A Grounded Theory Study of Open Data

From the Perspective of Developers Using Stack Exchange

Charlotte Buch

Thesis submitted for the degree of
Master in Informatics: Design, Use, Interaction
30 credits

Department of Informatics
Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

May 2019

A Grounded Theory Study of Open Data

From the Perspective of Developers Using Stack Exchange

2019

A Grounded Theory Study of Open Data From the Perspective of Developers Using Stack Exchange

Charlotte Buch

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Abstract

This thesis presents a two-dimensional study of which the first is a grounded theory study of open data from the perspective of developers, based on secondary data collected from content at the Stack Exchange network. The second is a study of the methodology through the implementation of Glaser's version of grounded theory, based on the first study.

The study of open data aims to understand the developers' relation to open data and to investigate how open data is used, and what it is used for. My findings demonstrate the scale of challenges associated with using open data, and show how the need for technical expertise and equipment in practice excludes the vast majority of people from being able to use open data. The results of the study indicate that open data is not inherently open to anyone even when it is shared openly for anyone to use for free for any purpose. The study also shows a tendency that developers are not particularly concerned with the concept of open data.

The study of grounded theory is intended to be a contribution to future students or researchers who are new to the methodology, and who are in search of an example of a grounded theory study based on secondary data that follows recommended guidelines for implementation and reporting. The thesis provides a thorough description of the entire research process and all the processes involved, from the preparation of the study to the iterative collection and analysis of data to the theoretical sorting. The study shows that it is possible to conduct such a study even in a short master's thesis without prior experience with the methodology.

Acknowledgements

This master's thesis marks the end of my time as a student at the University of Oslo. Through more than seven years of studies at the university, several of my dreams have been fulfilled. I have had the opportunity to experience a large university from within, and to get acquainted with the culture and ambience at no less than four faculties and departments. I have acquired knowledge above all expectations in several subjects from talented lecturers who have given their time to new students each year. However, what has made the biggest impression during my time being a student is all the great young students I have had the pleasure to meet. Despite my age, exceeding the age of their own parents, they have still welcomed me into their communities, and have taught me things about life that are at least as valuable as the purely academic.

I want to thank Tone Bratteteig for encouraging me to complete the master's degree by writing this thesis and not letting the chance pass me by when the motivation had weakened. I also want to thank Gisle Hannemyr for being my supervisor and following me through the process. I have appreciated our meetings and conversations. Finally, I would like to thank my close family and friends for the incredible patience and understanding they have shown me throughout this period. My semester has been challenging for everybody, I know, so now it is my turn to be supportive and helpful in every way possible.

Table of Contents

1 Introduction

1.1 Motivation and Choice of Theme

For nearly two decades I have been working within the financial sector, in IT departments where data constitutes the main component of all work being done. The amounts of data have increased each year, as well as the resources and knowledge needed to collect, administrate, secure, develop, and create value of the collected data. I have seen the consequences of the growing volumes of data from two different angles, from the inside and outside of work. Over the years, I have experienced an emerging distance, a growing gap between the two worlds. Inside the workplace, the consequences of the growing volumes of data have been perceived more as a challenge than an opportunity. While at the outside of work, the new services made possible by the collection of more and more data almost exclusively have been perceived as positive innovations, creating value for users. The distance between the experience of complexity on the one hand and the demand for simplicity on the other seems to have developed as technology has been put into use. Is it the technology itself that creates this divide, this gap?

In the context of the cities in which we live, the expectations for the development of new innovative services and smart solutions in our everyday life are also rising. The politicians have high ambitions, not the least for our own capital, Oslo, which aims to be “a smarter, greener, more inclusive and creative city for all citizens – a smart city that innovates with the citizens’ interest and well-being at the core.” (‘Smart Oslo Strategy’, 2019). This future vision for urban development to improve the lives of the citizens is driven almost exclusively by the use of information and communication technologies and the exploitation of data. By being open and connected, by sharing data openly and freely for everyone to use, businesses and entrepreneurs are invited to participate in building the services and infrastructures we need for our future urban environments. The prospects for open data are bright and promising, but are they aligned with the perspectives of the developers who are to develop these services based on open data? Or are we witnessing a similar distance here, between vision and reality?

In this master’s thesis, I set out to investigate open data from the developers' point of view. In order to gain insight into the developers' universe and to understand their position towards the concept of open data, it seems appropriate to study data created by the developers, content published for fellow developers on an open network. Therefore, I will carry out the

data collection through the Stack Exchange network and thus conduct the study on open data using open data and open metadata produced by developers themselves.

My stance and perspective on the topic of this thesis are critical. I am concerned about the ethical and political implications of the increasing digitisation of society. The developers have a crucial role in building our future and make the digital society happen. Therefore, I believe it is of great importance to understand their view of the world and the reality they are experiencing. We need to create a common understanding of what we want our society to be. For this to happen, we must have a way to communicate, a common language that bridges the gap between the technical and non-technical world. If we are not able to speak with the developers, and to understand their way of seeing the world, I am afraid we will be spectators and not participants in shaping our future.

1.2 Research Questions

1.2.1 The Role of Research Questions in a Grounded Theory Study

The methodology chosen for this study has a great influence on how the research questions should be interpreted and evaluated against the findings of the study. When using grounded theory, the initial research questions have a different role than in other types of studies. Because the insight in the topic is often low when the study is initiated, and it is advisable not to study literature in advance, the research questions will be based on neither theory nor domain knowledge or practice. Consequently, the research questions are purely indicative, and will work exclusively as a starting point for the entry into the topic and as an idea of how to approach the first data collection.

I choose to present my preliminary research questions from the initial project description both in this introduction and in the chapter describing the findings for two reasons. Firstly, I have been particularly committed to follow the recommended guidelines for a well-conducted and documented grounded theory study (Stol, Ralph, & Fitzgerald, 2016). Preferably, both the initial research questions and the questions that have emerged from the study should be stated, according to these guidelines. Secondly, I have conducted the study based heavily on each of the preliminary research questions, one by one. Through data collection, analysis and theoretical sorting with the question in mind, I have eventually arrived at answers to revised or new questions, brought forth as findings. As such, the preliminary research questions constitute an important element of the implementation and the structure of my thesis.

1.2.2 Preliminary Research Questions

The main topic of this thesis is open data. The purpose has been to investigate how and by whom open data is being shared and used, and how open data can contribute to innovation and development of smart cities. In the preliminary problem definition, the following research questions were set up as the foundation for the study.

- Which types of data are shared and made openly available?
- Who is using open data?
- What is open data used for?
- How can open data contribute to innovation and development of smart cities?

1.2.3 Revised Research Questions

As the study progressed, unknown and unforeseen contexts and interrelations within the landscape of open data were discovered. New insights changed continually the direction of the data collection and analysis, which in turn opened new interesting issues and relationships guiding the subsequent data collection and investigation. Where the questions turned out to be imprecise or less relevant, for various reasons, the process nonetheless uncovered new contexts that led the process forward. The research questions that are answered by the actual findings that have emerged during the study therefore deviate somewhat from the preliminary research questions.

The revised research questions are as follows.

- Which types of data are shared, and where is the data made openly available?
- Is open data open for anyone to use?
- How do developers relate to open data?
- How do developers relate to the concept of smart city?

1.3 Scope of Study

1.3.1 Preconditions

The fact that this is a short master's thesis of 30 credits has had a crucial impact on the scope and implementation of this study. An absolute requirement for short theses is that the study period extends over exactly 17 weeks from the time the project description is given until the completed thesis is to be delivered. This leaves very little time to each of the various elements that are included in the production of a full master's thesis. Compared to a regular master thesis of 60 credits, which in practice extends over a year and a half, 4 times as long as a short thesis, the administration of time is an additional challenge, especially when the subject being studied is not very well known in advance.

The research group I belong to at the Department of Informatics seldom offers topics or assignments that are suitable for short theses. It is two years since the last time a student from the research group wrote and delivered a short thesis. Consequently, the experience with supervising students conducting studies for short theses must be limited. Students are in fact advised not to write short theses, perhaps because it is equally demanding for the supervisor as for the student?

1.3.2 Perspective

As a starting point for the study of open data, I have chosen to take the perspective of the developers, which are supposedly one of the main user groups of open data. To understand what they are concerned with and what challenges they face when using open data, the questions and answers published at the Stack Exchange network seem a good place to start. A special site for developers interested in open data was created by the network in 2013, and this will be the obvious place to start the data collection. How many and what questions to investigate will depend on what they reveal, and at what time a certain point of saturation will be reached. Narrowing the study to include only secondary data, content on the Stack Exchange network has been a necessary limitation of scope, considering the time available.

1.3.3 Preparations

Before entering the actual study of Stack Exchange that will constitute the basis of my master thesis, I will use a couple of weeks on a broad investigation of concepts and conditions important for the understanding of open data. I will look into all types of available

sources from different actors and stakeholders to get an overall picture of the ecosystem. Throughout the study period, I will also attend relevant events, seminars and meetups organised by public and private institutions to be inspired by contemporary perspectives, ideas, assumptions and approaches to the topic of open data.

1.3.4 Process and Methodology

The principles of grounded theory will guide my analysis of the data through a process of constant comparison of codes, categories and notes. Although it seems intangible and highly unstructured for a newcomer, the methodology promises the opportunity to develop well-founded hypotheses, or at best, theory. However, examples of how to document and describe the evolution from data to theory seem to be non-existing for a grounded theory study without interviews, with only secondary data available. As much as this is a study of open data I therefore consider this thesis to be a study of grounded theory in a new context not documented before. Glaser states that it requires experience to be able to use grounded theory to its fullest (Glaser, 1978), but I hope nevertheless to be able to make a contribution that will prove useful to future students.

1.4 Structure of the Thesis

The thesis is structured as follows. The opening section, chapter 2, explains important concepts necessary for the understanding of open data and the study of data published at the Stack Exchange network. Chapter 3 presents the methodology of grounded theory and the rationale for the choice of version. In the proceeding section, chapter 4, the research process is described in detail, from the approach to getting started with grounded theory to conducting the study itself through the various phases, until theoretical sorting and development of theory. Next, the findings from the study of Stack Exchange are described in chapter 5 followed by a discussion of the implications and contributions of the findings in chapter 6. Finally, the thesis is rounded off with a brief conclusion in chapter 7.

2 Background

As a background for the study, this chapter gives a general presentation of the main concepts that are analysed and discussed in the subsequent chapters. First, the concept of open data is introduced from a historical perspective and then described through definitions from various actors. Next, the concept of smart city is discussed briefly, before the Stack Exchange network is finally explained in more detail to provide an overview of the use, the users and the structure of the network.

2.1 Open Data

Data has been collected and stored digitally for more than half a century. However, in recent years we have been witnessing an unprecedented explosion of data, founded by the rapid development of information and communication technologies (Kitchin, 2014). We have entered an era in which vast volumes of dynamic, varied digital data are easily collected, stored, distributed, and shared. The accelerating speed at which data is generated across networks, often in real time, enables the development of new innovative solutions and services not seen before. The volume, velocity and variety of data have the potential to transform business, government and society, especially if the data is shared openly (Janssen, Charalabidis, & Zuiderwijk, 2012).

The request for data to be open, and research results to be freely accessible and shareable, came initially in the 1940s from researchers who urged fellow researchers “to give up intellectual property rights to allow knowledge to move forward.” (Chignard, 2013). However, the term open data did not occur in a scientific document before 1995, and it took 15 more years before the open data movement began to gain traction. The movement to open up data for wider reuse, and to provide easier-to-use tools for analysing the data, accelerated in the late 2000s with initiatives in Europe and the US to open and publish government data (Kitchin, 2014). When the G8 leaders signed the Open Data Charter in 2013, promising to make governments “open by default”, the open data movement gained global traction (‘Open Data Barometer’, 2017). As of May 2019, Norway has not adopted the Open Data Charter.

2.1.1 Definition of Open Data

The definition of what characterises open data has gathered a large degree of consensus, but there are still small variations in the level of detail depending on perspective, whether the focus is governmental, legal or technical.

The International Open Data Charter defines open data as “digital data that is made available with the technical and legal characteristics necessary for it to be freely used, reused, and redistributed by anyone, anytime, anywhere.” (‘International Open Data Charter | Principles’, 2019).

Difi - the Norwegian Agency for Public Management and eGovernment - define open public data as “.. information that is made available so that it can be read and interpreted by both machines and people, and that anyone can access, use and share.“ (Difi, 2019). Difi recommends further that open data from the public sector should be well documented, linked and traceable, and be available in a machine-readable, standardised and structured format with guarantees of accessibility and reliability.

The EU, however, defines open data slightly different, referring to the definition of open data from Open Knowledge: “Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.” (‘Open Data in a nutshell - European Data Portal’, 2019). This definition applies to open data in general and is not limited to government data. According to the EU, there are two aspects to openness, that the data is: 1. technically open, machine-readable and in a nonproprietary file format, and 2. legally open, published under an open license where conditions for re-use are limited to attribution. An open license is a standardised agreement describing the properties that apply to all users and all purposes. The Norwegian license for public data (NLOD) is an example of such a license (‘NLOD | Difi’, 2019).

Two non-profit organisations based in the UK promoting open data - the Open Data Institute (ODI) and the Open Knowledge Foundation (OKF) – defines open data in exactly the same way as the EU. OKF refers to Open Knowledge, as the EU (‘What is Open Data?’, 2019), while ODI defines open data in a more popular language, but with the same content (‘What is “open data” and why should we care?’, 2019).

The W3C however, provides a comprehensive recommendation rather than a definition with best practices “related to the publication and usage of data on the Web designed to help support a self-sustaining ecosystem.“ (W3C, 2017). The introductory part of this document describing open data is slightly more directed towards the actual users, the developers and scientists, than the previous definitions: “Data should be discoverable and understandable by humans and machines. Where data is used in some way, whether by the originator of the data or by an external party, such usage should also be discoverable, and the efforts of the data publisher recognised.”. Interestingly, the word "discoverable" has been added to the definition.

2.1.2 Benefits of Open Data

Open data – data shared mainly from the public sector but also from the business world - holds the potential to provide benefits and to create value in a variety of areas such as transportation, health care, energy, education, just to mention a few (Janssen et al., 2012). Public sector and open data non-profit organisations are unconditionally positive and enthusiastic about the prospects of open data, which is expected to unleash economic value and benefits to governments, businesses, individuals and society at large. The commercial actors, on the other hand, do not seem to see the same potential in open data, and are not so noticeable in their communication and projection of the benefits and values that are supposed to lie in open data. Their focus appears to be directed more towards the smart city concept, where open data is a possible component.

Norwegian authorities consider the benefits of sharing data to relate mainly to the development of business and society. Difi highlights the following three main reasons why access to open data is important (‘Åpne data | Difi’, 2019):

1. **Efficiency and innovation** - Sharing public data with the private sector and within the public sector is expected to facilitate interaction to be more efficient, which in turn will lead to better public services.

2. **Business development** – Access to public information is believed to strengthen the private sector by providing opportunities for the development of new services, products and business models.

3. **An open and democratic society** - transparency and insight in public data will give the opportunity to verify priorities and decisions taken by the public administration.

The EU focuses on the same three main groups of benefits of open data (‘Benefits of Open Data - European Data Portal’, 2019). However, the EU also emphasises benefits such as sharing of knowledge, participation and inclusion in society, social welfare, and leaves special attention to unnecessary spending and economic growth. Through case studies, it is exemplified how open data can save lives, and how open data can help achieve environmental benefits. It is estimated that there will be 100,000 people working with open data within the 28 EU countries in 2020, an increase of 25.000 people within a four-year period.

The NGOs, OKF and ODI, have approximately the same promising expectations for the benefits of making data open and available for others to use. OKF highlights the three areas

transparency, social and commercial value, and lastly, participation and engagement ('Why open data?', 2019). ODI addresses a slightly different audience than OKF, seemingly younger people, and has therefore adapted the communication to the target audience with a more oral tone of voice. ODI lists all the same benefits mentioned above, democracy, improvement of services, efficiency, accountability, participation, and so on, but puts them in a more user-focused context. Open data, "How does it fit in?" ODI asks their audience, and answers: "Used for all sorts of things by all sorts of people" and exemplifies it with a multitude of examples from everyday life, such as preventing bicycle theft, helping people with allergies, getting fastest possible from A to B, visualisation of air traffic, and much more ('What is "open data" and why should we care?', 2019).

