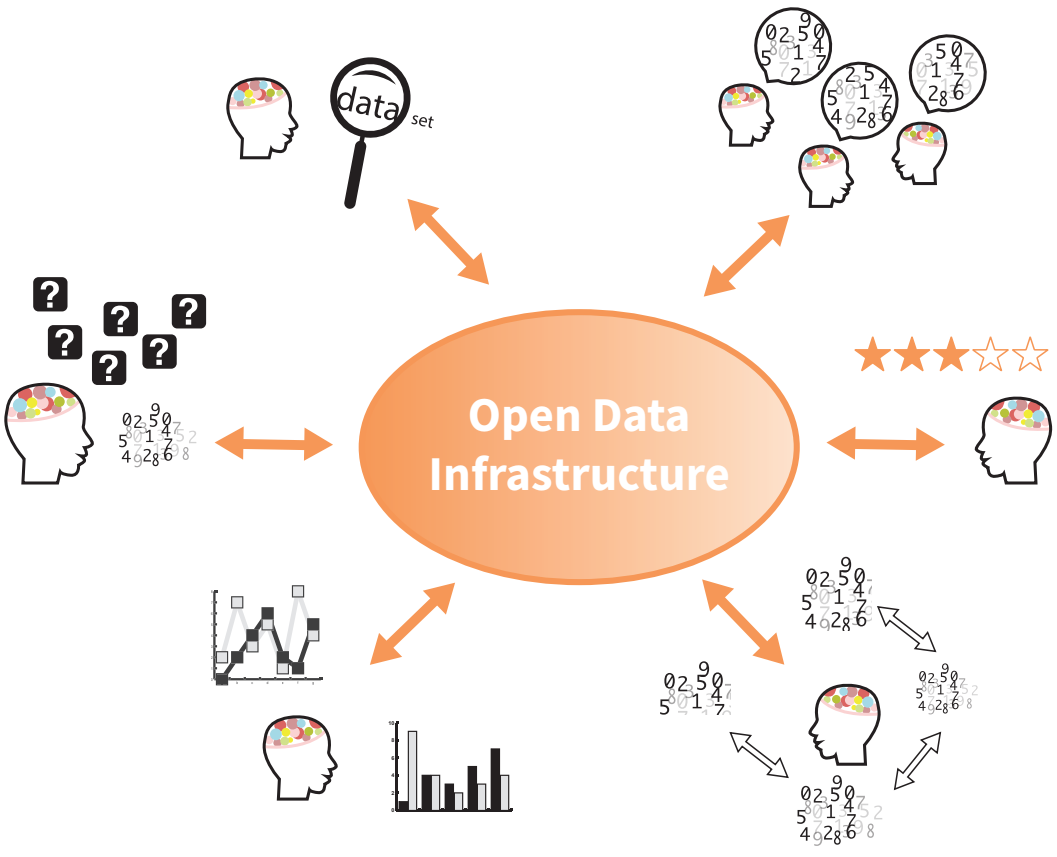
Open Data Infrastructures

The design of an infrastructure to enhance the coordination of open data use



Anneke Zuiderwijk

OPEN DATA INFRASTRUCTURES
The design of an infrastructure to enhance the coordination of open data use

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.Ch.A.M. Luyben;
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
donderdag 29, oktober, 2015 om 15:00 uur

Door

Anne Maria Gerarda ZUIDERWIJK – VAN EIJK

Master of Science in de Criminologie, geboren te Leidschendam, Nederland

This dissertation has been approved by the promotor: Prof. dr. ir. M.F.W.H.A. Janssen

Composition of the doctoral committee:

Rector Magnificus	chairman
Prof. dr. ir. M.F.W.H.A. Janssen	Delft University of Technology

Independent members:

Prof. dr. Y.H. Tan	Faculty of Technology Policy and Management, Delft University of Technology
Prof. dr. J. Stoter	Faculty of Architecture and the Built Environment, Delft University of Technology
Prof. dr. M. Thaens	Erasmus University Rotterdam
Prof. dr. K.G. Jeffery	Cardiff University
Dr. E. Loukis	University of the Aegean
Dr. ir. R. Choenni	Wetenschappelijk Onderzoek en Documentatie Centrum

Keywords: open data, open government data, use, infrastructures, coordination, metadata, interaction, data quality

This research was funded by the European Commission through the Seventh Framework Programme ENGAGE Project.

ISBN: 978-94-6295-351-2
Printed by Proefschriftmaken.nl | Uitgeverij BOXpress
Published by Uitgeverij BOXPress, 's-Hertogenbosch
Cover design by Proefschriftmaken.nl | Uitgeverij BOXPress

Preface and acknowledgements

I was only six years old when I decided that I wanted to become a writer. The type of writer that I had in mind was somewhat different from the type of writer I am now. While I imagined that I would write children's books, I ended up writing this dissertation. Nevertheless, I have never regretted saying 'yes' to the offer of the PhD position that has resulted in this dissertation. Conducting the PhD research and writing it down in my dissertation was not an easy job, yet the many challenges that I faced and the difficulties that I overcame made me a stronger person and taught me a lot about myself. I also got the chance to collaborate with many inspiring people who helped me to further develop myself. I am very grateful for getting the chance to conduct this PhD research at Delft University of Technology and I would like to make use of this opportunity to thank the many people who contributed to my research.

First and foremost, I want to thank Marijn Janssen. Marijn, from the beginning you were confident that I would succeed, and you motivated me considerably by always staying optimistic, no matter what happened. You taught me how to work as an independent researcher and you supported me in seizing many opportunities that positively influenced my career as a scientist. In addition, I want to thank Yao-Hua Tan for his valuable feedback on my research. After I had dived into the details of my research, you helped me to take a broader perspective again that I needed to reflect on my work. Moreover, I would like to express my gratitude to Jantien Stoter and Marcel Thaens for being part of my doctoral committee and for providing useful feedback on my research.

I also learned a lot from the international group of people that I worked with for the ENGAGE-project (funded by the European Commission). The project allowed me to collaborate with many highly esteemed people, including two of my committee members: Keith Jeffery and Euripidis Loukis. Keith and Euripidis, I very much appreciate your presence at my doctoral defence as well as your feedback that helped me to improve the quality of my research. I would also like to thank Yannis Charalabidis, Charalampos Alexopoulos, Spiros Mouzakitis and Evangelos Argyzoudis for the various collaborations within and outside the ENGAGE-project.

There are also many employees of the Research and Documentation Centre (WODC) that I am grateful to. Before I started my PhD research, the environment of the Research and Documentation Centre made me aware of the large amounts of data that governmental organisations were collecting and of the difficulties that they faced in the release of these data to researchers. One could say that this environment challenged me and made me realise that opening governmental data is actually a very relevant topic for both science and practice. Sunil Choenni, thank you for offering me the opportunity to work at the Research and Documentation Centre and for recommending me to Marijn when he was looking for a PhD candidate. Without your support I could never have started my PhD at Delft University of Technology. Also many thanks for your useful comments during the four years of my PhD project. Furthermore, I would like to thank Frans Leeuw for his support during the entire PhD project and for his critical comments that helped me further improve my work. A big 'thank you' also goes to Ronald Meijer, Roexsana Sheikh Alibaks and the members of the Data Archiving Working Group, as well as to my former colleagues of the Statistical Data and Policy Analysis Division (SIBa).

Besides conducting a part of my study at the Research and Documentation Centre, I also got the opportunity to study the release and use of governmental data at The Netherlands Institute for Social Research (SCP) and at the Data Archiving and Networked Services (DANS). I very much appreciate the discussions that I had with various data experts: Jan Spit, Patty Adelaar, Ineke Stoop, Jetske van der Schaaf, Marion Wittenberg, and many other SCP and DANS employees. You assisted me in obtaining the data that I needed for my research.

Many thanks go to those who have helped to conduct the quasi-experiments. Iryna Susha, you were an excellent facilitator and tester of the evaluations, and you helped me considerably by providing the right comments on the quasi-experimental design at the right time. Wally Keijzer, you did not only come up with helpful ideas during our informal discussions, you also participated in testing the quasi-experimental design. Iryna, Wally, Fatemeh, Jolien, Jie, Wen, Ebrahim, Ying, Kostas, Wei and Armin, you were great observers! I also want to thank the students and the professional open data users who participated in the quasi-experiments. In addition, I am grateful to the many students of Delft

University of Technology who have participated in testing the developed prototype, and who have contributed to the final design of the infrastructure.

I would also like to express my gratitude to my colleagues of the ICT section of the faculty of Technology, Policy and Management. Wally Keijzer and Fatemeh Nikayin, you were great office mates and friends. It was always a pleasure to go to the office knowing that I would see you again. Our time together was invaluable and I learned so much from you. Anne Fleur van Veenstra, Ameneh Deljoo, Potchara Pruksasri and Yuxin Wang, thank you for the informal discussions about work and anything else during our time as office mates. Jolien Ubacht, it was great to supervise students together and to collaborate in many other ways. Iryna Susha, thank you for the many interesting discussions about our (open data) research. Many other TU Delft colleagues also contributed to my research, intellectually or in other ways. Just to mention a few, Bram Klievink, Chris Davis, Bastiaan van Loenen, Joris Hulstijn, Mark de Reuver, Sélinde van Engelenburg, Klara Pigmans, Yiwei Gong, Jie Jiang, Nitesh Bharosa, Ying Li and Ebrahim Rahimi, thank you for your support! Additionally, I am grateful to the helpful secretaries of the ICT section. Jo-Ann Karna, Eveline Zeegers, Karin van Duyn, Laura de Groot, Laura Bruns and Diones Supriana, thank you for your help!

I owe much gratitude to my friends. Ellen, Robin, Nienke, Daphne and Daniel, thank you for reminding me that a PhD is also ‘just’ work and that there is more beyond it! Benny, Linda, Vincent and Esther, cooking delicious dinners together once in a while helped me to relax! Marjolein, thanks for the many amusing lunches that we had together and for providing me with insight in how one can perform PhD research at TU Delft in a totally different way than I did. Maaike, Marsha and Annouk, thanks for your support. Additionally, I thank all the TPM PhD candidates with whom I had much fun at the monthly PhD drinks and at the other events organised by the PhD Council.

Finally, I am greatly indebted to my family. Mum and dad, you were always interested in the peculiarities of the scientific world and you followed each and every step of my PhD research. Esther, Vincent, Sophie, Tom, John, Willy, Mandy, Dick, Dennis and Samantha, thank you for your personal support! Special thanks go to my beloved husband. Patrick, after you had talked to Marijn in the first month of my PhD, I found out that he had given you a special task that you took very

seriously. For better or worse, you were dedicated to your task to considerably encourage and motivate me on this PhD journey. I could not have done this without you.

Table of Contents

1. Introduction

Researchers are able to access more and more data opened by the government. Open Government Data (OGD) refers to *structured, machine-readable and machine-actionable data which governments and publicly-funded research organisations actively publish on the internet for public reuse and which can be accessed without restrictions and used without payment* (European Commission, 2011, 2013; Geiger & von Lucke, 2012; Gurin, 2014; Open Knowledge Foundation, 2015). For many years, governments and publicly-funded research organisations have been making data available to researchers. OGD has the potential to lead to benefits, such as gaining new insight for data-driven research (Krotoski, 2012), allowing the generation of new datasets, information, and knowledge when data from various sources are combined (Uhlir & Schröder, 2007), and permitting in depth public scrutiny by making it easier to analyse, process and combine data (Yu & Robinson, 2012).

After OGD providers have disclosed governmental data to the public, researchers outside the government can find and use these data. However, OGD use activities are often not coordinated (we define coordination as *the act of managing dependencies between and among activities performed to use OGD,* see section 3.2.4), and tools for using OGD are fragmented and hardly integrated. In addition, both the literature and practice focus on the publication of OGD, whereas the use of the data is also needed to obtain the benefits. Because of the lacking coordination of the activities of researchers using OGD and because of the lack of integrated tools, OGD are not yet showing their full potential.

An open data infrastructure can enhance the coordination of OGD use by researchers. Such an infrastructure can be defined as *a shared, (quasi-) public, evolving system, consisting of a collection of interconnected social elements (e.g. user operations) and technical elements (e.g. open data analysis tools and technologies, open data services) which jointly allow for OGD use* (see section 3.2.2). OGD infrastructures are internet-based and are usually owned and maintained by governmental organisations. Users and social elements play an important role in OGD infrastructures, since an OGD infrastructure can function as

a central place where researchers can find and use the data published by OGD providers, where they can use integrated tools, and where they can interact with OGD providers and policy makers to discuss their findings from open data use. For example, through the user interface of an OGD infrastructure, a researcher working at a university may find datasets concerning employment. The researcher may use the tools of the infrastructure to analyse and visualise the data and to combine them with other open employment data. This may lead to new insights, which the researcher may discuss with other researchers and with governmental policy makers through the infrastructure. As such, the infrastructure can lead to enhanced coordination of the activities of OGD users. It can reduce fragmentation of open data use activities, and a premise is that it can subsequently be used by governments to improve policy making.

The objective of this study is *to develop an infrastructure that enhances the coordination of OGD use.* Outside the scope of this study are the OGD providers and the policy makers. This study is focused on a specific type of OGD use through infrastructures, namely the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government (see section 1.2). The following section of this chapter provides background information regarding the actors involved in OGD-related activities, followed by a section that offers insight in our focus on researchers as OGD users, and by a section that discusses the dependencies of OGD use for research purposes. Subsequently, the problem statement is provided, including the contributions of this thesis. Thereafter, an overview is given of the research objective and the questions that will be answered with this research. Finally, this chapter provides an outline of the dissertation.

1.1 Open data actors and activities

This study makes a distinction between three types of actors that are involved in OGD, namely 1) OGD providers, 2) OGD users and 3) policy makers (see Figure 1-1). We concentrate on the OGD users, and more specifically on researchers as OGD users. Although OGD providers and policy makers can also be OGD users, this study does not focus on OGD providers and policy makers in the role of OGD

users (see section 1.2). The activities that the three actors involved in OGD-related activities perform are explained below.

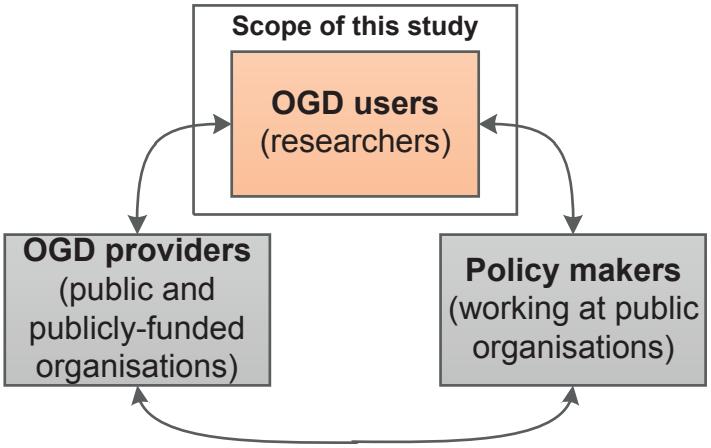


Figure 1-1: Open data actors.

Generally, governmental agencies produce and collect large amounts of data in order to fulfil their daily tasks, or they fund other organisations to produce and collect data for them. Some of these data are obtained through research. For example, through a study of a Ministry of Justice, data concerning numbers of crime victims may be collected in order to formulate the ministry's safety and security policies, or this Ministry may fund a university or another research organisation to carry out research and collect the data. After the data have been collected, representatives of the public agency or of the publicly-funded (research) organisation may decide to release the data to the public by making them publicly available on the internet. In this study we consider both the data collected by government agencies and the data collected by publicly-funded research organisations to be OGD, and we mainly consider data collected through research. The first actor involved in OGD, namely the OGD providers, refers to governmental agencies and publicly-funded research organisations that provide their (research) data to the public.

After governmental data have been released, a second actor – OGD users – can reuse the data. OGD can be used for many different purposes by different types of users. For example, a researcher may use OGD for a scientific study, a

journalist may use OGD to write a news article, and a citizen may use OGD to obtain information about his or her neighbourhood. This study focuses on researchers as OGD users (see section 1.2). Examples of data use by researchers include detecting and correcting records in a dataset, analysing data (e.g. studying a dataset and deriving useful information from this activity, or performing a statistical analysis by using software), visualising data, enriching and curating data (e.g. adding information that was derived from the statistical analysis or visualisation) and linking, comparing and integrating data (see section 3.2.3).

The third actor involved in OGD includes policy makers that work for governmental agencies. Policy makers may use the insights that researchers outside the government obtained from open data use as input for the policies that they develop. For instance, policy makers may use insights that were obtained with the use of open crime data by non-governmental researchers to develop governmental policies about security measures or police surveillance, or they may use insights from external OGD use about epidemic diseases to develop governmental vaccination policies. Policy makers can work on many different types of policies, such as policies in the field of social security, economy, justice, elections, agriculture, transport, health, energy and welfare.

1.2 Open data use and users

This study focuses on the coordination of the use of OGD. OGD users encompass a heterogeneous group of actors that use OGD for different purposes. The needs of each type of user can differ, for example depending on whether their open data use is strategic, tactic or operational, whether it takes place in an international or a national context, whether it takes place inside or outside the government, and whether it focuses on a particular domain (e.g. geographical or social data). In this study we concentrate on *the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures.* We focus on the operational use of structured research OGD, since this is a complicated process that requires data with a high level of detail. Although one may argue that the use of structured data is less complex than the use of unstructured data, even for structured OGD the semantics are often not clear, and they change over time.

This study focuses on the domains of social sciences and humanities, because data from these domains are important for identifying and solving various societal issues, such as poverty, social exclusion, (un)employment, education, social security, integration and immigration. While OGD providers and policy makers can also be OGD users and they can use OGD directly without intervention of actors outside the government, this study focuses on OGD users outside the government. This focus is in line with the PSI directive (European Commission, 2013), which emphasises the use of OGD outside the government. Although our study focuses on OGD use that takes place outside the government, the results of this external data use can thereafter be used within the government. The results may contribute to governmental policy making.

We focus on researchers who use open data for scientific and non-scientific research. This focus leads to the study of a very specific target group of OGD users, namely only those people who are interested in using and who can use research data. This type of OGD use is different from other types of OGD use, such as the use of OGD by citizens or by entrepreneurs. Finally, we focus on OGD use through infrastructures, since an open data infrastructure can function as a central place where researchers can find and use OGD, where they can use integrated tools, and where they can interact with OGD providers and policy makers to discuss their findings from open data use. Subsequently, this can enhance the coordination of OGD use by researchers. When we refer to OGD use and OGD users in the remainder of this dissertation, we refer to the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures.

1.3 Open data use dependencies

Researchers using OGD conduct various activities for which they depend on a variety of tools (that also depend on other tools), on each other, and on other actors (see Figure 1-2). For example, researchers as OGD users depend on OGD providers for the provision of the data, they use different tools for finding, analysing and visualising OGD, and they depend on other OGD users for discussing the outcomes of OGD use. In the context of open data, *dependence* can be defined as

the extent to which open data activities require the actors and tools to work with one another. This section discusses open data use dependencies.

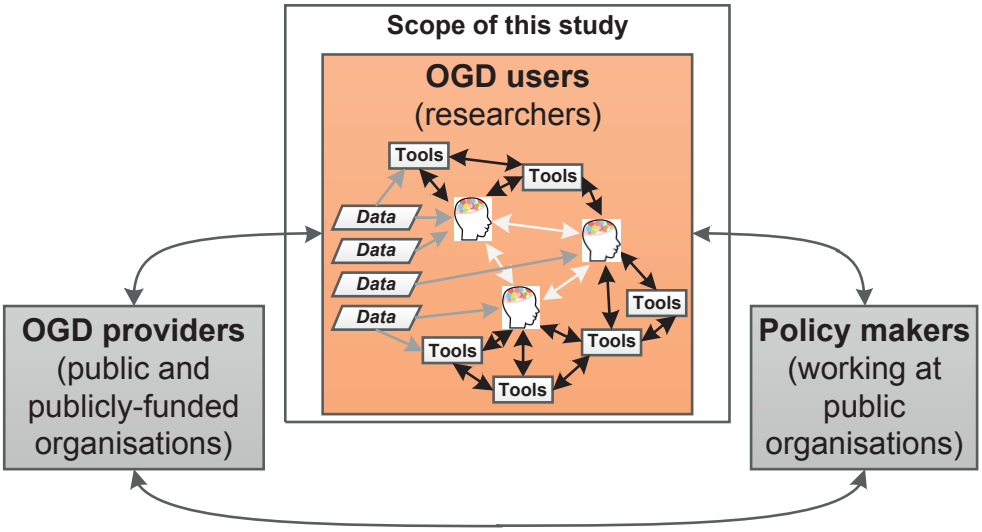


Figure 1-2: Open data use dependencies.

A first dependency results from data publication and is related to the activity of searching for and finding OGD (see the light grey arrows in Figure 1-2). Researchers using OGD depend on OGD providers for obtaining the data that they are interested in. Not only the availability of the data, also the way in which the data are provided and the way that they can be reused (through tools) leads to a dependency between OGD users and OGD providers. The way that the data are published can strongly affect the way that people can access and use the data (Braunschweig, Eberius, Thiele, & Lehner, 2012b). For instance, if open datasets are incomplete or inaccurate, researchers may not be able to use the data. Hence, OGD users depend on OGD providers for the usability of the data.

A second dependency concerns the dependence of researchers on the tools that they need in order to use OGD, such as tools for analysing datasets (e.g. Open Refine, Nesstar or the Microsoft Excel Web App) and tools for visualising datasets (e.g. IBM Many Eyes, Map Designer or Google Charts) (see the black arrows in Figure 1-2). Different tools can be used for each OGD use activity, and the tools also depend on each other for their interoperability. Currently, the tools

that can assist OGD use are provided at many different places on the internet, and they are hardly integrated in OGD infrastructures. Moreover, they often do not interoperate with other tools. The lack of integration and interoperability of tools complicates OGD use. For instance, a researcher who wants to use OGD now needs to search for OGD use tools at many different places on the internet, which is time-consuming and requires expert knowledge regarding which tools are available for which OGD use activity, and where these tools can be found.

A third dependency can be found among OGD users (see the white arrows in Figure 1-2). Researchers using OGD depend on other OGD users for discussing what can be learned from OGD use. These kinds of discussions are important, since OGD use results may be open to multiple interpretations. Researchers may discuss the way that they have used open datasets with their peers, as well as the way that the findings from the data use can be interpreted. They may discuss the findings that they derived from the data use with other researchers to advance their understanding. Existing OGD infrastructures barely support those types of discussions.

Although this is outside the scope of this study, OGD providers and policy makers also depend on researchers using OGD. OGD providers depend on OGD users to obtain feedback regarding data publication that can be used for future data supply. For example, a governmental agency that releases cadastral data may wonder whether the released data are of interest to OGD users, which other (currently closed) datasets OGD users would like to use, and whether the opened data are provided to OGD users in a useful format. Moreover, since OGD use may lead to new insights that can be used for governmental policy making (Napoli & Karaganis, 2010), policy makers depend on OGD users to obtain information that can be used in the development of policies. For instance, a policy maker in the area of crime and justice may use the insights that researchers obtained from combining OGD regarding crimes, police observations and recidivists to develop crime prevention policies (e.g. to determine in which neighbourhoods most crimes are committed, whether it would be useful to increase police observation in these neighbourhoods and how this might affect the number of crimes).

The foregoing shows that researchers using OGD conduct various activities for which they depend on different tools (that also depend on other tools),

on each other, and on other actors. Malone and Crowston (1990, p. 361) refer to "the act of managing interdependencies between activities performed to achieve a goal" with the term *coordination*. They state that coordination is needed to map goals to activities, to relate activities performed by different actors and to manage the interdependencies between these activities (Malone & Crowston, 1990; Malone & Crowston, 1994). The challenges resulting from dependencies between open data related activities can be seen as coordination challenges. The following section discusses the key coordination challenges for open data, and describes how this thesis contributes to solving them.

1.4 Problem statement

At the start of this PhD research, most open data studies were oriented towards data provision (Conradie & Choenni, 2012; Huijboom & van den Broek, 2011; Meijer & Thaens, 2009). Although some research on (closed) data use in general and on OGD use had been conducted (e.g., Braunschweig et al., 2012b), OGD use had received less attention than OGD publication. The lack of attention for open data use was not only reflected in the literature, but also in practice. Governments focused on the publication of OGD, whereas the actual use of the data (which is necessary to gain the benefits) was often neglected. This dissertation contributes to the literature concerning infrastructures that facilitate the coordination of OGD use by researchers outside the government. In the following sections we discuss the key coordination challenges that hinder the coordination of OGD use, and we describe how this study contributes to solving these challenges.

First, at the beginning of this study, open data was an upcoming field and there was hardly any research available. There were some tools available that assisted in making use of open data (e.g. Google Refine and IBM Many Eyes), however, these tools were fragmented and there was no infrastructure enabling the integration of existing tools and enabling the coordination of the activities of OGD users. There was no comprehensive overview of factors that influence OGD use through infrastructures, nor was there an overview of barriers that hinder OGD use. Whereas various studies had been conducted on factors influencing OGD use (Davies & Bawa, 2012; Gurstein, 2011) and on OGD use barriers (Böhm et al., 2012; Braunschweig et al., 2012b), the factors were often defined on a high level of

abstraction or they did not focus on OGD use by researchers. They often did not focus on the barriers related to the dependence of OGD users on different tools, on each other, and on other actors. The contribution of this study is to provide a comprehensive overview of the factors and the barriers that need to be taken into account when one wants to improve the coordination of OGD use by researchers.

