago Tribune Data Journalism Model 2012 (p. 30) is presented for the interviewees. How does your team's processes compare to it?

    1.1. Find story lead?

    1.2. Data manipulation?

    1.3. Story creation?

1.4. Story visualization?

2. Which data journalism process phase is the most work-intensive in terms of time or resources used?

3. Which phase is the most challenging?

4. Do you use any automation for any of the phases?

5. Are all the phases done in-house or are some of them outsourced?

6. Which phases you think that could be outsourced / automated? Which ones and why?

7. Which phases you think cannot be outsourced /automated? Which ones and why?

8. Which phases you hope could be outsourced? Why?

IV. Source of data

1. How much do you use the open data that can be downloaded from the government's website?

2. What other data you use in your projects?

2.1. From where and how you acquire the needed data?

3. What is the biggest problem in using the open governmental data?

4. Is use of the open data going to increase in your projects in the future?

V. Future of data journalism

1. The streamlined process model for the future data journalism (p.54) is presented for the interviewees.

1.1. What do you think about outsourcing other phases except the story creation phase?

1.2. Is there anything that would prohibit outsourcing of the phases?

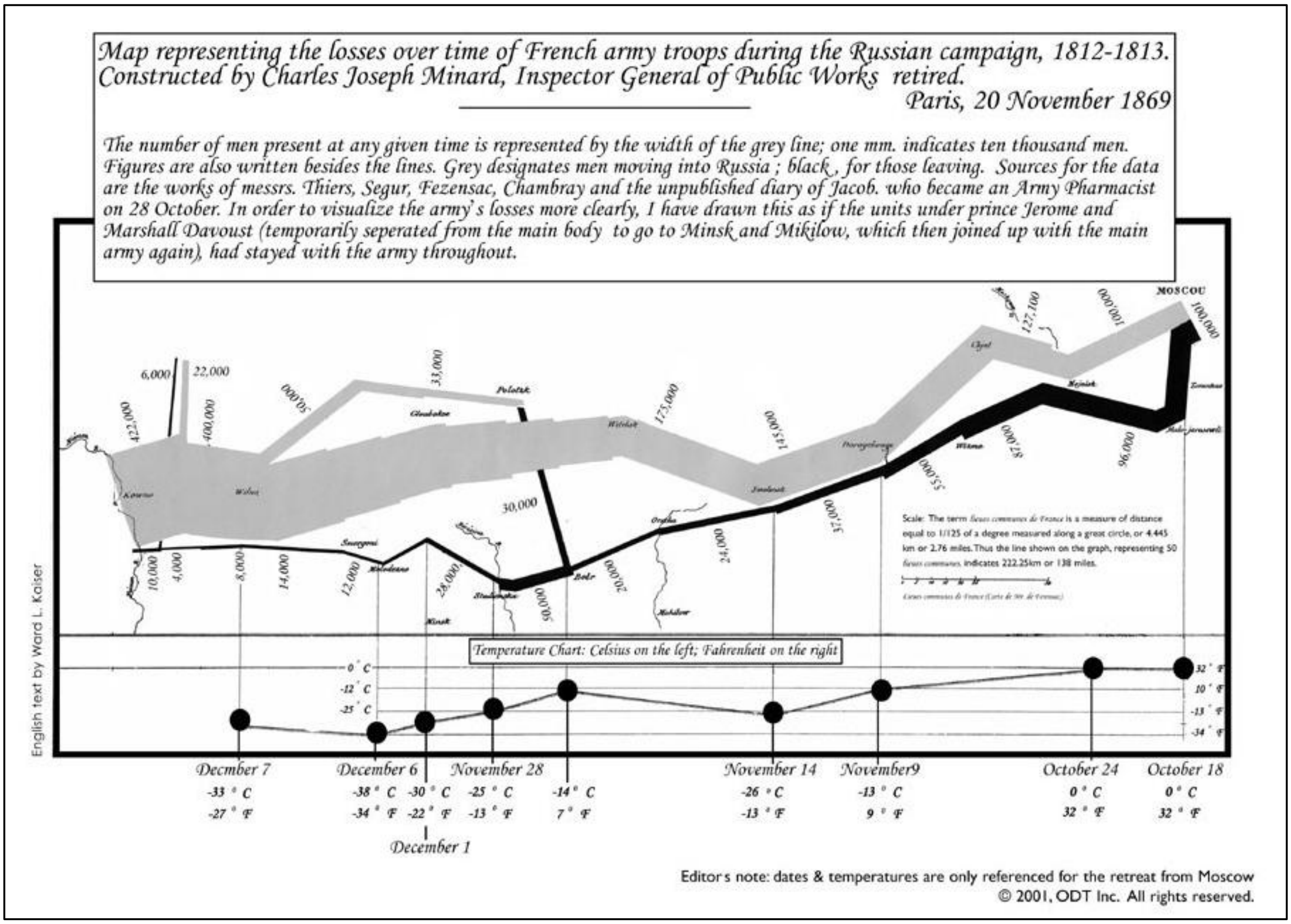
1.3. Does the streamlined process model for the future of data journalism look feasible?