2.1.3 Open Data Trends

The open data movement is said to have gained global traction after 2013 ('Open Data Barometer', 2017), but how has it developed? The Open Data Barometer is "A global measure of how governments are publishing and using open data for accountability, innovation and social impact." (World Wide Web Foundation, 2017). Findings in their fourth edition of the global report from 2017 state that commitment to open data among the participating 115 countries is slowing down, in some cases even being undone. "Government data is usually incomplete, out of date, of low quality, and fragmented." and that "Procedures, timelines, and responsibilities are frequently unclear among government institutions tasked with this work. This makes the overall open data management and publication approach weak and prone to multiple errors." Even the Nordic countries seem to have lost momentum in recent years, as shown in Figure 1. For some reason, Norway has not been included in this list, but is at the top of the Nordic countries with a 10th place in the fourth edition of the barometer, up from the 17th place in the third edition from 2015.

Governments	Ranking ODB 4th ed.	3rd ed.	2nd ed.	1st ed.
Denmark	13th ▼	5th ▲	9th ▼	5th
Finland	20th ▼	11th ▲	12th ▲	14th
Iceland	36th ▼	22nd ▲	27th ▼	13th
Sweden	14th ▼	9th ▼	3rd ◀	3rd

Figure 1. Ranking of the Nordic Countries in the Open Data Barometer from the first edition (2013) to the fourth edition (2017). https://opendatabarometer.org/4thedition/report/

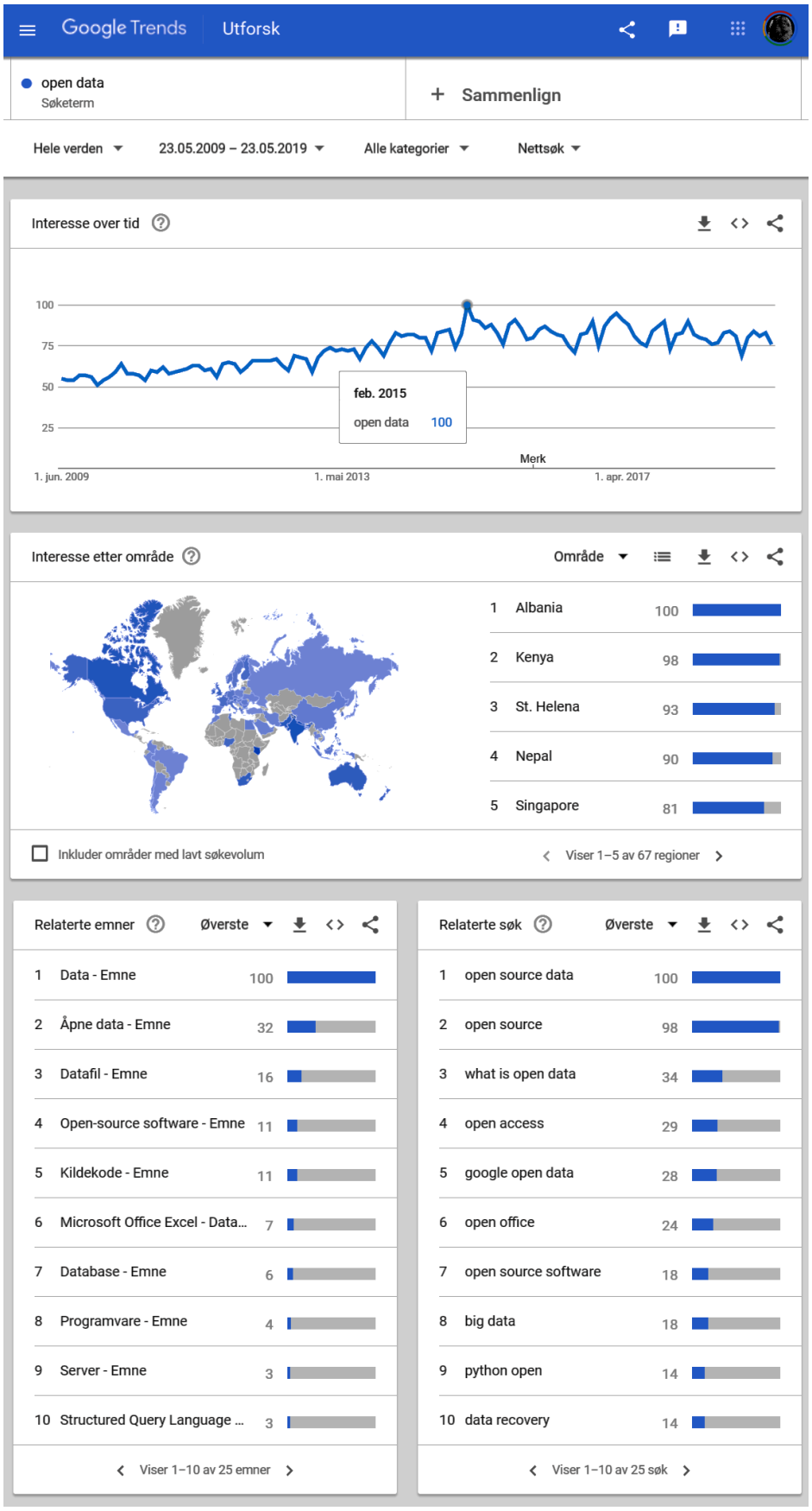


Figure 2. Google trends. Search term "open data". https://trends.google.com.

Google Trends is a useful tool that shows the frequency, or trend, of the use of a specific search term entered into Google over a chosen period of time. A search for "open data" on Google Trends over a period of 10 years up to today shows a growing trend curve peeking in February 2015. After this, the curve has stagnated and even appears to be slightly declining. See Figure 2. As such, the curve indicates a similar pattern as were found by the Open Data Barometer, that the global development of open data seems to be slowing down.

The interest in open data comes from all regions of the world, even though most of the countries in Africa are almost entirely unrepresented. Surprisingly, most of the searches for open data come from as diverse countries as Albania, Kenya, Nepal, Singapore and St. Helena, which is one of the most isolated places in the world. Less surprising is it to find high frequencies from countries like India, Pakistan, Canada, South Africa and Australia, where the data literacy is perceived to be much higher. Japan, Turkey, Russia, Vietnam and most of the countries in South America is at the bottom of the list.

Another significant finding from Figure 2 is that the two search terms open data and open source are so tightly linked together. Most of the users who searched for open data also searched for open source or open source data. This indicates a degree of relation between the two concepts, and could be an interesting topic for further studies.

2.2 Smart City

While there is an overall agreement on the definition of the concept of open data, the understanding of what constitutes and define smart cities is much more unclear. Since the term smart city was first used in the 1990s, many definitions and meanings have been given to it. Still, it is considered to be a “fuzzy” concept that is used ambiguously and inconsistently, creating confusion about what elements characterise a smart city. (Albino, Berardi, & Dangelico, 2015).

The general perception of what the concept entails has changed over time, from being particularly focused on the significance of information technologies, to increasingly emphasising the human side as well. While previously associated mainly with the technical infrastructure of urban environments and the diffusion of new technologies, social inclusion and quality of life are now recognised as much more prominent elements of the concept. It has become evident that the development of smart cities must be user-driven and focus on social values and needs in order to be successful. Instead of people adapting to technology, technology must adapt to people’s needs (‘What is a Smart City?’, 2014).

More than half of all people on earth are now expected to live in urban areas. In Europe alone, it is estimated that nearly 80% of the population already lives in cities (Albino et al., 2015). The demands for the input of resources and output of waste increase when the population grows, setting the administration of the city under increasing pressure to solve the new challenges that arise. New efficient and innovative services are needed within all areas of city planning, administration and urban infrastructure to find solutions to these issues, preferably in a way that ensures sustainment and circularity. Here, the notion of a smart city comes into play, embraced by governments and public agencies as a universal solution to the multiple challenges that lie ahead. At the same time a term that communicates well to the people.

Key components of a smart city include “hard” urban infrastructure in areas such as construction and building, energy, water and waste management, mobility, logistics and so forth. Included in a smart city is also “softer” components in areas such as culture, education, health care, social inclusion and more. However, technology and data have a crucial role in all these areas, enabling the smart city to happen. To increase and unleash the full value and benefits inherent in the data, the data needs to be shared publicly as open data. (Albino et al., 2015).

2.3 The Stack Exchange Network

This study will be based on secondary data mainly in the form of questions and answers published on the Stack Exchange network. This network is “the largest, most trusted online community for developers to learn, share their knowledge, and build their careers.” according to stackexchange.com. The network has staggering 1.3 billion page views per month. To date, more than 20 million questions have been published at the network, and it receives approximately 10 thousand questions per day on average.

Since the launch in 2010, the network has grown to consist of a total of 175 different communities within the main categories of technology, Culture/Recreation, Life/Arts, Science, Professional and Business. Each of the communities is focused on a specific topic. Stack Overflow is the oldest and by far the largest of the Stack Exchange communities, aimed at professional and enthusiast programmers. Each month, 50 million people visit this community (‘All Sites - Stack Exchange’, 2019).

2.3.1 Open Data Site

The Stack Exchange network contains a community gathered around Open Data - for developers and researchers interested in open data. It was created in May 2013, proposed by an American who is also interested in open government and open source (‘User Dmitry Kachaev’, 2019). As of May 2019, the site has gathered around 17.500 users and contains 4.600 questions of which 73% have been answered. Due to the low number of published questions per day - 2,3 questions on average - and the rate of answered questions, the site is still in beta. Figure 3 presents the metadata of the site.

Figure 3. Metadata of the Open Data site at the Stack Exchange network. https://area51.stackexchange.com/proposals/51674?phase=beta

2.3.2 Users of the Network

The “Stack Overflow Annual Developer Survey 2019” gives valuable information about the profiles of the users, their roles, education and details about how they work and what they think about different topics. The survey received nearly 90.000 responses in a three-week period in January/February 2019, and presents new information about and from a wide range of respondents from all over the world. This survey thus provides background information about the developers which is highly relevant for the study when I am to take the developers' perspective. For the sake of simplicity, I will consider the developers to be a relatively homogeneous group with the average characteristics found in the survey, which are as follows.

- **Demographics** - Most of the respondents are white or of European descent (71%), male (92%) under the age of 35 (75%).
- **Education** - Approximately 90% have a university degree or have studied at university level. 25% are still full or part-time students.
- **Employment** - 74% are employed full-time while 10% are self-employed. They work in a variety of industries, in companies of all sizes. 60% have changed jobs within the past 2 years.
- **Experience** - 41% have less than 5 years of professional coding experience. Academic researchers and data scientists have the fewest years of experience.

Location	Amount	%
Europe	36 073	40,58
North America	25 526	28,72
Asia	18 273	20,56
South America	3 459	3,89
Africa	2 850	3,21
Australia/Oceania	2 434	2,74
Other (country not listed)	268	0,30
Total	**88 883**	**100**

Figure 4. Residency of respondents of the Stack Overflow Developer Survey 2019.

The respondents of the survey come from 179 different countries, but most of them live in countries located on the three continents Europe, North America and Asia. Figure 4 shows the numbers. The percentage of respondents coming from the other three continents is very limited. The survey offers no answers to why there is such a difference in numbers of users between the continents.

2.3.3 Participation Reward System

Stack Exchange is a self-moderating network through a reputation award process that motivates to participation (‘Tour - Stack Exchange’, 2019). Users earn reputation by providing content, thus increasing their own privileges to edit and vote on other users’ questions and answers. Voting is the central mechanism of the process. It raises quality content over poorer content, and grants reputation and privileges to the users who are perceived to have produced the best content. For special achievements, users can also be awarded badges in gold, silver or bronze which in turn increases their reputation.

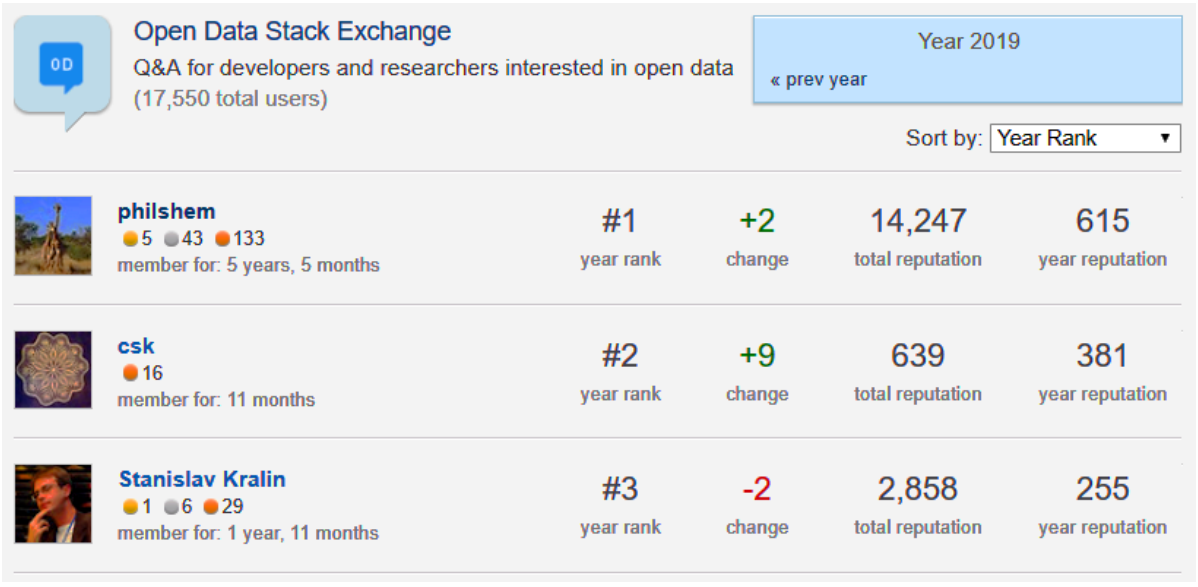


Figure 5. Top ranked users at the Open Data site 15.05.2019. https://stackexchange.com/leagues/498/year/opendata

Stack Exchange maintains extensive statistics and metadata on the users of the network and displays their rank, reputation and badges sorted at different time intervals. Figure 5 shows an example of a list of top ranked users from the Open Data site. Appearing at the top of these lists of highly-regarded and trusted users will most certainly provide high status within the communities. This reward system motivates and incentivises many users to produce more content of good quality and high relevance, and thus increases the use and credibility that makes it the large and worldwide network it is today.

3 Methodology

3.1 Choice of Methodology

The main topic of this master's thesis is open data in general and the issues of sharing and using open data from the perspective of the developers. On the recommendation of my supervisor grounded theory was chosen as the research methodology for the study. Grounded theory is a methodology that allows construction of theory from data through a systematic approach to data collection and analysis while being flexible in the choice and application of methods. It is especially suitable when trying to understand "what is going on" (Glaser, 1978) in an empirical context not known in advance. It allows to follow new interesting paths as they become apparent, and to select appropriate data sources and methods as the study develops. These characteristics suited the context and my purpose with the thesis well. In addition, it seemed appropriate in relation to my theoretic and thematic starting point, which was very limited.

3.2 Grounded Theory

Grounded theory was developed in the social sciences by Glaser and Strauss in 1967 through their book "The discovery of Grounded Theory". The authors argued that scientist needed a method that allowed to move from data to theory in order to create new and alternative theories based on contextual observations rather than predefined constructs from established theories.

Glaser and Strauss later disagreed on how grounded theory should be applied, which resulted in a split of the methodology in two versions, the Straussian version and the Glaserian version. The divergence between the two versions of grounded theory arose when Strauss published "Qualitative analysis for social scientists" (Strauss, 1987) in which he suggested a more prescriptive approach and a stronger focus on following procedures in the data analysis process. The Straussian version of grounded theory emphasises deduction and verification against known facts, by testing theory against predefined hypotheses. In contrast, Glaser's approach has remained true to the classical inductive form of grounded theory in which the process of data collection and simultaneous analysis of the data are directing the study forward towards the emergence of theory. Studies using the Glaserian version start out with a general area of interest, a problem, or simply a question rather than a specific research question. It is less formalised, considers everything to be data and

emphasises variation in methods and techniques. While Strauss recommends following guidelines, Glaser focuses on constant comparison and theoretical sensitivity, and less of a recipe.