Second, although the need for taking a user perspective was acknowledged in the literature, there was a lack of insight in the user requirements for an infrastructure that enhances the coordination of OGD use. Only recently some studies have been conducted on how OGD use can be improved (e.g., Jurisch, Kautz, Wolf, & Krcmar, 2015), yet at the start of this study limited research had been published on user requirements for OGD infrastructures. This study contributes to the existing literature by offering a comprehensive overview of user requirements for enhancing the coordination of OGD use based on practical case studies. Furthermore, while most open data research was focused on the perspective of the OGD provider (Conradie & Choenni, 2012; Huijboom & van den Broek, 2011; Meijer & Thaens, 2009), this research studied functional requirements from the perspective of the open data user.

Third, although the literature suggested a number of functional elements for the development of an open data infrastructure, such as social media (Bertot, McDermott, & Smith, 2012) and access to metadata (Braunschweig et al., 2012b), these elements were often described on a high level of abstraction and they were not described in such a way that they could be used for generating OGD infrastructures that enhance the coordination of OGD use. This study is among the first to describe the design of an OGD infrastructure, including the functional elements it encompasses. This study builds on the existing literature regarding metadata (e.g., Gilliland, 2008; Jeffery, Asserson, Houssos, & Jörg, 2013; Vardigan, Heus, & Thomas, 2008) and the literature regarding the other proposed OGD infrastructure elements, and contributes to the literature by proposing a combination of functional elements that can be used to enhance the coordination of OGD use through an OGD infrastructure. In this study, metadata, interaction mechanisms and data quality indicators are combined in one infrastructure, and existing open data metadata models are refined beyond existing standards.

Fourth, whereas some studies had described architectures for the development of OGD infrastructures (e.g., Charalabidis, Ntanos, & Lampathaki, 2011) when we started this study, research had barely shown what such an infrastructure should look like. This research contributes to the literature for developing an OGD infrastructure by providing a description of what the designed OGD infrastructure should look like and how it can be developed.

Fifth, at the start of this study, there was no insight in how OGD infrastructures can be evaluated to identify their strengths and weaknesses. It was not clear how one can evaluate to which extent functional OGD infrastructure elements can enhance the coordination of OGD use. This study contributes to the literature by showing how quasi-experiments can be used to investigate the effects of developed OGD infrastructures on the coordination of OGD use.

Finally, many studies on coordination have been conducted (Crowston, Rubleske, & Howison, 2004; Malone & Crowston, 1990), and insights from these studies can be used to enhance coordination of open data use activities. However, the literature on coordination is mainly focused on improving processes (e.g., Malone & Crowston, 1990) and none of this work is in the domain of OGD. Our study shows that open data use does not only involve processes, yet it also requires a technical perspective including the integration of tools, a social perspective including interaction between researchers, OGD providers and policy makers, and the interaction between the social and technical perspective. Both the technology and its use are needed to enhance coordination. Coordination literature does not provide guidance regarding how OGD technology is intertwined with OGD use and processes. This study builds on the coordination literature (Crowston et al., 2004; Gittell, 2011; Lu, Xiang, Wang, & Wang, 2011; Malone & Crowston, 1990) and shows that coordination of OGD use does not merely require a focus on processes, but additionally requires the integration of technology and social aspects into these processes.

1.5 Research objective and research questions

In the foregoing it was stated that an OGD infrastructure can potentially enhance the coordination of OGD use by researchers. At the same time, coordination

challenges exist for researchers using OGD (see section 1.4). Taking into account the identified challenges, the objective of this study is as follows.

The objective of this study is to develop an infrastructure that enhances the coordination of open government data use.

To attain the research objective, five research questions have been defined (see Figure 1-3). This study aims to develop an artefact, namely an OGD infrastructure, and therefore we use a design science research approach (Hevner & Chatterjee, 2010; Hevner, March, Park, & Ram, 2004; March & Smith, 1995) for the formulation of the research questions. The research has been divided into five design science research phases, corresponding to the common elements of design science research. Each of the five design science research phases, namely 1) the identification of the problem and related factors, 2) the definition of objectives of a solution, 3) the design of the artefact, 4) the development of the prototype, and 5) the evaluation of the prototype, is addressed by one research question (for more information about the design science research phases see section 2.4).

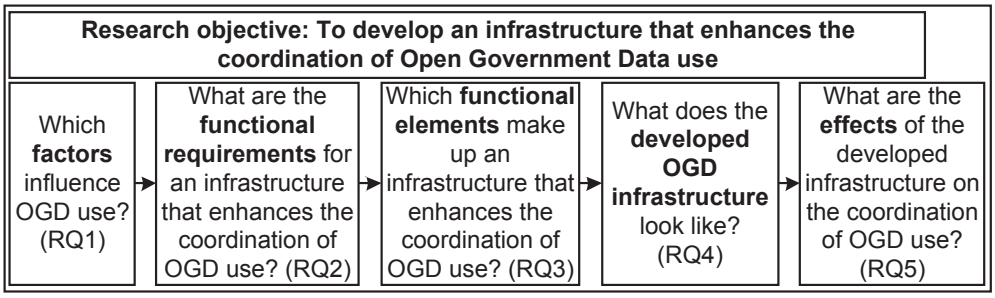


Figure 1-3: Overview of this study's research objective and research questions.

The first research question (RQ1) explores which factors influence OGD use. Factors influencing OGD use are studied by conducting a literature review, as this is expected to provide an overview of the existing knowledge base, so that we can build on the research that has already been performed in the field of open data. OGD is expected to be influenced by social factors, such as the interaction of and collaboration between open data providers and users, as well as by technical

factors, such as the format in which data are presented and tools for monitoring data quality. The identified factors influencing OGD use are clustered.

Within each of the identified clusters of factors influencing OGD use (RQ1), we search for functional requirements for an infrastructure that enhances the coordination of OGD use (RQ2). Functional requirements are identified through case studies that focus on a specific type of open data, namely structured open judicial and social data. Reasons for focusing on these types of data include that they are already disclosed by governmental organisations, and that they are important for identifying and solving various societal issues (see section 4.1.3). Requirements can be defined as detailed descriptions of "what is wanted from the design by the client and by potential users" (Dym & Little, 2004, p. 20). Functional requirements are the requirements that define the specific functionality that shows how a system can be used, while non-functional requirements refer to requirements "which impose constraints on the design or implementation (such as performance requirements, quality standards, or design constraints)" (Stellman & Greene, 2005, p. 110). Since this study aims to improve the functional use of OGD, it focuses on the functional requirements for the OGD infrastructure. While focusing on the functional requirements, an assumption of this study is that the non-functional requirements are met (see section 8.4.2 for a discussion on this topic).

The functional requirements that are identified through the second research question contribute to the identification of functional elements of the OGD infrastructure that enhances the coordination of OGD use (RQ3). Elements are defined as parts of a larger whole, namely parts that together provide the complete infrastructure. The functional elements of the OGD infrastructure will meet the functional requirements that are identified through the second research question. Coordination theory and literature regarding metadata, interaction and data quality underlie the design of the OGD infrastructure. The OGD infrastructure design incorporates the system design, the coordination patterns and the function design. The system design describes the structure and the behaviour of the system. A three-tier metadata model is developed incorporating discovery metadata, contextual metadata and detailed metadata. Two types of interaction mechanisms are designed, namely feedback mechanisms and collaboration and discussion mechanisms. A data quality indicator model is developed which incorporates

different quality dimensions that can be assessed through structured data quality rating (e.g. accuracy and completeness), and also takes into account the purpose of open data use (e.g. through free text quality reviews and evaluator information), since OGD quality depends on the fitness for use. The patterns define the reusable parts of the design with their benefits and an explanation of how they can be applied, and the relation between them. With regard to the coordination patterns, it is explained how the functional elements of the OGD infrastructure can together enhance the coordination of OGD use by researchers. Finally, the function design outlines the functions of the infrastructure.

To be able to evaluate the developed OGD infrastructure, a prototype is developed and described as part of the fourth research question (RQ4). The prototype is constructed as part of the ENGAGE-project, which is a combination of a Collaborative Project and Coordination and Support Action (CCP-CSA) funded by the European Commission under the Seventh Framework Programme. In this project various universities, research organisations and companies collaborate to construct the prototype. The prototype is called 'ENGAGE', which refers to its functions related to engaging OGD users, OGD providers and policy makers. The prototype allows for further refining and testing the user requirements. The answer to the fourth research question reports on the results of the prototype creation, and shows what the developed OGD infrastructure looks like.

The ENGAGE prototype is accessible for the public via a website (www.engagedata.eu). The prototype allows for searching for open datasets in different ways (e.g. entering data in a search bar, filtering, sorting, ordering, categorisation, multilingual search). For each dataset an overview of basic information is provided (e.g. contextual metadata, general data quality assessment score, main content and resources, the options for viewing, downloading and visualising data, comments and remarks on the dataset) as well as more detailed information (e.g. detailed metadata). Users can analyse datasets by exploring the various options provided in the dataset overview (e.g. viewing a dataset without downloading it, viewing which other users had extended or amended the dataset). The prototype allows for using different tools to create tables, charts and maps of open datasets. Interaction mechanisms can be used to give feedback on datasets and processes related to data provision and use, and they can discuss what could

be learned from the use of the data. Various data quality indicators are available, including rating the quality of datasets by assessing the accuracy, completeness, consistency and timeliness of a dataset, by writing a review of the dataset in an open text box (e.g. to elaborate on the purpose of data use), and by viewing information about the data evaluator. These elements and functions together comprise the prototype.

The evaluation of the artefact, i.e. the evaluation of the developed OGD infrastructure, is central to the fifth research question (RQ5). The artefact is evaluated by conducting quasi-experiments that provide insight in the effects of the designed infrastructure on the coordination of OGD use. In the evaluations the participants complete scenario tasks that prescribe them to use various tools, to interact with other OGD users and to use tools that allow for interaction with OGD providers and policy makers. This means that they use OGD in a way that corresponds to our definition of coordination (see section 3.2.4). In the quasi-experiments we examine to which extent the ease and the speed of OGD use was improved by the developed OGD infrastructure, and we examine the coordination of OGD use by including the management of dependencies between and among activities performed to use OGD in the evaluation scenarios. The evaluation indicates to which extent the designed OGD infrastructure can enhance the coordination of OGD use, and it provides insight in how the functional elements of the OGD infrastructure can be used by end-users. The evaluation of the OGD infrastructure also results in suggestions regarding how the OGD infrastructure can be used in the future and which improvements can be made.

1.6 Outline of this dissertation

Figure 1-4 provides an outline of this dissertation and shows the relationship between its chapters.

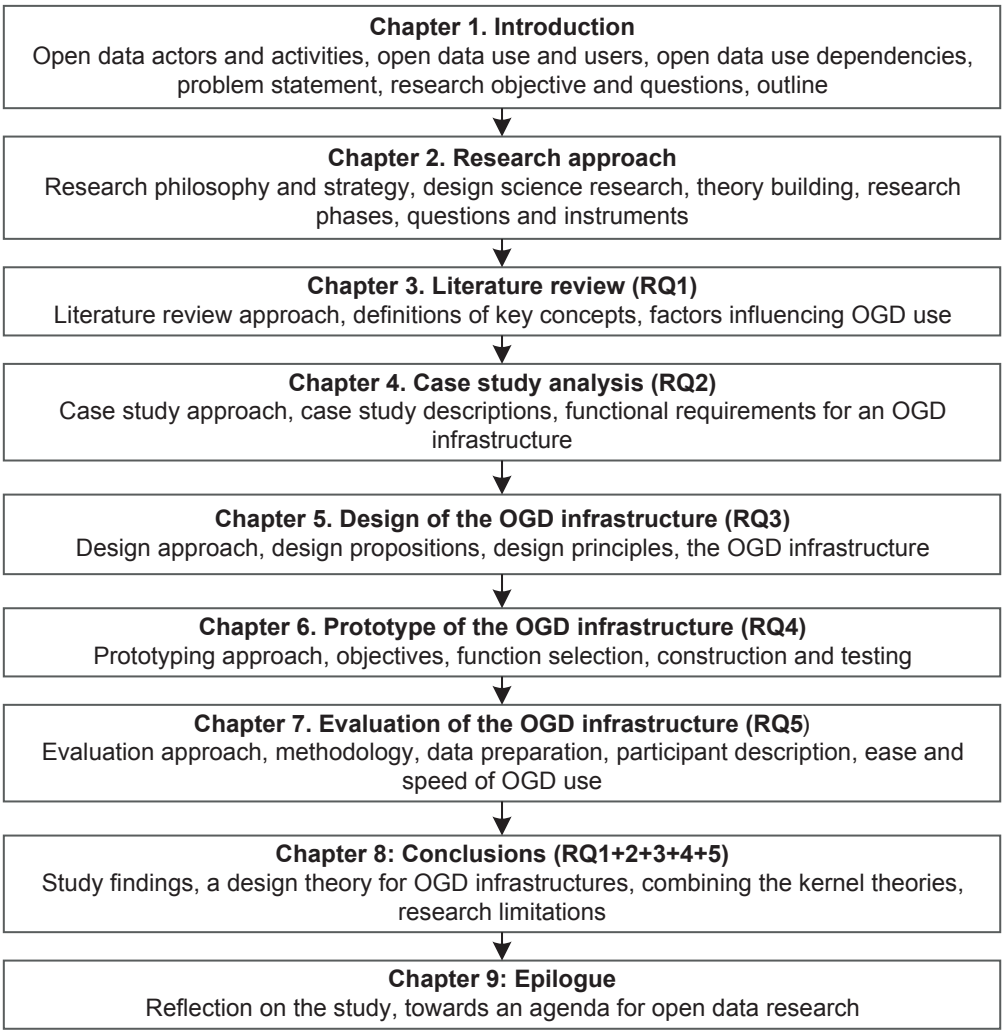


Figure 1-4: Outline of this dissertation.

2. Research approach

The objective of this study is to develop an infrastructure that enhances the coordination of OGD use. This study focuses on the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government. Outside the scope of this study are the data providers and the policy makers, and a premise is that enhanced coordination of OGD use will support policy making. This chapter describes the approach that is used to attain the research objective. It starts with a description of the adopted research philosophy, followed by an explanation of the chosen design science research paradigm. Subsequently, it is described how this study aims to contribute to theory building. Finally, the research phases that will be used to attain the research objective are presented and related to the research questions and the research instruments.

2.1 Research philosophy and strategy

OGD can be investigated from a number of philosophical perspectives. A research philosophy or research paradigm guides the decisions for the research strategy and the selection and use of appropriate research methods (Altinay & Paraskevas, 2008; Easterby-Smith, Thorpe, & Lowe, 2002). A research philosophy can lead to a research strategy of how research is conducted (the methodology), as opposed to strategies that are developed based on the actual research outcomes (e.g. economic or political strategies). The research philosophy comprises the researchers' "assumptions about the nature of the social world and the way in which it may be investigated" (Burrell & Morgen, 1979, p. 1), as well as their assumptions about the physical world (Hirschheim & Klein, 1989). It consists of assumptions about reality (ontology), knowledge (epistemology) and the relationship between human beings and their environment and their extent of free will (human nature) that underlie researchers' intellectual endeavour (Burrell & Morgen, 1979). A research philosophy determines the boundaries of knowledge that a study can result in (idem) and the results and conclusions that it can lead to (Hovland, 1959). The selection of a set of assumptions influences which research

methodologies can be used (Burrell & Morgen, 1979). The research philosophy guides the research strategy, which refers to the "general orientation to the conduct of social research" (Bryman, 2012, p. 35). This study uses a social science perspective rather than, for instance, an engineering perspective. This conditions the selected research strategy. Two types of research strategies dominate the social science literature, namely positivism and interpretivism (Gibbs, 2005). In the following sections the positivist and interpretivist perspective are explained, the motivation for the research paradigm chosen in this study is given, and it is explained how this study deals with the criticisms on interpretivistic research.

2.1.1 Positivism

According to the positivist paradigm, reality is probabilistic and the 'truth' is universal (Vaishnavi & Kuechler Jr, 2008). Relationships within phenomena are fixed and knowledge is obtained through structured instruments (Orlikowski & Baroudi, 1991). A positivist researcher can objectively observe 'the truth', collect data and test hypotheses and theories (Walsham, 2001), which may subsequently contribute to theory generation and development. Positivist researchers often do not intervene in the studied phenomenon and aim to play a passive, neutral role (Dubé & Paré, 2003). Positivism assumes that natural science methods are applied to social science studies and beyond (Bryman, 2012). However, positivism has been criticised for not appropriately accounting for humans' free will. It has been stated that positivism does not take into account that human behaviour does not always conform to certain social 'laws' or rules. Other criticisms are that science is not as objective as positivism claims, and to-date universal positivist laws have not yet been created (Macionis & Plummer, 2005).

2.1.2 Interpretivism

The interpretive research paradigm advocates that multiple realities exist, and that realities are socially constructed by human actors (Vaishnavi & Kuechler Jr, 2008; Walsham, 2001). From this perspective, the interaction between researchers and the world around them results in subjective knowledge. Phenomena are studied from the perspective of the meaning that research participants assign to them (Orlikowski & Baroudi, 1991), and interpretivist research aims to acquire meaning and understanding (Kroeze, 2012). It has also been stated that 'objectivity' in

interpretivism refers to what people agree is objective, and objectivity is therefore a social agreement (Smith, 1983). Interpretive research methods are mainly qualitative and participatory, aimed at understanding situations (Vaishnavi & Kuechler Jr, 2008). Interpretive research is suited for situations in which problems are not completely understood or emotionally charged, or for politicised organisational contexts (Trauth & Jessup, 2000).

Due to the nature of interpretive research, it has been criticised for not having objective evaluation criteria (Chen & Hirschheim, 2004). Interpretive research does not follow that pre-determined criteria can be applied in a mechanistic way, and this type of research cannot be judged by standards (Klein & Myers, 1999). There is no consensus among interpretive researchers on which categorising schemes and scaling justifications should be applied. As a consequence, interpretive research may result in different outcomes (Chen & Hirschheim, 2004). In addition, the interpretive perspective is generally mainly focused on producing general theoretical knowledge through the *generation* of new knowledge (Gregg, Kulkarni, & Vinzé, 2001). In general, the interpretive paradigm does not seek to obtain knowledge from the development and creation of new systems and software (idem). In the following section we explain the motivation for the research philosophy chosen for this study, including the way that this study handles the above-mentioned criticisms.

2.1.3 Motivation for chosen research philosophy and strategy

This study has been conducted from the interpretivistic paradigm, and uses a design science research approach. Design scientists claim that there are multiple, contextually situated alternative world-states which are socio-technically enabled (Vaishnavi & Kuechler Jr, 2008). Design scientists believe that knowledge can be obtained through the controlled construction of artefacts. The design of such artefacts is determined by its context and develops through a number of steps (idem). By creating new and innovative artefacts, design science can widen the limits of human and organisational capabilities (Hevner et al., 2004).

We follow Iivari and Venable (2009) in the sense that we see design science research (also commonly referred to as the design science paradigm or design science) as a type of research that can be based on positivistic or

interpretivist assumptions, rather than a separate research paradigm that contrasts positivism and interpretivism (as argued by Vaishnavi & Kuechler Jr, 2008). Although much design science research is epistemologically oriented towards positivism (Iivari & Venable, 2009), Niehaves (2007) and Iivari and Venable (2009) claim that the interpretive epistemology is also highly relevant in design science research, especially for the evaluation of developed artefacts. Since this research aims to develop and evaluate an artefact (i.e. an OGD infrastructure), the interpretive paradigm may be relevant to this study.

Our major reason for choosing the interpretivist paradigm is that we attempt to understand how the coordination of OGD use can be enhanced through an infrastructure in which humans play a role. We develop an artifact and evaluate how it is used by humans. Open data use is studied from the perspective of the meaning that OGD users assign to it, which is an interpretivistic perspective, rather than testing theories or confirming hypotheses, which is typically done in positivist research. This perspective is taken because the behaviour of OGD users can be capricious, since researchers can use open data for different purposes (e.g. to verify results, to create new datasets, to test hypotheses), they may have different requirements and desires, and they may disagree with each other or change their minds based on the context in which they function. Using a positivistic approach by testing theories or confirming hypotheses would be less applicable for this study, since this would not account for the free will of the actors involved in the use of OGD infrastructures.

Moreover, when we started this study the development of theory for the design of OGD infrastructures was still in a starting phase, and previous research had not provided theory or hypotheses regarding the coordination of OGD use through infrastructures. The study was exploratory and the key variables and the way that they were perceived by the examined actors were unknown. The interpretivist paradigm is often used for exploring new phenomena with unspecified variables, actors and relationships.

Furthermore, to obtain the research objective, this study uses research methods in a way that can be considered interpretivistic. While one may argue that the used research methods can also be conducted from a positivist perspective, we use these methods to acquire meaning and understanding of OGD use from an

interpretivist perspective. For instance, case studies are used to examine how OGD stakeholders perceive functional requirements for an OGD infrastructure from their social reality, participant observations are used to examine the effects of the developed OGD infrastructure from the viewpoint of the observers, and surveys are used to evaluate the meaning that the quasi-experiment participants assign to the developed OGD infrastructure from their own perspective. Rather than seeking to confirm or disconfirm hypotheses as is common in positivism, we try to understand the meaning that people assign to the developed artefact. This meaning is important, since different OGD users may value the elements of the artefact differently. Using a positivistic approach would be less applicable here, since it would prescribe the objective observation of the coordination of OGD use without considering the free will of the OGD users, and without considering the exploratory nature of this research.

2.1.4 Dealing with the criticisms on interpretivistic research

This study handles the criticisms on interpretivistic research as follows. In section 2.1.2 we wrote that interpretivism has been criticised for not having objective evaluation criteria (Chen & Hirschheim, 2004). Although there is no set of agreed criteria for judging the quality of interpretivist research (Oates, 2006), Lincoln and Guba (1985) have proposed alternative and parallel criteria to those for positivist research (e.g. internal and external validity). These criteria include trustworthiness (how much trust can be placed in the research?), confirmability (can we judge how the findings flow from the data and experiences in the setting?), dependability (has the research been recorded and the data been documented?), credibility (is the study's subject accurately identified and defined?) and transferability (to which extent can the findings of the study be transferred to other contexts?) (Lincoln & Guba, 1985; Oates, 2006).

With regard to the above-mentioned criteria, this study addresses trustworthiness by examining different perspectives (e.g. by speaking to different case study participants and by using a variety of evaluation measures). Regarding the confirmability, we tried to make the process that led from data and experiences to findings as transparent as possible. Each chapter of this dissertation starts with an in-depth explanation of the research approach and defines how we reached the

answers to the research questions. The dependability of this study is enhanced by documenting the collected data and by presenting them where relevant in the thesis. With regard to the credibility, we clearly identified and defined the study's subject and the key constructs (see chapter 3). Finally, as far as transferability is concerned, we took various measures to allow for replicating this study, so that generalisations become possible. For instance, protocols have been developed for the case studies and for the participant observations.

Another criticism on interpretive research is that it generally focuses on producing general theoretical knowledge through the *generation* of new knowledge, rather than seeking to obtain knowledge from the development and creation of new systems and software (Gregg et al., 2001). We handled this criticism by using a design science approach. The design science approach emphasises the important role of generating knowledge from design processes and products. In the following sections we elaborate on the design science approach.

2.2 Design science research

Design science is a type of research that does not contrast positivism or interpretivism, but that can be based on positivistic or interpretivist assumptions and that complements these paradigms. In this study we use a design science research approach based on interpretivist assumptions. Design science is concerned with "producing and applying knowledge of tasks or situations in order to create effective artefacts" (March & Smith, 1995, p. 253). According to Simon (1996, p. 114), "design […] is concerned with how things ought to be, with devising artefacts to attain goals", and with "creating something new that does not exist in nature" (Vaishnavi & Kuechler Jr, 2008). Design can be defined as "shaping artifacts and events that create more desirable futures" (Orlikowski, 2004, p. 92). Design science aims to create effective artefacts to "create things that serve human purposes" (March & Smith, 1995, p. 253), or, in other words, to solve problems (Hevner et al., 2004, p. 76), or "turn things into value that people use" (Hevner & Chatterjee, 2010, p. 1).

While design in industries is mainly concerned with the creation of an artefact itself, design science research is also focused on the production of new knowledge (learning through building) that is interesting to a community (Vaishnavi

& Kuechler, 2004). In this way, design science research contributes to a general class of problems and to a broad variety of organisational and societal settings, rather than on a unique design problem of one organisation in one setting (Venable, Pries-Heje, Bunker, & Russo, 2010). The adoption of the design science perspective in this study allows for contributing to solving various challenges for the coordination of OGD use. Hevner and Chatterjee (2010) define design science research as follows.

> *"Design science research is a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence. The designed artifacts are both useful and fundamental in understanding that problem"* (Hevner & Chatterjee, 2010, p. 5).

Thus, Hevner and Chatterjee (2010) emphasise that the designed artefact should be useful, as design science aims to contribute to solving human problems. Design science consists of two major activities: 1) building; the process of constructing an artefact for a specific purpose, namely to generate a design solution to solve a problem; and 2) evaluation; the process of assessing the performance of the artefact (March & Smith, 1995; March & Storey, 2008). The construction and application of the artefact are essential for design science research, as they enable the acquisition of knowledge and understanding of a design problem and its solution (Hevner & Chatterjee, 2010).