VI. Other questions or comments emerged during the interview

1. Can you think of anyone else that should be interviewed for this thesis?

Appendix 2 Infographic examples

Figure 1 Minard’s map of Napoleon’s Russian campaign. This map has been translated from French to English and modified to most effectively display the temperature data.



Map representing the losses over time of French army troops during the Russian campaign, 1812-1813. Constructed by Charles Joseph Minard, Inspector General of Public Works retired.

Paris, 20 November 1869

The number of men present at any given time is represented by the width of the grey line; one mm. indicates ten thousand men. Figures are also written besides the lines. Grey designates men moving into Russia ; black , for those leaving. Sources for the data are the works of messrs. Thiers, Segur, Fezensac, Chambray and the unpublished diary of Jacob. who became an Army Pharmacist on 28 October. In order to visualize the army's losses more clearly, I have drawn this as if the units under prince Jerome and Marshall Davoust (temporarily seperated from the main body to go to Minsk and Mikilow, which then joined up with the main army again), had stayed with the army throughout.

Scale: The term *lieues communes de France* is a measure of distance equal to 1/125 of a degree measured along a great circle, or 4.445 km or 2.76 miles. Thus the line shown on the graph, representing 50 *lieues communes*, indicates 222.25km or 138 miles.

Temperature Chart: Celsius on the left; Fahrenheit on the right

Editors note: dates & temperatures are only referenced for the retreat from Moscow

Figure 2 Example of Isotype (International System Of Typographic Picture Education) invented by Otto Neurath.