3.3 Rationale for Choice of Version

The choice between Glaser and Strauss was determined by several aspects. First and foremost, it was the short time available that made Glaser the most relevant alternative. I had 17 weeks from the topic of the thesis was given until the thesis was to be submitted. Strauss requires the research questions to be formulated up front to set boundaries around the area of study, thus assuming that the researcher has a basis of prior knowledge of the phenomenon under study acquired through literature review. Glaser, on the other hand, believes that literature should be avoided before the study: “When the theory seems sufficiently grounded and developed, then we review the literature in the field and relate the theory to it through integration of ideas.” (Glaser, 1978). My professional experience as a developer using open data was limited to none, and my insight into the universe of open data was initially at an overall level. These reasons made Glaser’s version of grounded theory the most appropriate option.

3.4 Grounded Theory Process

Both versions of grounded theory aim to develop theories grounded in data, but the process and the methods involved are somewhat different depending on whether you follow Strauss or Glaser. When using Glaser’s approach, the process often starts with a question or a collection of data, as in my case. He considers everything to be data (Glaser, 1978) whether it is qualitative or quantitative data, semi-structured data, images, videos, audio files, or other types of data. This gives the freedom to develop perceptions and hypotheses about relationships from a broad basis of data sources.

Theoretical sampling is the continuous process of collecting and analysing data while simultaneously extracting concepts and categories. When the process progresses, the categories can be refined to a more abstract level where they can hopefully be integrated into a theoretical framework. Which data to collect and analyse further after the initial phase cannot be planned in advance. The emerging hypotheses, or theories at best, will decide which data to choose next.

Open coding is the process where data is broken down into distinct units of meaning, named to describe concepts and grouped into categories when patterns begin to emerge. The first codes will then be used to encode additional data, iteratively analysed and compared with concepts and categories that have emerged previously. This is a key component of grounded theory, the constant comparative analysis, where new codes and concepts are constantly being analysed against the ones that have already been found.

Selective coding is the process of identifying the central core category to which all other categories relate, and on which the theory is to be built. Once the core category is selected, it should guide subsequent coding selectively so that only concepts of importance to this category are included. Theoretical saturation is reached when the analysis of newly collected data no longer provides new insight or knowledge.

Memoing describes the process of taking notes and writing down ideas and questions during data collection and analysis, or whenever and wherever relevant thoughts suddenly come to mind. Memoing takes place throughout the entire grounded theory study. From being random and unstructured, the memos will become increasingly theoretical as they are continually complemented, sorted and structured. Ultimately, the memos will constitute the components of the theory, uncovering the main concerns and problems within the area of study.

4 Research Process

4.1 Approach to Getting Started with Grounded Theory

For this study, grounded was chosen as the methodology to be used. No other research methods were considered, partly because this is a short master's thesis in which the assignment and the topic are given and not chosen by the student, and partly because it was determined by the supervisor that this was the methodology to be used. I had no practical experience with grounded theory prior to this study, and I had only just heard it mentioned on a mandatory methodology course seven years earlier. At that time, little time and emphasis were given to grounded theory in this course compared to other methodologies. Consequently, it has been a study in the grounded theory itself, gaining insight into the methodology, its methods and variations, in addition to understanding how it can be used and applied in a context like the one I am facing. Finding relevant and useful sources of information for my purpose has proved challenging.

The original work of Glaser and Strauss describe the theoretical foundation of grounded theory and the actual processes and methods involved, on a general level (Glaser & Strauss, 1967). Understanding what the different phases entail is relatively simple, but finding descriptions and examples of how the coding, categorising and theoretical sorting is actually done in practice, down to the tools and techniques used for reporting the process, is not that easy. Although grounded theory is gaining popularity in software engineering, the methodology is not well understood by practitioners in this research area (Hoda, Noble, & Marshall, 2011a). A review of grounded theory used in software engineering research reveals the same problem (Stol et al., 2016). The authors of the review hold that many researchers have misconceptions of grounded theory and that they are unable to report and document the research process properly, despite this being of crucial importance for the evaluation of their studies. To address this issue, they have provided a set of guidelines which I have followed closely as far as relevant.

Furthermore, the design of grounded theory was developed to create theory through primary data, transcriptions of interviews with or without additional sources of data. However, within my scope there was not time to carry out enough interviews to arrive at an acceptable quality of the study. Instead, it was decided that I should examine secondary data, questions and answers published by developers and users at the Stack Exchange network.

There are no previous grounded theory studies based solely on secondary data to be found. This was a surprising discovery, to realise that I would probably be the first student to conduct such a study, having searched through the following sources.

- **Research articles about studies using grounded theory** in software engineering are few, and they are all based on developing theory through transcripts of interviews as the main element. However, there are good examples of grounded theory studies using interviews to learn from, among others the study of the daily stand-up meetings (Stray, Sjøberg, & Dybå, 2016) and the study of agile teams (Hoda, Noble, & Marshall, 2011b). They provided insight into the grounded theory process and how the study can be structured and reported.

- **Master theses and PhD theses using grounded theory** in the DUO Research Archive from UIO are even more sparse. A total of 12.222 documents have been registered under the community of the Faculty of Mathematics and Natural Sciences. However, a search for “grounded theory” in this collection of documents gives only 4 results, whereof one is actually not using grounded theory, and the remaining three are based on interviews (‘UIO : DUO Research Archive’, 2019).

- **Research articles about GT methodology** in software engineering are also based on primary data and on the assumption that the initial coding is carried out from transcripts of interviews (Coleman & O’Connor, 2007; Hoda et al., 2011b).

Because there were no previous studies or examples that could support me on how to perform the important reporting part of the study when I had performed no interviews, I decided to create an example for future students myself. This is the reason I have been so passionate about following the recommended guidelines for the implementation of this grounded theory study.

To begin the process of gaining insight into grounded theory, I followed the practical advice presented in “Grounded theory for geeks” in chronological order (Hoda et al., 2011a). As recommended, I started out reading the most important books by Glaser and Strauss. Then I examined several examples of grounded theory studies performed in the software engineering research area, in addition to studies of grounded theory itself. Subsequently, I conducted a minor literature review on relevant topics to set the proper terminology before entering the contemporary scene of open data. Although I was aware that literature review should be avoided up front, I found it necessary to have seen examples of how grounded theory studies were structured, reported and illustrated before starting up myself.

4.2 Grounded Theory Research Process

The overall topic of this thesis and possible approaches to the study were discussed and chosen in consultation with my supervisor a few days prior to the assignment was formally given. The project description was quickly prepared, and the preliminary research questions were formulated to get ready for the launch of the study 29 January 2019.

Figure 6 shows an overview of the different phases of the research process that I have followed. The first and initial phase of preparations was the most comprehensive both in terms of scope and time. The subsequent phase, the actual study of the content in Stack Exchange, was far more straightforward, and took place in an iterative process for each new collection of data. In the final phase, the findings were summarised, analysed and discussed, and the work on writing the thesis was completed. The remainder of this chapter describes the entire research process in detail.

4.2.1 General Overview

In the initial phase, I worked for several weeks to get a general overview of the topic of open data and related subjects. My prior knowledge of open data was limited, it soon turned out. I began searching for information from all types of sources and channels, from different communities, institutions, industries, actors and sectors, public and private, local and worldwide. At first, the search for information was perceived as being random and unstructured, but after a while, when relationships and patterns began to emerge, the search became gradually more targeted. The open data ecosystem revealed itself in increasing scope and complexity.

To keep track of relevant sources of information for later, I took notes in a spreadsheet. As the spreadsheet grew, I started categorising the sources of information into types, and certain structures began to emerge. However, the spreadsheet soon became large and unmanageable. Even though I created new sheets and started to sort the notes and links to sources of information on numerous spreadsheets, it eventually gave no further value to continue. I consider this to be my first experience of coding and categorising through constant comparison. Rather than saturation, a sense of information exhaustion had been reached. The experience and lessons learned from this very first attempt of coding showed that the reporting part of the study would probably be the main challenge to overcome.

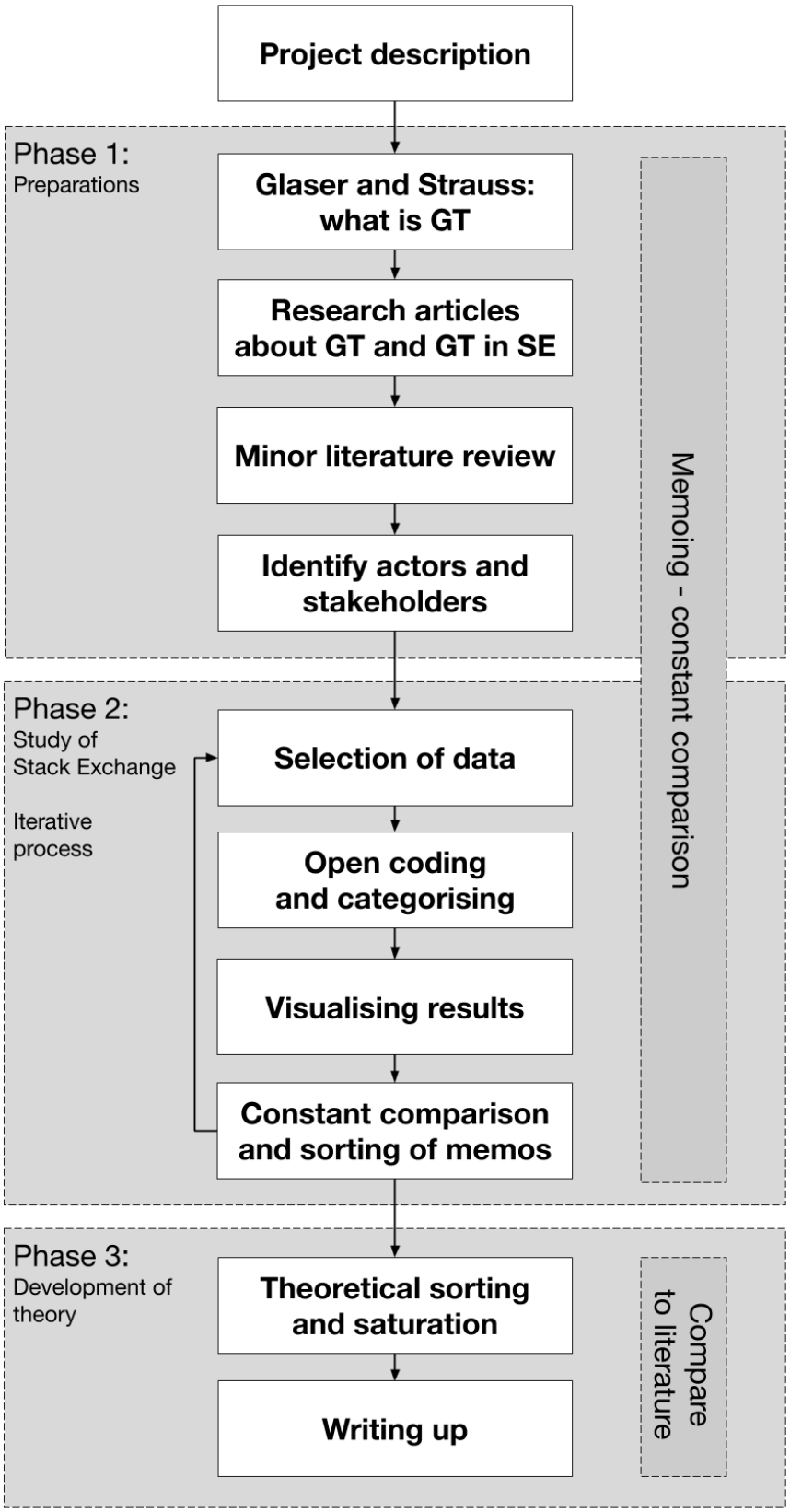


Figure 6. Grounded theory research process

4.2.2 Limitation of the Study

One of the preliminary research questions concerns a possible interrelation between open data and the concept of smart city. Having performed a light literature review on smart city and having consulted other information on the topic from public and private sources, I realised that I had to limit the scope of the study to focus on one single concept, open data. At first glance, smart city seemed to be a very diffuse concept with varying definitions depending on perspective. It would be far too time-consuming to include the subject in the study in order to add value to the thesis. In addition, when I did not find any clear support in the preliminary research questions for how to look at the relation between the two concepts, the decision was obvious, to continue with open data only. However, I still touched on the smart city concept in the last iteration of the data collection and made an interesting observation that will be discussed later.

4.2.3 Actors and Stakeholders

In the initial process of information gathering without having an exact point of departure, it was inevitable to constantly lose focus on which path to explore further. I tried continuously to return to the developers' perspective, but this gave little guidance because I did not know what roles the developers would have or could have in the system. In order to continue, it became necessary to first get an overview of all the actors and stakeholders involved in the open data ecosystem. Only then would it be possible to categorise them by type, and hopefully, to be able to identify which positions the developers possess in the ecosystem.

It turned out that developers, scientists and other IT specialists have many roles and can hold several positions in the open data ecosystem. Their expertise is valuable in most of the roles involved in open data. Then, based on my notes in all the spreadsheets, I made several attempts to visualise how the groups of actors and stakeholders relate and collaborate. I tried to group the actors from different perspectives, but failed to produce a map that provided useful information for my purpose.

When searching for inspiration from the world wide web, I found virtually no two similar illustrations of groups of actors in this space. The vision and the incentive in relation to open data appear to differ significantly dependent on perspective. On one level or another, most of the actors are both data owners, providers and users of open data, and make use of their own and other actors’ open data. Consequently, I simply chose to list all the actors without attempting to group them in any way, as shown in Figure 7. Although this figure perhaps

does not provide increased insight directly, the process of designing it has added to my insights and has been absolutely necessary for the continuing phases.

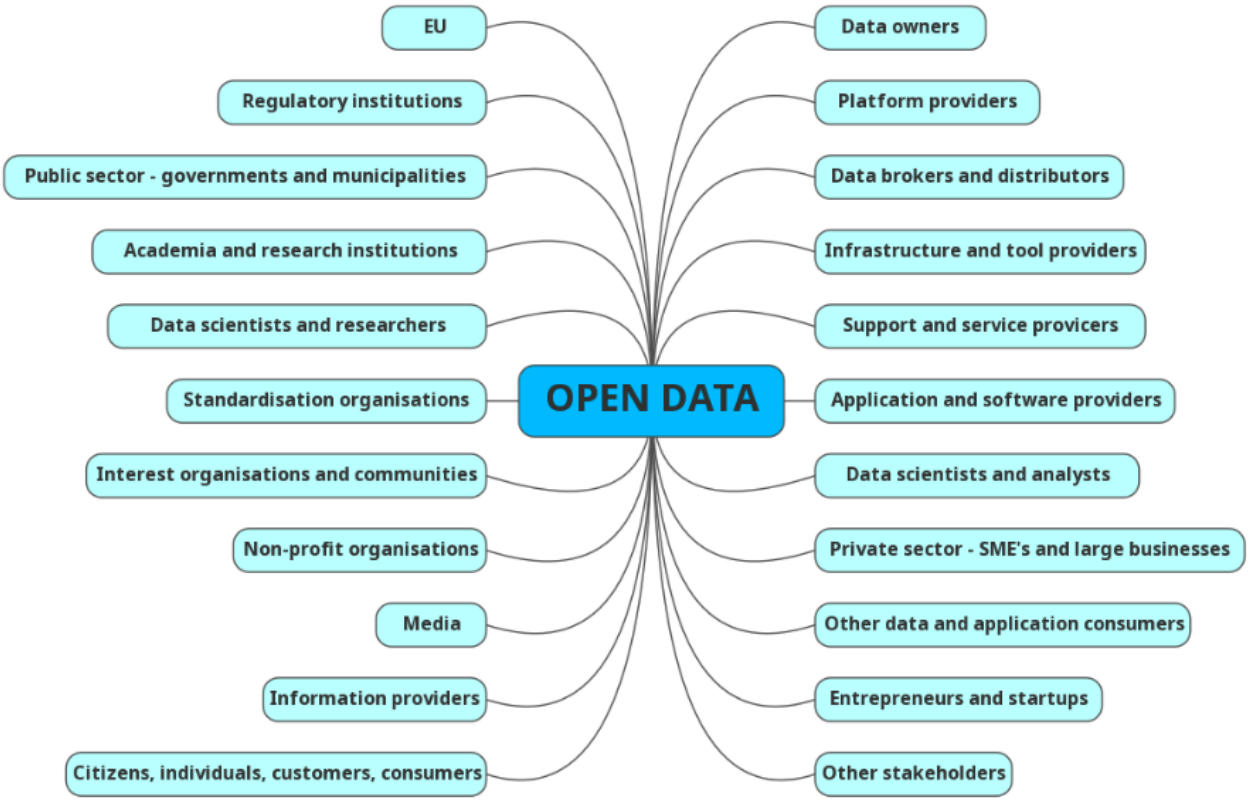


Figure 7. Actors and stakeholders in the open data ecosystem.