Although artefacts are usually represented in a structured form (Hevner et al., 2004), there is no common understanding of what comprises an Information Technology (IT) artefact (Offermann, Blom, Schönherr, & Bub, 2010). IT artefacts are usually made up of various interconnected components (Orlikowski & Iacono, 2001). Based on a literature review, Offermann et al. (2010) derived a typology of IT artefacts. They identified eight types of IT artefacts: system design, methods, languages/notations, algorithms, guidelines, requirements, patterns and metrics. In this study we design several artefacts of the typology of Offermann et al. (2010) (see section 2.4.3).

Design and the creation of artefacts can be supported or aided by kernel theories (Hevner et al., 2004; Pries-Heje & Baskerville, 2008). This study takes a broad view on what comprises a kernel theory, since we want to make use of the relatively limited research that has been conducted on theoretical foundations for the coordination of OGD use. We endorse the idea that kernel theories are "the underlying knowledge or theory from the natural or social or design sciences that gives a basis and explanation for the design" (Gregor & Jones, 2007, p. 322). Thus, kernel theories are "underlying an IS design theory" (Markus, Majchrzak, & Gasser, 2002, p. 181), and they can assist in analysing, explaining or predicting both the design product (the artefact) and the design process (Gregor & Jones, 2007). The advantage of adopting a broad view on kernel theories is that we do not limit the knowledge base that we can use for the design of the OGD infrastructure only to what other scholars have referred to as theories, and that we can make use of a broader range of studies and underlying knowledge that provide directions for the design of the OGD infrastructure. In our study four types of kernel theories are used to analyse and explain the design of the OGD infrastructure, including coordination theory and underlying knowledge originating from literature concerning metadata, interaction and data quality.

As this study aims to enhance the coordination of OGD use, it seeks for improving an existing situation and solving relevant, human, real-world problems. Besides design science, various other methodologies are action-oriented and seek to solve problems and improve existing situations. Examples of related methodologies are Action Research (AR), Action Design Research (ADR) and Soft Systems Methodology (SSM). AR aims at intervening in a particular problem situation and obtaining situation-specific insights by learning from improvement processes (Babüroglu & Ravn, 1992; Lewin, 1947). ADR has in common with AR that it addresses "a problem situation encountered in a specific organisational setting by intervening and evaluating" (Sein, Henfridsson, Purao, Rossi, & Lindgren, 2011, p. 40) and has in common with design science that it constructs and evaluates "an IT artefact that addresses the class of problems typified by the encountered situation" (Sein et al., 2011, p. 40). SSM provides models of 'purposeful action' that allow for exploring existing worldviews on situations to

discuss how one can take action to improve it through a structured process (Checkland, 1981; Checkland & Poulter, 2010).

Both AR and ADR focus on the organisational context. However, our study crosses organisational borders. OGD users are not all affiliated to the same organisation. Moreover, AR and SSM do not integrate the design of an innovative IT artefact into the improvement process, while the development of an innovative IT artefact is part of our research objective. Design science allows for creating an innovative artefact, an OGD infrastructure, which could contribute to the improvement of the coordination of OGD use. Using design science, knowledge can be obtained from the creation of the OGD infrastructure, as well as from its application to the OGD domain through its use. A design science approach is therefore appropriate for attaining the objective of this research. In the following section it is explained how this research contributes to theory building.

2.3 Theory building

There are different views on what a 'theory' comprises. For instance, theories may be prescriptive (statements that indicate how something should be done in practice), end products (statements providing a lens for viewing or explaining the world) or testable (statements of relationships among constructs that can be tested) (Gregor, 2006). Gregor (2006) identifies five interrelated types of theory that are relevant for the IS domain: 1) theory for analysis, 2) theory for explanation, 3) theory for prediction, 4) theory for explanation and prediction, and 5) theory for design and action. These types of theories build on each other and are complementary. Gregor (2006) postulates that the fifth type of theory prescribes 'how to do something'. A theory for design and action prescribes how an artefact can be created, including the methods, techniques and principles for the development of the artefact (Gregor, 2006).

What Gregor (2006) promotes as a 'theory for design and action' is not always recognised as a theory (Gregor & Jones, 2007). There is debate within the design science community about whether design theories and a science of design are even possible (Hevner & Chatterjee, 2010). It has been posited that a theory of design is problematic, because design is a practice, defined by assigned tasks (Hooker, 2004). It is not as obvious how one can organise one's knowledge of

design practice in a systematic way and derive a theory from this as it is for, for instance, a theory of chemistry (idem). The different views regarding theory may partly be related to semantics (Gregor & Jones, 2007). Some researchers take a narrow view on what encompasses a theory, and prefer to use the term 'theory' for natural science or social science types of theory (Gregor & Jones, 2007). From a positivist perspective, requirements of a theory are generally that they provide explanations and predictions, and that they are testable (Gregor, 2006). From an interpretivist perspective, theories do not need to be testable in a narrow sense, but they need to be usable to understand complex situations constructed by social actors (idem).

The use of the term 'theory' in this study is different from the use of the term 'theory' in natural sciences, where theory is commonly generated through observations, hypotheses and experiments. In social sciences, a number of researchers adopt a broad view on the term 'theory', and they also refer to 'theory' with what others might call models, frameworks, bodies of knowledge (Gregor & Jones, 2007), methods or implementations (March & Smith, 1995). For instance, Walls, Widmeyer, and El Sawy (1992) and Simon (1996) refer to design-type knowledge as a theory. Walls et al. (1992, p. 37) state that "a design theory is a prescriptive theory based on theoretical underpinnings which says how a design process can be carried out in a way which is both effective and feasible". Design theories need to be based on theory as well as provide guidance to practitioners (idem). Gregor (2006) integrates the different perspectives on theories by arguing that theories in general can be defined as "abstract entities that aim to describe, explain, and enhance understanding of the world and, in some cases, to provide predictions of what will happen in the future and to give a basis for intervention and action" (idem, p. 616).

Markus et al. (2002) add to this that IS design theories should be both prescriptive and evaluative, instead of merely descriptive, explanatory or predictive. IS design theories both provide guidelines to developers (predictive) and work in practice (evaluative) (idem). In line with this, Vaishnavi and Kuechler (2004) argue that design science research contributes to theory building in two ways: 1) the creation of an artefact could examine whether a certain method is useful, and 2) the relationship between elements of the artefact can be identified by designing

and evaluating the artefact. Previously theorised relationships can be falsified or verified and expanded. In this study case studies and evaluations are carried out to falsify, verify and expand the relationships theorised in the literature and in practice (see section 2.4.5). The evaluation activities of design science research may reveal weaknesses in theories or artefacts and they make it possible to refine and reassess the theories and artefacts (Hevner et al., 2004). Design theories provide knowledge support to design activities, and kernel theories are viewed as part of design theories (Goldkuhl, 2004).

The design artefact can be seen as "an embodiment of the design theory, and its operating influence on its setting is the phenomenon of interest" (Pries-Heje & Baskerville, 2008, pp. 749-750). Gregor and Jones (2007) argue that an IS design theory encompasses 1) the purpose and scope, 2) the constructs, 3) the principles of form and function, 4) the artefact mutability, 5) testable proposition and 6) justificatory knowledge (kernel theories), 7) principles of implementation, and 8) an expository instantiation. In section 8.2 we discuss the eight elements of the IS design theory developed in this study.

In sum, there are different views on what a 'theory' comprises (Gregor & Jones, 2007). Even though there is debate about whether 'theories for design and action' should actually be referred to as 'theories', design-type knowledge is ideally the result of design science research. Regardless of whether it is called 'theory' or not, in line with Gregor (2006) and other researchers we adopt a broad view on theory to refer to design-type knowledge, since we want to take the existing knowledge into account in the design of the OGD infrastructure. Our study strives to contribute to a theory for design and action by providing appropriate design-type knowledge for constructing an infrastructure that enhances the coordination of OGD use. Such a theory can discuss whether the method used to construct the OGD infrastructure is useful, and can assist in identifying the relationship between the functional elements of the constructed OGD infrastructure. Based on these outcomes, such a design theory can provide prescriptions (e.g. methods and techniques) for designing an infrastructure that enhances the coordination of OGD use. Our design theory is therefore both prescriptive and evaluative.

2.4 Research phases, questions and instruments

Various guidelines and methodologies for conducting design science research have been proposed (for example, Hevner et al., 2004; March & Smith, 1995; Peffers, Tunanen, Rothenberger, & Chatterjee, 2008). These studies provide common elements of design science research, and in essence they suggest to start design science research with the identification of a problem, and subsequently to identify objectives of a solution (Peffers et al., 2008). The design and development of an artefact (building), as well as the evaluation of the artefact are other elements that design science research commonly incorporates (March & Smith, 1995). The methodology of this study encompasses these common stages of design science research. Figure 2-1 depicts the five research phases of this study and their relation to the research questions and the applied research instruments. Research instruments can be defined as "the specific methods that are used to execute a particular research strategy" (Gonzalez, 2010, p. 17). Note that this study does not follow the classical natural science stages of observation, hypothesis, experiment and theory, since we strive to contribute to a theory for design and action. The construction of an artefact, namely an infrastructure that enhances the coordination of OGD use, is important to generate the appropriate design-type knowledge. The five research phases of our approach will be explained in the following sections.

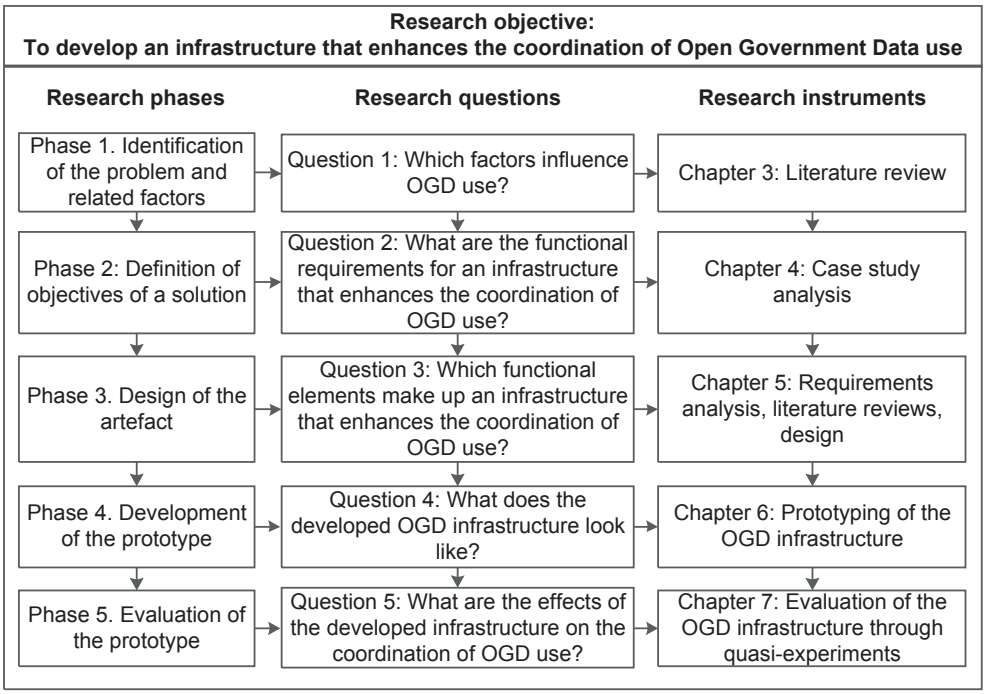


Figure 2-1: Research design.

2.4.1 Research phase 1: identification of the problem and related factors

The first phase of this study comprises the identification of the research problem and related factors. In the first phase we answer the first research question, which seeks for factors influencing OGD use. An introduction to the research problem and related factors has been described in chapter one, while we will elaborate on this in chapter three. The research instrument used in the first research phase comprises a literature review. Starting with the generation of a broad literature overview allows for building on the existing knowledge base, so that this study benefits from the findings from existing studies, rather than starting from scratch. A review of existing literature is critical for academic research (Levy & Ellis, 2006; Webster & Watson, 2002), since it provides the basis for the advancement of knowledge, and it allows for theoretical and conceptual progress (Webster & Watson, 2002). The goal of the literature review is threefold. First, the literature review allows for defining and operationalizing the key constructs of this study, including OGD, OGD infrastructures, OGD use and the coordination of OGD use. Second, the literature helps to identify factors that influence OGD use and thereby to answer the first research question. Third, the literature assists in the generation of a framework for

the identification of functional requirements for an OGD infrastructure in the second research phase. The broad overview of factors influencing OGD use can be used to investigate functional OGD infrastructure requirements in practice in a structured manner.

For the completion of the literature review this study follows key steps as proposed by various other studies (e.g., Danyer & Tranfield, 2009; Levy & Ellis, 2006; Webster & Watson, 2002). The literature review starts with the so-called input phase (Levy & Ellis, 2006), which means that the focus of the literature review is determined (Danyer & Tranfield, 2009), and that the motivation of the study's subject and the explanation of the contributions of the literature review are presented (Webster & Watson, 2002). Subsequently, the key concepts included in the literature review are described and ordered through concepts (Webster & Watson, 2002). Thereafter, the boundaries of the research are made explicit, as determined by the databases that have been searched, and by the criteria that are used to select articles from these sources. The second phase of the literature review concerns the processing of results from the literature search (Levy & Ellis, 2006), in which the applicability of articles from the literature review is determined. Finally, the third step of the literature review comprises the output phase (Levy & Ellis, 2006), in which conclusions are drawn from the literature review and the implications for researchers and managers are explained (Webster & Watson, 2002). The literature review approach will be described in detail in chapter three.

This study is among the first to investigate the coordination of OGD use through infrastructures. When we started this study, there was no literature regarding kernel theories that could be used to develop an open data infrastructure. Therefore, it was not clear at the beginning of this study which kernel theories we could use to attain our research objective. For this reason, the first part of our literature review did not include the kernel theories. An analysis of functional requirements derived from case studies in the third research phase was used to explore which kernel theories might be useful for the development of the infrastructure. The literature review was extended in the third research phase and this second part of the literature review also included a study of the kernel theories. This means that a literature review of the kernel theories can only be found in chapter five.

2.4.2 Research phase 2: definition of objectives of a solution

The second phase of this research employs case studies to define objectives of a solution. Objectives of a solution are sought for in the form of functional requirements for the OGD infrastructure, and they provide the answer to the second research question. We follow Dym and Little (2004, p. 20) by defining requirements as detailed descriptions of "what is wanted from the design by the client and by potential users". Functional requirements define certain functionalities which show how a system can be used (Stellman & Greene, 2005). In the scope of the research objective, the enhancement of OGD use coordination is central to this study. This study searches for an infrastructure that OGD users (i.e. researchers) find functional and usable. Rather than investigating how the infrastructure operates and looking for requirements regarding, for example, its maintainability, sustainability and scalability (i.e. the non-functional requirements), it aims to find out what the users want from the infrastructure (i.e. the functional requirements). While focusing on the functional requirements for the OGD infrastructure, an assumption of this study is that the non-functional requirements are met (see section 8.4.2 for a discussion on this topic).

Functional requirements are identified by carrying out explorative case studies. A case study can be defined as "an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident" (Yin, 2003, p. 13). Case studies are appropriate for investigating a set of broad and complex real-life contemporary events which require a holistic and in-depth examination (Dubé & Paré, 2003; Yin, 2003), and for phenomena that do not allow for studying them outside the context in which they occur (Dubé & Paré, 2003). In line with our definition of OGD use in section 1.2, the interviews focused on the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government. Case studies are conducted because they can be used to investigate the dynamics of single settings (Eisenhardt, 1989) in their natural environment (Benbasat, Goldstein, & Mead, 1987). Moreover, case study research is useful for examining problems in which research and theory are at an early and developing stage (Benbasat et al., 1987; Roethlisberger, 1977).

These characteristics make case studies an appropriate method to identify functional requirements for the OGD infrastructure.

The cases are studied using the framework of factors influencing OGD use as derived from the first research phase. Functional requirements are identified through iterative processes in which several rounds of data collection and data analysis take place. Moreover, the case studies take into account the case study guidelines as described by Yin (2003), Dubé and Paré (2003), Eisenhardt (1989) and other prominent case study researchers. The case study design, the relevance and applicability of case study research, criteria for the selection of these cases and the case study methodology will be described in detail in chapter four.

2.4.3 Research phase 3: design of the artefact

The third phase of this research consists of designing the artefact and answers the third research question. The artefact created in this research is the infrastructure that aims to enhance the coordination of OGD use. The design approach has been divided into three key steps, namely 1) the development of design propositions, 2) the development of design principles, and 3) the development of the system design, the coordination patterns and the functional design of the OGD infrastructure. The first two steps provide input for the design of the OGD infrastructure, whereas the final step provides the actual OGD infrastructure design. Various iterations between these steps took place.

First, building on the requirements identified in the second research phase and on a literature review, design propositions (i.e. assumptions) are developed for the design of the infrastructure. A *design proposition* can be defined as "a general template for the creation of solutions for a particular class of field problems" (Denyer, Tranfield, & van Aken, 2008, p. 395). The propositions are abstractions, and various mechanisms underlie these abstractions (see chapter 5). Whereas the design propositions are defined on a relatively high level of abstraction, the design principles further develop the design input on a more detailed level.

In the second design approach step, design principles are derived from an extended literature review. Design principles can be defined as "'normative' and 'directive' guidelines, formulated towards taking action by the information system architects" (Bharosa, 2011, p. 153). Design principles are identified from four types

of literature. Firstly, since we aim to improve OGD use by enhancing coordination, we derive coordination design principles that prescribe how coordination can be enhanced from the literature on coordination theory and coordination mechanisms. Coordination theory has been used by, for example, March and Simon (1958), Thompson (1967), Malone and Crowston (1990) and Gosain, Malhotra, and El Sawy (2004). Nevertheless, none of the existing studies had used coordination theory in the domain of OGD. This study is the first to use coordination theory and coordination mechanisms in the context of OGD, and to identify principles for the design of OGD infrastructures from coordination theory (see section 5.3.1).

Coordination design principles are overarching and can be used for all three functional infrastructure elements (i.e. metadata, interaction mechanisms and data quality indicators). Secondly, metadata design principles, interaction design principles and data quality design principles are derived from the literature on metadata, interaction, and data quality. The combination of design principles from different types of literature intends to enhance the coordination of OGD use processes and thereby to improve OGD use.

In the third step of our design approach, we describe the artefact, i.e. the OGD infrastructure design. While Offermann et al. (2010) identify eight types of artefacts, this study focuses on three types in particular, namely the system design, coordination patterns and function design. Even though we also develop requirements, methods, guidelines and metrics in this study, which Offermann et al. (2010) consider to be artefacts, we do not consider them to be at the core of this study. Requirements, methods, guidelines and metrics assist in the design of the system, the patterns and the functions, yet are not considered to be the key artefacts of this study. We only view the system design, the coordination patterns and the function design as the core artefacts. The system design can be defined as a "structure or behaviour-related description of a system, commonly using some formalism […] and possibly text" (Offermann et al., 2010, p. 83). A pattern defines the "reusable elements of design with its benefits and application context" (idem, p. 83). The function design translates the system design and patterns to concrete functions that can be implemented in the OGD infrastructure. Subsequently, the function design is used for the development of a prototype of the OGD infrastructure in the fourth research phase. The infrastructure is created in

collaboration with a design team of the EU FP7 ENGAGE project. It is developed through various iterative processes, in which case study participants were also involved and were provided with the opportunity to give feedback. The infrastructure design approach will be described in detail in chapter five.

2.4.4 Research phase 4: development of the prototype

The fourth research phase aims to answer the fourth research question. This research phase reveals how the design of the OGD infrastructure is implemented in a prototype. Prototyping refers to building a working version of various aspects of a system (Bernstein, 1996). The prototyping itself is divided into four prototyping stages. The first prototyping stage comprises identifying what exactly the prototype aims to achieve, while the second stage involves the selection of functions that need to be prototyped (Ince & Hekmatpour, 1987). The third phase concerns the development required to produce the prototype (idem), and, finally, the fourth prototyping phase consists of testing the prototype. The prototyping stages are described linearly, yet much iteration between these stages took place and in practice the stages are not followed in a linear manner. Each stage is expected to lead to new insights, which sometimes requires going back to a prior stage. For instance, testing the prototype may lead to new knowledge which requires modifications in the development of the prototype. The prototyping stages will be described in detail in chapter six.

2.4.5 Research phase 5: evaluation of the prototype

In the fifth research phase we evaluate the created prototype of the OGD infrastructure and answer the final research question. Evaluation of the artefact is of significant importance to design science research (Hevner & Chatterjee, 2010; March & Smith, 1995; March & Storey, 2008). In the process of designing an explicit role needs to be given to evaluation (Verschuren & Hartog, 2005). Evaluation can be defined as “the systematic determination of merit, worth, and significance of something […] or someone” (Hevner & Chatterjee, 2010, p. 109). It refers to the activity “to compare separate parts of a designing process with selected touchstones or criteria (in the broadest sense of the word), and to draw a conclusion in the sense of satisfactory or unsatisfactory” (Verschuren & Hartog, 2005, p. 738).

Quasi-experiments are conducted to evaluate the developed OGD infrastructure. Quasi-experiments encompass 1) a treatment and a control condition, 2) a pre-test and a post-test, and 3) a model that reveals the treatment and the control group effects over time, given no treatment effects (Kenny, 1975). In line with Peffers et al. (2008, p. 56), we use quasi-experiments to "observe and measure how well the artefact supports a solution to the problem" (p. 56). We examine the effects of the developed infrastructure on the coordination of OGD use. Corresponding to our definition of coordination (see section 3.2.4), the quasi-experiment participants complete scenarios that prescribe them to use various tools, to interact with other OGD users and to use tools that allow for interaction with OGD providers and policy makers. The perceived infrastructure's usefulness and usability are evaluated, as well as the conditions for the successful deployment of the infrastructure. A detailed description of the organisation of the prototype evaluation is provided in chapter seven.

3. Literature review

This chapter encompasses the first phase of our study, namely the identification of the research problem and the related factors. It aims to answer the first research question: *Which factors influence open government data use?* The research instrument used in the first research phase comprises a literature review. This chapter starts with a description of the approach that was used to conduct the literature review. Subsequently, the key constructs of this research are defined and the factors influencing OGD use are described. Finally, the findings of this chapter are summarised and the first research question is answered. Parts of this chapter have been published in Zuiderwijk, Janssen, Choenni, Meijer, and Sheikh Alibaks (2012), Janssen, Charalabidis, and Zuiderwijk (2012), Zuiderwijk, Janssen, Meijer, et al. (2012), Zuiderwijk and Janssen (2013a), Zuiderwijk and Janssen (2014a), Zuiderwijk, Helbig, Gil-García, and Janssen (2014), Zuiderwijk, Janssen, and Davis (2014), Zuiderwijk and Janssen (2015) and Zuiderwijk, Klievink, et al. (2013).

3.1 Literature review approach

Figure 3-1 repeats the research design of this study that was presented in chapter two, and extends this figure with the literature review approach used in this study. A review of existing literature is essential for academic research (Levy & Ellis, 2006; Webster & Watson, 2002). It provides the basis for the advancement of knowledge and allows for theoretical and conceptual progress (Webster & Watson, 2002). Hart (1998, p. 1) defines a literature review as "the use of ideas in the literature to justify the particular approach to the topic, the selection of methods, and demonstration that this research contributes something new".

Several scholars have proposed ways to conduct a literature review (e.g., Levy & Ellis, 2006; Webster & Watson, 2002). Levy and Ellis (2006) write that a literature review process consists of three key steps, namely inputs, processing and outputs. With regard to the inputs of the literature review, the focus of the literature review should be determined (Danyer & Tranfield, 2009). Webster and Watson (2002) propose to start with the motivation of the study's subject and the

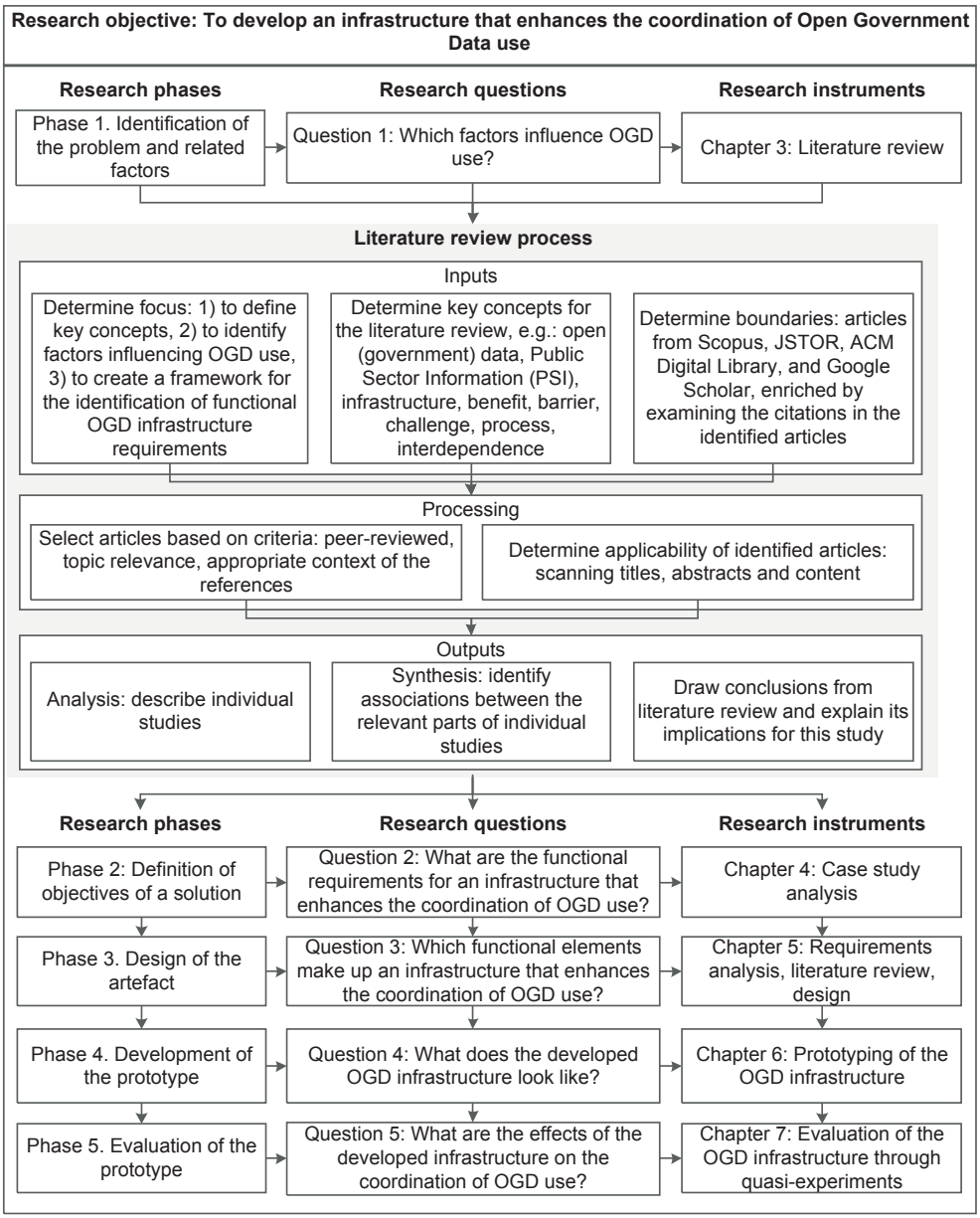


Figure 3-1: Research design including the literature review approach.

explanation of the contributions of the literature review. In addition, the literature review should provide “a firm foundation for advancing knowledge”, as noted by Webster and Watson (2002, p. xiii). Our literature review provides the state-of-the-art regarding the factors that influence OGD use. The goal of our literature review

was threefold, namely 1) to define the key constructs of our study, 2) to identify factors which influence OGD use, and 3) to create a framework that can be used to identify functional requirements for the OGD infrastructure.