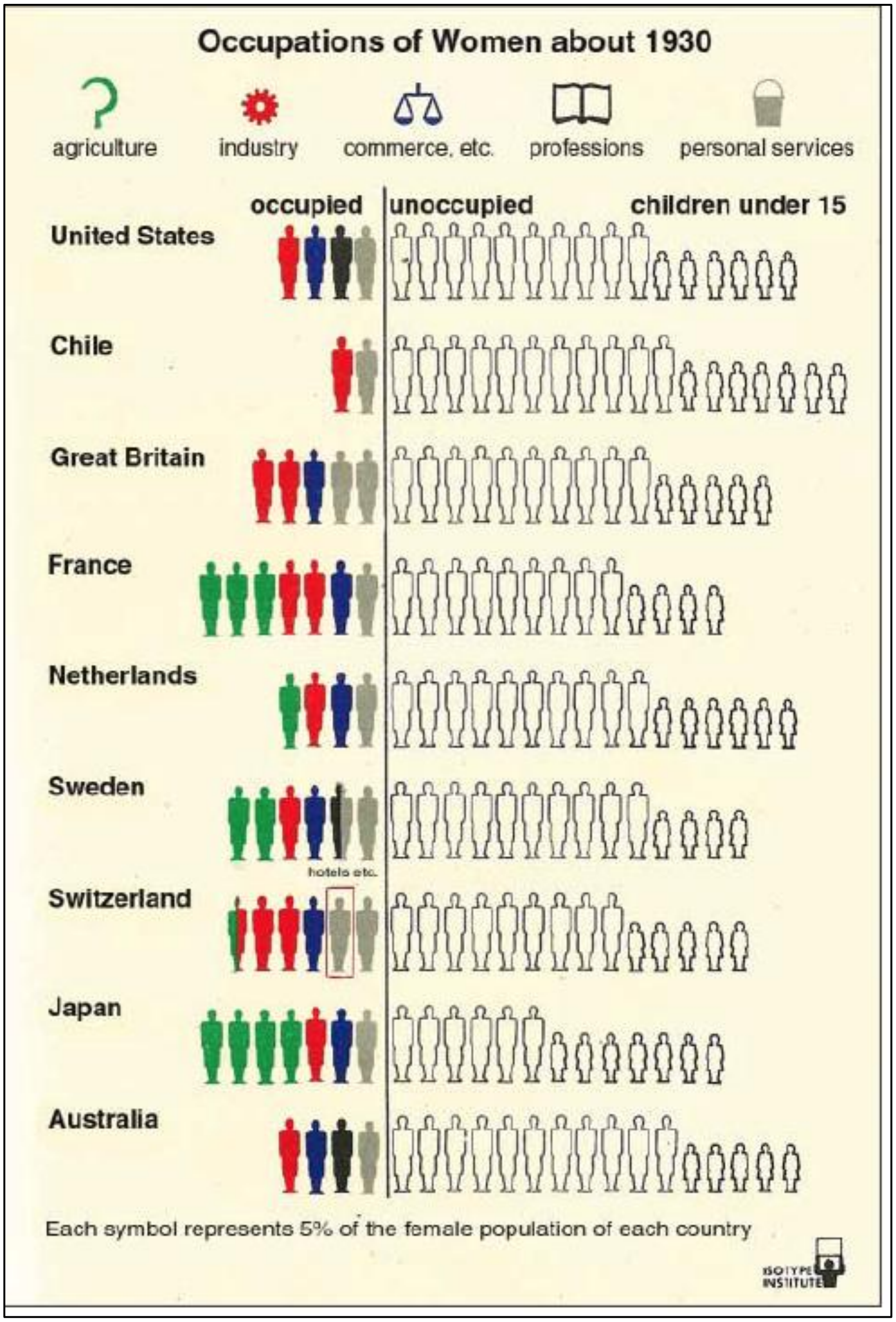


Figure 3 Map of London's underground designed by Henry Beck in 1938.