4.2.4 Selection of Data

In the project description, the selection of data for the study was defined to be questions and answers published at the Stack Exchange network, without further specification. Stack Exchange is a comprehensive and active network consisting of 175 sites and communities focusing on various topics. The Open Data site described as the “Q&A for developers and researchers interested in open data” appeared to be the most relevant place to start. As the analysis of the questions at this site progressed, and a sense of saturation was achieved, it became relevant to increase the selection to contain the entire Stack Exchange network.

4.2.5 Data Collection

The Open Data site contained about 4,600 questions during the period I visited it. The numbers were changing slightly up and down from day to day when questions were edited, moved or deleted by moderators. As the very first sample, I selected the most recent questions with the assumption that these questions would best reflect the initial issue of the questioner, before they were edited by the moderators. For each question, I recorded the type of data involved and the actual problem that was addressed. Subsequently, I selected a different collection of questions, the highest voted questions instead of the most recent. The idea was to compare and validate the previous findings against a more mature set of questions that were considered important and useful by many developers. The two sets of questions showed approximately the same pattern. The differences were minimal. I then continued to analyse how all the questions across the site were tagged by recording how many times the most popular tags were used. The result was consistent with the previous findings. Most of the tags described the type of data that the developer who had published the question was looking for.

The next sample of questions was extracted from the entire Stack Exchange network through a search for “open data”. The search generated just over 3.000 results. Almost 1.000 of these questions were published at the Open Data site. The remaining 2.000 questions were published mainly on Stack Overflow, but quite a few were also published on various other sites. The goal was to investigate how the questions were tagged and whether there had been a change in topics over the years. I started recording the tags in a table, registering also the year each question was asked. Unfortunately, the pagination on the search results page had a bug, with the result that I only was able to get access to the first 50 questions. It was not possible to verify that the next pages contained unique questions that had not occurred before. However, the result of the first 50 questions showed a clear pattern that might also have applied to a larger selection of questions. The variation was very large, and the tags mainly described technical terms, as it is shown in the table in Figure 8.

Finally, a search for the keyword "smart city” resulted in a sample of 29 questions, of which only 3 were published on the Open Data site. Considering that the network contains more than 20 million questions in total, a number of only 29 hits is a strong and significant result. A search for "smart cities" gave almost the same result, 28 hits, and the majority of these questions were the same as for the search term "smart city". Only one single question contained both the terms "smart city" and "open data".

Tags	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	Total
taxes								1				1
budget								1				1
online-data-source								1				1
government								1				1
science						1						1
duplicate						1						1
copyright				1								1
data				1								1
parsing										1		1
wikipedia										1		1
yql			1							1	1	3
c#				1	1							2
linq				1								1
ckan	1		1			1						3
discussion							1					1
stackexchange							1					1
area51							1					1
technology							2					2
android					2							2
android-intent					2							2
android-settings					1							1
opendata	1			1		1	1			1		5
bit.ly											1	1
sparql						1						1
semantic-web						1				1		2
wcf						1						1
hl7-v2					1							1
algorithm								1				1
graph								1				1
json				1			1					2
dojo							1					1
ibm-mobilefirst							1					1
swift		1										1
video		1										1
avplayer		1										1
avplayerviewcontroller		1										1
facebook					1							1
facebook-graph-api					1							1
r				1		1						2
rjsonio				1								1
xml			1									1
odk			1									1
amazon-web-services										1		1
amazon-s3										1		1
amazon-iam										1		1
amazon-sagemaker										1		1
python									1			1
odata									1			1
support							1			1		2
status-completed										1		1
data-explorer										1		1
api				1			1			1		3
ios					2							2
swift					2							2
rss					1							1
swift2					1							1
nsxmlparser					2							2
watchkit					1							1
apple-watch					1							1
odk								1				1
metadata	1											1
dkan	1											1
dkat	1											1
haskell						1						1
socrata				1								1
soda				1								1
javascript				1								1
node.js				1								1
rest				1								1
rgdal						1						1
internet-explorer									1			1
firefox									1			1
ftp									1			1
filezilla									1			1
asp.net					1							1
wpf					1							1
mvvm					1							1
mahapps.metro					1							1
ftp									1			1

Figure 8. Selected tags for 50 questions published across different sites at the Stack Exchange network containing the search term "open data".

To keep track of the data collection and to document the process as it progressed, all information was recorded in a master document, shown in Figure 9. For each collection of questions that were analysed, a new spreadsheet with codes, categories and findings were created and linked up from the master document. In addition to the six data collections listed in Figure 9, I examined several other collections of questions across the Stack Exchange network. However, for these collections I did not reach the same level of completeness and structure as for the collections listed in Figure 9. At this time where almost half of the study period had passed, it was unclear how long it would take to complete the remaining tasks with the completion of the thesis. I therefore prioritised to continue with the insights and findings I had already acquired, to have a better time for analysis, memoing and writing up the theory, and especially for the discussion and conclusion of the thesis.

	Source	Selection	Description of collection	Findings	Link
1	**StackExchange - Open Data site**	Most recently asked questions	**Topics / Types of data** 153 questions of 4.605 in total (All questions published in 2019: 01.01.2019 - 10.03.2019)	Most questions: search for specific datasets **Categories of topics** **Types of data**	https://opendata.stackexchange.com/
2	**StackExchange - Open Data site**	Questions with the most votes	**Topics / Types of data** 50 questions of 4.605 in total with 18 - 139 votes (11.03.2019)	Similar categories as above, but more focus on data quality and maintenance of data, and a couple of new categories **Categories of topics** **Types of data**	https://opendata.stackexchange.com/
3	**StackExchange - Open Data site**	Most used tags	**Tags** 36 most used tags for all 4.605 questions (10.03.2019)	More than half of all questions is tagged [data-request] **Tags**	https://opendata.stackexchange.com/tags
4	**StackExchange - all sites**	Search for "open data"	**Tags** 50 first questions of 3.084 in total (16.03.2019)	The variation of tags is very large **Distribution of tags**	https://stackexchange.com/search?q=%22open+data%22
5	**StackExchange - Open Data site**	Search for "smart city"	3 questions of a total of 4.603 (09.03.2019)	No direct relation to the concept of Smart city	https://opendata.stackexchange.com/
6	**StackExchange - all sites**	Search for "smart city"	**Topics / Tags** 29 questions of more than 20 million questions (16.03.2019)	Most questions have no relevance to "Smart City". A few questions refer to a Smart City context. **Categories of topics**	https://stackexchange.com/search?q=%22smart+city%22

Figure 9. Excerpt from the master document that provides a full overview of the data collections.

At the time the project description was prepared, I had no practical experience with grounded theory, and anticipated that the data sources to be used in the study had to be clearly delimited. In this regard, the questions on Stack Exchange were an appropriate choice for the selection of data to be studied. Later I realised that the methodology allows for all other imaginable data sources to be used to elucidate the subject under study from new

dimensions and perspectives. Using many different sources of information throughout the study period has been invaluable to my study of open data.

In addition to the questions on Stack Exchange, the Stack Overflow Developer Surveys from 2018 and 2019 have obviously been important sources in understanding the background of the developers. It has also been useful to look at the structure of the Stack Exchange network itself and the additional services it provides. Beyond that, I have used unstructured and semi-structured data like videos, podcasts, text documents, reports, surveys and news from national and international media. Participation in relevant seminars, meetups and conferences have also been important elements in the continuous information gathering and memoing towards - what may perhaps be described as - the development of theory.

4.2.6 Open Coding and Reporting

Grounded theory refers to open coding as the initial form of abstraction from the data allowing concepts and categories to emerge through constant comparison and analysis. In theory, relatively straightforward, but in practice more challenging, especially in relation to reporting. Digital tools like NVivo can be used in qualitative research to automate manual tasks associated with the analysis such as classification and organisation of data and codes (Universitetet i Oslo, 2017). These tools are mentioned to be considerably time-consuming to set up and also complex to use. Consequently, given my time frame and scope of study, using NVivo was not an option. However, images of diagrams and figures created by the software published openly on the Internet were helpful in the further process.

In the literature on grounded theory, I have found little information about how the codes should be recorded, in which formats, and how the categorisation process is to be organised and documented. Figures showing the iterative process of grounded theory, demonstrate the open and selective coding solely as two boxes associated with an arrow. For some reason, I had a strong assumption that special techniques would apply to this process. But which? After several attempts to write the codes in different ways on coloured cardboard and post-it notes, structuring them on the kitchen table, on the walls and on the floor, I eventually gave up. Finally, I understood that an ordinary Google spreadsheet would be the most appropriate solution to my challenge, being editable, portable, easy to copy, available anytime, anywhere on every device, having diagrams and charts nearby. The coding could begin.

Understanding how to code, from what perspective to encode the questions, was also a learning process. Glaser's version of grounded theory suggests several framing concepts that may be used, so-called coding families (Glaser, 1978). I selected the coding family that

appeared to be the most relevant for my content, the C-family, looking for causes, contexts, consequences, and conditions within the questions. This endeavour proved to be far too extensive for my purpose. Inexperienced as I was, it only increased the complexity. Ultimately, I chose to start coding with only one dimension in mind, context, investigating what was the issue being addressed. At this time, I already had a good overview of the actors and roles involved in the open data ecosystem, as described and visualised in Figure 7 above. This knowledge influenced the coding, and made the categories begin to stand out. The categories were renamed and regrouped several times as new codes changed the scenery, until a stable framework appeared to have been established where new codes no longer gave additional insight. This process was repeated for each new collection of data. While coding and categorisation took place, increased insight raised new questions, thereby opening possible directions of which data to analyse next, and from what perspective to proceed.

In parallel with the coding and categorisation of the questions in Stack Exchange, figures and diagrams were developed to visualise the results. Various types of charts were tried out for each collection to find the best representation. Transforming the codes and categories from appearing in rows and columns in a spreadsheet to materialise as visual shapes and figures had a profound impact on my understanding and interpretation of the results.

The proceeding step in grounded theory is selective coding and the process of theoretical sorting where the core category that relates to all other categories is to be found. The core category is supposed to constitute the fulcrum of the study results. In this modest project, I experienced that the selective coding became an integral part of the memoing process, impossible to distinguish as a separate phase.

4.2.7 Memoing and Theoretical Sorting

The memoing proved to be a very important part of the entire study. It all started with some very simple notes in the margin here and there and everywhere, and ended up being a very large document constituting the most important components of the thesis. This was where the real work took place, where the actual insights and foundation of the findings matured and developed.

Memos were written continuously right from the beginning of the project period, at any time of the day, during coding, reading, writing, searching, or whenever an idea or relevant thought came to mind. The memos could be any kind of short or long notes, ideas, or questions that could prove to be valuable later. Experiences of the actual process of

implementing grounded theory were also described and documented. I realised early on that the thesis would be a study of grounded theory in itself as well as a study of open data. Therefore, it was of particular importance to comment on how I experienced the different phases while they were still going on. Otherwise, I would probably have forgotten what challenges I had faced at the various stages, when I later was to write up the thesis.

As the number of memos increased, they were collected in a common Google document to make it possible to work with the notes in a more structured manner. This memo document gained increased importance as it was filled with content and the analysis of the content progressed. The memos were thoroughly reviewed, reformulated, and repeatedly sorted until an overall framework began to come into view. From being a chaotic document of nearly 40 pages, it gradually changed characteristics to consist of ordered sections consisting of all the concepts and categories of themes that had emerged during the entire study.

Next came the challenging but intriguing process of keeping the memos up against the actual findings from the data collection, the theoretical sorting. It involves circulating deep down in all the information that has been gathered, structured in different ways, from different dimensions, and starting to see new connections and explanations develop. Separate elements of the study suddenly appear to be interconnected and coherent at a higher level, and to constitute a system, a tangled network of causes and conditions.

4.2.8 Writing up the Theory

In preparation for writing up the theory, the memo document was then again restructured and combined with the findings from the data collection. The content and categories were further divided into components that more or less matched the formal requirements for the structure and contents of master theses. These components thus worked as a first draft for several of the sections and subsections of the thesis.

Now that the basic framework was well established, it was time to perform the review of relevant literature. Only when the study is completed, and the core category and the framework of the study have been established, similarities and convergences with existing literature should be investigated (Glaser, 1978). In practice, however, I experienced that it was necessary to study relevant literature earlier in the process, even before I started writing the first chapters of the thesis. The literature created an important counterweight to the commercial views on the subjects I were studying, and should not be isolated to be used at the end of the study, in my opinion.

Another aspect that was perceived as challenging was the use of the standard template for theses. Regardless of the methodology chosen for the implementation of a study, the standard template is to be used for the final thesis. The structure requires the theoretical background and methodology chapters to be presented before the findings and discussion, which unquestionably requires a review of literature up front. In Glaser's version of grounded theory, the process is reversed. Literature should not be reviewed until after the study. When following the methodology strictly, writing can only begin when the entire study is completed. The problem is that when writing, the theoretical sorting and creation of theory continues until the end, according to Glaser. My point is that Glaser's version does not fit the template, or opposite, the template does not fit to Glaser. This mismatch causes an extra iteration of the study to be added to the end of the research process, enlarging the writing phase significantly.

Consequently, the memoing and theoretical sorting continued throughout the entire process of writing up the thesis. New patterns and relationships emerged constantly as the sections were written, then again requiring finished chapters to be rephrased and restructured. Theory is obviously a momentary construct, a continuous process, developing until the very last sentence has been written and the thesis is to be delivered. This is especially true when following Glaser's version of grounded theory, I will argue.

4.3 Reliability and Validity of Data

The amount of data I have been able to collect and analyse during this short period of time has been limited, both when it comes to the number of data collections and in terms of questions in each collection. This raises the question of whether the findings would have been different if the amount of data under study had been larger, or wider, and whether the findings are representative of a larger selection. I am not in doubt that a more comprehensive study of the data in Stack Exchange, cross-examined from different dimensions, supplemented by interviews or surveys, would have led to findings of a completely different depth and quality. Nonetheless, despite the limited amount of data, I still believe that the findings I have made illuminate aspects that were not very well known in advance.

Then there is the question of whether Stack Exchange has been the right place to seek answers to the preliminary research questions about open data. Considering that I was determined to take the developers' perspective on open data, and that Stack Exchange is the world's largest programming community for developers and researchers and even have an open data community, I am convinced that this has been a relevant choice. The idea of

having conducted a study on open data, based entirely on openly available collections of questions and answers, supplemented with open statistical data and metadata about users and their use of the network, makes good sense. Conversely, there is also the question of whether I asked the right preliminary research questions, having the developers' perspective in mind. However, since this is a grounded theory study that is not based on theoretical knowledge up front, it is expected that the research questions must be revised during the study, and

Just as the size of the data collection may have influenced the findings, the method used may also have had significance. The outcome of studies using grounded theory is to a large extent dependent on the researcher who undertakes the study. Demographic background, core values, scholarly education, motivation, background knowledge of the subject being studied, are just some of the elements that influence the researcher's position to how the data is perceived and analysed, and how theory is created. Two grounded theory studies based on the same data will most probably produce different results. Consequently, the documentation of the entire process is all the more important. To be able to trace and evaluate how the results have been reached, increases the credibility and reliability of the study. Special attention has therefore been paid to reporting of the research process.