A second step in the input phase of the literature review is to describe the key concepts included in the literature review, and to order the literature review concept-centric rather than author-centric (Webster & Watson, 2002). Levy and Ellis (2006) posit that appropriate literature can be found by conducting a literature review for each stream of theory or for each construct that the study is concerned with. The following keywords were used in various combinations to find literature relevant to this research: open data, open government data, linked open data, Public Sector Information, PSI, open data use, digital infrastructure, information infrastructure, infrastructure, socio-technical, open data ecosystem, benefit, advantage, disadvantage, open data impediment, open data barrier, open data challenge, open data problem, open data restriction, coordination, coordination theory, management, interdependencies, process, and requirements. The search on these terms in the various databases resulted in a rich collection of articles. The literature overview was categorised through clusters of factors influencing OGD use.

The literature review focused on keywords that are important for attaining our three literature review objectives, i.e. defining the key constructs of our study, identifying factors influencing OGD use, and creating a framework that can be used to identify functional requirements for the OGD infrastructure. When we started this study, there was no literature regarding open data kernel theories that could assist in attaining these literature review objectives, and thus kernel theories were not included in the keywords. An analysis of functional requirements derived from case studies in the third research phase (chapter 5) is used to explore which kernel theories might be useful for the development of the OGD infrastructure. In the third research phase, the literature review was extended with additional keywords (see section 5.1), where we used the literature to search for possible functional infrastructure elements that might meet the functional user requirements elicited in the second design science research phase.

In the final step of the input phase the boundaries of the literature study ought to be made explicit. The boundaries of our literature review were determined

by the databases that were searched. Papers were sought for in the following databases: Scopus, JSTOR, ACM Digital Library, and Google Scholar. Scopus includes Elsevier (ScienceDirect), Springer, Taylor & Francis, Wiley Blackwell, Institute of Electrical and Electronics Engineers (IEEE), Sage, Emerald, Cambridge University Press and many other sources. As proposed by Webster and Watson (2002), the citations in the identified articles were also examined to find additional relevant literature and to enrich the literature base.

The second phase of the literature review concerns processing the literature search results (Levy & Ellis, 2006). Criteria were defined for the selection of articles in the literature review. First, since peer-reviewed articles can improve the quality of a literature review (Levy & Ellis, 2006), one selection criterion was that articles needed to be peer-reviewed. The selected peer-reviewed articles in this study included journal articles, conference proceedings, books and a few peer-reviewed reports as well as peer-reviewed papers that were published at workshops and websites. A second criterion for selecting articles was that they described information relevant to the topic of this study (e.g. OGD use and infrastructures). A third selection criterion was that the context of the references appeared appropriate for citing them in this study. The applicability of articles from the literature review was determined by scanning their titles, abstracts and content.

Finally, the third step of a literature review concerns the output (Levy & Ellis, 2006). The output phase consists of analysis, i.e. a description of individual studies and how they relate to each other, and synthesis, which is focused on identifying associations between the relevant parts of individual studies (Danyer & Tranfield, 2009). Moreover, in this phase conclusions are drawn from the literature review and the implications for researchers and managers are explained (Webster & Watson, 2002). One should search for patterns in the review results and present what has been learned from the literature review. In our study we create an overview of the identified factors which influence OGD use from the selected relevant publications. Conclusions regarding these factors will be drawn in section 3.4 (also see Appendix A).

Different types of literature reviews exist, including systematic (literature) reviews, meta-analysis and narrative reviews (Petticrew & Roberts, 2006). Systematic (literature) reviews are reviews that strive “to comprehensively identify,

appraise, and synthesize all the relevant studies on a given topic" (p. 19). A meta-analysis is "a review that uses a specific statistical technique for synthesizing the results of several studies into a single quantitative estimate (i.e., a summary effect size)" (p. 19). Narrative reviews are "the process of synthesizing primary studies and exploring heterogeneity descriptively, rather than statistically" (p. 19). The literature review that we undertake in this chapter can largely be characterised as systematic. Following Petticrew and Roberts (2006), we clearly define the research question that needs to be answered with the literature review and determine the types of studies that we need to answer this question. Moreover, we carry out a comprehensive literature search to find the studies and screen the search results. While we do not evaluate the methodology used in each study (which is a limitation of our literature review), we do exclude studies that were not peer-reviewed. Finally, we synthesise the studies and describe them in this chapter.

3.2 Definitions of key constructs

The first objective of the literature review was to define the key constructs of this study. In the following sections we discuss the constructs OGD, OGD infrastructures, OGD use and coordination of OGD use.

3.2.1 Open Government Data (OGD)

To be able to define Open Government Data, one needs to know what is meant with 'data'. The Data Information Knowledge Wisdom (DIKW) model is often used to define data and to explain how it differs from 'information', 'knowledge' and 'wisdom' (and sometimes also 'understanding'). The DIKW model dates back to the 1960s, and is generally accepted "as one of the most well-known models in information science" (Ma, 2012, p. 720). Data is at the bottom of the model, and one needs to ascend from data to information, from information to knowledge, from knowledge to understanding, and from understanding to wisdom (Ackoff, 1989). Data, information, knowledge and understanding deal with the past or with what is already known, while wisdom is at the top of the model, and concerns the construction of a future vision (idem). According to Ackoff (1989, p. 3), "data are symbols that represent properties of objects, events and their environments" and "they are products of observation". Information can be defined as something that "is extracted from data by analysis in many aspects of which computers are adept"

(idem, p. 3). Knowledge then applies data and information and answers “how” questions, while understanding is an appreciation of “why” questions. Finally, wisdom refers to evaluated understanding (Ackoff, 1989). In this study we focus on data, while we assume that open data can be used to generate information, knowledge and understanding.

The European Commission uses the term Public Sector Information to refer to OGD and defines it as “all the information that public bodies in the European Union produce, collect or pay for” (European Commission, 2011, p. 1). The Open Knowledge Foundation (2015) writes that open data or open content refers to the situation in which "anyone is free to use, reuse, and redistribute it - subject only, at most, to the requirement to attribute and/or share-alike”. Lindman, Kinnari, and Rossi (2014, p. 740) refer to open data as “data, which is legally accessible through the Internet in a machine-readable format”. Geiger and Von Lucke (2012, p. 269) define OGD as "all stored data of the public sector which could be made accessible by government in a public interest without any restrictions for usage and distribution". This definition excludes the release of governmental data which should remain confidential, are private or contain industrial secrets.

Existing OGD definitions have in common that they refer to governments and publicly-funded research organisations as the collector and the provider of the data (European Commission, 2003, 2011), and that they indicate that OGD are published on the internet with the aim to have them reused by the public (European Commission, 2003; Janssen, 2011; Open Knowledge Foundation, 2015). Moreover, various definitions show that public access to OGD should be provided without restrictions (Geiger & von Lucke, 2012; Open Knowledge Foundation, 2015) and that the data should be usable free of payment (Gurin, 2014; Open Knowledge Foundation, 2015). Furthermore, there is a general notion that OGD are preferably structured and machine-readable (Lindman et al., 2014; Martin, Foulonneau, & Turki, 2013). Building on these common elements of existing OGD definitions, this research defines OGD as follows.

> *Open Government Data are structured, machine-readable and machine-actionable data that governments and publicly-funded research organisations actively publish on the internet for public reuse and that can be accessed without restrictions and used without payment.*

This study is oriented towards data released by governments and by publicly-funded research organisations. We refer to both types of data with the term OGD. An important reason for focusing on governmental data is that public agencies have increasing volumes of data (Karr, 2008, p. 504), and that much of the available open data is provided by governments. Governmental data providers may have produced or collected the data themselves, or they may have paid for the production or collection of the data by external organisations. Moreover, this study addresses data that are published actively on the internet, and that can be accessed without restrictions and used without payment. This means that data which are provided passively (i.e. on request of a data user, e.g. a Freedom of Information Request), or which the data consumer needs to pay for are considered to be outside the scope of this study. This does not mean that there are no restrictions for usage or (re)distribution of the data, since in reality there can be many open data related barriers.

With regard to the data themselves, this dissertation is directed specifically at OGD which are structured, machine-readable and machine-actionable. The European Commission also focuses on the release of documents (European Commission, 2013). Although documents have structure and are increasingly machine readable (e.g. documents can be analysed with machines), this study does not focus on documents as OGD since they may not be machine actionable. Textual data (e.g. PDFs) and audio and video files (e.g. recorded interviews) are considered to be outside the scope of this study. Data that are within the scope of this study are eXceL Spreadsheets (XLS-files) and Comma Separated Value files (CSV-files) (see Berners-Lee, 2009). The reason for excluding non-machine readable, non-structured data and non-machine actionable data (e.g. PDFs) is that these data are more difficult to reuse. For instance, the analysis of textual data is complex and they may have less potential to facilitate the reuse of open data by the public at large.

As far as the data are concerned, this study focuses on OGD that can be reused. If data cannot be reused, they have little potential to increase gaining new insight for data-driven research and contribute to other possible open data advantages. The focus is on OGD for which the way that they are interpreted is important in their reuse. The semantics of OGD may be unclear outside the context of the governmental agency that produced the data. Therefore, interpretation is often important to understand OGD. Finally, this study is oriented towards OGD that can contribute to or influence policy-making, since open data can then be used for additional purposes than the ones that they were created for initially.

3.2.2 OGD infrastructures

There is no common understanding of what an OGD infrastructure comprises. In this section we develop a definition of OGD infrastructures based on literature regarding digital infrastructures and literature regarding information infrastructures. The term digital infrastructures refers to "a collection of information technologies and systems that jointly produce a desired outcome" (Henfridsson & Bygstad, 2013, p. 912). Tilson, Lyytinen, and Sørensen (2010) add to this that digital infrastructures may also consist of organisational structures and related services, and facilities which are necessary for the functioning of an enterprise or industry. Both social and technical elements and the interactions between such elements play an important role in digital infrastructures (Henfridsson & Bygstad, 2013; Janssen, Chun, & Gil-Garcia, 2009). Digital infrastructures can be characterised as public and quasi-public utilities and facilities (Janssen et al., 2009). They typically have large numbers of users which may vary, and the usage and type of users of infrastructures may evolve over time (idem). Digital infrastructures may be focused on a certain industry, or they may be corporate, regional, national or global (Tilson et al., 2010). Sharing information by a large number of users is often a necessary condition for the existence of the infrastructure (Janssen et al., 2009). Examples of digital infrastructures are the internet and information exchange networks (idem).

With regard to information infrastructures, Braa, Hanseth, Heywood, Mohammed, and Shaw (2007) claim that such infrastructures include technological and human components, networks, systems and processes that contribute to the functioning of a specific information system. Hanseth and Lyytinen (2010, p. 4)

define an information infrastructure as "a shared, open (and unbounded), heterogeneous and evolving socio-technical system (which we call installed base) consisting of a set of IT capabilities and their user, operations and design communities" (p. 4). Information infrastructures are seen as shared universally and across multiple IT capabilities, because they are a shared resource, a foundation, for a community (Hanseth, 2004). The sharing aspect refers to the web of integrated applications and networks as an infrastructure, rather than single applications by themselves (Bygstad, 2010). Monteiro et al. (2012) add to this definition that information infrastructures consist of interconnections of a large number of modules or systems, such as a multiplicity of purposes, agendas and strategies. It has been argued that information infrastructures are open in the sense that they allow for unlimited connections to user communities and new IT capabilities, and that they are heterogeneous because they consist of different components (e.g. technological and non-technological), layers, and sub-infrastructures (Hanseth, 2004). Information infrastructures are evolving continuously and applications can be integrated with others and may be included in a network of applications. The evolvement is unlimited by time or user community and are evolution path dependent (Hanseth, 2004).

In sum, the literature shows that digital and information infrastructures are commonly defined as shared systems (Hanseth, 2004; Hanseth & Lyytinen, 2010), that can be public or quasi-public (Janssen et al., 2009), and that evolve over time (Janssen et al., 2009). Another common element of digital and information infrastructures is that they encompass social and technical elements that interact and that are connected (Braa et al., 2007; Henfridsson & Bygstad, 2013; Janssen et al., 2009). Building on these common elements identified from the literature with respect to digital and information infrastructures, this study adopts the following definition of an OGD infrastructure.

> *An OGD infrastructure is a shared, (quasi-)public, evolving system, consisting of a collection of interconnected social elements (e.g. user operations) and technical elements (e.g. open data analysis tools and technologies, open data services) which jointly allow for OGD use.*

OGD infrastructures are shared since they are controlled by multiple actors, such as data providers and different types of data users. A necessary condition for the existence of the OGD infrastructure is that data and information are shared by a large number of users. At the same time OGD infrastructures are usually owned and maintained by governmental organisations. Such infrastructures are to a large extent open, which allows for their use all over the world. They are internet-based and can be provided through cloud computing. Most parts of the infrastructure are public and can be accessed and used by anyone, while other parts are quasi-public as they can only be used by certain user groups (for example, some modules related to data publication are intended for data providers only). The OGD infrastructure does not only include the backbone of the system, but also the interface. OGD infrastructures evolve over time, have different types of users with different needs, and information and data are shared by large numbers of users. This definition reflects the socio-technical focus adopted in this study. Modules of OGD infrastructures are both social and technical, and only through their interaction OGD use is enabled. The different types of modules jointly allow for OGD use.

3.2.3 OGD use

This section clarifies what is meant with OGD use in this study. This type of insight also provides directions in which we should search for functional infrastructure elements that may contribute to better OGD usage. Davies (2010) makes a distinction between five types of data use, namely:

1) data to fact – the extraction of particular facts from datasets;
2) data to information – generating a representation of a dataset, interpreting it, and reporting on the interpretation;
3) data to data – extending an original dataset by combining it with other data, changing its format or manipulating it otherwise, and subsequently sharing the extended dataset;
4) data to interface – providing an interface to interactively access and explore data;
5) data to service – integrating datasets to produce new products or services.

The first three types of data use are fundamental. Without the first three types of data use, the last two are not possible. For instance, the development of a service requires that certain information is extracted from datasets. The fourth and the fifth data use types refer to the development of interfaces and innovative services, which refer to the use of OGD by developers and entrepreneurs and which demand considerable skills and experience from the data user. Since this study focuses on OGD use by researchers rather than developers or entrepreneurs, this study is scoped towards Davies' (2010) first three data use types.

The data to fact, data to information and data to data types of data use refer to various OGD use activities, such as downloading datasets and viewing them online. We divided these OGD use activities into five categories. Table 3-1 depicts the main OGD use activities that we derived from the literature in the first three categories of Davies' (2010) OGD use categorisation. The table shows that various OGD use activities can be identified within the three categories of Davies (2010), such as data querying (Auer, Lehmann, Ngomo, & Zaveri, 2013), data cleansing (Alexopoulos, Spiliotopoulou, & Charalabidis, 2013) and statistical analysis of data (Kuk & Davies, 2011). We divided the identified OGD use activities into five categories, namely OGD searching, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. These OGD use categories help to scope the literature review towards the type of information that we aim to acquire.

OGD use categories	OGD use activities	Source
Searching for and finding OGD	Data search (e.g. simple text search, advanced search)	Auer et al. (2013), Charalabidis et al. (2011), Kuk and Davies (2011), Petychakis et al. (2014)
	Data navigation	Charalabidis et al. (2011)
	Data browsing	Auer et al. (2013), Kuk and Davies (2011), Petychakis et al. (2014)
	Data extraction, querying and exploration	Auer et al. (2013), Lindman et al. (2014)
	Data sorting and data requests	Charalabidis et al. (2011)
OGD analysis	Data manipulation and contextualisation	Kuk and Davies (2011)
	Online view	Petychakis et al. (2014)
	Statistical analysis	Charalabidis et al. (2011), Kuk and Davies (2011)
	Data download (e.g. through API)	Alexopoulos, Spiliotopoulou, and Charalabidis (2013), Kuk and Davies (2011), Alexopoulos, Spiliotopoulou, et al. (2013)
	Data integration	Kuk and Davies (2011), Lindman et al. (2014)
	Detecting and correcting records in a dataset ('cleansing')	Alexopoulos, Spiliotopoulou, et al. (2013)
	Data curation, semantic annotation, data and knowledge mining (e.g. clustering, regression, association rule discovery)	Charalabidis et al. (2011)
	Data enrichment, create new structured information and correct and extend existing information	Auer et al. (2013)
	Linking and connecting to other entities	Auer et al. (2013), Behkamal, Kahani, Bagheri, and Jeremic (2014)
	Data transformation	Charalabidis et al. (2011), Lindman et al. (2014)
OGD visua-lisation	Data visualisation (e.g. plots, maps, graphs)	Charalabidis et al. (2011), Kuk and Davies (2011), Lindman et al. (2014)
Interaction about OGD	Take feedback and input from end users as training input (i.e. as positive or negative examples, e.g. correct data and engage the public in agency operations)	Auer et al. (2013), Bertot et al. (2012)
	Collaboration (e.g. discussion forums, messaging, user groups)	Charalabidis et al. (2011)
OGD quality analysis	Quality analysis	Auer et al. (2013), Charalabidis et al. (2011)

Table 3-1: OGD use categories and activities derived from the literature.

3.2.4 Coordination of OGD use

In order to enhance the coordination of OGD use, this section explains what is meant with coordination. Coordination can be defined in different ways. Some scholars have examined coordination to study dependencies on a high level of analysis, such as dependencies between two organisations or two divisions within one organisation. For example, Van de Ven, Delbecq and Koenig (1976, p. 322) write that "coordination means integrating or linking together different parts of an organisation to accomplish a collective set of tasks". Others have studied dependencies on a more detailed level, focusing on dependencies between and among activities performed by actors. For instance, Malone and Crowston (1990, p. 361) write that coordination refers to "the management of interdependencies between activities to achieve a goal", and Crowston et al. (2004) emphasises that dependencies also arise among tasks or among resources. According to this view, coordination maps goals to activities, relates activities performed by different actors, and manages the interdependencies between these activities (Malone & Crowston, 1990; Malone & Crowston, 1994). This study focuses on the coordination of OGD use activities conducted by different actors, rather than the coordination of departments within an organisation, or the coordination between organisations. Building on the coordination definition of Malone and Crowston (1990), we define coordination of OGD use as *the act of managing dependencies between and among activities performed to use OGD.*

3.3 Factors influencing OGD use

The previous sections addressed the first goal of our literature review, namely to define the key constructs of our study. This section addresses the second literature review goal, namely to identify factors which influence OGD use. In the following sections the clustered factors are described using the categories of OGD use as identified in section 3.2.3, since these OGD use categories already indicate the direction in which we need to search for factors influencing OGD use. A comprehensive overview of influencing factors is provided in Appendix A.

3.3.1 Factors influencing searching for and finding OGD

Four clusters of factors which influence searching for and finding OGD were derived from the literature (see Table 3-2), namely 'data fragmentation',

'terminology heterogeneity', 'search support', and 'information overload'. A first cluster of factors influencing searching for and finding OGD refers to the *fragmentation of datasets*. The literature shows that it is often difficult to locate released OGD (Cowan & McGarry, 2014; Ding, Peristeras, & Hausenblas, 2012), since data are offered at many different places (Braunschweig, Eberius, Thiele, & Lehner, 2012a; Conradie & Choenni, 2014; De Vocht et al., 2014). Numerous websites have been developed which all provide different types of OGD on different topics. Open data are fragmented by default (De Vocht et al., 2014).

Clusters of factors influencing searching for and finding OGD	Cluster description
Data fragmentation	Data users find it difficult to locate the datasets that they want to use, since the data are offered at many different places.
Terminology heterogeneity	Heterogeneous terminologies are used to describe datasets, so that users often do not know which terms they should use to search for the data that they need.
Search support	Most OGD infrastructures provide simple search functionalities and there is a lack of more advanced multilingual, data query functionalities.
Information overload	Amounts of OGD can at a certain point become overwhelming which complicates finding the OGD that a user needs.

Table 3-2: Clusters of factors influencing searching for and finding OGD.

A second cluster of factors influencing OGD use concerns *terminology heterogeneity*. Each discipline has its own terminologies which leads to heterogeneity (Reichman, Jones, & Schildhauer, 2011). In general, there are differences in the way that programs and organisations define datasets (Dawes & Helbig, 2010). This may also apply to OGD, since OGD originate from many different organisations. In addition, controlled vocabularies, which are "formally maintained list[s] of terms intended to provide values for metadata elements" (Duval, Hodgins, Sutton, & Weibel, 2002, http://www.dlib.org/dlib/april02/weibel/04weibel.html), are often not used for describing OGD, while controlled vocabularies can be used to make the use of terminology more consistent. Different terms and vocabularies are used to describe open datasets (Yannoukakou & Araka, 2014; Zhang, Dawes, & Sarkis, 2005). Such heterogeneity complicates searching for and finding OGD, since users may not know which terms to use for finding a dataset on a particular topic.

Third, factors related to *search support* appear to influence searching for and finding OGD. Petychakis et al. (2014) state that the search options for open datasets are limited. Most open data portals allow for a simple text search or browsing through categories, yet they often do not provide more advanced search functionalities (idem). Moreover, searching for OGD in multiple languages is often not supported (idem), since both the metadata and the data are often provided in only one language, which makes it difficult to search for OGD published in another language.

The fourth cluster of factors influencing OGD use encompasses *information overload*. On the one hand, research in general shows that up to a certain point individuals perform better when they receive more information, i.e. the quality of their decisions and reasoning improves. However, beyond this point their performance rapidly decreases and the provided information will no longer be used for decision-making (Eppler & Mengis, 2004). This situation is typically referred to as information overload (idem). In research on amounts of web information, Ho and Tang (2001) found that available data and information may become overwhelming. This may also be the case for OGD. More and more governmental datasets are becoming available for public reuse (Kulk & Van Loenen, 2012; Magalhaes, Roseira, & Manley, 2014), and this may lead to the situation in which open data users receive too much information. The availability of increasing amounts of OGD complicates their effective use (Magalhaes, Roseira, & Strover, 2013) and people are limited by their ability to curate, search, analyse and visualise open data (Cowan & McGarry, 2014). In combination with a lack of search support, augmenting numbers of datasets make it difficult to search for and find the OGD that a user needs.

3.3.2 Factors influencing OGD analysis

The four clusters of factors which influence OGD analysis are summarised in Table 3-3 and include ‘data context’, ‘data interpretation support’, ‘data heterogeneity’, and ‘data analysis support’. First, the *context of datasets* was identified as a cluster of factors which influence OGD analysis. Alexopoulos, Spiliotopoulou, et al. (2013) note that open data infrastructures traditionally do not add contextual information to the datasets that they provide. Dawes and Helbig (2010) write that the ease of

finding and understanding a dataset depends on the availability of data about the dataset and the contextual information. Dawes, Pardo, and Cresswell (2004) note that definitions of key terms can be lacking for datasets. This may also be the case in the open data domain. This poses a problem, since a large part of the population lacks knowledge of the context of these data (Foulonneau, Martin, & Turki, 2014). The lack of data about the data may hinder the adequate use of these datasets, and, more specifically, the lack of contextual information may make it difficult to analyse and interpret the data.

Clusters of factors influencing OGD analysis	Cluster description
Data context	Open data providers often do not provide extensive contextual data about a dataset, which complicates the analysis and interpretation of the data.
Data interpretation support	There is potential for the misuse, misunderstanding and misinterpretation of open data, since open datasets often lack extensive contextual data, which complicates data analysis and interpretation.
Data heterogeneity	The heterogeneity of data with regard to, for instance, their format and semantics complicates OGD analysis.
Data analysis support	Open data use requires tools that support data analysis, while these are often not provided on open data infrastructures.