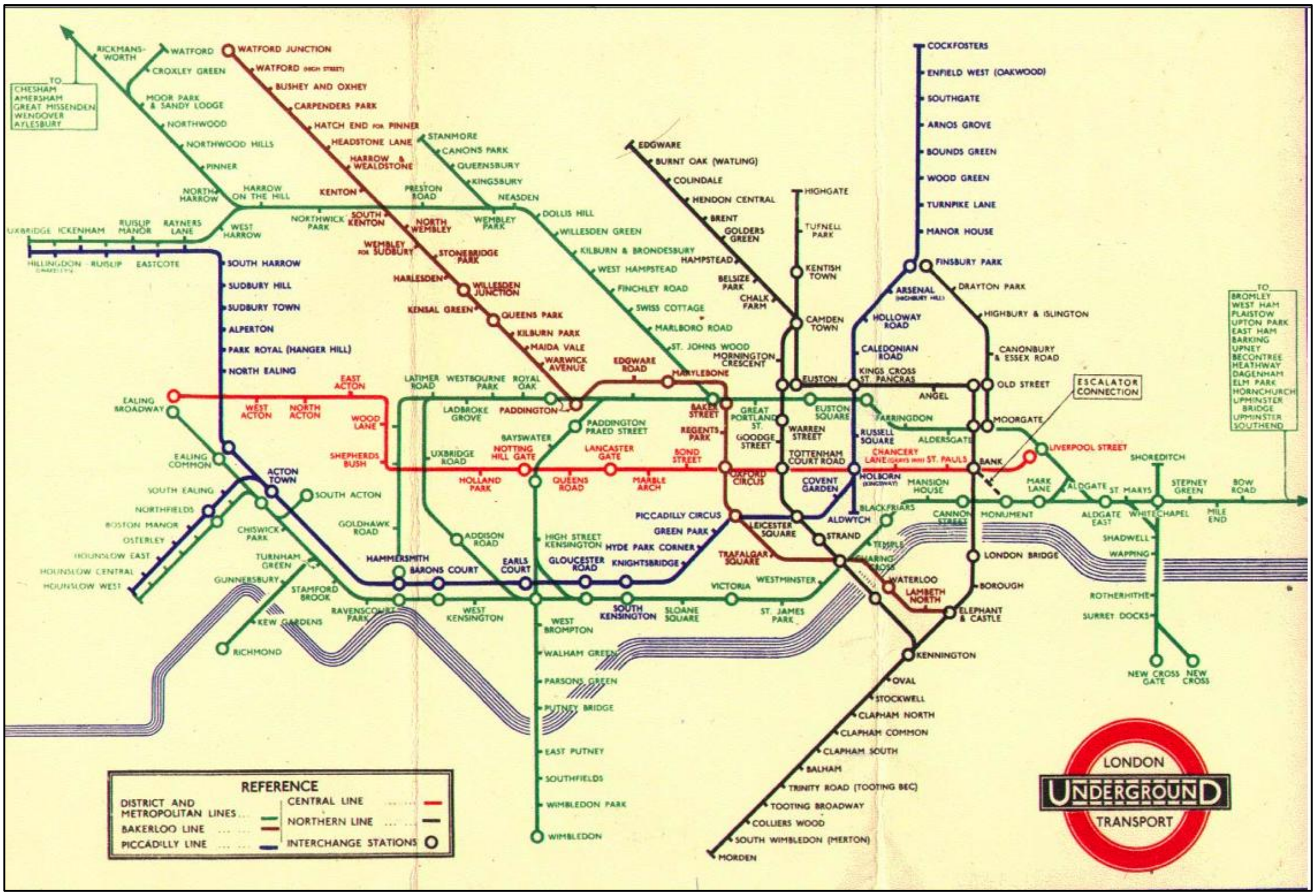


Figure 4 Example of Nigel Holmes's cartoonish infographics.