My personal background and life experience have influenced the outcome of this study in several dimensions. First, I am an adult student with a long professional career in the private sector, first as a developer, later with other roles cooperating with developers every day. I know their world from within and feel their struggles. I have lived life for a while, and remember the day the TV was turned on for the first time. I have seen society transform, but still I am excited to see new technology change the day. What worries me is who masters the technology, and for what purposes. However, what frightens me the most is to see how political and financial agendas destroy our globe and the very basis for human life.

4.4 Strengths and Weaknesses of Using Grounded Theory

Grounded theory has been an appropriate choice for this study of open data for many reasons explained in the previous chapters. In my context, the methodology has proven its strength by allowing me to go straight to the data collection without the same amount of preparations as in other types of studies. However, this is only due to the fact that there have been no interviews involved. I have saved time by not having to develop a thorough theoretical basis and well-founded hypotheses in the first phase of the study. Instead, I have

been able to use the preliminary research questions from the project description as a guide to enter the topic, and quickly get started with the study once I had gained insight into the methodology.

Understanding how to use grounded theory was much more time-consuming than anticipated, however. To get acquainted with the methodology for the very first time in a master's thesis is not to be recommended, especially not in a short thesis. It leaves too little time for the study itself. With less time to investigate the data, there will be less material to analyse and discuss, which further implies that the training in using the methods and techniques through the different phases will be limited. This became evident very early in the project period. I could have chosen to spend more time on studying the actual data in Stack Exchange, and less time on trying to understand the actual methods and techniques of grounded theory properly. I am quite convinced, however, that the consequence of such a choice would have been a study and a thesis without the required structure to be considered credible. Therefore, I considered it to be absolutely critical to spend the necessary time to be confident with the methodology prior to the study of the data. By making this a study of the chosen methodology as well as open data, I was able to acquire sufficient understanding of the processes involved to carry out the study in a more controlled manner. It made it clearer how the important reporting of the findings could be carried out.

Is grounded theory appropriate for a study of the chosen format of secondary data, random technical questions published by people within a context I know little about? Is it possible to extract any qualitative knowledge from a text that strives to contain only technical language? As for this question, I am more in doubt. I experienced the coding process of the questions more like a statistical categorisation than an abstraction of the real issues addressed by the questioners. This may be a consequence of my inexperience with the methods though. Perhaps I could have conducted the initial coding at a different and deeper level and have reached other types of results? Or maybe grounded theory can be used more successfully with some minor adjustments to the methods, customised to the use of secondary data?

In the new data-driven era, we will see changes in the way we live happen in the very near future. The digitisation happening around us will affect our social lives to a degree we are not even able to imagine. Phenomena and environments will emerge that cannot be explained from any previous knowledge. In these new contexts, grounded theory may prove to be a useful methodology that will have the potential to provide valuable insights about what is really happening, without being dependent on previous research or theories. This requires that the guidelines must be followed, and that grounded theory is not only used in words.

5 Findings

Now the time has come to present the findings from the study of the questions at the Stack Exchange network, and to explain how the preliminary research questions have led to these findings. For each of the findings, both the preliminary and the revised research question are listed before the explanation and discussion of the finding itself.

5.1 Finding 1: Open Data Can be Hard to Find

Preliminary research question: *Which types of data are shared and made openly available?*

Revised research question: *Which types of data are shared, and where is the data made openly available?*

Having studied content published on the Stack Exchange network for a few days, it became clear that the first research question is possible to answer only at a very high level, for several reasons. First of all, there are no comprehensive lists or services that provide information about which types of data is openly available at any given time at any given location. Open data is scattered across numerous platforms, databases and networks, provided by the public and private sector, nationally and globally, and it is changing rapidly and constantly. This is exactly the problem developers seem to face. First, how to determine which data is shared openly? And next, where are the datasets or APIs providing this data to be found? In which formats? How is the quality, validity, access, licensing and maintenance of the shared data? The developers' questions about open data reveal challenges in multiple dimensions.

What I found through the review of the questions was that developers are asking their peers for help to find the data they need. In fact, the majority of the questions posed and published at the Open Data site are asking where to find data of a special type or format. More than 2.500 questions of a total of 4.600 were tagged with data-request as the main tag. The second most commonly used tag was geospatial with only 550 appearances, even though geospatial data is highest in demand, as we shall see later. Figure 10 shows the distribution of the 36 most selected tags across all the 4600 questions published at the Open Data site. The figure indicates that the site is mainly used to request data. The same pattern appears when reviewing the actual content of the questions. Developers seem to use the site mainly as a reference, by asking their peers where to find specific types of data in a certain format,

quality, range, or with other characteristics. Based on this, I will hold that **even though data is shared openly, it is a challenge even for experienced developers and researchers to find it.**

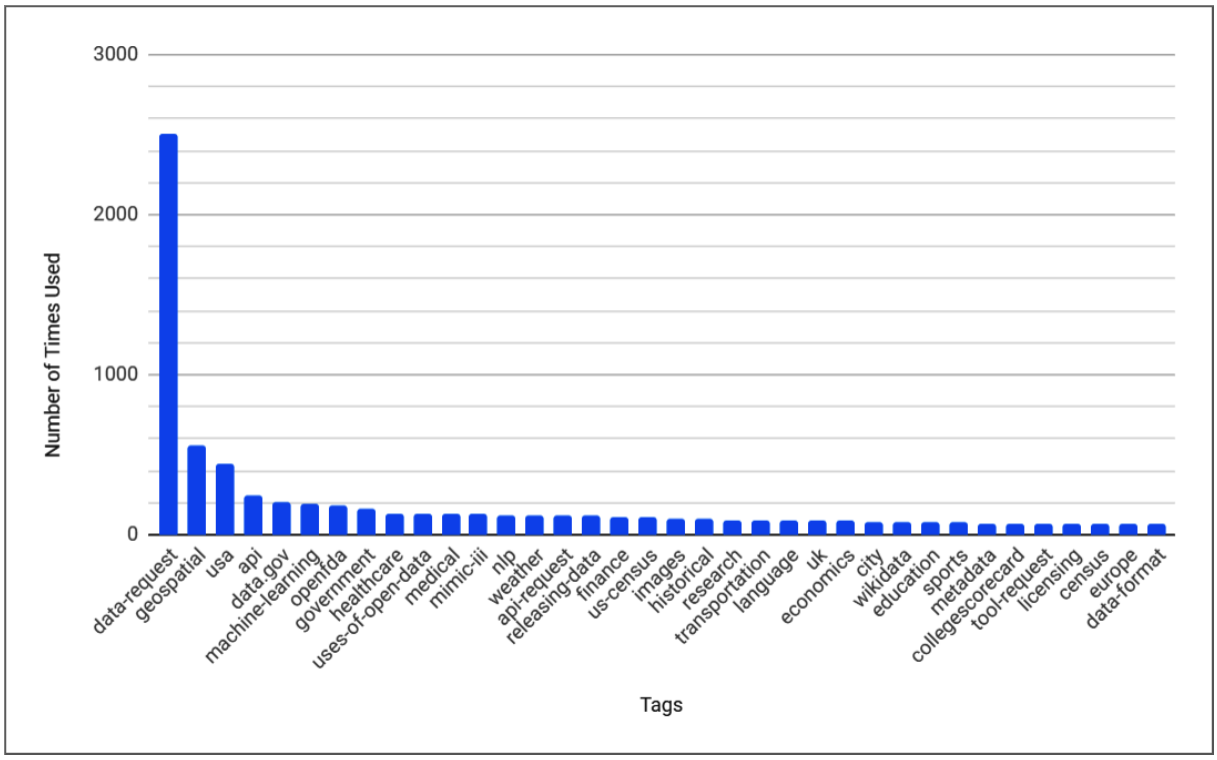


Figure 10. Distribution of the 36 most selected tags at the Open Data site 10.03.2019. opendata.stackexchange.com

Another finding from the review of the questions was that the **developers are searching for a surprisingly large variety of data types**. By registering which types of data were discussed in the 153 most recent questions published at the Open Data site, and categorising the types by subject, a very large variety of categories emerged. See Figure 11. The aggregation of subjects could, of course, have been done at a higher level to reduce the number of categories. However, the categorisation already carried out proved to be difficult enough due to the great variety of subjects requested in the individual questions. There are almost no limits to what data the users are requesting, right down to very detailed data on various subjects. Very few users ask for the same types of data, but this is not so surprising since there are many answers to be found in previously published questions. As such, the Open Data site functions as a form of reference consisting of a large and continually updated base of information about where to find specific types of open data.

Geospatial data is unquestionably the type of data that is most in demand, we see clearly from both Figure 10 and Figure 11. The reason for this may be that geospatial data is often used to visualise the geographical distribution of other types of data.

It is interesting to find that the developers seem to be quite confident that the very special data they are looking for exists and is shared openly, somewhere. Answers provided by other users of the network often give good suggestions or solutions to where the data may be found or obtained. I have come across only a few questions that have been given a purely negative answer, that the data does not exist or are unavailable. The reason for this may be that the questions are simply left unanswered if there is no good answer to the question. 27% of the questions on the Open Data is left unanswered as of May 2019 according to the numbers in Figure 3. I have not investigated the reason for this any further, but it would be an exciting question to pursue.

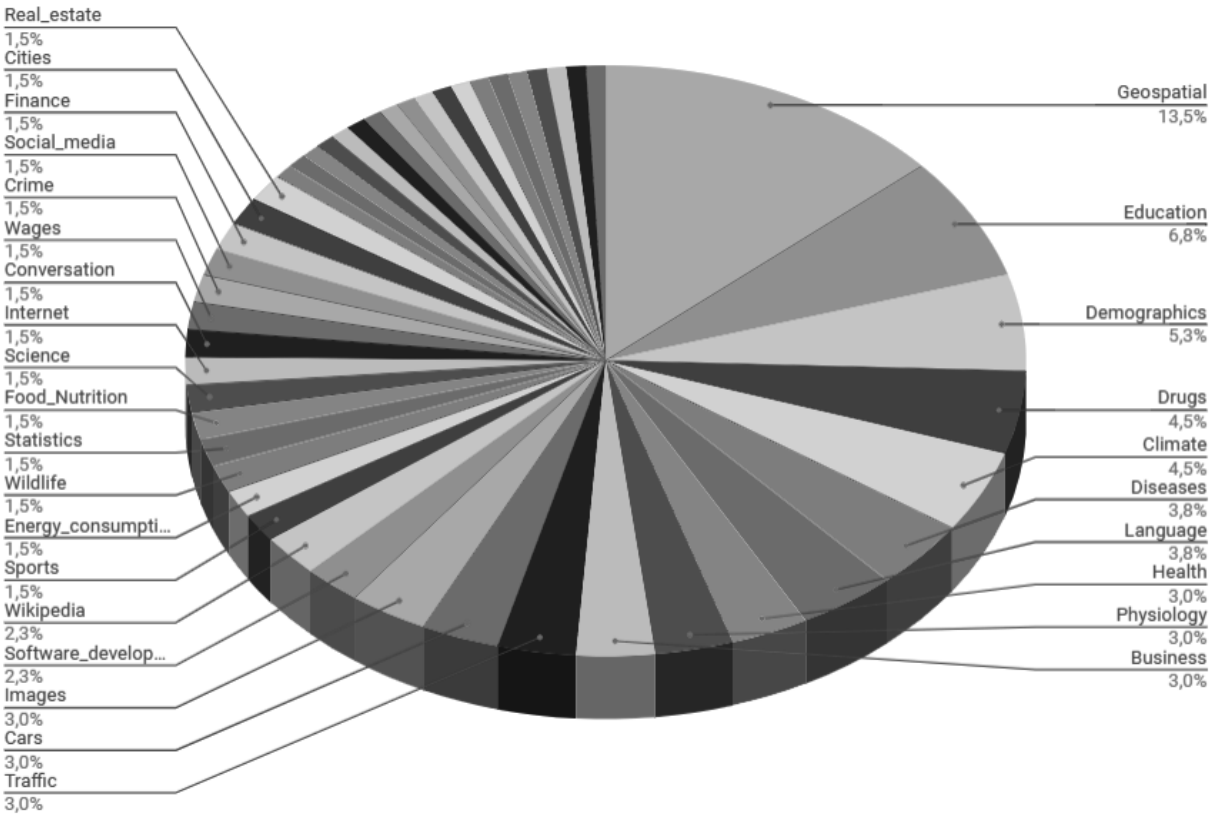


Figure 11. Requested data types in questions published at the Open Data site between 01.01.2019 and 10.03.2019.

5.2 Finding 2: Open Data is Not Inherently Open to Anyone

Preliminary research question: *Who is using open data?*

Revised research question: *Is open data open for anyone to use?*

This research question immediately triggered a new question. What does it mean to use open data? This must be specified before it is possible to determine who the user might be. I see two main types of use. Using open data can either mean using applications or services that are based on open data. Users - that I would call end-users - will then be individuals or organisations that do not relate directly to the underlying open data. On the other hand, it can mean working with data, processing data to develop these applications and solutions for the end-users. Users will now be the developers and researchers. Because I have chosen to take the developers' perspective for this study, I will assume the latter type of use in the continuation.

Based on the sheer fact that the open data community exists leaves no doubt that the users of Stack Exchange are using open data. However, this only partially answers the question of who is using open data. Seen from a larger perspective, the users of Stack Exchange are a relatively small group of individuals with special characteristics. Who are they? Is it possible to find out more about who they are from the data published on the Stack Exchange network? Further, can this information tell anything about developers and researchers in general? I have investigated what characterises the users of Stack Exchange in an attempt to find answers to these questions.

The Stack Overflow Developer Survey (2019) reveals that the respondents, mainly developers, work in a diverse range of industries, in the public and private sector, in NGO's and within education and research, as shown in Figure 12. By comparing this figure with the map of actors and stakeholders involved in the open data ecosystem in Figure 7, certain interrelations emerge. Many of the actors can be found in the list of industries that the developers work within, and vice versa. Consequently, it may be reasonable to assume that developers are involved as actors in the open data ecosystem themselves, not merely as developers or technical advisors for decision makers, but probably also in other roles with high demands on technical knowledge and expertise. My aim with this is to demonstrate how some of **the respondents of the survey may possess several of the important positions in the open data ecosystem.**

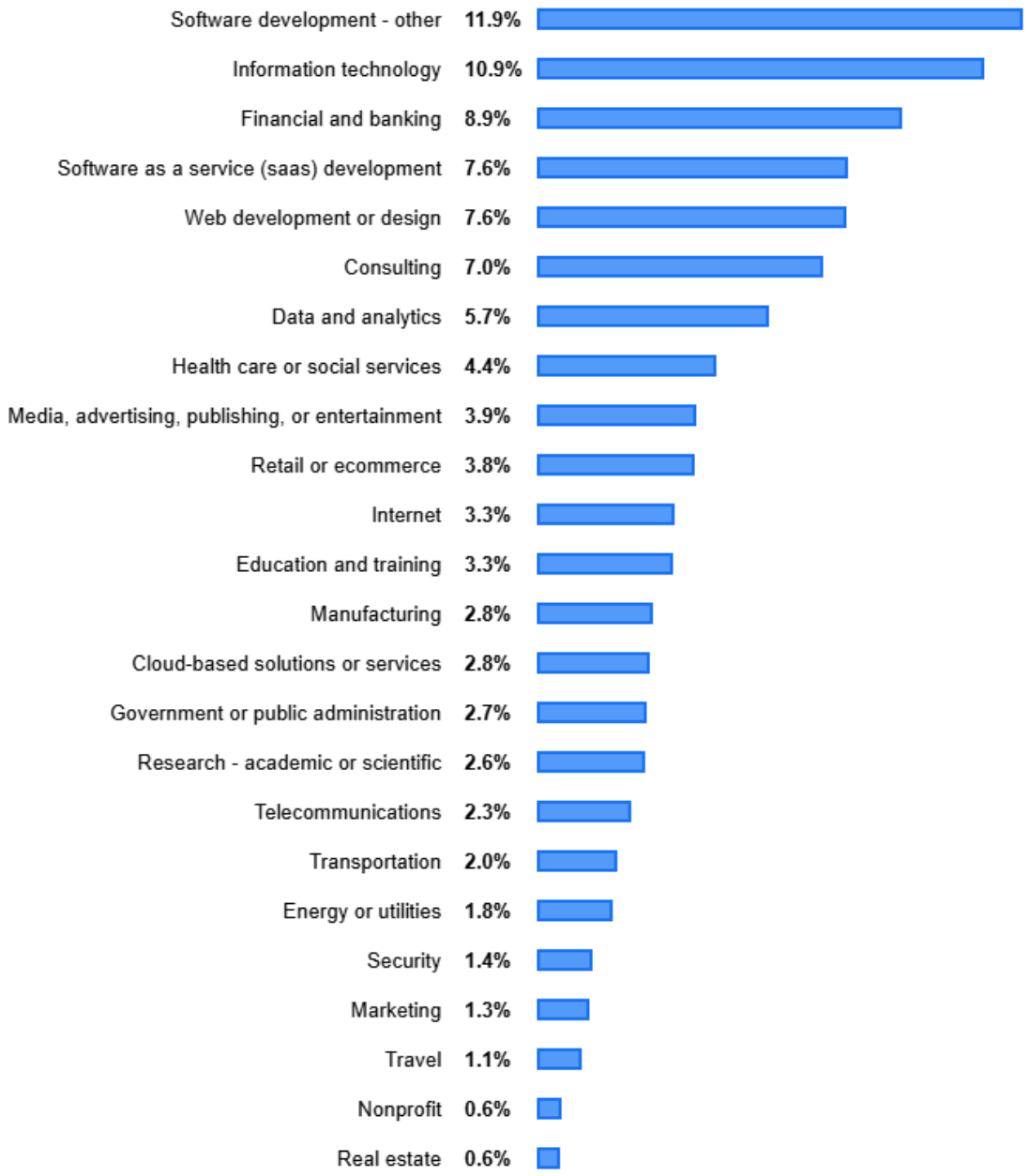


Figure 12. Industries in which the respondents of the Stack Overflow Developer Survey 2019 work. https://insights.stackoverflow.com/survey/2019.