Table 3-3: Clusters of factors influencing OGD analysis.

Secondly, the literature shows the importance of *data interpretation support* for OGD analysis. Research conducted by Conradie and Choenni (2014) shows that the fear of drawing false conclusions from open data use is commonly heard. This is not surprising, as data users might (either intentionally or unintentionally) misinterpret datasets (Kucera & Chlapek, 2014). Dawes et al. (2004) found that reusing information for a particular purpose while they were collected for another purpose potentially leads to misuse, misunderstanding and misinterpretation of datasets. Data may be completely inappropriate for certain purposes (Dawes, 2010), yet they may still be used for these purposes and this might lead to false conclusions. This equally applies to the open data field, as open data can be reused for other purposes than they were collected for originally.

A third cluster of factors, *data heterogeneity*, refers to differences between open datasets that complicate the analysis of these data. The literature shows that

the use of open data through applications requires the interpretation and combination of heterogeneous data from a variety of sources (Mora Segura, Sanchez Cuadrado, & De Lara, 2014). Various articles refer to the disclosure of open datasets in heterogeneous formats (Jeffery, Asserson, Houssos, Brasse, & Jörg, 2014; Mora Segura et al., 2014; Yannoukakou & Araka, 2014). Moreover, the semantics of open datasets may be ambiguous (Conradie & Choenni, 2014).

Fourth, ODG is influenced by factors related to *data analysis support*. Braunschweig et al. (2012a) posit that the analysis of data requires the use of different tools. Mora Segura et al. (2014) state that open data use requires the development of applications with rich data analysis and visualisation tools. At the same time, Novais, Albuquerque, and Craveiro (2013) point at the lack of tools to generate information with open data that can easily be understood by the population. Moreover, it has been argued that most traditional open data infrastructures only supply basic data download and upload functionalities instead of more advanced data analysis tools (Alexopoulos, Spiliotopoulou, et al., 2013; Charalabidis, Loukis, & Alexopoulos, 2014). The lack of support for data analysis might influence the extent to which OGD can be analysed effectively.

3.3.3 Factors influencing OGD visualisation

One cluster of factors which influence OGD visualisation was identified, namely: 'data visualisation support'. *Data visualisation support* refers to the necessity of visualisation tools for making sense of OGD. Several authors have stated that visualisation tools are useful (De Vocht et al., 2014) or even necessary for using open data (Shadbolt et al., 2012). In general, data visualisations can be used to make information more visible, to tell stories and to simplify, clarify and analyse data (De Vocht et al., 2014; Stowers, 2013). This may also apply to open data. Open data visualisation, such as graphs, may be used to discover links between resources and identify interesting new information (Dadzie & Rowe, 2011). Open data visualisations may facilitate the processes in which non-expert users discover and analyse data, find links between them and obtain insights (Dimou et al., 2014). Open data visualisations may reduce information overload. O'Hara (2012) and Alani et al. (2008) specifically point at the importance of maps for making sense of data. However, the literature also shows that OGD visualisation functionalities are

barely provided to OGD users by existing OGD portals (Liu, Bouali, & Venturini, 2014; Sayogo, Pardo, & Cook, 2014). The literature has mainly described single tools and services for OGD use (e.g., Volpi, Ingrosso, Pazzola, Opromolla, & Medaglia, 2014) instead of a set of visualisation tools integrated in OGD infrastructures.

Clusters of factors influencing OGD visualisation	Cluster description
Data visualisation support	While visualisation support is important for obtaining insight in and understanding OGD, visualisation support is often not integrated in OGD infrastructures.

Table 3-4: Clusters of factors influencing OGD visualisation.

3.3.4 Factors influencing interaction about OGD

Table 3-5 depicts the clusters of factors which influence interaction about OGD, namely 'lack of interaction' and 'interaction support and tools'. The first cluster refers to the *lack of interaction* regarding OGD use. In one respect the literature shows that data providers can use information about OGD use to make more informed future investment decisions concerning the supply of open data (Davies, 2010). Participation (as a form of interaction) may take place, for example, by allowing citizens to contribute to discussions on how to better address their needs (Kassen, 2013). However, access on itself is not enough to generate active participation (Alani et al., 2008). The literature shows that interaction related to OGD use is limited, for instance because conversations about released data are lacking (Lee & Kwak, 2012) and because many OGD providers do not know who their external users are (Archer, Dekkers, Goedertier, & Loutas, 2013).

Clusters of factors influencing interaction about OGD	Cluster description
Lack of interaction	Interaction regarding OGD use is limited, for instance because conversations about released data are lacking.
Interaction support and tools	Interaction support and tools influence the extent to which individuals can interact about OGD publication and use, yet there is a lack of interaction support and tools at OGD infrastructures.

Table 3-5: Clusters of factors influencing interaction about OGD.

The second cluster of influencing factors concerns *interaction support and tools*. The type of participatory tools that users are provided with influence the extent to which open data users can interact regarding OGD publication and use. For instance, social media technologies allow for access to and interaction with government operations, programs and data (Bertot et al., 2012). Social media can be used to stimulate participation (Veljković, Bogdanović-Dinić, & Stoimenov, 2014) and interaction (Mora Segura et al., 2014), and to engage people in open data (Garbett, Linehan, Kirman, Wardman, & Lawson, 2011). Moreover, interactive communications (blogging, micro blogging, tagging, photo and video sharing) may be used for this purpose, and feedback from users may be used for updating resources (Lee & Kwak, 2012). Yet, interaction support is often lacking for OGD. Most governmental agencies do not offer feedback mechanisms for open data (Alexopoulos, Spiliotopoulou, et al., 2013; Archer et al., 2013). In addition, most open data infrastructures traditionally do not facilitate the improvement of opened data (e.g. through cleaning and processing) (Alexopoulos, Spiliotopoulou, et al., 2013). Lee and Kwak (2012) and Whitmore (2014) posit that the delivery of open data is characterised by a lack of opportunity for public participation.

3.3.5 Factors influencing OGD quality analysis

Three clusters of factors affecting OGD quality analysis were identified, including 'dependence on the quality of open data', 'poor data quality', and 'quality variation and changes'. First, there is strong *dependence on the quality of open data* for successful open data use (Behkamal et al., 2014). Data quality plays an essential role in the use of government portals (Detlor, Hupfer, Ruhi, & Zhao, 2013), and a certain level of data quality is essential for OGD use (O'Hara, 2012). To be able to assess the quality of datasets in general, data users need to have information about the nature of the data (Dawes & Helbig, 2010).

Clusters of factors influencing OGD quality analysis	Cluster description
Dependence on open data quality	Successful OGD use strongly depends on the quality of the data.
Poor data quality	Open datasets may suffer from poor data quality.
Quality variation and changes	The quality of data varies (e.g. per source, after reuse and over time).

Table 3-6: Clusters of factors influencing OGD quality analysis.

Second, the *poor quality* of open data can be a major issue (Karr, 2008; Whitmore, 2014). Users may be unrealistically optimistic about the quality of government data, believing that such data are, for example, objective and neutral (Dawes, 2010; Radin, 2006). On the other hand, users may also be concerned about the quality of open data (Martin, 2014). Kuk and Davies (2011) state that open data often suffer from poor quality, such as inconsistency in terms used in datasets and a lack of granularity. It is, however, difficult to measure the quality of the data.

Finally, the cluster of *quality variation and changes* encompasses differences in open data quality over time, reuse and over sources from where they were obtained. In general, the quality of data that flow through different information systems can quickly degrade over time without control of the processes and information input (Batini, Cappiello, Francalanci, & Maurino, 2009). Open data specifically may be reused over time, which can easily affect the quality of the data (Oviedo, Mazon, & Zubcoff, 2013). The quality of data on the web in general varies widely (Auer et al., 2013). This may also be the case for open data, since open datasets are produced by many different organisations. The literature shows that the quality of open data varies, for example, per country and per data provider (Petychakis et al., 2014).

3.4 Summary: overview of factors and answer to the first research question

The first chapter of this dissertation started with a high-level description of open data dependencies. This third chapter provided more detailed insight in open data dependencies, and the first research question was answered: *which factors influence open government data use?* Based on the literature, the main activities of OGD use were divided into five categories: searching for and finding OGD, analysing OGD, visualising OGD, interaction about OGD and OGD quality analysis. Figure 3-2 extends figures 1-1 and 1-2 regarding open data actors and dependencies, and shows how the five identified OGD use activities are related to OGD use.

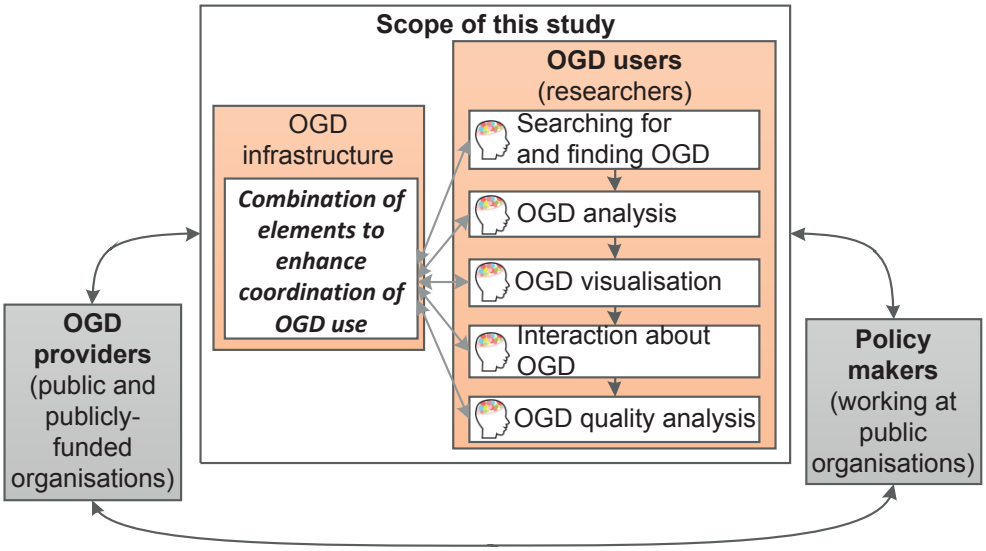


Figure 3-2: Open data dependencies related to the five identified OGD use categories.

The distinction of OGD use categories was used to identify factors influencing OGD use. From the literature it can be concluded that each OGD use activity was influenced by various factors. A comprehensive overview of influencing factors is provided in Appendix A, while Table 3-7 summarises the clustered factors described in this chapter. The table shows that most factors influencing OGD use are both technical and social. For example, data quality refers to the technical parameters of data quality, yet data quality also influences the use of the data and may complicate its interpretation, which can be considered a social aspect.

OGD use category	Clusters of factors influencing OGD use (RQ1)
Searching for and finding OGD data	Data fragmentation
	Terminology heterogeneity
	Search support
	Information overload
OGD analysis	Data context
	Data interpretation support
	Data heterogeneity
	Data analysis support
OGD visualisation	Data visualisation support
Interaction about OGD	Lack of interaction
	Interaction support and tools
OGD quality analysis	Dependence on open data quality
	Poor data quality
	Quality variation and changes

Table 3-7: Overview of cluster of factors influencing OGD use as identified through the literature review.

The third goal of the literature review was to create a framework that can be used to identify functional requirements for the OGD infrastructure. In the following chapter, the clusters of factors influencing OGD use will be used for this purpose.

4. Case study analysis

This chapter describes the second phase of this research; the definition of objectives of a solution. It aims to answer the second research question: *What are the functional requirements for an infrastructure that enhances the coordination of open government data use?* Case studies are used as the main research instrument to gather functional requirements. The case study data collection and analysis were guided by the framework of factors which influence OGD use as derived from the literature in chapter three. This chapter starts with an overview of the approach of the case studies, followed by the case study descriptions. Subsequently, the requirements for the OGD infrastructure are derived from the cases. The requirements then provide the foundation for the design of the OGD infrastructure in the third research phase. Parts of this chapter have been published in Zuiderwijk, Janssen, Choenni, and Meijer (2014), Zuiderwijk, Janssen, Meijer, et al. (2012), Zuiderwijk and Janssen (2013b) and Zuiderwijk and Janssen (2014b).

4.1 Case study approach

This section discusses the case study approach (see Figure 4-1). We start with a discussion on the relevance and the applicability of case study research for this study, followed by a discussion of criticisms on case study research. This will show the advantages and disadvantages of case study research, and will provide the argumentation for using case studies to elicit OGD infrastructure requirements. Thereafter, typical topics relevant for case study research are described, including the case study selection procedure. The selection procedure will show that the cases are selected based on theoretical sampling, a multiple-case design and six selection criteria. The two selected cases concern open judicial data use and open social data use. An overview of the cases is provided. Then the developed case study protocol is outlined, encompassing the field procedures and interview topics, and a guide for the case study report. Subsequently, the five information sources for the case studies are discussed, namely documents, archival records, open and semi-structured interviews, direct and participant observations, and dataset

analysis. Sections 4.2 and 4.3 then describe the cases and the functional requirements for the OGD infrastructure. Finally, this chapter provides an overview of the requirements and the answer to the second research question.

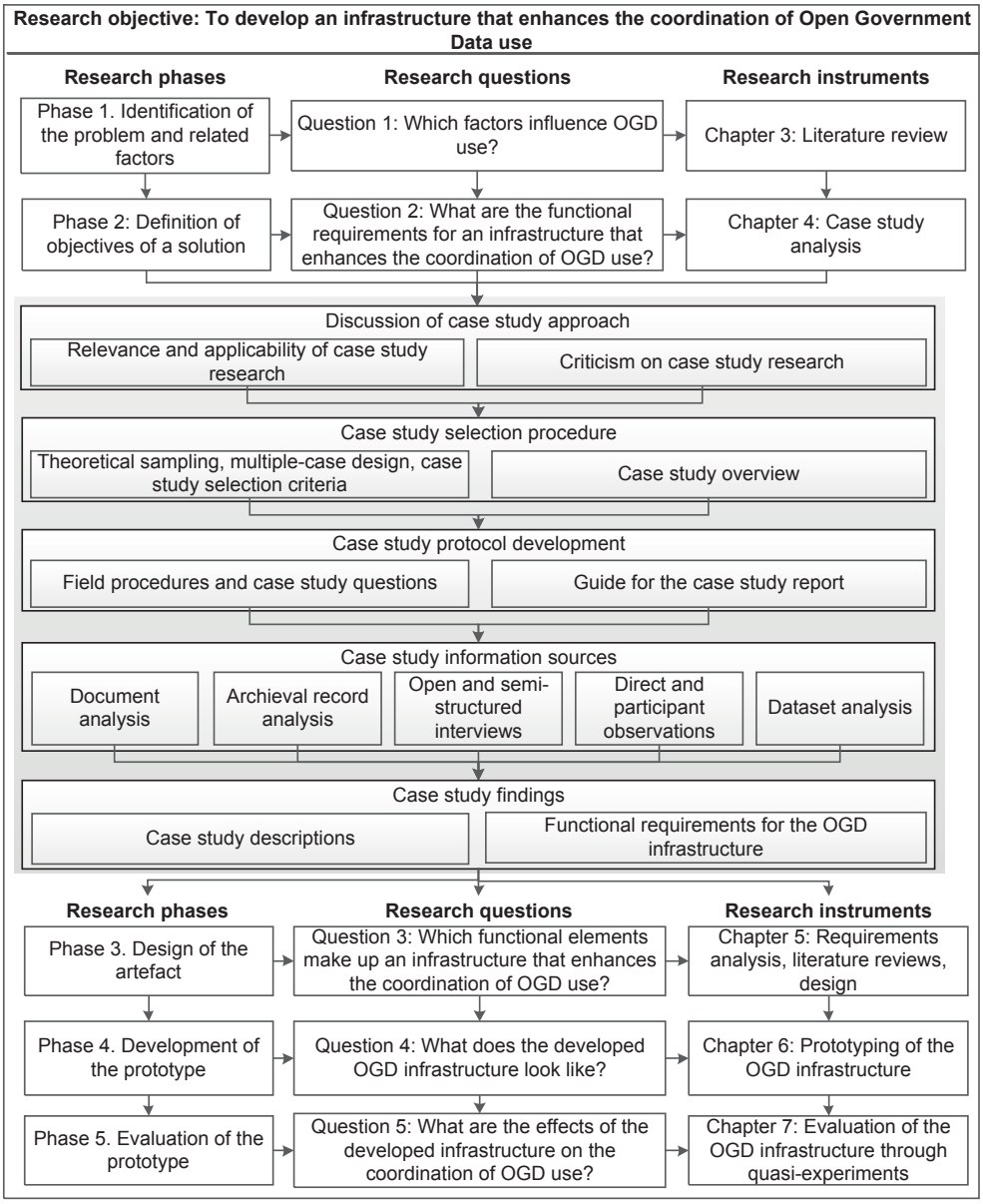


Figure 4-1: Research design including the case study approach.

4.1.1 Relevance and applicability of case study research

The relevance and applicability of a research design depends on the type of research question that is posited. Yin (2003) states that case studies can be used to answer questions that handle operational links rather than frequencies or incidence. A case study can be defined as "an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident" (Yin, 2003, p. 13). Case studies can be used to examine the dynamics of single settings (Eisenhardt, 1989) in their natural environment (Benbasat et al., 1987). They are appropriate for investigating a set of broad and complex real-life contemporary events which require a holistic and in-depth examination (Dubé & Paré, 2003; Yin, 2003), and for phenomena that do not allow for studying them outside the context in which they take place (Dubé & Paré, 2003). Case studies are therefore very valuable to explorative research in order to investigate contextual factors over which the researcher has no or little control (Yin, 2003). The holistic perspective facilitates the research of "complex and ubiquitous interactions among organisations, technologies, and people", and is therefore particularly suited for studying Information Technology (IT) (Dubé & Paré, 2003, p. 598). Moreover, case study research is useful for problems in which actor experiences and the context of action play a crucial role (Benbasat et al., 1987; Bonoma, 1983), as well as for problems in which research and theory are at an early and developing stage (Benbasat et al., 1987; Roethlisberger, 1977). It has been postulated that interpretive case studies can contribute to IS theory and practice (Walsham, 1995). Case research can be used with any philosophical perspective (e.g. positivist or interpretivist) (Dubé & Paré, 2003), and is by far the favourite research method for e-government research (Yildiz, 2007).

Various case study characteristics make case study research appropriate for this study. First, since this chapter examines functional requirements for an OGD infrastructure, it is focused on relations between objects and activities rather than on frequencies. For example, it considers the relation between barriers for OGD use and their influence on OGD use activities. Second, the boundaries between OGD use and its context are not clearly evident, and the literature shows that OGD use consists of complex and dynamic activities for which the actors

depend on each other (see chapter 1 and 3). It is therefore important to take a holistic perspective and to study OGD use in the context in which it takes place. Third, the experiences of actors are important in OGD use, since they are the ones who contribute to realising benefits with OGD. OGD actors need to be taken into account when one wants to improve the status quo of OGD use, as can be done through case studies. Fourth, research and theory in the field of open data are at an early and developing stage (see chapter 1), and case study research can be used to contribute to open data theory generation. For these reasons case studies are carried out to acquire functional requirements for the OGD infrastructure.

4.1.2 Criticism on case study research

Even though case study research is relevant and appropriate for the objectives of this research, there is also criticism on the case study approach. Yin (2003) argues that case studies can be criticised for a lack of rigor, their long duration, and a lack of basis for scientific generalisation of case study research. Yin (2003) argues that the criticism is misdirected, because it is mainly due to not having followed systematic procedures, not having followed specific methodological procedures, generalising in an inappropriate way (to populations or universes rather than to theoretical propositions) and due to incorrect examples of the application of case studies that are available in literature.

The quality of the design of case study research is affected by four conditions, namely construct validity, external validity, reliability and internal validity (Yin, 2003). First, *construct validity* refers to the establishment of correct operational measures for the investigated constructs (idem). Case study research has been criticised for failing to create a sufficiently operational set of measures and for the subjectivity of case study data collection (idem). Since the study of cases in real-world settings does not allow for laboratory controls or experiments, this raises the question how controlled observations can be made (Lee, 1989). Lee (1989) proposes to respond to this problem by incorporating natural controls in case studies. He suggests holding most variables in case studies constant, while only varying with the essential variables that one wants to examine in the case study. Moreover, Yin (2003) suggests to use multiple sources of evidence, to

generate a chain of evidence, and to let key stakeholders review the case study report.

In this study *construct validity* is optimised by investigating constructs in the cases based on the operational measures that were described in chapter three. In addition, the framework of factors influencing OGD use provides guidance for the study of infrastructure requirements. With regard to construct validity we follow Lee's (1989) suggestion to hold most variables in case studies constant, while only varying with a limited number of variables. A sub section of section 4.1.3 outlines which variables in the cases were similar and which varied. In addition, multiple sources of evidence were used to examine the cases (see section 4.1.5) to establish a chain of evidence. Finally, key stakeholders reviewed the case study report and were given several opportunities to provide feedback.

Second, *external validity* refers to the establishment of the domain to which the findings of the research can be generalised (Yin, 2003). External validity is concerned with the generalisation of the case study findings beyond the single case. The problem of generalizability refers to the incapability of case studies to provide generalizable conclusions (Dubé & Paré, 2003). As a response to this problem, Lee (1989) posits that additional case studies should be carried out to test whether theories are confirmed in the circumstances of other cases. For case studies as well as for various other methods the generalizability of theories should be tested through and confirmed in various situations (Lee, 1989). This means that sufficient information needs to be provided to allow for the replication of the case studies. This research addresses external validity by studying multiple cases instead of one, so that the findings from one case study can be examined in the context of the other case. External validity is also enhanced by providing sufficient information about the design of the case study to allow for their replication. This makes it possible to carry out additional case studies to test the findings from this research and to investigate to which extent they can be generalised. Nevertheless, even though a multiple case study approach is used, it is important to note that only two cases are studied. The cases focus on a specific type of open data in a particular context. The functional infrastructure requirements that will be elicited are important in the context of these cases, yet the findings from the case studies may not be generalizable beyond this context.

Third, *reliability* is concerned with demonstrating that the repetition of the operations of the study (e.g. data collection procedures) is possible and that it would provide the same results (Yin, 2003). This means that if a case study methodology is applied to a case study multiple times, this should result in similar conclusions (idem). Lee (1989) notes that although it may not be possible to replicate the observations of a certain case study, it is possible to test the same theory in a different set of initial conditions. In this way the findings from case studies can still be replicated. The optimisation of case study reliability requires the documentation of the procedures followed for the research in a case study protocol (Yin, 2003). In this study the reliability is optimised by clearly defining the case study methodology and by developing a case study protocol. Insight is provided in which data and documentation are gathered that lead to the findings of the case studies.

Finally, *internal validity* refers to the establishment of a causal relationship, showing that certain conditions lead to other conditions (Yin, 2003). Internal validity is important for explanatory or causal research, since it focuses on whether a particular event leads to another event. However, internal validity is of less concern for descriptive and exploratory research. Since our case study research is not concerned with inferring about causal relationships, this type of validity is not discussed here.

4.1.3 Case study selection

An important aspect of case study research is the selection of the cases, as the selection of cases helps to determine the limits for generalising the research findings (Eisenhardt, 1989). This section discusses the selection of the case studies, including the selection criteria, and an overview and comparison of the cases.

Selection criteria

Cases can be selected in two ways, namely by statistical sampling or by theoretical sampling (Eisenhardt, 1989). Statistical sampling refers to acquiring accurate statistical evidence on the distribution of variables within a specific population and selecting cases based on this information. Theoretical sampling refers to the situation in which cases are chosen because they are expected to replicate

previous cases, extend emergent theory, fill theoretical categories, or provide examples (idem). Since this research aims to contribute to theory building in the field of open data rather than to test theories in this field, the selection of cases for this research was based on theoretical sampling. Open data is a relatively new field in which existing theory often seems inadequate, which shows the need for theory building and which makes theoretical sampling an appropriate approach for the case study selection in this research.

Even though there may be good reasons to perform a single case study, such as the unique or extreme circumstances that the case represents, multiple-case designs are preferred over single-case designs (Yin, 2003). Investigating multiple cases provides more compelling evidence, as the analytic conclusions which arise independently from multiple cases are more powerful than when they come from only one case. Furthermore, if common conclusions can be derived from multiple cases, which usually have different contexts, this expands the external generalizability of the research findings compared to a single case study (Yin, 2003). This research therefore opts for studying multiple cases.

Selection criteria were produced to define explicitly which characteristics the cases needed to have. The following six criteria were defined.

1. *The cases involve open data provided to the public by Dutch governmental organisations.* Our definition of OGD in chapter three showed this study's focus on governmental data. Although open data can be used world-wide as the internet is not hindered by country borders, we focus on open data which are produced and published by organisations in one country within an open data infrastructure in that same country. This is done to keep cultural influences on OGD use as equal as possible, since cultural influence on OGD use is not the main topic under investigation in this study. For practical reasons, the Netherlands was selected as the country where the case studies would be conducted. The researcher who carried out this study was based in the Netherlands and spoke the language, which supported easy access to information from the cases. Additionally, the cases involve data which are made available to the public by research organisations that are part of Dutch ministries. We selected research organisations that are part of ministries because these organisations are

expected to produce many datasets as input for governmental policy making. Moreover, because of the obligation for various Dutch governmental organisations to open their data resulting from policy-oriented research, these organisations were expected to be releasing governmental data already for a number of years.