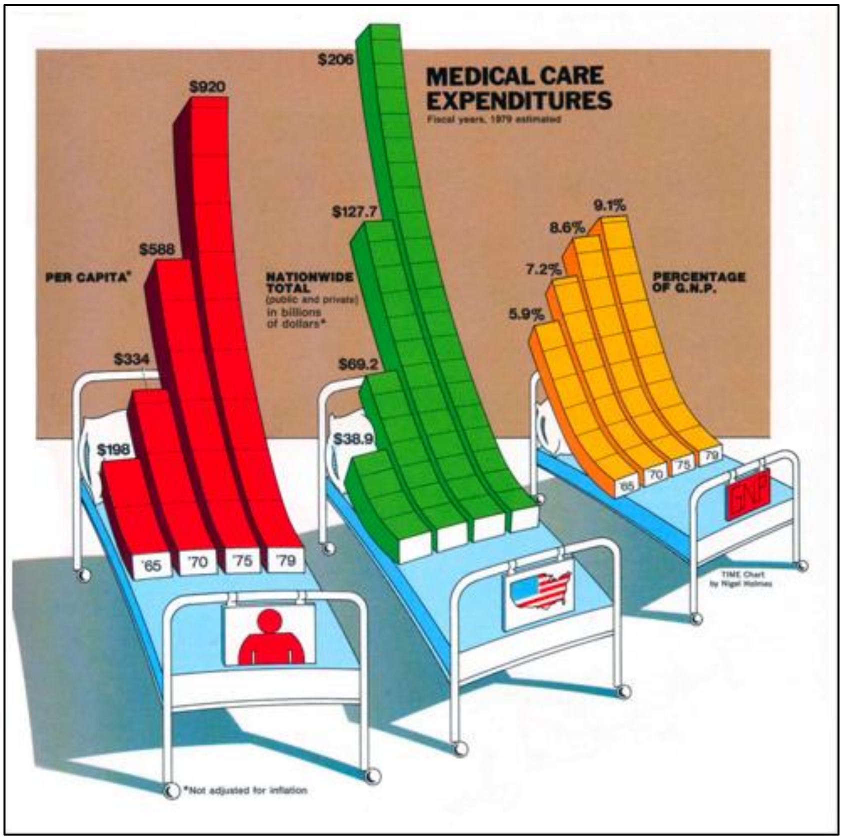
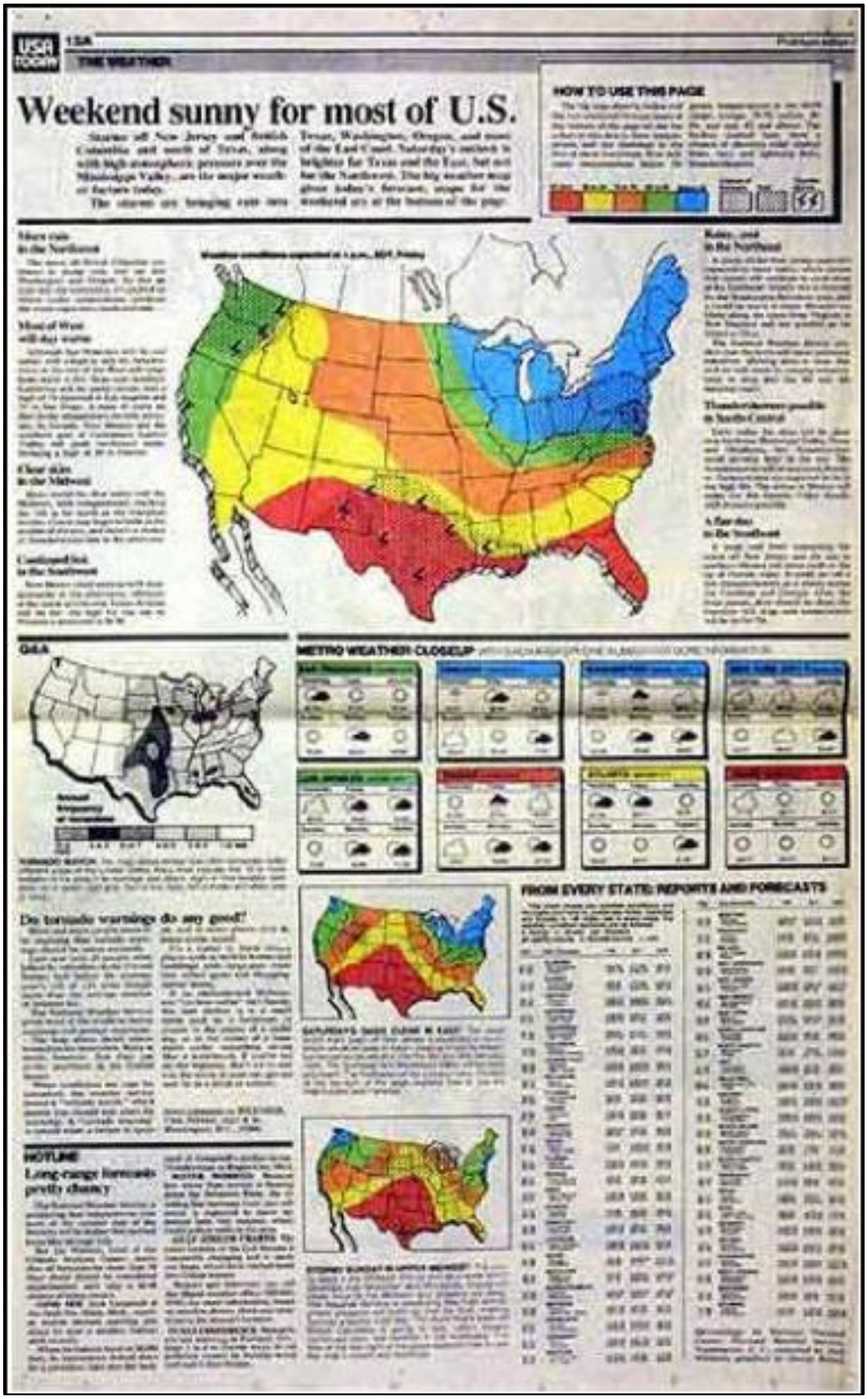


Figure 5 George Rodrick’s colored weather map featured in USA Today in 1982.



USA TODAY 13A

THE WEATHER

Weekend sunny for most of U.S.

HOW TO USE THIS PAGE

Q&A

METRO WEATHER CLOSEUP

FROM EVERY STATE: REPORTS AND FORECASTS

Do tornado warnings do any good?

HOTLINE

Long-range forecasts pretty chancy

Figure 6 Example of a visualization created through Visual.ly. Joint creation between General Electric and designer called Jess.