Now, turning to the actual data, the questions at Stack Exchange concerned with open data, can the content in these questions add insight and knowledge about the background of the users? The first finding grounded in the data was that **the issues and challenges the developers addressed in the questions varied significantly.** The rationale for this statement is based on results from the investigation of two different sets of questions. First, collection 3 concerned with tags, as described in the master document in Figure 9, and next, a combination of collection 1 and 2 concerned with topics.

In the first collection, the investigated data consisted of the first 50 questions, of about 3000, resulting from a search for the phrase “open data” across all the 175 sites. The results page did not provide any information on the criteria used for sorting the questions, but it was obviously not the date of publication. The publication date for the 50 questions stretched from 2009 to 2019. This is in itself interesting all the time that Stack Exchange claims that the network was launched in 2010. However, the tagging of each of the 50 questions was recorded and counted. A total of 79 different tags were used. Of these tags were 64 used only once, 14 tags were used 2 or 3 times, while opendata was the most used tag with a count of 5 times. Figure 13 shows the distribution of the 79 tags visualised in a word cloud. The variation of tags represents the developer's perception of **the issues in question, down to the specific technology, tool, format, application or language**. Issues related to their work with open data, in other words.

Figure 13. Selected tags in 50 questions containing “open data” at Stack Exchange sites. 16.03.2019.

Initially, I wanted to investigate the tagging of a much larger number of questions, and to record which year each tag was used to reveal possible changes or interrelations over the years. The spreadsheet used for this purpose is shown in Figure 8. Unfortunately, the pagination on the search page had an error that made it impossible to continue to the next 50 unique questions after the first page. Manual identification of each of the following questions would be very time consuming if I were to continue. I found the trend from the first 50 questions to be sufficiently clear. Investigating a larger collection of questions would probably not have changed the overall finding significantly, only supplemented the outcome.

In the next collection of data, more than 200 questions from the Open Data site were investigated, the 153 most recent questions, and the 50 highest voted questions. The issue addressed in each of the questions were recorded and categorised through a process similar to coding. The categories were then compared with relevant memos, renamed and regrouped, until the main categories seemed to be stable and consistent. Here too, I found that **the developers addressed a large variety of topics and challenges**. The main categories are shown in Figure 14, supplemented with a more detailed specification of the topics in question. Which topics or categories the developers find to be the most challenging were not investigated. The content of the questions was not perceived to be suitable to determine this issue within my scope of the study.

The data collections show two different dimensions of the variety of problems and challenges that may arise when using open data from the developers' perspective. The first challenges facing the person who is going to work with open data, beyond finding it, is to understand what the dataset or API contains, what the data is, the structure of it, and from what context it has been extracted and collected. Secondly, the person must have the necessary technical skills within several areas to be able to use the data for the intended purpose. Whether it is making data openly available, developing or managing open data portals, developing software using open data, managing systems and technologies on which the open data reside, developing standards for open data, it all **requires a high level of expertise in a range of technical areas to be able to use open data**. The same applies to analysis and research based on open data. Even the regulators in this area need technical and digital capabilities (Open Data Institute, 2019).

Another important aspect in this context is the pace of the technological development in the era of digitisation of society. The digital landscape changes so rapidly that the people working within these areas need to continually update their expertise. 90% of the developers using Stack Exchange have studied at university level as their formal education ('Stack

Overflow Developer Survey 2019', 2019). 25% are still studying, 60% have taken courses, and almost 90% have taught themselves a new language, framework, or tool recently. **Working with data requires both a high degree of technical skills and the ability and motivation to continually keep up-to-date on new tools and technologies.**

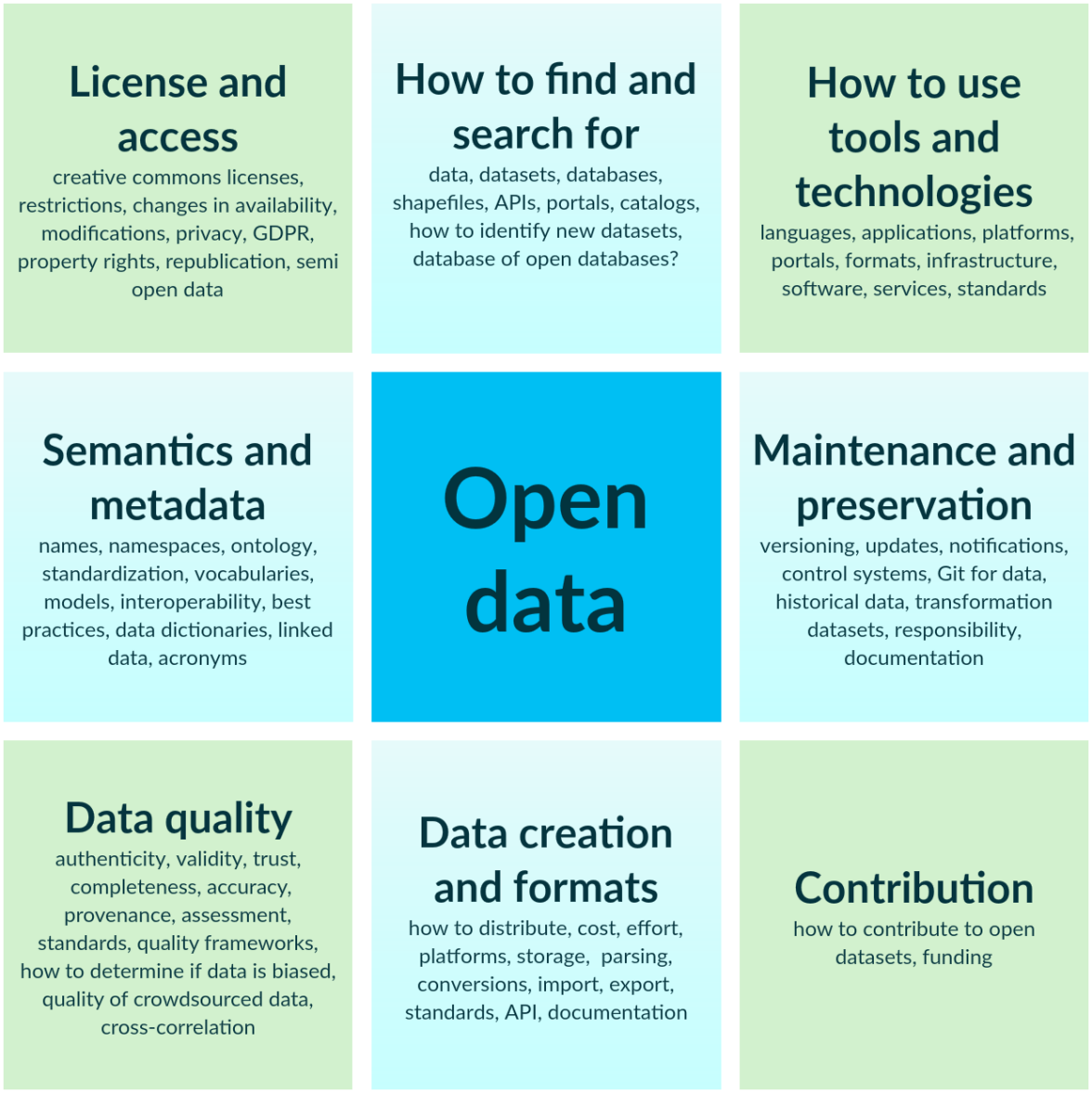


Figure 14. Challenges addressed in questions about open data on Stack Exchange. 11.03.2019.

In the above, I have demonstrated the scale of challenges associated with using open data and the amount of expertise required to master these technical challenges, which again impacts the need for education, training and continuous updating. I have also shown that many, if not most, of the roles in the open data ecosystem are possessed by developers or people with technical expertise. On these conditions, I will argue that most of **the average citizens are excluded from being able to use open data**, other than from using applications and services based on open data as an end-user.

These arguments have led me to claim that **open data is not inherently or implicitly open to anyone simply because it is shared and made openly available**.

5.3 Finding 3: Developers Are Not Particularly Concerned with the Concept of Open Data

Preliminary research question: *What is open data used for?*

Revised research question: *How do developers relate to open data?*

The data available, content published at different sites at the Stack Exchange network, provides no clear answer to this open-ended research question. Apart from the fact that geospatial data is relatively high in demand, there is no obvious pattern to be found in the data, indicating specifically what open data is used for. Conversely, it seems like open data may be used in most types of digital applications and solutions in a wide range of sectors and industries. This is shown by the great variety in what types of open data the developers are asking for, as visualised in Figure 11, and how they have tagged their questions at the Open Data site, shown in Figure 10.

Open data is used, or is at least expected to be used, in many types of applications and solutions. From this, it can be assumed that many developers must have some experience working with open data already. The number of questions about open data in Stack Exchange shows the breadth of the work they do, and the challenges it entails, listed in Figure 14, many of which apply explicitly to data that is shared openly. The use of open data can thus cause additional challenges to the developer compared to using data that is not shared openly. But beyond these technical issues, how do the developers relate to open data? How do they relate to the open data movement at large? How do they value the

benefits of open data against the additional challenges it entails to use it? Can Stack Exchange provide any answers to these questions?

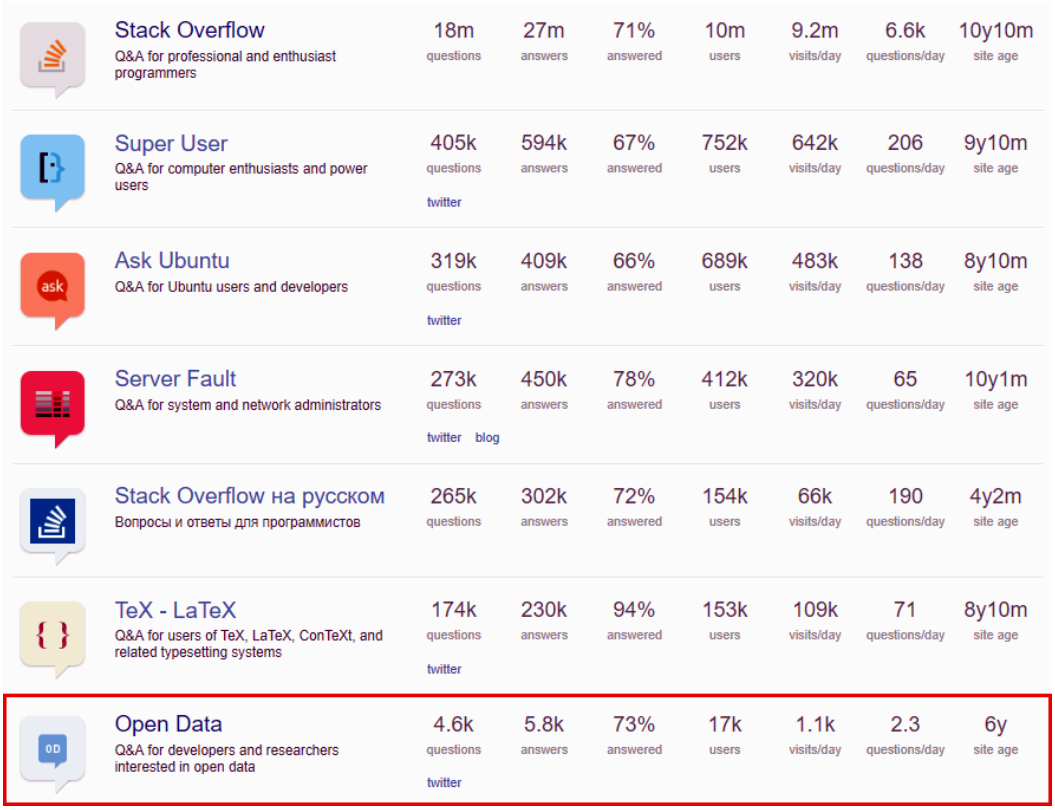
From the size of the Open Data site compared to all the other sites within the area of technology at the Stack Exchange network, **open data does not attract that much attention** according to the numbers in Figure 15. Turning six years as of May 2019, the Open Data site has gathered just over 17.000 users, and is ranked number 122 of a total of 175 sites when counting the total number of published questions on each of the sites (‘All Sites - Stack Exchange’, 2019). Again, it can be discussed to what extent the users find answers in previously published questions, so that they do not need to publish new questions of their own. However, as seen in Figure 15, the average number of visits per day at the site is also not very high, especially if the numbers are compared to the Stack Exchange network as a whole, which has more than 10 million users worldwide and gets well over 10 million visits per day.

The fact that open data has got its own site and community for six years, although still in beta, is an indication of a certain interest and commitment over time. It is reasonable to assume that open data is incorporated in the general technical vocabulary, and that is well established in the consciousness of developers and researchers. Does this indicate that open data is perceived as being an important concept of and in itself by the developers, different from data that is not shared openly? The distribution of questions containing the search term "open data" across all the sites at the network indicates that the **developers are not especially concerned with the concept of open data in itself** even though they may use open data in their daily work. Of more than 3.100 questions containing “open data”, around 1.000 were published at the Open Data site. What is surprising is not the 1.000, but rather the 2.100 questions that were not published at the Open Data site. 500 of these questions were published at the Stack Overflow site, 280 were found at the Geographic Information Systems site, and the remaining 1.300 questions were scattered across a large number of sites. This may imply that some **developers relate data more to type or characteristics of use, and less to whether it is open or not.**

Based on the results of the Stack Overflow Developer Survey (2019), it is reasonable to assume that most developers are employed and work in a company or within a project with predefined tasks. Much of the data they need will probably be provided by the employer, reducing the need for using open data, unless they are working specifically within the open data ecosystem of course, or are freelancers, entrepreneurs or independent developers or researchers. As discussed above, using open data will in many cases be more demanding

than using internal proprietary data. Special conditions regarding licensing, personal data, access and maintenance, among other issues, come into play when using open data. Therefore, it is reasonable to assume that developers will avoid these obstacles if they have other alternatives that are equally good. Unless a special requirement or other reasons apply, there is reason to believe that **developers are likely to prioritise internal data over open data** if they have the choice.