2. *The cases involve OGD which have already been made available to researchers outside the government for at least several years.* This criterion is in line with our definition of OGD use in section 1.2, since it considers the use of open data by researchers outside the government. This criterion was expected to result in cases in which OGD use has already been established, and a relatively mature OGD infrastructure had been developed with limited growing pains. Furthermore, this second criterion was expected to lead to the involvement of organisations with considerable experience with how their data can (potentially) be reused.
3. *The cases involve open data provided through an OGD infrastructure that meets the non-functional requirements.* The cases need to involve open data provided through an OGD infrastructure, since this research is focused on the development of such an infrastructure. Moreover, since this study does not focus on the non-functional requirements, and a premise of this study is that the non-functional requirements are met, the selected cases should provide OGD through an infrastructure that meets the non-functional requirements.
4. *The cases involve different types of structured OGD from the domains of social sciences and humanities.* This criterion corresponds to our definition of OGD use in section 1.2, since it considers the use of structured OGD from the domains of social sciences and humanities. One heterogeneous variable in the cases concerned the type of OGD involved in the cases. The reason for selecting cases which involve different types of OGD was that this allows for investigating whether the results from one case are also applicable in another context. If common conclusions can be derived from different cases, which usually have different contexts, this expands the external generalizability of the research findings compared to a single case study (Yin, 2003). The first case was focused on open judicial data,

whereas open social data was central to the second case. The cases focus on judicial data and social data because the selected research organisations involved in the cases (the WODC and SCP) already disclose these types of data through an infrastructure. In addition, social and judicial data are important for identifying and solving various societal issues. For example, judicial data can be used to obtain insight in the number of crimes committed in a certain area, for estimations of the future required capacity of prisons, and the effectiveness of anti-recidivism programmes for convicted individuals. Social data may be used to obtain insight in poverty and social exclusion among certain groups in the population, and issues related to (un)employment, education, social security, integration and immigration.

5. *The cases allow for investigating functional requirements for an OGD infrastructure which aims at enhancing the coordination of OGD use (i.e. searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis).* Since we aim to investigate functional requirements in a structured manner, the cases should allow for examining functional requirements related to the five types of OGD use that we identified in the literature in chapter three.
6. *The cases involve organisations and people who are willing and ready to cooperate in the research and to share information that is required to conduct this research.* This sixth criterion mainly concerns the willingness and readiness of the data publishers involved in the cases, and of the employees of the organisation that maintained the infrastructure on which the data were published.

We refer to the first case study as the open judicial data use study, whereas the second case study is referred to as the open social data use study. The case studies will be explained further in the following sections.

Overview of the cases

The case study design needs to clearly define the unit of analysis of the case (Dubé & Paré, 2003; Yin, 2003). In this section an overview is given of the two case studies, and their key characteristics are compared.

Case 1: Open judicial data use

The first case study encompasses the use of open judicial data. Figure 4-2 depicts the boundaries of the case study and clarifies the unit of analysis. The figure shows that a holistic view is adopted that goes beyond the limits of individual organisations, a single infrastructure or a single OGD user group, since functional requirements for the OGD infrastructure may come from the complex interaction between these units. The boundaries of the first case study are determined by the publication of open judicial data by a particular organisation, the infrastructures that these data are published on, the organisations that provide the infrastructures and the usage of these particular open judicial data.

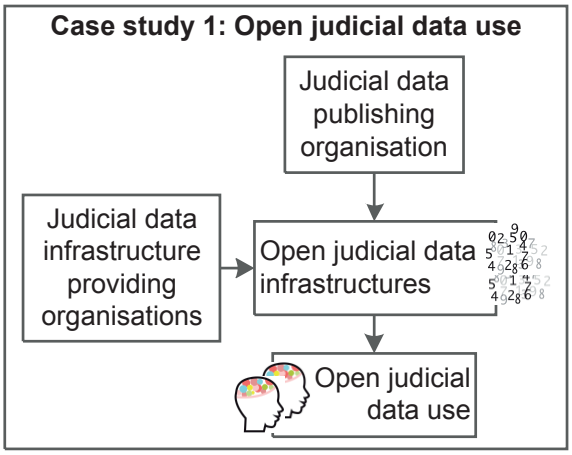


Figure 4-2: Unit of analysis of the first case study.

The governmental provider of the judicial data selected for the first case study is the Research and Documentation Centre (In Dutch: *Wetenschappelijk Onderzoek- en Documentatiecentrum*; WODC). The WODC is a semi-independent criminal justice knowledge centre, which is part of the secretary general cluster of the Dutch Ministry of Security and Justice (Staatscourant, 2011, nr. 22848). The WODC started in 1949 when the Study and Documentation Centre (Studie- en Documentatiecentrum) was founded, and exists in its current form since 1975. The WODC operates in a field located between policies, science and politics (Visitatiecommissie WODC, 2014). It aims “to be a leading scientific research and knowledge centre for the broad field of Security and Justice” (Wetenschappelijk Onderzoek en Documentatie Centrum, 2014, http://www.wodc.nl/organisatie/). The

research conducted by the WODC is of high quality and can be used for policy making regarding justice, safety and security (Visitatiecommissie WODC, 2014). In January 2013, the scale of the WODC was 92,6 fulltime employees and 105 employees in total. Approximately 25 per cent of the research is performed by the employees of the WODC and about 75 per cent is subcontracted to universities and commercial research organisations (Wetenschappelijk Onderzoek en Documentatie Centrum, 2013). The WODC has the following tasks:

- Conducting and commissioning research by other organisations, including the evaluation of policies and policy programs. This task concerns planning, execution and production of research by internal research and external research institutions, combining knowledge and subsiding external initiatives that enable the production or dissemination of knowledge;
- Advising about intended policies and policy programs. This task refers to advising about policy relevant cases from a research perspective, especially advising about research and scientific support for answering policy questions, and questions from the Second Chamber;
- Developing and maintaining data and making data accessible;
- Disseminating knowledge that is available within the WODC. This includes bringing together knowledge, knowledge storage and knowledge building, distributing knowledge in the form of data and documentary information, (scientific) publications, presentations and participations in all kinds of national and international initiatives;
- Documenting scientific publications in the field of Security and Justice (Staatscourant, 2011, nr. 22848; Wetenschappelijk Onderzoek en Documentatie Centrum, 2013, 2014).

The WODC has been publishing judicial data to stimulate external and internal transparency, to disseminate knowledge, to let people reuse the judicial data that were collected for other purposes, and possibly to reduce the workload of the organisation. External transparency refers to accountability, showing to individuals outside the organisation that they can acquire information about what kind of research is conducted by the WODC, and that the organisation is open and transparent. Internal transparency relates to informing employees within the

organisation about which research other departments conduct, as well as managing data storage and preservation, and to avoid fraudulent data usage and storage. Knowledge dissemination is one of the key tasks of the WODC, and might be stimulated through judicial data publication. Data reuse refers to the reuse of data which have already been collected by the WODC for one project. Data that are gathered by the WODC are usually only used for one project. After the project ends, the data are often not reused by WODC-researchers, but they might be reused by individuals outside the organisation. The possible decrease of workload may be realised by replacing the many individual data requests that the WODC currently receives with the single publication of the datasets as open data. Instead of putting effort, time and costs in complying with various individual requests for a certain dataset, the WODC could publish the dataset once and refer to it in case that more individual requests for this dataset will be made. On the other hand, the publication of the data may also lead to more information requests, which may increase the workload.

The WODC has been publishing judicial data since 1975. The WODC never aimed to provide and maintain an infrastructure for the publication and use of its judicial data by itself, since various organisations already developed initiatives in this regard. Data have always been uploaded to the open data infrastructures that were maintained by other organisations. Firstly, data were offered at the social science Steinmetz institution (Steinmetz Stichting). Steinmetz and various other data archiving organisations merged in 2009 and together founded the Data Archiving and Networked Services (DANS) (Wittenberg, 2009). DANS is an organisation of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO) and provides access to thousands of Dutch digital research datasets, e-publications and other research information. DANS also gives training and advice and performs research about sustained access to data (DANS, 2013). DANS provides access to judicial data through the Electronic Archiving SYstem (EASY). EASY provides access to datasets from various disciplines, including the social sciences, humanities and other disciplines. EASY can also be used to publish research data by the governmental agencies themselves (Data Archiving and Networked Services, 2014b). The data publishers decide whether data users will have open access,

restricted access or no access to the data that they upload. Metadata that are stored at DANS are also stored at the portal of the National Academic Research and Collaborations Information System (NARCIS). NARCIS contains information about researchers, (open access) publications, current and completed research projects in the Netherlands, and other work of Dutch researchers.

The DANS infrastructure has not been created as part of this study, but is already used by many Dutch organisations to make their data publicly available for many years. The DANS infrastructure is well-accepted for publishing data from social sciences, humanities, behavioural sciences and geospatial sciences. In addition, the Dutch government has created general conditions for contracting orders to conduct services (Algemene Rijksvoorwaarden voor het verstrekken van Opdrachten tot het verrichten van Diensten; ARVODI), which obliges organisations to publish data derived from policy focused research that has been conducted by order of the Dutch federal government and that provides data that are appropriate for reuse at the DANS infrastructure (Data Archiving and Networked Services et al., 2008; Ministry of the Interior and Kingdom Relations, 2008). The DANS infrastructure meets the non-functional requirements of accessibility, availability, sustainability, maintainability, usability, and compliance with legislation (e.g., privacy, security).

Finally, the infrastructure offered by DANS allows for connecting the WODC to the users of its data. Since DANS maintains this infrastructure, DANS is expected to have information about the ways that governmental data are used by the public and about the functional user requirements for OGD infrastructures. In sum, the use of open judicial data is studied by adopting a holistic perspective on judicial data collecting and publishing organisations, open judicial data infrastructures, organisations which provide and maintain these infrastructures, and the use of the open judicial data through these infrastructures.

Case 2: Open social data use

The second case study focuses on the use of open social data. Figure 4-3 depicts the boundaries of the case study and defines the unit of analysis. Equal to Figure 4-2, this figure shows that the second case study goes beyond the limits of single organisations, infrastructures or groups of people. The units of analysis for the

second case study are similar to those of the first case study. The boundaries of the second case study are determined by the publication of open social data by a particular organisation, the infrastructures on which these data are published, the organisations which provide the infrastructures and the use of these particular open social data.

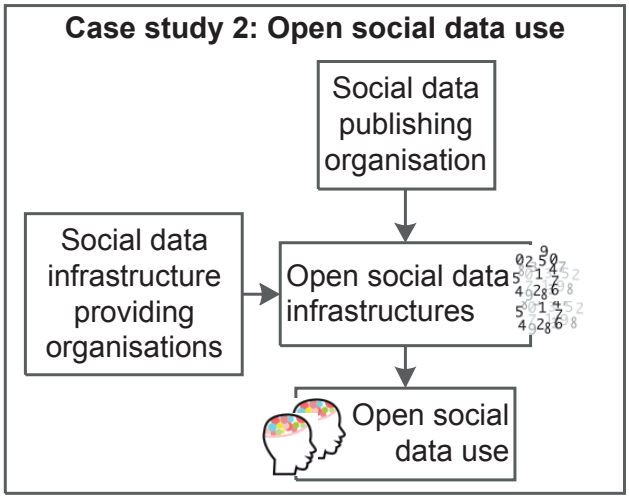


Figure 4-3: Unit of analysis of the second case study.

The data publishing organisation involved in the second case was The Netherlands Institute for Social Research (In Dutch: *Sociaal en Cultureel Planbureau*; SCP), which is part of the Dutch Ministry of Health, Welfare and Sport. The Netherlands Institute for Social Research (Sociaal en Cultureel Planbureau; SCP) has been founded by Royal Decree ("Koninklijk Besluit") on 30 March 1973 (Overheid.nl, 2012). The SCP is an interdepartmental scientific institute which performs research "into the social aspects of all areas of government policy" (Sociaal en Cultureel Planbureau, 2013c). Major research topics of the SCP are health, welfare, social security, the labour market and education, with a particular focus on the interfaces between these fields. The SCP examines the developments of governmental policies in relation to the daily life of the Dutch population. In November 2013 the personal occupation of the SCP was 82 full-time employees and 92 persons are employed by the SCP in total. Key tasks of the SCP are as follows.

- "To describe the social and cultural situation in the Netherlands and outline anticipated developments;

- To provide the information needed for a well-considered choice of policy objectives and resources and for the development of alternatives;
- To evaluate government policy, especially interdepartmental policy, for example concerning the elderly, young people, ethnic minorities […]. SCP also publishes reports on several other topics" (Sociaal en Cultureel Planbureau, 2013b, www.scp.nl).

Opening data is not one of the main tasks of the SCP, but it is seen as an additional task. The SCP deposits data for reasons of decreasing its workload, internal and external transparency, and giving data back to the public that they paid for through tax revenues. First, data are made available to the broad public to prevent numerous individual data requests and thus to decrease the workload. Second, as far as internal and external transparency is concerned, the SCP aims to be transparent about the scientific foundations of the results that the SCP concluded upon in its reports and articles by revealing the underlying data. Third, the data that were produced by the SCP have indirectly been financed by citizens via the taxes that they pay. When these data are made available to the public again, citizens receive the results of the financial funding that they contributed to in order to conduct the research. The dissemination of knowledge or the reuse of data by external users are not essential reasons for making SCP-data available.

The SCP has been publishing social data since 1974, although data publication in the pre-internet age was of course different from data publication using the current technological advancements. Just like the WODC, the SCP did not develop its own open data infrastructure, but uploaded data to infrastructures maintained by other organisations. The SCP firstly made social data available to the public via the predecessor of DANS, namely the social science Steinmetz institution (Steinmetz Stichting). Thereafter, most of the data opened by the SCP were made available via DANS through EASY, and its metadata through NARCIS. More information about DANS, EASY and NARCIS can be found in the description of the first case study. Additionally to the data access through EASY, DANS offers access to social data through the NESSTAR software system for data disclosure. The content that DANS offers via NESSTAR has mainly been developed before

DANS was founded, and DANS does not add new datasets to NESSTAR anymore. Yet, NESSTAR can still be used for several social datasets.

Moreover, in the period of 1975 to 2012 the SCP deposited eight datasets to the Time Use Archive. This publication concerns social data from international research about time use (Tijdsbestedingsonderzoek), and the data are offered at the Internal Time Use Archive (Sociaal en Cultureel Planbureau, 2013a)[1]. The Time Use Archive contains data about time use in various European countries. Additionally, the SCP makes these data available via DANS. The Time Use Archive is provided and maintained by the Centre for Time Use Research.

In sum, the use of open social data is studied by adopting a holistic perspective on social data collecting and publishing organisations, open social data infrastructures, organisations which provide and maintain these infrastructures, and the use of the open social data through these infrastructures.

Comparison of the cases

Although the overview of the cases showed that they had several similar characteristics, there were also various differences. It is important to be aware of these similarities and differences, since they may have influenced the case study results. Table 4-1 shows the key characteristics of the two cases. It demonstrates that a first similarity is that both data providing organisations operate as part of a Ministry, yet they are both to a large extent independent of this ministry. At the beginning of each year the responsible minister needs to approve the work programme of the WODC and the SCP. Thereafter, these organisations do not need to give account to the minister anymore. Moreover, both data providing organisations focus on the disclosure of research data, and both organisations reuse data collected by other organisations and are not always the owner of the data. The datasets that the organisations do not own often cannot be disclosed. Furthermore, both organisations mainly collect data on a micro level, such as data about persons and households. Both data providing organisations maintain an embargo period. At the WODC, data were not disclosed within two years after a

[1] The Time Use Archive can be accessed via http://www-2009.timeuse.org/information/studies/data/netherlands-2011-2012.php

report concerning the data had been published, while the SCP maintained a one year embargo period.

A difference between the cases concerns the types of data that they encompassed, namely judicial and social data. In addition, the sensitivity of the data collected by the data providing organisations differed. While the judicial data provider owned and maintained considerable privacy sensitive data, the social data provider mainly collected non-sensitive datasets. Moreover, whereas the SCP conducts research and collects data by order of all Dutch ministries, the WODC mainly works by order of the Ministry of Security and Justice. Next to this, for the SCP all fieldwork (i.e. the actual data collection) is outsourced to external research organisations. The WODC also outsourced a part of its data collection, yet not all of it. Finally, at the judicial data providing agency multiple persons were responsible for data publication, while at the social data providing agency, only one person was held responsible for this. The type and the sensitivity of the data, the authority commissioning the research, the level of outsourcing data collection, and the organisation of data provision by the governmental agencies may have influenced the functional requirements for OGD use. For instance, these differences may have influenced the quality of the data and the extent to which they can be analysed.

	Case study 1: judicial data case	**Case study 2: social data case**
Organisational position	Semi-independent: part of Ministry yet largely independent	Semi-independent: part of Ministry yet largely independent
Data ownership	Many of the collected datasets are not owned by the organisation	Many of the collected datasets are not owned by the organisation
Unit of analysis	Mainly micro data	Mainly micro data
Embargo period	Data are not disclosed within two years after a report about the data has been published	Data are usually disclosed one year after a report about the data has been published
Data type	Judicial research data	Social research data
Data sensitivity	Considerable amount of sensitive data	Relatively limited amount of sensitive data
Organisational research orders	Mainly by order of the Ministry of Security and Justice	By order of various ministries
Data collection procedure	A part of the fieldwork (data collection) is outsourced to external research organisations	All fieldwork (data collection) is outsourced to external research organisations
Data disclosure procedure	Multiple persons in the governmental agency are responsible for data publication	One person in the governmental agency is responsible for data publication

Table 4-1: Key characteristics of the cases.

4.1.4 Case study protocol

Yin (2003) advices to use a protocol to increase the reliability of case studies, especially when multiple-case studies are conducted. In this section we elaborate on the case study protocol by discussing the field procedures and interview topics, as well as a guide for the case study report.

Field procedures and interview topics

One important aspect of field procedures concerns how the researcher analyses and collects the data. Many case studies maintain an overlap between data analysis and collection (Eisenhardt, 1989), which also occurred in our case studies. One advantage of this overlap is that the researcher remains flexible during the data collection. Simultaneously collecting data and analysing them makes it possible to make adjustments during the process on the basis of early analysis, such as adjustments to the focus on specific themes, as they were unexpectedly relevant (Eisenhardt, 1989). For example, based on our case study findings the clusters of factors influencing OGD use derived from the literature may be adapted.

Eisenhardt (1989) shows that the overlap between data analysis and data collection can be stimulated by making field notes. Field notes report on what is happening in the research; they are a running commentary to oneself or to a research team. More specifically they describe what is observed and what is analysed. Subsequently, patterns within individual cases can be analysed by describing the case early in the phase of data collection. Then, a cross-case search for patterns can be done, forcing the researcher to go beyond initial impressions by using structured and diverse lenses on the data. Field notes were used to describe the findings during interviews and observations.

The cases were examined through the eyes of the case study inquirer rather than through the eyes of the actors involved in the case. The case studies started from a broad perspective, and a broad list of topics to discuss in the interviews was created in advance (see Table 4-2). The data collected with the interviews were analysed during the first half of the case study. Based on this analysis, the topics discussed in the case studies were narrowed down, and the case study data collection and analysis were then guided by the framework of factors which influence OGD use as derived from the literature in chapter three. The framework determined the boundaries of the second half of the case studies

Research questions	Topics discussed during the interviews
Initial broad list of discussed topics	- The institutional context of the governmental data publishing agencies (e.g. origin of the organisation, history of data publication, type of data that is collected, the way of publishing data, institutional aspects such as the departments, number of employees) - Information about the context in which the judicial/social data are created and collected - Objectives and benefits of, and reasons for publishing open judicial/social data - Infrastructure for data publication and topics of datasets published - Attention for data publication within the governmental agencies - Perceived reasons for and benefits of using open judicial/social data - Political, economic, social and technical barriers of and challenges for publishing open judicial/social data - Perceived political, economic, social and technical barriers of and challenges for using open judicial/social data - Sensitive data and data which cannot be made available to the public - Processes to publish and use open judicial/social data, actors involved in the processes, internal data storage, data flows, and other interdependencies in the publication and use of open judicial/social data - Mechanisms that are used to conduct the processes of data publication and use, management of processes and interdependencies - Organisational open data policy and regulations - Lessons that can be learned from and recommendations for making judicial/social data available to the public
Narrowed down list of discussed topics	- The terminology and vocabularies used to describe the data - The data to describe the collected and published crime/social data - Support for finding data that users are searching for - The contextual and other information provided to potential data users - Provision of data that describe the datasets - The interpretation of the opened judicial/social datasets by data users - The type of support offered for the interpretation and analysis of opened data - A helpdesk or other service to support open judicial/social data use - Differences and similarities between disclosed datasets - Tools offered by the infrastructure to support OGD use (e.g. visualisation, analysis) - The extent to which the published judicial/social data can be used for statistical analyses, visualisations, linking data, and policy making (e.g. data quality) - The way that the published data are reused (e.g. number of published datasets, number of downloads, feedback from users, specific types of reuse, and other information about the use of judicial/social data) - Insight of the data provider in the data reuse - Conditions and licenses for data reuse - Contact, meetings and discussions between the data provider and user and tools/mechanisms to support this - (Potential) collaborations and interactions between data publishers and users - Lessons that can be learned from and recommendations for using judicial/social data

Table 4-2: Overview of the topics discussed during the interviews.

and prescribed the case study aspects that needed to be investigated. Finally, the results of the case studies as well as different versions of the design of the OGD infrastructure (see chapter 5) were presented to a number of case study participants, who were provided with the opportunity to give feedback on the findings. The requirements from the cases were elicited through iterative processes, which means that the above-mentioned steps were repeated many times. After the case study participants provided feedback, another round of data collection and analysis took place.

Both case studies started with research at the organisations that produced and published judicial and social data. This was done because OGD use is highly interdependent with OGD publication. Chapters one and three showed that the way that OGD are published influences to which extent the data can be used. Thus, to obtain insight in OGD use, one should also investigate the way that the OGD are opened to the public. Second, starting the case studies from the perspective of OGD users would be complicated, since there is no central register of OGD users, and it is complicated to find large homogeneous groups of individuals who use the same type of OGD. It was expected that starting from the perspective of OGD providers could help to obtain easy access to information sources regarding OGD use, and that they may provide insight in how their data were being reused by the public. Third, it was expected that starting the research with examining the data provider would provide a more systematic approach to investigate OGD use than starting from the perspective of the OGD user. Involving judicial and social data providers from specific governmental organisations allowed for clearly defining the case study boundaries. In this way, we could systematically investigate factors which influenced OGD use in the particular context of the judicial and social data. Therefore, starting the case studies with searching for information at the OGD publishing organisations was found to be a feasible approach which provided considerable information relevant for studying OGD use.

Guide for the case study report

The case study report was guided by the framework of factors which influence OGD use that was developed in chapter three. The framework demonstrates which OGD use activities need to be investigated (i.e. searching for and finding OGD,

OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis), and it reveals which clusters of factors need to be examined in the cases, such as data fragmentation, interaction support and tools, and the context of the data. For each of the cases the clusters of factors from chapter three were investigated. Building on the clusters of factors identified in the literature the requirements for the OGD infrastructure were elicited.

4.1.5 Case study information sources

Case studies typically combine various data collection methods (e.g. archives, interviews, questionnaires and observations) (Eisenhardt, 1989). Multiple sources are expected to provide more comprehensive results than a single source, and may help to maximise construct validity and reliability (Yin, 2003). The data collected from multiple sources of evidence is then often integrated in a triangulating fashion (idem). Yin (2003) states that the main sources of evidence that are used in case studies are documents, cultural and physical artefacts, direct observations of the events being studied and interviews with the persons involved in the events. In this research five sources of evidence were combined for each of the case studies (see Table 4-3).

Case study	Documents studied	Archival records studied	Open and semi-structured interviews	Direct and participant observations	Datasets analysed
1. Judicial data	21	21	24 (with 12 persons)	During 27 meetings	45
2. Social data	36	11	9 (with 8 persons)	During 11 meetings	7
Total	47*	27*	32*	37*	52

Table 4-3: Overview of the case studies that were performed for this research and the information sources that were used in the case studies (**Note that various information sources were used for both case studies (e.g. documents about the DANS infrastructure))*.

The following sources of information were used to conduct the case studies.

- Documents. The desk research encompassed the study of policy documents, public and non-public governmental reports, research reports, contracts, websites, scientific articles and other documents. Appendix B offers an overview of the studied documents.

- Archival records. These records included internal e-mails and PowerPoint presentations regarding OGD, internal PDF files, texts on the organisations' intranet, overviews of datasets which would and would not be appropriate for publication as open data, overviews of how many published datasets were downloaded and how they were used, as well as minutes of meetings concerning OGD that had taken place in the past at the data providing organisations or the infrastructure maintaining organisation.
- Open and semi-structured interviews. Interviews were conducted with key persons who had considerable knowledge of the publication and use of open judicial and social data (see Table 4-4). Although no actual open judicial and social data users were interviewed, interviews took place with individuals who were concerned with the management of open data use and who were regularly in contact with open data users. In line with our definition of OGD use in section 1.2, the interviews focused on the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government. The interviewees worked at the data publishing organisations and the organisation that maintained the OGD infrastructure, and most interviewees had more than ten years of experience with data publication and/or use. Within each organisation, multiple individuals were interviewed to avoid the bias of reporting findings based on the opinion of a single person. For each case study persons with similar types of functions were interviewed, although no other researchers were interviewed for the first case study, whereas two interviews with other researchers took place for the second case study (see Table 4-4). These people were interviewed because of their expertise regarding OGD.
- Direct observations and participant observations. Activities of the open data process and stakeholders that were involved in the open data process were observed. This information source comprised observations which took place during discussions of teams and working groups, and of

conversations between and with stakeholders within the organisations. Notes that were made during these discussions were studied.