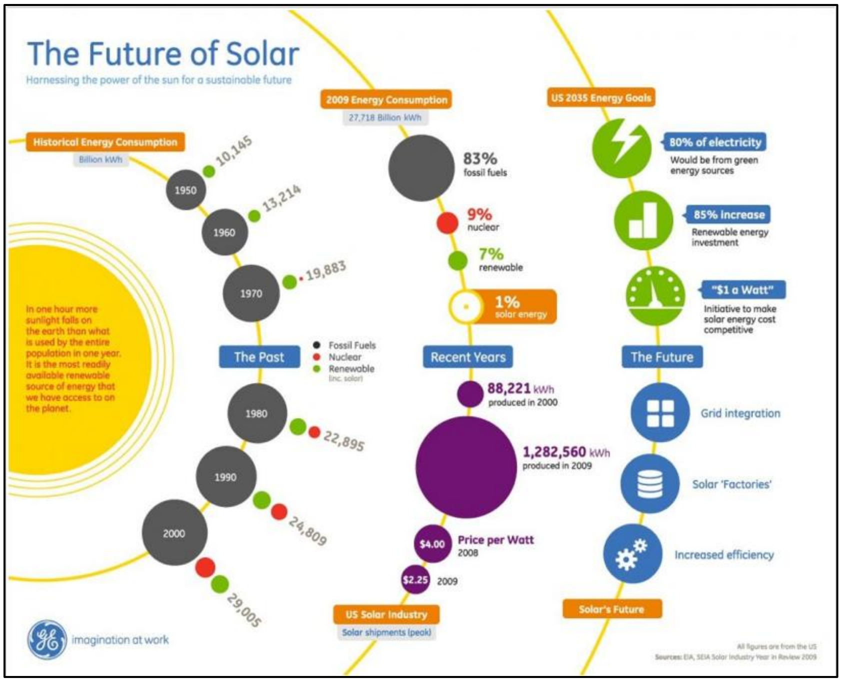
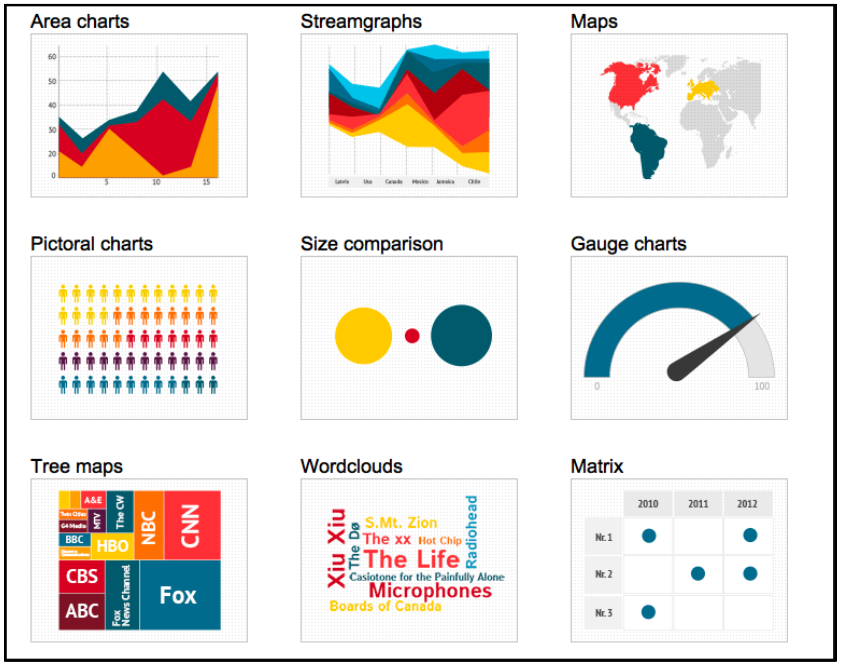


Figure 7 Readymade template collection available at Infogr.am.


Area charts
Streamgraphs
Maps
Pictoral charts
Size comparison
Gauge charts
0
100
Tree maps
A&E
MTV
The CW
NBC
CNN
BBC
HBO
CBS
ABC
Fox News Channel
Fox
Wordclouds
Xiu Xiu
S.Mt. Zion
The xx
Hot Chip
The Life
Radiohead
Casiotone for the Painfully Alone
Microphones
Boards of Canada
Matrix
2010
2011
2012
Nr. 1
Nr. 2
Nr. 3