The actual content of the questions on **the Open Data site does not leave the impression that a strong community around open data has been established**. Nor is there any indication of a strong ideological stance among the 17.000 users towards opening data and sharing it with the world. There seem to be dedicated users to be found at the site, definitely. But still, the findings show that the site is mainly used as a reference, and to assist new members of the community to find the data they are looking for. Consequently, based on the above-mentioned findings and conditions I dare to state that **developers at large do not seem to be particularly concerned with the concept of open data** if it is not an implicit part of their work, or they have a special interest in the Open Data movement.



Stack Overflow Q&A for professional and enthusiast programmers	18m questions	27m answers	71% answered	10m users	9.2m visits/day	6.6k questions/day	10y10m site age
Super User Q&A for computer enthusiasts and power users twitter	405k questions	594k answers	67% answered	752k users	642k visits/day	206 questions/day	9y10m site age
Ask Ubuntu Q&A for Ubuntu users and developers twitter	319k questions	409k answers	66% answered	689k users	483k visits/day	138 questions/day	8y10m site age
Server Fault Q&A for system and network administrators twitter blog	273k questions	450k answers	78% answered	412k users	320k visits/day	65 questions/day	10y1m site age
Stack Overflow на русском Вопросы и ответы для программистов	265k questions	302k answers	72% answered	154k users	66k visits/day	190 questions/day	4y2m site age
TeX - LaTeX Q&A for users of TeX, LaTeX, ConTeXt, and related typesetting systems twitter	174k questions	230k answers	94% answered	153k users	109k visits/day	71 questions/day	8y10m site age
Open Data Q&A for developers and researchers interested in open data twitter	4.6k questions	5.8k answers	73% answered	17k users	1.1k visits/day	2.3 questions/day	6y site age

Figure 15. Traffic and engagement at the Open Data site compared to the largest Stack Exchange sites within technology. Retrieved 01.05.2019.

5.4 Finding 4: Developers Do Not Use the Term Smart City – at Stack Exchange

Preliminary research question: *How can open data contribute to innovation and development of smart cities?*

Revised research question: *How do developers relate to the concept of smart city?*

In this final phase of the study, the goal was to investigate how the developers relate to the concept of smart city. Initially, I assumed that the data collection and findings from the first three research questions would help to answer the question of possible connections or interrelations between the two concepts of smart city and open data. However, the content of the questions in Stack Exchange did not provide any direct information on this possible relationship.

Considering my perspective, the viewpoint of the developers, and my focus for the study, open data, I decided to simply investigate the questions in which the term “smart city” or “smart cities” appeared. Among the 4.600 questions at the Open Data site, I found only 3 questions containing the phrase “smart city”. Two of these questions referred to a data set where smart city was part of the name in the url. The third, a two-year-old open question asking the community of known data sets for the smart city domain, were unanswered. At a higher level, across the entire Stack Exchange network with more than 20 million questions, I found additional 26 questions containing the term “smart city”. Most of these questions were completely irrelevant to the actual concept of smart city. Three questions were not accepted or closed by the community for being off-topic, too broad or not adequate. Of the remaining two questions, one was questioning privacy in smart cities, and the other, repeated three times, asked for opinions on blockchain and IoT in a smart city context.

The data did not offer any information about the relation to smart city, other than that **developers do not use the term smart city** in the questions they publish at Stack Exchange. The overall finding is that the term smart city is very close to being non-existent in the vocabulary of the developers when they discuss solutions to issues across the entire Stack Exchange network. This finding is significant in itself, but it does not give any hints about the cause, and it is therefore not possible to deduce any ideas that can explain the finding.

There was no direct connection between the concepts of open data and smart city to be found in the data. However, the data suggests, but let us not conclude, that an indirect

connection exists. First, the term smart city is not particularly relevant to developers looking for open data, even if the data is going to end up in applications and information systems destined for a smart city project. The developer would be much more specific in the description of the requested data, to increase the probability of getting a good answer

Next, comparing the areas listed by Albino et al (2015) as crucial for enabling the smart city to happen - building, energy, water and waste management, mobility, culture, education, health care – with the listed tags and categories reported in Figure 10 and Figure 11, we find that developers request open data about the same areas – such as education, demographics, health and health care, transportation and more. While there is no one-to-one relationship, it may still suggest that developers relate to the concept of smart city indirectly, by requesting types of data suitable for developing solutions and services for smart cities. However, to verify the existence of such an indirect connection between the two concepts is not possible without a more thorough study performed through interviews or other methods.

6 Discussion

6.1 What Does the Study Say?

The findings from the study of the contents at Stack Exchange has, in my view, led to new insights into two main areas which I will discuss in the following.

6.1.1 Open Data – Accessible to the Few?

First, the analysis of the findings has shown that open data is not inherently accessible to the public simply because it is made openly available. If people of the public are actually able to find the data they need, then using it is will be the next major challenge. The people who actually make use of open data today describe a whole series of challenges they face in working with openly shared data. These challenges are of such a nature that they require the person working with open data to have completed a certain level of technical education and to continuously keep up-to-date on new technologies. The average person does not have such an educational background and is therefore - to some extent at least - excluded from being able to use open data. (Janssen et al., 2012).

Without a certain degree of technical understanding, it is uncertain whether the people at all are able to understand what data really is, how it is collected and how it is stored, what is collected and what is stored, or how this data reflects the reality they experience to be a part of. When digital data from the analogue world is to be collected and stored digitally, it must be determined which details or units are to be registered and how these units should be linked together in databases. This structure is determined by the few who have sufficient technical expertise, but it reflects the physical reality and behaviour of the many.

I have uncovered how the few – exemplified through the developers using Stack Exchange - are demanding digital data that reflects all possible aspects of human life and environment, with the purpose, presumably, of using it for the development of innovative and smart solutions for the many. These solutions are based on the developers' perception of societal structures of today and will reflect their visions and desires for the society of tomorrow. They will build the infrastructure of our future and will decide how we will interact and form communities when they are in place.

In the previous, I have demonstrated how open data - for various reasons – is open and available to be used only by the few. I have also shown how it is the developers that will

decide the constituents of the digital future on behalf of the many even though data is made openly available for everyone to use. In this context, I question how open data can contribute to increase inclusion, participation and empowerment of the many, as prescribed benefits of the open data movement.

6.1.2 Open Data Movement – Engaging the Few?

The second main finding concerns the developers' attention and commitment to the open data movement. The analysis of the data indicates that developers at large are not particularly concerned with the concept of open data. In relation to the Stack Exchange network's overall size and global reach, the Open Data site has attracted relatively limited attention, beyond being used as a pure reference. The site leaves no impression that a strong community around open data has been established. When asking their peers about open data, the developers often publish their questions on other sites than the Open Data site. Do they relate data more to type or area of use than to whether it is open or not?

The findings have shown the extent of additional challenges facing the developers when using open data in relation to internal proprietary data that they themselves control. These challenges will require the use of extra time and resources, which in turn will provide additional costs for the employer. If there are no specific incentives or benefits to use open data, then it is reasonable to believe that developers will choose the easiest way to fulfil their goals and choose internal data over open data. For other groups, such as independent developers, freelancers, students and employees in the public sector, this scenario may be different. However, these groups make up a minority compared to the professional developers employed in private businesses and public agencies, according to the Stack Overflow Developers Survey 2019. This survey does not tell the definitive truth in any dimension, but it has been the basic source of data for my study, and has thus given rise to my findings. An overall analysis of these findings gives the impression that the ordinary developer – active on the Stack Exchange network - is not particularly concerned with the promotion of the Open Data movement.

Some of the developers may themselves be actors or hold positions in the open data ecosystem, as discussed in the section about actors and stakeholders. Some of the developers may also have strong ideological perspectives, political interests or general concerns about open data and the open data movement. However, none of these groups appears to be very active or present with their views or practical questions on the Stack Exchange network.

Perhaps Stack Exchange is just not the place where the developers would address challenges in relation to the prevalence of open data and the promotion of sharing data openly? Perhaps they only use Stack Exchange to discuss technical issues related to formats, tools, languages, quality, and the like? It is probably well known to the developers who use the network that "The site is all about getting answers. It's not a discussion forum. There's no chit-chat. Just questions .. and answers." ('Tour - Stack Exchange', 2019). At the same time, the network promotes itself as a place for "people passionate about a focused topic" and that it facilitates the creation of "communities built by narrowly focused groups of experts.". The possible passionate people about open data at the network, are they just not questioning the problematic issues that the open data movement faces? Or could it be that my data collection, or time, simply have been too limited to have noticed their questions? I have read through the contents of nearly 400 questions about open data though. I have not been able to trace any particular levels of ideological considerations in favour of the Open Data movement.

6.1.3 Incentives for Open Data

In the following, I will continue with the assumption – developed from my findings - that the general developers do not seem to be particularly engaged in the prospect of sharing data openly for everyone to use for whatever purpose. Why do they not seem to embrace the idea of open data? Or maybe they do, but perceive it to be other actors' responsibility to drive the process up and forward? Perhaps there is so much politics associated with open data that the developers at large do not perceive themselves as having any role in this game? But who are the actors that drive the open data movement then?

The definition of open data - freely available for everyone to use for whatever purpose – sounds promising. The benefits of open data are diverse and are susceptible to affect many dimensions of government, business, the individual and society ('Åpne data | Difi', 2019; 'Benefits of Open Data - European Data Portal', 2019; Janssen et al., 2012). According to these sources, open data has the potential to solve many of the major challenges facing society today. Open data will strengthen democracy, transparency, accountability, citizen rights and efficiency of governments. It will enhance inclusion, self-empowerment, participation and engage people. It will facilitate innovation, development of new products and services, new business models, it will provide new jobs and economic growth. Open data also has the potential to save lives, save energy, and have a positive impact on the environment ('Creating Value through Open Data', 2019). The overall message is: the more we share, the better.

The opening of data for everyone to use for whatever purpose, free of charge, is presented as an inherently good thing in and of itself (Kitchin, 2014). The long revelation of positive prospects and future scenarios are not questioned in any dimension. The benefits of open data are most often presented in a one-dimensional and simplistic manner, idealised without any assessment of the potential unfavourable or negative consequences (Janssen et al., 2012).

There may be contradictions that are not immediately visible in the predicted benefits of open data which are not evaluated or even considered. Here are a few examples. Improving the efficiency of the public sector could reduce manual routines and the need for physical labor, when tasks are taken over by digital solutions, managed by technical staff. The requirement for technically educated people will increase and leave less educated people without a job. Open data will provide new jobs, for some yes, but it may be at the expense of social inclusion and self-empowerment of somebody else.

Another example concerns how accountability is challenged by open data in itself (Janssen et al., 2012). It is inherent in open data that it can be used for anything without the data owner's consent, or knowing. When public data is used, and the processing of this data leads to incorrect results causing serious problems, it is likely that the public sector will be held responsible and will meet expectations of solving the problems. It is likely that the public sector must expect such scenarios to arise when sharing data, and that preventive measures must be implemented in advance. This challenges the perspective that open data unambiguously contributes to more openness and transparency and simultaneously increasing accountability.

6.1.4 Economic Incentives

In my view, there are two main incentives that drive the open data movement, the ideological forces and the much stronger economic forces. I will discuss the latter first. The prospects for economic savings, gains or growth permeates most of the benefits of sharing data, directly or indirectly.

Sharing data internally within the public sector, between agencies and departments, is obviously fostered by a financial incentive. Sharing data with the private sector are supposed to save the public sector both time and resources and thus, public money, in the way that the development of smart and innovative services are outsourced to the industry. At the same time, it is an important task for the public sector to ensure that the country's businesses and companies thrive and can offer jobs to the citizens. For these reasons, it makes sense that

the public sector drives and pushes the open data movement forward, using public money to publish government data. There is a catch, however. Open government data does not come without a cost (Haraldsen, 2019).

Continuing the argumentation, I will cite a critical voice who claim that we are witnessing a marketisation of public services (Kitchin, 2014). This author argues that strong capitalist interests have been heavily involved in pushing the open data movement forward ".. to get access to expensively produced data for no cost, and thus to a heavily subsided infrastructural support from which they can leverage profit.". He also refers to how public data and related services, which were previously susceptible to fees, are now free and can no longer fund internal employees in public agencies. As a result, the services must be outsourced to the private sector, which in turn sells their services back to the public sector. With this, he shows how open data has increasingly become a financial gain for the big companies, at the expense of the taxpayers. Sharing public data have the potential to change the balance of power in society (Haraldsen, 2019).

This development also affects the developers in the way that their jobs are moved from the public to the private sector. The work they do is basically the same, but they are often better paid in the private sector. As such, open data can also have a positive financial consequence for the developers, albeit indirectly.

6.1.5 Ideological Incentives

The ideological incentives for opening data are intended to provide benefits and values both socially for the individual person and for society at large. The premise is that the person can make use of open data. Being able to use open data requires not only technical knowledge and skills, but it also requires access to technical equipment such as hardware and software suitable for processing open data and connection to the internet. In addition, the person must have varying skills to be able to interpret and analyse the data. My findings have shown that the ordinary citizen does not have the necessary background to be able to extract the promised values from open data.

When a large part of the population cannot actually make use of open data, and are excluded from taking advantage of the benefits it provides, it will inherently create a divide to those who can benefit from open data. Contrary to the intention, open data can reduce social inclusion and have a negative impact on democracy when a large part of the population cannot participate. It can cause inequalities in society to increase rather than decrease. Kitchin is even more critical on behalf of society in his future vision of open data: "Open data

can work to further empower the empowered, and to reproduce and deepen power imbalances." (Kitchin, 2014).

6.1.6 The Developers' Position and Perspective

In my view, open data has the potential to affect many of the fundamental questions in our society and can prove to have a decisive impact on how our future will be shaped. There is power in controlling data, and this power is redistributed when the data is shared openly. Data has no value in itself, it is only when the data is used that it provides value and benefits for people or the environment, advantages for some and perhaps disadvantages for others.

I will return to the developers. Where are the developers in this scenario? How are they affected when the data they need to perform their job is being shared openly? From my findings, it seems that using open data causes them a lot of extra challenges. The benefits of using open data are hardly mentioned. But again, maybe Stack Exchange is just not the place where this is discussed. The benefits of open data are possible sufficiently described elsewhere.

The digitisation of society requires an ever-increasing number of developers to create the innovative services we need. Hence, I deduce that the number of developers is steadily increasing. This pattern is supported by data on the users of Stack Exchange. The Stack Overflow Developers Survey from 2019 reveals that many of the developers that have responded to the survey are young – 75% are under the age of 35 - and have a relatively short experience as developers – 41% have less than 5 years of coding experience. 92% are male and 71% are of European descent. Perhaps the developers' young age has an influence on their involvement in the Open Data movement? The survey shows that they mainly live in parts of the world that have very good living conditions in relation to the under-represented continents. I think it is well known that developers are often well paid in relation to other groups in society. I have experienced through my studies at the University of Oslo how skilled developers are increasingly popular in the market, in the private sector especially.

I will argue that the developers are one of the most privileged groups of people in society today. They are high in demand and get well paid for their work. The whole world is their workplace through the Internet. Wherever they live or are educated, they can do work in whatever part of the world they want. They master the tools and technologies needed to participate in society, even to create society themselves through the products and services they are developing. They are most definitely self-empowered and most probably have some

sense of social inclusion. So, how do the economic and ideological incentives for the concept of open data appeal to the average developer? Perhaps my findings have actually revealed a tendency, that they are not really that concerned about the prospect of sharing data openly with the world?

6.1.7 Mind the Gap

If my findings prove to bring even the slightest degree of truth, that the developers at large do not engage much in the concept of open data, it is quite remarkable. The assumption that they do not actively support the ideological incentives for open data can be partly because they will not benefit from them themselves, or alternatively, that they do not see how the predicted benefits can be achieved by sharing data openly. Or perhaps the developers just don't care, and find other issues to be more urgent? Whatever the reason, they do not seem to participate in the discourse. They develop our future solutions and infrastructures from their computers in offices worldwide while the battles for power and influence are going on outside.

We are about to enter an age where artificial intelligence becomes increasingly embedded in the solutions we surround ourselves with. It is the developers who code the actual lines that decide how machines are learning and how the systems will behave and react to the physical world. It is powerful forces they are launching, forces that we – who har not developers - do not understand. If we also do not understand the developers' language, or even recognize that we need to talk, then we enter a world where the battles in the digital world can be far more powerful and take place at a completely different level than in the physical.