- Datasets. Datasets were studied to investigate factors influencing whether data would become available as open data to the public and to which extent these datasets would then be appropriate for reuse by the public.

The following sections present the results obtained with these information sources.

Organisation	Function	Judicial data case	Social data case
Functions of interviewed persons working at the data publishing organisations (WODC and SCP)	Data disclosure coordinator	Interviewee 1	Interviewee 13
	Organisational managers	Interviewees 2 and 3	Interviewee 14
	Metadata expert	Interviewee 4	Interviewee 15
	Person(s) responsible for collecting the potentially publishable datasets	Interviewees 5, 6, 7 and 8	Interviewee 13
	Person(s) responsible for preparing datasets for publication	Interviewees 8 and 9	Interviewee 13
	Other researchers	-	Interviewees 16 and 17
Functions of interviewed persons working at the data infrastructure providing organisation (DANS)	Person responsible for data user support and data management	Interviewee 10	Interviewee 10
	Project manager	Interviewee 11	Interviewee 11
	Metadata expert	Interviewees 11 and 12	Interviewees 11 and 12
	Long term data preservation expert	Interviewee 12	Interviewee 12

Table 4-4: Overview of persons interviewed for the case studies.

4.2 Case study descriptions

This section describes the findings from the case studies. The cases are described for each of the clusters of factors influencing OGD use as identified in chapter three, since these clusters guide the directions in which we can search for OGD infrastructure requirements. The functional requirements for the OGD infrastructure will be based on these case study descriptions and will be provided in section 4.3.

4.2.1 Searching for and finding OGD in the two case studies

Factors influencing searching for and finding OGD can be divided into the clusters of data fragmentation, terminology heterogeneity, search support, and information overload (see section 3.3.1). Case study findings are described for each of these clusters.

Data fragmentation

The framework for studying the cases derived from chapter three shows that data fragmentation refers to difficulties for open data users to locate the datasets that they want to use, since data can be offered at many different places. In both case studies it was found that the EASY online archiving system of the DANS infrastructure was the key location where the investigated governmental organisations published their judicial and social data. To quote from the data disclosure coordinator in the social data case: *"DANS should be seen as the most important archive to deposit SCP-data"*. The importance of DANS was confirmed by the judicial data case, where the data disclosure coordinator expressed that at the time that the governmental organisation started publishing data, *"DANS was the national organisation for archiving research files"*. Both the SCP and the WODC published data at DANS and its predecessors for many years. A few datasets were also supplied to the public via other infrastructures, but this was exceptional. For example, an interviewee in the social data case mentioned that "*a few times time use data was released to the Time Use Archive, but this occurred only occasionally*". Even when datasets were offered at other infrastructures, they were usually also offered at the DANS infrastructure, which means that almost all the opened data could be found at a single place. In sum, from the perspective of the data provider data fragmentation did not seem to be a problem.

From the perspective of the OGD user, the case studies showed that besides the DANS infrastructure many other OGD infrastructures existed, such as the national infrastructure and infrastructures provided by the police, Statistics Netherlands, ministerial organisations, and municipalities. These OGD infrastructures also provided judicial and social data and other types of data. The publication of open judicial and social data on different infrastructures complicates the process of finding the data that OGD users are looking for, since users do not always know which organisation produces the data that they are looking for and on

which infrastructure this organisation releases its data. From a data user perspective the case studies confirmed that data fragmentation could hinder searching for and finding OGD.

Terminology heterogeneity

The literature review from chapter three pointed at the potential influence of terminology heterogeneity on OGD use, which means that different terminology may be used to describe datasets on the same topic or variable, and that there may be a lack of common data definitions. In the cases it was found that the examined judicial and social data used terminology that was typical for research from the judicial and sociology domains, and some open datasets were described through controlled vocabularies. For instance, the judicial data case showed that various judicial datasets used a classification of criminal offences developed by Statistics Netherlands. The use of controlled vocabularies made it possible for OGD users to use terminology in open judicial data consistently, since the use of terminology from this controlled vocabulary could clarify the meaning of terms from the judicial domain. The use of controlled vocabularies facilitated the process of searching for open datasets, since OGD users familiar with the vocabularies could use the controlled terms to search for open datasets, and they could easily identify the meaning of these terms.

Moreover, the case studies showed that various standards can be used by OGD infrastructures to facilitate searching for and finding OGD and to counteract terminology heterogeneity. The DANS infrastructure integrates a number of standards to describe the data (i.e. metadata standards). Datasets archived in DANS' EASY are described in Dublin Core fields with additional options from Qualified Dublin Core[2], although not all fields were mandatory to complete (The Data Seal of Approval Board, 2013). The data infrastructure also stimulated the consistent use of terms by providing pre-defined options for describing datasets in the uploading menu. For example, the DANS data infrastructure provided pre-defined drop-down menus for efficiency and consistency whenever possible (The Data Seal of Approval Board, 2013). Avoiding free text fields reduced inconsistencies between the terms that were used to describe datasets.

[2] http://dublincore.org/documents/dcmi-terms/

Although the judicial data provider often provided data according to the standards used by DANS, it did not use metadata standards to store its metadata internally. The lack of standards used by the judicial data provider complicated opening datasets. When uploading a dataset to the DANS infrastructure, the judicial data provider had to search for the required metadata in different systems. This sometimes led to the publication of limited metadata. The social data provider integrated several standards into its metadata system to describe the research data. The metadata expert of the social data case indicated that the description of social data in well-accepted standards by the governmental organisation makes it easier to supply data to DANS, since the metadata described in the metadata system corresponded to the metadata that DANS demands.

Search support

Search support refers to advanced, multilingual search functionalities to locate datasets. The investigated DANS infrastructure provides search support by facilitating dataset search through keywords and by browsing certain data categories, such as the creator of the dataset, the date of submission, and the type of access provided. In addition, the EASY data archive of DANS can be used in English. An OGD user can enter a search term in English, and various metadata fields and metadata are available in English. It was found that users of open judicial and social data could not search for data in other languages than English.

Various other search functions, such as searching by querying large numbers of datasets, were also not supported by the DANS infrastructure. The case studies showed that the users of the investigated judicial and social data could not express a search query themselves to search throughout diverse data sources. Research among OGD users by DANS showed that the search functionality can be improved. One OGD user said: *"it takes just a bit too much time to find the […] data. The search functionality works rather mediocre"* (Grootveld & Egmond, 2011, p. 8). Another OGD user said that "*what is currently used works reasonably to good, but the Google search engine is faster and works better*" (Grootveld & Egmond, 2011, p. 8). The provision of advanced search support may make it easier for the user to locate only the most relevant data and exclude datasets irrelevant to the user's search query.

The cases also demonstrated that there is a need for comparing data from multiple countries in different languages. For example, the case study of social data showed that data on time use from multiple countries are shared, so that they can be compared internationally. Furthermore, an interviewee from the judicial data case explicitly expressed the importance of a multilingual infrastructure: *"multilingualism is a very strong selling point"*. The case study on judicial data also showed that judicial data are exchanged internationally, so that crime rates in different countries can be compared.

Information overload

The literature showed that an information overload may occur when overwhelming amounts of open data are provided. From the perspective of the data providers, an overview of the published datasets showed that the number of opened datasets was not enormous. Less than one hundred datasets were offered by the judicial data provider, and the number of social dataset published was somewhat higher. Yet, the number of opened datasets was steadily increasing over time, and from the perspective of the OGD user, thousands of datasets were available on the DANS infrastructure. Since judicial and social data are also offered at various other OGD infrastructures, the number of datasets reaches beyond these thousands of datasets. Combining and integrating datasets from different OGD infrastructures results in enormous numbers of datasets on judicial and social data. These findings confirmed the literature, and showed that it can be difficult for users to locate the most relevant data among all large numbers of datasets, especially because of the lack of dataset search support (see previous section).

4.2.2 OGD analysis in the two case studies

Chapter three showed that OGD analysis is influenced by the data context, data interpretation support, data heterogeneity and data analysis support. The following sections provide insight in the case study findings regarding these four clusters.

Data context

In both cases it was investigated which data were collected about the context in which the OGD were created, since chapter three showed that these 'data about the data' (i.e. metadata) are important for the reuse of open data. The case studies showed that the DANS infrastructure did require the provision of a certain amount

of information about the context in which a dataset was created. For instance, a description of the instruments used to gather the data, the variables included in the study, a report of the fieldwork and the sample survey method needed to be provided by the governmental organisations which supply their data. However, this contextual information was not provided in a machine-readable format, but mainly in PDF or Word documents, which complicates the process of finding this contextual information by OGD users. OGD users needed to access the documents, and search for the contextual metadata within them. Often this required searching for contextual information in long research reports.

The provision of contextual metadata was also examined from the perspective of the data provider. It was found that the metadata system used by the judicial data provider contained several metadata fields that are useful for OGD users to comprehend a better understanding of the context in which a dataset has been created. For example, it consisted of metadata fields about data sources, which may help OGD users to determine where the data come from and whether they 'fit' their use purposes. Nevertheless, while various metadata fields were included in the metadata system, at the moment of the case study only few fields related to the produced dataset and to the context in which it was created. At the same time the judicial data provider needed to provide certain types of contextual information with the published OGD as demanded by DANS. Considerable effort and time investments were required from the judicial data provider to upload contextual metadata, since this information often could not be derived automatically from the WODC's metadata system.

The social data provider used a relational database system which described both administrative and non-administrative (descriptive) metadata. The system contained relatively much data about the context in which data were collected, such as to which time period and population the data applied, which sample entity they concerned, how the sample was taken, how the data were collected, who ordered the research, and who collected and deposited the data. Moreover, the system integrated existing standards for describing the data, including Standard Study Descriptions (SSD), Dublin Core (Dublin Core Metadata

Initiative, 2010), the Data Documentation Initiative (DDI[3]) and other international metadata standards, which made it easier for the SCP to supply these metadata together with the datasets that they uploaded to DANS.

Metadata is seen as a very important aspect of making judicial and social data available to the public. The data disclosure coordinator of the judicial data case expressed his concerns about the restricted metadata provision for their open judicial datasets and about the considerable effort that metadata provision by the judicial government agency required: *"[challenges] from the viewpoint of the data user: metadata. Insufficient documentation to be able to use the dataset well"*. The provision of metadata in general was also seen to be important for the use of social data. When the metadata expert of the SCP was asked which recommendation she would give to other governmental organisations that desire to disclose their data, she answered that maintaining metadata is important when one wants to deposit governmental data. The data disclosure coordinator of the SCP also expressed the significant importance of metadata by arguing that "*data without metadata is worthless*". This confirmed the finding from the literature that contextual information is important for the ability of OGD users to analyse, interpret and otherwise reuse datasets.

Data interpretation support

The literature review in chapter three suggested that there is potential for the misuse, misunderstanding and misinterpretation of open data, since open data infrastructures traditionally do not add contextual information to the datasets that they provide, and this complicates the analysis and interpretation of the data. In both case studies the potential for misuse, misunderstanding and misinterpretation of open data was a concern. The data disclosure coordinator from the judicial data case said: *"In practice the documentation of datasets is imperfect and may contain mistakes. This leads to the risk of misinterpretation",* and *"reputation damage is of importance […], damage resulting from misuse of data does not lead to improvements of our society. Then we should put effort in taking the edge of inaccurate conclusions"*. The way that datasets are interpreted was also found to

[3] A metadata standard for describing social and behavioural sciences data, see http://www.ddialliance.org/

be important in the social data case. For one particular complex dataset an interviewee from the social data case also said that *"the interpretation of this dataset is of great importance; this dataset is not meant for a random citizen"*.

To support the interpretation of datasets, the DANS infrastructure requires the provision of metadata using the Dublin Core standard with additional fields from Qualified Dublin Core[4]. These metadata standards are mainly focused on discovery metadata and provide limited information about the context of datasets (Zuiderwijk, Jeffery, & Janssen, 2012b). Fields that were obligatory to complete when a dataset was uploaded to the data infrastructure were the title of the research, the creator of the dataset, the date that the dataset was created, the description of the dataset, access rights, the date that it became available, and the audience of the dataset. Not all fields were mandatory to complete, and the number of obligatory fields was kept as low as possible to stimulate researchers to release their data (The Data Seal of Approval Board, 2013). A number of optional fields are recommended by DANS to fill (contributor(s), subject, spatial coverage, temporal coverage, source, identifier), and a number of fields are presented as additional (format, relation, language, remarks). The provision of limited metadata complicates the interpretation and use of OGD, and makes it difficult to find out for which purposes a dataset can be reused.

Data heterogeneity

Data heterogeneity refers to differences between open datasets that complicate the analysis of these data, such as the provision of open datasets in many different formats. In the judicial data case, most datasets were provided in the machine-readable Extensible Markup Language (XML), or as a SAV-file which can be used to save data in the Statistical Package for the Social Sciences (SPSS). In addition, most datasets were accompanied by a report or publication in a PDF-file or Word Document, yet these are not machine-actionable. Many open judicial datasets were also complemented with other types of files (e.g., DIC, FRQ, LAB, POR, XLSX, CMD, SRN, DAT). In the social data case most data were released as SAV-files, complemented with a PDF-file and one or more documents (DOC-files), although other types of files were also disclosed, such as XML, XLS, POR, DTA, LAB, DIC

[4] http://dublincore.org/documents/dcmi-terms/

and FRQ. The use of many different data formats complicates OGD use, since users need to be able to work with the different formats, and the infrastructure needs to support all these formats.

Data analysis support

The literature review demonstrated that various tools can be used for data analysis. For both cases the analysis support by the data infrastructure was studied. The infrastructure allowed only governmental data providers to upload their data to the DANS infrastructure and it was not possible for OGD users to do this. Therefore, OGD users could not upload datasets from other infrastructures to this infrastructure to combine and analyse them on the infrastructure.

Furthermore, for most datasets the DANS infrastructure provided no tools to support their analysis, which may hinder OGD analysis. The infrastructure offers the EASY Online Analysis Tool for the analysis for nine specific datasets. This tool facilitates the creation of tables, and makes it possible for data users to calculate correlations between variables, and to perform a regression analysis. In addition, some datasets can be analysed by examining a quality assessment (see section 4.2.5). For other datasets besides those nine, the infrastructure that was used in the cases obliged OGD users to first download the data to their own computer, and then to use the data in their own personal environment. This means that each OGD user needs to have the facilities and tools to reuse the data on their personal computer, while an infrastructure could provide the facilities and tools for all its users at once. Whereas some OGD users may have access to these tools on their own personal computers, other users may not, which complicates OGD analysis.

Moreover, an important barrier for uploading a dataset to another infrastructure by OGD users to use the infrastructure's data analysis tools concerned the license used by the DANS infrastructure. The infrastructure provider developed General Conditions of Use based on Copyright, Database rights, the Personal Data Protection Law and other laws. According to the General DANS Conditions of Use, users are neither allowed to distribute DANS datasets further or to make them publicly available for other people without prior written consent of the data provider, nor to sell datasets for commercial purposes. This limits the openness and the reusability of the datasets. Interviewees of DANS explained that

this data licence was developed to avoid the publication of one dataset at different places, since that might lead to a lack of overview on where the data are published. Moreover, a metadata expert and project manager at DANS stated that publishing data at different infrastructures may result in changes to the dataset that cannot be traced back anymore, which may lead to a decrease of trust in the dataset by the users. The data disclosure coordinator of judicial data mentioned that *"the WODC supports the use of this license. DANS is only the maintainer of the data […]. DANS will never become the owner of the data. DANS should therefore mention in its licence that permission to also publish the data at other places besides DANS should be requested at the data provider […]"*. This may be explained by the relatively high sensitivity of the judicial data. The social data provider was less concerned with the publication of their data on other infrastructures besides the DANS infrastructure, as their data were less sensitive than the judicial data.

4.2.3 OGD visualisation in the two case studies

In our literature review, factors which affect OGD visualisation were found in the cluster of data visualisation support. In this section we describe the case study findings for this cluster.

Data visualisation support

The literature review indicated that OGD users may profit from data visualisation support to make sense of OGD. The case studies confirmed that the judicial and social datasets may be appropriate for visualisations, although the social data provider was more optimistic about this than the judicial data provider. The data disclosure coordinator of the social data case said that *"[the opened] datasets contain enough geographical information to visualise datasets, for example, on a map"*. The data disclosure coordinator of the judicial data case said: *"we do not know for each dataset, but the motivation to open datasets is that the WODC believes this is possible"*. Yet, the cases showed that the OGD infrastructure did not allow for visualising most of the studied judicial and social data on maps or in other ways. For a small selection of all the datasets the EASY Online Analysis Tool could be used to visualise datasets in tables. No other visualisation tools were available for these nine datasets, and for the other datasets there was no tool to visualise them in tables or in other ways. Furthermore, after downloading the

judicial and social data from the cases to their own personal computers, the case studies showed that OGD users needed to search for visualisation tools at other places than the offered data infrastructure. While most computers provide some basic data visualisation tools, OGD users may need to search for more advanced data visualisation tools at different websites. The OGD infrastructure used to publish judicial and social data in the cases did not combine and integrate a variety of visualisation tools, which complicates OGD use.

4.2.4 Interaction about OGD in the two case studies

The literature in chapter three showed that a distinction can be made between two clusters of factors which influence interaction about OGD, namely a lack of interaction, and interaction support and tools. The findings from the case studies concerning the two clusters are as follows.

Lack of interaction

Interaction about OGD can be used to obtain feedback from other stakeholders involved in the open data process. The case studies showed that although interaction between OGD providers, policy makers and OGD users sometimes took place, this type of interaction did not occur regularly, and it was not facilitated by the infrastructure. At the same time, the case studies showed that OGD providers, policy makers, and OGD users could benefit from their concerted interaction in the open data process. From the perspective of the data providers, the cases showed that insight in how the data of the WODC and SCP were reused could be of interest to them. Especially in the judicial data case, more comprehensive insight in how the judicial data were being reused was desired, so that the employees involved in data disclosure could justify the effort and time that they spent on data publication. Furthermore, both in the judicial and social data case it was mentioned that it is interesting for the data providers to know whether people outside their organisation are reusing their data, and that the governmental agencies were interested in how often the data that they made available to the public were downloaded from the DANS infrastructure.

In addition, the judicial data disclosure coordinator said that interaction between the WODC and the users of its data could help to improve decision-making: *"based on experiences with learning moments about certain datasets, one*

can say which types of data the users find interesting and which are less interesting to them. For efficiency considerations one can then, for instance, give less priority to particular data". Employees of the social data provider mentioned that they were not very interested in considerable discussions with OGD users and that they did not want to collaborate with users of their data. Nevertheless, the social data provider would respond to feedback and questions from OGD users if these would be delivered to them through the infrastructure. Both the WODC and SCP already respond to questions and feedback of users via e-mail (one-to-one to data users), and if these questions and feedback would be provided to them in an automated way they would also respond to the data users. From the perspective of the data user, it was found that users may benefit from interaction with the WODC and SCP. They could ask questions that they have about a certain dataset, or about the context in which such a dataset can be reused. Mistakes in datasets may also be reported to the data provider.

In sum, the case studies confirmed the literature by showing that interaction regarding OGD publication and use processes was not a common phenomenon, both from the perspective of the OGD provider and from the perspective of the OGD user. Although the data providers sometimes had one-to-one conversations with OGD users via e-mail, these conversations did not occur regularly, and recurring interaction did not occur.

Interaction support

With regard to the second interaction about OGD cluster – interaction support – it was found that various types of feedback were important for the judicial data provider. The data disclosure coordinator of the judicial data provider stated explicitly that *"feedback is a very important element […]: feedback about privacy violations […], feedback when data are requested, who download the data and for which purpose, feedback ex post when the data have been used"*. The desire to learn from the use of open data was emphasised several times in the judicial data case (*"one just needs to learn from it"*). The judicial data provider also mentioned that insight in who had used the judicial datasets and in which ways could help to find out whether datasets had been misused and misinterpreted. This type of insight was probably desired more by the judicial data provider than by the social

data provider because of the higher sensitivity of the judicial data. It was mentioned in the judicial data case that *"we want to know what happens with the used data"*.

However, the OGD infrastructures on which the judicial data and the social data were published barely provided automated support for interaction between open data providers and open data users, or between open data users themselves. The infrastructure did not support the supply of feedback from OGD users to the data providers regarding how the publication of the data could be improved. The OGD providers could not participate in discussions on how their datasets were reused by the public, or what the needs of the public were for the release of new datasets. For the OGD providers it was also complicated to derive information about the number of dataset downloads from the DANS infrastructure, and, as stated by the data disclosure coordinator of the judicial data case: *"Information about the data users given by DANS is sparse"*. As a result, there was a lack of information that the data providers could use to make more informed future investment decisions for the supply of governmental data. Although the governmental data providers could guess which types of users would potentially be interested in reusing their data (i.e. mainly researchers and students), it was not clear whether these groups of people actually used the OGD and how often this was done.

The literature review in chapter three showed that interaction may take place through discussions with data users. From the perspective of the data user, it was found in both case studies that the data infrastructure did not provide an environment to request datasets from governmental agencies, to view how other users had reused datasets and for which purposes, and to discuss questions and conclusions about what other users had learned from the use of the judicial and social data. There were no automated mechanisms for conversations supported by the infrastructure. Moreover, the infrastructure on which the judicial and social data were published did not offer any automated mechanisms to keep track of changes that different data users can make to datasets, such as changes regarding removed or added variables, added analysis results and added metadata. The cases confirmed the literature finding that the type of interaction tools that open data users can utilise influence to which extent interaction may occur. The lack of

interaction tools appeared to negatively influence the extent to which individuals could interact regarding OGD use.

4.2.5 OGD quality analysis in the two case studies

Chapter 3 showed that OGD quality analysis is influenced by factors in the cluster of dependence on the quality of the data, poor data quality and quality variation and changes. Case study findings in each of these three clusters are identified below.

Dependence on open data quality

The literature demonstrated that successful open data use strongly depends on the quality of the data. From the perspective of the data providers, the case studies showed that both data providing organisations paid considerable attention to the quality of the datasets that they produced. Both governmental organisations employed individuals experienced with setting up high-quality research, and they contained various working groups and divisions that were responsible for carefully watching the quality of the conducted research and the data produced by them. Quality audits and visitations were also conducted regularly. Moreover, in both organisations the quality of the data was examined again before datasets were released to the public. In the judicial data case several persons checked the quality of each dataset that seemed appropriate for publication, including the researcher and producer of the dataset, and at least two researchers from other divisions. It was stated by one of the interviewees that "*data quality issues are important, also metadata quality*". In the social data case the quality of datasets was also checked before data disclosure. The coordinator of social data publication said that datasets of insufficient quality do not become available as open data on the data infrastructure. This was also the case in the judicial data study.

From the perspective of the data user it was found that the data providing infrastructure supplied some information on the quality of the data. Although data quality is subjective and depends on the purpose of data use, insight in various quality aspects and the purposes for which the datasets have been used may help potential data users to decide whether a dataset is appropriate for their own data reuse purposes. The infrastructure on which the judicial and social data were published had recently started to explore possibilities for gathering information

about the quality of datasets by adopting peer review processes. Individuals who had downloaded data from the infrastructure were asked to answer a number of questions about the quality of the data. First the data quality in general was assessed, and subsequently a number of data quality aspects were rated, including the quality of the documentation, data completeness, data consistency, data structure, and usefulness of the file formats. In addition, reviewers were asked whether they recommended the use of a dataset, whether they published articles using the dataset and whether they intended to use the data for a publication in the future. According to the infrastructure providing organisation, "*it should then be visible in EASY who has assessed a dataset and what that assessment resulted in*" (The Data Seal of Approval Board, 2013, p. 5). For a number of datasets data reviews were gathered in this way[5]. While no data quality reviews were provided for the judicial data at the time of this research, a few reviews were available for the social data from our case studies.

The data quality reviews provide OGD users with information about various quality dimensions. Nevertheless, these reviews did not provide OGD users with the possibility to freely discuss any quality aspect of the data. It was not possible for data reviewers to add data quality information in a text box, and users of the investigated judicial and social data could not discuss with other users or with data providers about the quality of the data.

Poor data quality

The poor quality of open data can be a major issue and may influence OGD use. The previous section showed that insight in the data quality can be obtained from data quality reviews and ratings. However, such data quality information does not provide any insight in the purposes of data reuse for which the dataset was insufficiently complete. For example, when a dataset receives a low average rating score regarding dataset completeness because it lacks the data that all the data reviewers needed, it may still contain those data that another OGD user needs. While a dataset may receive a low score on the quality dimension 'timeliness', for certain reuse purposes it may not be necessary that the dataset is timely and current. Therefore, data quality information on its own appeared not to be sufficient

[5] See http://datareviews.dans.knaw.nl/index.php?l=en for the overview of reviews.

for the reuse of OGD. The case studies showed that it is important for OGD users not just to have information about the quality of a dataset, but also to be able to relate data quality information to a description of the context in which a person aimed to reuse a dataset, including the purpose of use for the particular dataset.

Quality variation and changes

The cluster of quality variation and changes refers to the differences in open data quality over time, over reuse and over sources from where they were obtained. The data infrastructure used in the investigated cases does not allow for comparing the quality of different datasets in one overview. It was also not possible to see whether there were differences in the quality of a certain dataset over time or between sources. For instance, social data on time use and on elections were collected over many years, while data quality aspects of these datasets cannot be compared. Furthermore, the data infrastructure did not allow for examining adjustments of datasets over data reuse. When a judicial or social dataset was reused, the DANS infrastructure did not provide mechanisms to see how the dataset had been reused, and whether this had resulted in any changes in the quality of the dataset. For example, an OGD user might improve the usefulness of a dataset by making it available in multiple formats, improve the structure by better organising the dataset, or improve the completeness by adding data or metadata derived from data analysis. This was not possible in the investigated cases.