6.2 Contributions

The starting point for this study was the preliminary research questions asking who is using open data, which types of open data are shared, for what use, and finally, how can open data contribute to the development of smart cities. I have tried to find answers to these questions by analysing the content of the questions and answers on the Stack Exchange network, content that has been produced for a very different purpose in a whole different context. I have made assumptions and developed hypotheses about conditions and relationships that are based on this content, without having had the time to talk directly to the persons who have created it. The chosen methodology has determined how the data has been collected and analysed, and my personal background has influenced both the entrance to the subject and the outcome of the study. There are many factors that have affected the findings and

analysis I have just presented, and it is, of course, an important question to what extent these findings apply in a larger context. I hope, nevertheless, that through my efforts with this study and thesis I have managed to shed light on some aspects that may not have been so obvious before.

My goal with this thesis has been twofold, and the contribution therefore extends in two dimensions. I have placed equal weight on both parts and have spent about the same amount of time on each of them.

The first contribution consists of findings from the study of developers' perspective on open data based on data collected from the Stack Exchange network. The main finding from this study, which constitutes the contribution, is the revelation that open data is not implicitly and inherently open and accessible to everyone even if shared openly. The study has uncovered the extent of technical challenges facing the person who is to make use of open data, which thus excludes the ordinary citizen from accessing open data.

The second contribution constitutes the thorough description and reporting of how the grounded theory study itself has been conducted. The methodology of grounded theory is not widely used within the research field of computer science. In fact, according to my review of several sources, there is not a single grounded theory study to be found that collects data from other sources without the use of interviews as the basis. Also, most of the studies that have actually been conducted with grounded theory as the methodology, and with data collected from interviews, are lacking a proper description of the implementation of the study and are poorly reported (Stol et al., 2016). Consequently, my aim with this study has therefore been to follow all recommendations and guidelines for a well-conducted and reported grounded theory study. I have been particularly concerned with describing and reporting each step and method used in the various processes throughout the study, from the moment I understood that grounded theory was to be used to the end of the study.

I hope this study of grounded theory will be a useful contribution for future students who need support and examples when setting out to conduct their own grounded theory studies.

6.3 Further Research

This study has had two branches. One is the study of the grounded theory methodology that has been used to study the other, open data from the perspective of the developers. Both parts have been equally interesting and have triggered my curiosity on many new contexts and conditions that I did not know the existence of before. The time has been too short to be

able to follow all the exciting tracks into the data and the literature, unfortunately. Yet, I have identified several interesting areas with potential for further research within grounded theory, open data and not the least, within the use and users of the Stack Exchange network.

6.3.1 Grounded Theory

Grounded theory is a methodology that requires some level of experience to be able to take full advantage of its methods in order to develop theory (Glaser, 1978). By adding extra focus and emphasis to the method part of this study, on the actual research process and the description of its implementation, I have shown that it is possible to carry out a grounded theory study even with little time and little experience. The value of the findings and contributions is probably limited, because precisely here lies the challenge, in using the methods and techniques so well that it facilitates the development of new hypotheses and theories of good quality and verifiability. There are many studies of grounded theory within the social sciences and other research areas, but very few in computer science and design. Researchers and students within these disciplines should absolutely study the opportunities and benefits of using this methodology more closely.

Grounded theory was initially developed to extract knowledge from interviews. As shown above, it has apparently not been used for studies without interviews in the field of computer science. In my view, however, the method can prove to be a promising alternative in the near future where data and digitalization increasingly will structure the way we live our lives. As we gradually become more digital than today, new digital communities and contexts will arise. Contexts in which grounded theory can be relevant, by looking into data that is not necessarily extracted from speech. Here, there is also a great potential for further research, in uncovering which adaptations and alterations of techniques and methods are appropriate in purely digital settings.

6.3.2 Open Data

The open data movement is still young, but many initiatives and projects have recently been launched to make data openly available to the public, so also in Norway. Difi and Brønnøysundregistrene have for a long time collaborated on establishing a national portal for publishing open data, and this work led to the launch of The National Data Catalog in November 2017 (Brønnøysundregistrene legger til rette for deling av data, 2017). After attending several workshops and seminars this spring, organised by Difi to increase knowledge about publishing public data on this portal, I began to realise the extent of the

challenges the public sector is facing. At one of these seminars, I overheard a conversation between two senior leaders in agencies in the public sector. They both expressed frustration over the technical level of the content of the full-day seminar, and one of them said: “We are not yet even aware of what data we actually possess.”.

An exciting area for further research would have been to study the Norwegian landscape of open data. There are several perspectives that could have been interesting in such a study. A study of The National Data Portal itself, from the perspective of decision makers, managers and directors, could reveal practical, ethical and political issues to open data from this level of the organisations. The challenges with open data are probably found as much here, as on the purely technical level. Another option would be to take the researchers' perspective. How do the researchers make their data available? What does the research institutions do to facilitate data sharing? What are the main challenges for researchers and research institutions? A third option is to examine the private sector's relation to data sharing and the use of open data. How do the companies perceive the prospects of values and benefits from open data?

There is an infinite number of topics suitable for future research in open data. Important topics that touch both politics and power distribution in society. It is a study in itself, looking at potential future research within the concept and movement of open data.

6.3.3 The Stack Exchange Network

Last but not least, I see many exciting opportunities for further studies of the Stack Exchange network. The data collection I have conducted has raised many questions about the users, about their incentives to use the network and about the importance it has to developers in a global and local context. I believe there is an infinite amount of insight buried, or grounded, in the vast amount of data that is constantly being added to this network.

Would it be possible to use tools like NVivo to extract new types of insight from the content of Stack Exchange? The tools used in grounded theory studies, have they been used for the analysis of data other than interview transcripts? Would it be possible to upload large amounts of questions from Stack Exchange into NVivo, and would the generation and aggregation of categories of this content add new insights compared to the manual coding and categorisation?

There is most certainly additional knowledge and insight that can be extracted from this network about various issues relating to open data. How can data owners engage users to

use open data at a much higher level than today? How can data providers and developers improve collaboration so the data that is shared is what developers want and need? How to engage developers in the ideological perspectives of open data? Or, conversely, how to bring developers' perspective into the prospects and predictions of what open data can bring to our planet and its population?

6.4 Limitations of the Study

Beyond the limitations already described in previous chapters, there are a few additional conditions that should be mentioned to set the study in an appropriate context. This thesis is the result of a 17-week concentrated and intensive study of the comprehensive and complex subject of open data, something that is happening right now and develops as we speak. The short time available has perhaps been the most prominent limitation and weakness of the study. The study has just scratched the surface and may be perceived as superficial without a sufficient theoretical foundation. If this had been a long master's thesis and not a short, the insight into the topic would have matured significantly. The study could have been conducted with the use of primary data, which would most certainly have produced a better result.

However, given the time and data available, it has been necessary to make certain simplifications and assumptions to be able to complete the thesis. First, I have simplified my understanding of the users of the Stack Exchange network to be a general developer with the average characteristics uncovered in the Stack Overflow Developer Survey 2019. This is, of course, a great weakness, but it was not considered relevant for the research questions to investigate this more closely. Throughout the thesis, I have treated and discussed open data without being specific about whether it was private or public data. This is unfortunate if it has added any confusion. I might also have been more precise about which actors I have been thinking of in my descriptions, but they have been based on the insights I have managed to establish.

7 Conclusion

This chapter marks the end of the two-dimensional study, where the first dimension has been a grounded theory study of open data based entirely on secondary data, while the second dimension has been a study of the use of grounded theory based entirely on the data from the first study. The study of open data was planned and prepared in a project description with preliminary research questions as an entry to the topic of open data. The study of grounded theory, however, was not planned but was prepared when it became clear that such a study did not exist before.

Grounded theory is said - by its founders - to be a methodology that requires experience to be able to take full advantage of it, even with primary data. I have shown, with dedication and stubbornness, that it is possible for an inexperienced master's student to use the methodology with secondary data, even in a short thesis, and produce a result that follows recommended guidelines for implementation, reporting and documentation. My experiences with the methodology, which may be considered as findings, show that it is a flexible method that offers great freedom to the researcher in virtually all dimensions. The opposite side of this freedom – when using secondary data as the only data source - is the lack of format and structure. You have to create this structure by yourself and to decide which formats and techniques to use. The reporting of the study in a way constitutes the entire framework. The more artefacts that can be created for this framework, the stronger the study. For future students or researchers who aim to use the method in a context of only secondary data, I highly recommend that great emphasis be placed on the reporting part of the study from the very beginning.

My original goal of the study, however, was to study open data from the developer's point of view by looking at content published on the Stack Exchange network. I wanted to understand how the developers perceive openly shared data, how they relate to the concept and the open data movement, how they use open data, and what they use it for. My personal experience with developers is that they often exist in their own sphere, in a community of their own, speaking a language that few really understand, if they are not developers themselves. The two separate worlds meet - the developer's world and the reality of the others - in the process towards the construction of services and solutions. Apart from that, they rarely meet, or speak. I think the developers have good insight into what is happening in the non-technical world, but I think the opposite is not the case, unfortunately.

This was the backdrop for my entry into the study of the open data ecosystem which I knew very little about in advance. It is a complex system with many different actors and stakeholders involved. Through my analysis of the data, I found that the developers are likely to possess many of the roles and positions in this system. It requires high technical expertise in a number of different areas, working with open data. The questions asked by the developers show a wide range of technical challenges they meet when working with open data, many of which only applies to open data. I also found that they are requesting virtually any type of data that can be imagined. As such, it testifies that open data is used for many different purposes and applications.

Stack exchange seems to work mainly as a reference tool to get help finding data, and less as a place where the more fundamental issues of sharing data openly are addressed. The findings also showed no direct evidence of the establishment of a large community built around open data. Having analysed and discussed a number of conditions, I then asked if the average developers are not that concerned about the prospects of open data beyond using it themselves?

The findings have shown that the developers face a large variety of technical issues in their work with open data. From this I deduce that the average citizen does not have the necessary technical skills or equipment to be able to understand the data, to know where to find the data, how to assess or access it, or how to make use of the open data. Hence, I hold that open data is not inherently open to anyone just because it is shared openly and promised to be free for anyone to use for whatever purpose. This finding further implies that the predicted benefits of open data, such as social inclusion, self-empowerment and participation, do not necessarily hold. Then open data will even have the potential to increase inequalities in society, contradicting the promised benefits.

Through the data, I expected to find a close interrelation between the concepts of open data and smart city. Surprisingly, I found that the developers do not use the term smart city whatsoever. However, a possible indirect connection between the concepts of open data and smart city were still found by comparing requested types of open data in Stack Exchange with the typical types of data used in smart city solutions. The fact that developers do not use the term smart city, which otherwise gathers so much enthusiasm and attention from both the public and the private sector is remarkable. Again, the gap between vision and reality!

References

Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology*, *22*(1), 3–21.

All Sites - Stack Exchange. (2019). Retrieved from https://stackexchange.com/sites#questions

Åpne data | Difi. (2019). Retrieved from https://www.difi.no/fagomrader-og-tjenester/digitalisering-og-samordning/apne-data

Benefits of Open Data - European Data Portal. (2019). Retrieved from European Data Portal website: https://www.europeandataportal.eu/en/using-data/benefits-of-open-data

Brønnøysundregistrene legger til rette for deling av data. (2017, November 1). Felles datakatalog. Retrieved from Brønnøysundregistrene website: https://www.brreg.no/nyhet/felles-datakatalog-legger-til-rette-for-deling-av-data/

Chignard, S. (2013, March 29). A brief history of Open Data - Paris Innovation Review. Retrieved from Paris Innovation Review website: http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data

Coleman, G., & O'Connor, R. (2007). Using grounded theory to understand software process improvement: A study of Irish software product companies. *Information and Software Technology*, *49*(6), 654–667.

Creating Value through Open Data. (2019). Retrieved from European Data Portal website: https://www.europeandataportal.eu/en/highlights/creating-value-through-open-data

Difi. (2019). Veileder for tilgjengeliggjøring av åpne data. Retrieved from https://doc.difi.no/data/veileder-apne-data/

Glaser, B. G. (1978). *Theoretical sensitivity: advances in the methodology of grounded theory*. Mill Valley, California: Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory : Strategies for Qualitative Research*. New York, NY, US: Routledge.

Haraldsen, A. (2019, April 29). Mens vi venter på regjeringens digitaliseringsstrategi – del 2. Retrieved from Digi.no website: https://www.digi.no/artikler/kommentar-mens-vi-venter-pa-regjeringens-digitaliseringsstrategi-del-2/463417

Hoda, R., Noble, J., & Marshall, S. (2011a). Grounded theory for geeks. *Proceedings of the 18th Conference on Pattern Languages of Programs - PLoP '11*, 1–17. Portland, Oregon: ACM Press.

Hoda, R., Noble, J., & Marshall, S. (2011b). The impact of inadequate customer collaboration on self-organizing Agile teams. *Information and Software Technology*, *53*(5), 521–534.

International Open Data Charter | Principles. (2019). Retrieved from International Open Data Charter website: https://opendatacharter.net/principles/

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

NLOD | Difi. (2019). Retrieved from Difi | data.norge.no website: https://data.norge.no/nlod/en

Open Data Barometer. (2017). Retrieved from Open Data Barometer website: https://opendatabarometer.org/?_year=2017&indicator=ODB

Open Data in a nutshell - European Data Portal. (2019). Retrieved from European Data Portal website: https://www.europeandataportal.eu/en/providing-data/goldbook/open-data-nutshell

Open Data Institute. (2019). *ODI Fridays: Regulating for responsible technology – is the UK getting it right?* Retrieved from

https://www.youtube.com/watch?v=HCZK9G95jpw&feature=youtu.be&mc_cid=3b6b040cbb&mc_eid=ee869d3b11

Smart Oslo Strategy. (2019). Retrieved from Oslo kommune website: https://www.oslo.kommune.no/english/politics-and-administration/smart-oslo/smart-oslo-strategy/

Stack Overflow Developer Survey 2019. (2019). Retrieved from Stack Overflow website: https://insights.stackoverflow.com/survey/2019/?utm_source=social-share&utm_medium=social&utm_campaign=dev-survey-2019

Stol, K., Ralph, P., & Fitzgerald, B. (2016). Grounded Theory in Software Engineering Research: A Critical Review and Guidelines. *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*, 120–131. Austin, Texas.

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. New York, NY, US: Cambridge University Press.

Stray, V., Sjøberg, D., & Dybå, T. (2016). The Daily Stand-up Meeting: A Grounded Theory Study. *Journal of Systems and Software*, *114*, 101–124.

Tour - Stack Exchange. (2019). Retrieved from Stack Exchange website: https://stackexchange.com/tour

UIO : DUO Research Archive. (2019). Retrieved from https://www.duo.uio.no/

Universitetet i Oslo. (2017, June 14). NVivo - Universitetet i Oslo. Retrieved from https://www.uio.no/tjenester/it/forskning/datafangst-og-analyse/nvivo/index.html

User Dmitry Kachaev. (2019). Retrieved from Stack Exchange | Area 51 website: https://area51.stackexchange.com/users/69883/dmitry-kachaev

W3C. (2017). Data on the Web Best Practices. Retrieved from https://www.w3.org/TR/dwbp/

What is a Smart City? (2014, May 29). Retrieved from Centre for Cities website: https://www.centreforcities.org/reader/smart-cities/what-is-a-smart-city/

What is Open Data? (2019). Retrieved from Open Knowledge Foundation | The Open Data Handbook website: http://opendatahandbook.org/guide/en/what-is-open-data/

What is ‘open data’ and why should we care? (2019). Retrieved from The ODI - Open Data Institute website: https://theodi.org/article/what-is-open-data-and-why-should-we-care/

Why open data? (2019). Retrieved from Open Knowledge Foundation website: https://okfn.org/opendata/why-open-data/

World Wide Web Foundation. (2017, May). Global Report | Open Data Barometer. Retrieved from https://opendatabarometer.org/4thedition/report/

Table of Figures