4.3 Functional requirements for an OGD infrastructure

In this section we aim to define how the OGD infrastructure is supposed to behave according to the case study findings. On the basis of the case study descriptions provided in section 4.2, this section elicits functional requirements for an OGD infrastructure that enhances coordination of OGD use. Whereas the previous section discussed the factors in each case study that influence OGD use, this section translates the case study findings to concrete functional requirements for the design of the OGD infrastructure. Functional requirements are the requirements which define the specific functionality that shows how a system can be used (Stellman & Greene, 2005). We focus on functional requirements, since this study aims to find out which infrastructure OGD users find functional and usable. This study does not search for infrastructure requirements regarding, for example, its

maintainability, sustainability and scalability (i.e. the non-functional requirements), yet searches for what users want from the infrastructure. This study therefore focuses on functional requirements for the OGD infrastructure. While focusing on the functional requirements for the OGD infrastructure, one assumption of this study is that various non-functional requirements, such as the maintainability and sustainability of the OGD infrastructure, are met.

While describing the requirements we again follow the distinction of types of OGD use as identified in chapter three. The functional requirements are described within each of the clusters of factors influencing OGD use. For instance, functional requirements are described in the clusters of search support and data context. This is done because the OGD use activities and the clusters of influencing factors within them already provide high level directions for the requirements. For instance, the first type of OGD use – i.e. searching for and finding OGD – already reveals on a high level the requirement that the OGD infrastructure should provide mechanisms to search for and find OGD. Some of the functional requirements that will be described in the following sections can be seen as subsets of other requirements. Yet, these requirements are still formulated as individual requirements because of the importance that was assigned to them in the cases. For each of the requirements that may be seen as a subset of other requirements it is explained how it relates to other requirements and which case study findings led to defining them as individual requirements.

4.3.1 Functional requirements for searching for and finding OGD

The requirements for searching for and finding OGD data can be focused on improving existing search and find functionalities of OGD infrastructures. Requirements regarding data fragmentation, terminology heterogeneity, search support and information overload will be discussed.

Data fragmentation

From a data user perspective, both case studies confirmed that data and metadata fragmentation could be a problem for OGD users who desired to find a certain dataset. This leads to *the first requirement that the OGD infrastructure should be a one-stop shop for datasets and metadata from a variety of OGD infrastructures*, so that OGD users can go to a single place to find all sorts of data and have access to

the OGD provided by various governmental organisations. Some of the case study interviewees noted that the development of a one-stop-shop has the drawback that considerable expert knowledge is needed for the integration of datasets, and that the creation of a one-stop-shop may demand considerable effort from the data and infrastructure provider. Nevertheless, this functional requirement was found to be important to enhance the coordination of OGD use. Since this study aims to enhance the coordination of OGD use, priority was given to the view of OGD users in the formulation of this requirement. The case studies showed that the provision of a single point of access to OGD may make it easier to find data that are offered on different OGD infrastructures through a single infrastructure.

A community of OGD users might help to maintain an overview of datasets and to counteract data fragmentation. The case studies showed that the existing OGD infrastructures where OGD are disclosed only allowed governmental organisations to upload data. It was not possible for OGD users to upload datasets themselves. This leads to *the second requirement that the OGD infrastructure should allow OGD users to integrate and to refer to datasets from various other OGD sources*, so that data fragmentation can be counteracted not only by the government, but by multiple stakeholders. This second requirement is related to the first requirement, yet it emphasises that the integration of data and metadata into a one-stop-shop should not only be possible for infrastructure developers and OGD providers, but also for OGD users.

Terminology heterogeneity

Both the judicial and social data case demonstrated that it is important to use consistent terminology, preferably using controlled vocabularies, which leads to *the third OGD infrastructure requirement, namely to use controlled vocabularies to describe OGD*. From the perspective of the OGD user, the case studies showed that the use of international well-accepted standards facilitated interoperability with other OGD infrastructures, since this stimulated the use of similar terms, and made it easier to search for datasets across infrastructures. *The fourth OGD infrastructure requirement was therefore to use interoperable standards to describe datasets*.

Search support

In the case studies it was found that the users of the investigated judicial and social data cannot express search queries themselves to search throughout diverse or large numbers of data sources. Especially if users want to flexibly search for data by using their own queries, rather than being bound to the traditional search functionalities provided by most existing OGD infrastructures, advanced search functionalities become essential. In addition, the use of OGD in semantic web applications requires a search functionality to express a query across various data sources. This resulted in *the fifth requirement that the OGD infrastructure should support data search through keywords, data category browsing and data querying*. Moreover, both case studies showed that research resulting in judicial and social data is carried out internationally, and that there is a need for comparing data from multiple countries in different languages. Since various languages are used in these countries, this leads to *the sixth requirement that the use of OGD needs be supported by the ability to search for data and metadata in multiple languages*.

Information overload

To counteract an information overload, both case studies showed that structured and ordered overviews of search results, as well and filtering and sorting functions may help to understand the search results provided by the infrastructure. These functions may help by finding relevant datasets and excluding irrelevant data from the search results. This leads to *the seventh requirement that the OGD infrastructure should facilitate filtering, sorting, structuring and ordering relevant search results*. This requirement is related to the fifth requirement since it also focuses on categorising data. However, the fifth requirement emphasises category browsing as a way to support finding datasets within a certain data category, while the seventh requirement focuses on ordering datasets that were already found (e.g. based on keyword search).

Summary

From the case studies we elicited the functional requirements for searching for and finding data in our OGD infrastructure (see Table 4-5). Not all of these requirements were mentioned by each of the case study interviewees, but sometimes only by some of them. Yet, functional requirements were also derived

from other information sources than the interviews (e.g. participant observations and dataset analysis). When we consider all the information sources it can be seen that all requirements were found in both case studies.

Clusters of factors influencing searching for and finding OGD	Functional requirements for the OGD infrastructure	Open judicial data use case	Open social data use case
Data fragmentation	1. The OGD infrastructure should be a one-stop shop for datasets and metadata from a variety of other OGD infrastructures.	X	X
	2. The OGD infrastructure should allow OGD users to integrate and refer to datasets from various other OGD sources.	X	X
Terminology heterogeneity	3. Use controlled vocabularies to describe OGD.	X	X
	4. Use interoperable standards to describe OGD.	X	X
Search support	5. The OGD infrastructure should support data search through keywords, data category browsing and data querying.	X	X
	6. The OGD infrastructure should support OGD use by the ability to search for data and metadata in multiple languages.	X	X
Information overload	7. The OGD infrastructure should facilitate filtering, sorting, structuring and ordering relevant search results.	X	X

Table 4-5: Functional requirements for searching for and finding OGD.

4.3.2 Functional requirements for OGD analysis

Four clusters of factors influence OGD analysis: data context, data interpretation support, data heterogeneity and data analysis support. In this section we discuss the requirements in these clusters as derived from the case studies.

Data context

When we compare the judicial data case and the social data case, it can be seen that more contextual metadata were collected in the social data case. The provision of these metadata to OGD users makes it easier to understand the context in which the social data were created and whether and how they can be used for other purposes. The case studies showed that it is of significant importance to supply metadata for the interpretation and analysis of OGD. One civil servant working at the social data provider mentioned that it is important that

governmental organisations maintain at least those metadata that the infrastructure provider requires from data providers. The organisation that maintains the data infrastructure asks for certain metadata, which are important for the use of open data, and the data that it asks for should be registered well by governmental organisations which aim to make their data available via the infrastructure. The interviewee stated that if these metadata are maintained well and are up-to-date, this makes it much easier to make data available to the public. This suggests that if the infrastructure would require the provision of more metadata for each uploaded dataset, the data providers would adapt to this situation, which can subsequently lead to offering more metadata and improving OGD use. This leads to two key requirements for our OGD infrastructure: *the eighth requirement that the OGD infrastructure should provide data that describe the dataset*, and *the ninth requirement that the OGD infrastructure should provide data about the context in which the dataset has been created*. These two requirements are related, since they both show the need for providing data about the dataset. While the eighth requirement focuses on general data that describe the dataset, such as its title and the topic, the ninth requirement focuses on data that describe the context in which it was created, such as the temporal granularity or the methods that were used to collect the data.

Data interpretation support

The infrastructure on which the judicial and social data were published in our cases allowed for the supply of metadata in only few metadata fields regarding the context in which the data were created. Moreover, many of the metadata fields were not mandatory to complete. As a consequence, limited metadata were provided for each dataset, which potentially leads to the misuse, misunderstanding and misinterpretation of datasets. The case studies showed that an OGD infrastructure should support the interpretation of open datasets. To avoid misuse, misunderstanding and misinterpretation by the OGD user as much as possible, contextual and domain knowledge about how to interpret and use the data can be provided. The case studies also showed the dark side of providing considerable metadata, since it takes much time to derive contextual and detailed metadata from researchers and research reports. It is expensive for governmental organisations to

offer considerable metadata. However, if governmental organisations wish that their data are reused and if they aim for the benefits of OGD, these user requirements need to be considered. The foregoing leads to *the tenth requirement that it should be clear for which purpose the data have been collected,* and to *the eleventh requirement that it should provide examples of the context in which the data might be used*. Whereas the tenth requirement focuses on the purpose for collecting the data, the eleventh requirement focuses on the potential purposes for reusing the data in the future. Moreover, the case studies showed the importance of *the twelfth requirement that the OGD infrastructure should provide domain knowledge about how to interpret and use the data*. Compared to the eighth and ninth requirement from the 'data context' cluster, this requirement focuses on detailed domain-related data that describe the dataset, whereas the previous requirements focus on general data or contextual data about the dataset.

Data heterogeneity

The case studies showed that many different types of formats were used to publish the examined open judicial and social data. This means that the OGD infrastructure should allow for the publication and analysis of datasets in these different formats to facilitate OGD use. The use of well-accepted standards for these formats allows for interoperability with other programmes and tools, which contributes to clarifying the semantics of the data. These findings lead to *the thirteenth requirement that the OGD infrastructure should allow for the publication of datasets in different formats*. Moreover, they confirmed the fourth requirement in the cluster of 'terminology heterogeneity' *that interoperable standards should be used to describe OGD*.

Data analysis support

The case studies showed that without data analysis support, it may be very difficult for OGD users to comprehend insight in the meaning of the data. For instance, the judicial data provider produced many datasets on the development of crime, security and justice over time. Without the proper support for the analysis of such data, it is complicated to identify patterns over time. The foregoing leads to *the fourteenth requirement that support should be provided in the form of tools that make it possible to analyse OGD* (e.g. tools that allow for identifying correlations

between variables). If uploading data by any user would be facilitated, this would also make it possible to use the tools provided by one OGD infrastructure for open datasets obtained from other infrastructures. Thus, the OGD infrastructure could then be used to upload datasets from other infrastructures to it, so that the tools of the OGD infrastructure can be used. The provided tools should take into account the format of the open datasets. Nevertheless, it may become problematic to provide appropriate tools to support OGD analysis if datasets are in many different formats. This would require the provision of a large variety of tools.

Additionally, we conclude from the case study descriptions that there may be sound reasons for applying conditions of use to the judicial and social datasets, yet this may also hinder the use of the data. Since the data infrastructure in the cases did not provide many tools for the analysis of the data, users may want to upload the dataset at another infrastructure which provides more tools for data analysis, while this was not allowed. This leads to *the fifteenth requirement that the conditions for data publication and use should be clear to the users*. If necessary, the infrastructure can make it possible for users to contact the data provider to ask for permission to also publish data elsewhere than at the currently used infrastructure.

Summary

In the previous sections, requirements for the OGD infrastructure related to OGD analysis were elicited. The requirements are summarised in Table 4-6. Several functional requirements were not mentioned by all the interviewees. Yet, functional requirements were also derived from other information sources than the interviews (e.g. participant observations and dataset analysis). When we consider all the information sources, it can be concluded that all requirements are found in both case studies.

Clusters of factors influencing OGD analysis	Functional requirements for the OGD infrastructure	Open judicial data case	Open social data case
Data context	8. The OGD infrastructure should provide data which describe the dataset.	X	X
	9. The OGD infrastructure should provide data about the context in which the dataset has been created.	X	X
Data interpretation support	10. It should be clear for which purpose the data have been collected.	X	X
	11. It should provide examples of the context in which the data might be used.	X	X
	12. Domain knowledge about how to interpret and use the data should be provided.	X	X
Data heterogeneity	13. The OGD infrastructure should allow for the publication of datasets in different formats.	X	X
	Confirmed 4: use interoperable standards to describe OGD.	X	X
Data analysis support	14. The OGD infrastructure should offer tools that make it possible to analyse OGD.	X	X
	15. The OGD infrastructure should provide insight in the conditions for reusing the data.	X	X

Table 4-6: Functional requirements for OGD analysis.

4.3.3 Functional requirements for OGD visualisation

In this section we describe OGD infrastructure requirements in the cluster of data visualisation support.

Data visualisation support

Even though not all open datasets from the case studies may currently be appropriate for visualisation, the cases showed that various judicial and social datasets generated by the examined governmental organisations contained geographical information, such as places where crimes were committed (crime, law and order research from the crime data case) and where people worked (life-situation survey from the social data case). These datasets may be opened for public reuse in the future. For those datasets that were appropriate for visualisation, users had to download the data to their own personal computer and visualise them in that environment, instead of being able to use the OGD

infrastructure for this purpose. The lack of visualisation support means that each OGD user needed to have the facilities and tools to visualise the data on their own personal computer, whereas an infrastructure could provide the facilities and tools for all its users at once and could coordinate their use. While some OGD users may have access to these tools, other users may not have this access. This complicates data visualisation, and makes it more difficult for them to comprehend insight in the meaning of the data. Two requirements for the OGD infrastructure were derived from the case studies, namely *the sixteenth requirement that the OGD infrastructure should provide and integrate visualisation tools*, and *the seventeenth requirement that it should allow for visualising data on maps*.

While the seventeenth requirement can be seen as a subset of the sixteenth requirement, it was formulated separately because the importance of visualisations on maps was explicitly mentioned in the case studies. Moreover, one interviewee stated that whereas several OGD infrastructures already provide a number of visualisation tools, the possibility to create maps is barely provided and would therefore be an improvement compared to existing OGD infrastructures. In addition, the sixteenth and seventeenth requirement are related to the fourteenth requirement concerning the provision of tools to analyse OGD, since data visualisation might be seen as a form of data analysis.

Summary

The case studies provided two functional requirements for OGD visualisation in our infrastructure, which were found both in the first and in the second case study (see Table 4-7). The requirements were both derived from various information sources of the case studies, such as participant observations and dataset analysis..

Clusters of factors influencing OGD visualisation	Functional requirements for the OGD infrastructure	Open judicial data case	Open social data case
Data visualisation support	16. The OGD infrastructure should provide and integrate visualisation tools.	X	X
	17. The OGD infrastructure should allow for visualising data on maps.	X	X

Table 4-7: Functional requirements for OGD visualisation.

4.3.4 Functional requirements for interaction about OGD

A distinction can be made between two clusters of factors influencing interaction about OGD, namely a lack of interaction, and interaction support and tools. Functional requirements for an OGD infrastructure concerning both of these clusters are as follows.

Lack of interaction

The case studies showed that although interaction between OGD providers, policy makers and OGD users sometimes took place, this type of interaction did not occur regularly, and it was not coordinated through the infrastructure. At the same time, the case studies showed that OGD providers, policy makers and OGD users could benefit from their concerted interaction in the open data process. For example, interaction between OGD providers, policy makers and OGD users may help to improve the processes of disclosing governmental data to users. *The eighteenth requirement for the OGD infrastructure is therefore that it should support interaction between OGD providers, policy makers, and OGD users.*

From the perspective of the OGD user, it was clear that there were limited opportunities to become involved in discussions with data providing organisations, or to discuss about data use with other OGD users. One finding in both cases was that the OGD infrastructure did not assist in conversations about released and used open datasets, neither between the data provider and user, nor between OGD users. This leads to *the nineteenth requirement that our OGD infrastructure should allow for conversations and discussions about released governmental data*, and to *the twentieth requirement that the infrastructure should assist in viewing who has used the dataset and in which way*. While the eighteenth requirement ensures that each of the three key stakeholders can interact concerning the OGD infrastructure, the nineteenth requirement ensures that interaction between these three stakeholders can take place. The case studies showed that data describing the dataset are not always sufficient for an OGD user to understand and interpret a dataset. Therefore, it is important that OGD users can discuss the peculiarities of a dataset with OGD providers and policy and decision makers through the infrastructure. The twentieth requirement should contribute to the users' understanding of a dataset by looking at previous uses of the data.

Interaction support

The judicial data case showed that the judicial data providers and the policy makers desired more insight in who had used their datasets and in which ways, since this might help in providing datasets that OGD users want to use. While this seemed to be less of a concern in the social data case, the social data provider would also be willing to respond to questions and requests from data users, and to interact about OGD use in a more restricted form. Based on the findings from the case studies we derived *the twenty-first requirement that the OGD infrastructure should provide tools for interactive communications between OGD providers, policy makers, and OGD users in OGD use processes*. For instance, governmental organisations can use tools to let potential data users request datasets concerning certain topics, or they can communicate about data needs and data reuse via social media, such as LinkedIn and Twitter.

While open data interaction needs to be supported, the case studies demonstrated that there is a lack of interaction tools in the OGD infrastructures where the investigated social and judicial data are published. This leads to *the twenty-second requirement that the OGD infrastructure should provide tools for interactive communications between OGD users*. For example, discussion forums and social media can be used to support the discussion of data reuse results, and social media can be used to contact other data users, which could further enhance the coordination of OGD use. Whereas the twenty-first requirement focuses on the interaction between OGD providers, OGD users and policy makers, the twenty-second requirement focuses on the interaction of OGD users with other OGD users.

In addition, the infrastructure on which the judicial and social data were published did not offer any mechanisms to keep track of the changes that different data users can make to a dataset. This may lead to a decrease in trust of other data users in the amended data, since they may not know which changes have been made to a particular dataset. From this infrastructure impediment we elicited *the twenty-third requirement that the OGD infrastructure should provide tools to keep track of amended datasets so that users know how the datasets have been changed*. Meeting this requirement should help users to obtain insight in the

provenance of datasets, which can be difficult when different versions, combinations and derivations of datasets have been created.

Summary

Table 4-8 offers an overview of the OGD interaction requirements. Some requirements were emphasised more in the first case study and less in the second case study. Despite the differences in emphasis on requirements from the perspective of the OGD providers, we argue that interaction about OGD needs to be improved in the OGD infrastructure. From the perspective of the data provider, interaction about OGD may lead to insight in errors of datasets, insight in how data publication can be improved, and insight in new findings from dataset reuse, which could enhance OGD use coordination. From the perspective of the OGD user, interaction about OGD is necessary to better understand how datasets can be reused and to learn from OGD reuse by other individuals, which is also expected to enhance the coordination of OGD use.

Clusters of factors influencing interaction about OGD	**Functional requirements for the OGD infrastructure**	**Open judicial data case**	**Open social data case**
Lack of interaction	18. The OGD infrastructure should support interaction between OGD providers, policy makers and OGD users in OGD use processes.	X	X
	19. The OGD infrastructure should allow for conversations and discussions about released governmental data.	X	X
	20. The OGD infrastructure should allow for viewing who used a dataset and in which way.	X	X
Interaction support	21. The OGD infrastructure should provide tools for interactive communications between OGD providers, policy makers, and OGD users (e.g. data request mechanisms and social media).	X	X
	22. The OGD infrastructure should provide tools for interactive communications between OGD users (e.g. discussion forums and social media).	X	X
	23. The OGD infrastructure should provide tools to keep track of amended datasets so that users know how datasets have been changed.	X	X

Table 4-8: Functional requirements for interaction about OGD.

4.3.5 Functional requirements for OGD quality analysis

Three clusters of factors were found to influence OGD quality analysis, namely dependence on the quality of the data, poor data quality and quality variation and changes. Infrastructure requirements in each of these clusters are elicited below.

Dependence on open data quality

The data providing organisations emphasised the importance of the quality of research and datasets, and various measures were taken to ensure high-quality research. On the one hand the data infrastructure offered data quality reviews, which may help OGD users to obtain more insight in various quality aspects of datasets. On the other hand, the data infrastructure provided limited opportunities to freely discuss data quality aspects with other users of the infrastructure, or with the data providers. The case study findings led to *the twenty-fourth requirement that the OGD infrastructure should provide insight in quality dimensions of OGD*, and *the twenty-fifth requirement that it should be possible to discuss these quality dimensions by OGD users, OGD providers and policy makers*. The twenty-fifth requirement also requires the storage of usage history. It relates to the nineteenth requirement since they both call for discussions about released OGD. While requirement nineteen focuses on conversations and discussions in general, requirement twenty-five focuses on discussions about quality dimensions in particular.

Poor data quality

Both case studies showed that it is important for OGD users not just to have information about the quality of a dataset, but also to be able to relate data quality information to a description of the context in which a person aims to reuse a dataset, including the purpose of use for the particular dataset. The twelfth requirement in the cluster of 'data interpretation support' already showed the importance of clarifying for which purpose the data can be used and in which contexts. While the twelfth requirement focused on envisioned purposes of OGD use in the future, it is also important for OGD users to know for which purposes the data have actually been used in practice, in order to coordinate OGD use. Such insight can help in detecting whether the quality of the data is adequate for the purposes of data use or whether the quality is too poor for a particular data reuse

purpose. This leads to *the twenty-sixth requirement that the OGD infrastructure should provide information on the context in which a person reused a particular dataset*.

Quality variation and changes

The data infrastructure used in the investigated cases does not allow for comparing the quality of different datasets in one overview. It was also not possible to see whether there were differences in the quality of a certain dataset over time or between sources. Furthermore, the data infrastructure did not allow for examining amendments of datasets over data reuse, while these operations were found to be useful for OGD use. These findings lead to *the twenty-seventh requirement that the OGD infrastructure should provide quality dimensions of datasets that are comparable with other datasets and with different versions of the same dataset*. Moreover, it shows the importance of *the twenty-eighth requirement that it should be possible to compare the quality of datasets over different data sources, over time and over data reuse*. Both requirements require the recording of usage history. Whereas the twenty-fourth requirement focused on having insight in the quality dimensions of a dataset, the twenty-seventh and twenty-eighth requirements emphasised that the quality dimensions of datasets needed to be comparable. The twenty-seventh and twenty-eighth requirement also relate to the twenty-fifth requirement that focuses on discussions of quality dimensions, yet these requirements differ from the twenty-fifth by emphasising comparable quality dimensions among datasets. One should note that such comparisons are subjective and depend on the purpose of OGD use.

Summary

Table 4-9 summarises the OGD infrastructure requirements for analysing the quality of OGD. The requirements were found in both case studies. Some requirements were mentioned by only some of the case study participants. Nevertheless, functional requirements were also derived from other information sources than the interviews (e.g. participant observations and dataset analysis), and when we consider all the information sources it can be seen that all the requirements concerning OGD quality analysis were found in both case studies.

Clusters of factors influencing OGD quality analysis	Functional requirements for the OGD infrastructure	Open judicial data case	Open social data case
Dependence on the quality of open data	24. The OGD infrastructure should provide insight in quality dimensions of OGD.	X	X
	25. It should be possible for OGD users, OGD providers and policy makers to discuss the quality of a dataset.	X	X
Poor data quality	26. The OGD infrastructure should provide information on the context in which a person reused a particular dataset.	X	X
Quality variation and changes	27. The OGD infrastructure should provide quality dimensions of datasets that are comparable with other datasets and with different versions of the same dataset.	X	X
	28. It should be possible to compare the quality of datasets over different data sources, over time and over data reuse on the data infrastructure.	X	X

Table 4-9: Functional requirements for OGD quality analysis.

4.4 Summary: overview of functional requirements and answer to the second research question

In chapter three we identified clusters of factors which influence OGD use from the literature and answered the first research question. In this fourth chapter the previously identified clusters of factors were used as a framework to study functional requirements for an infrastructure which intends to enhance the coordination of OGD use. The framework was applied to two cases on open judicial data use and open social data use. First, a description of the case studies was provided within each cluster of factors influencing OGD use. Second, functional requirements for the OGD infrastructure were elicited and the second research question, *what are the functional requirements for an infrastructure that enhances the coordination of OGD use*, was answered. The requirements from the cases were elicited through iterative processes. The case study participants provided feedback on a preliminary description of the case study findings and some of them on different versions of the infrastructure design (see chapter 5), and then another round of data collection and analysis took place. The twenty-eight elicited requirements are summarised in Table 4-10.

OGD use category	Clusters of factors influencing OGD use (RQ1)	Requirements for the OGD infrastructure (RQ2)
Searching for and finding OGD data	Data fragmentation	1. The OGD infrastructure should be a one-stop shop for datasets and metadata from a variety of other OGD infrastructures.
		2. The OGD infrastructure should allow OGD users to integrate and refer to datasets from various other OGD sources.
	Terminology heterogeneity	3. Use controlled vocabularies to describe OGD.
		4. Use interoperable standards to describe OGD.
	Search support	5. The OGD infrastructure should support data search through keywords, data category browsing and data querying.
		6. The OGD infrastructure should support OGD use by the ability to search for data and metadata in multiple languages.
	Information overload	7. The OGD infrastructure should facilitate filtering, sorting, structuring and ordering relevant search results.
OGD analysis	Data context	8. The OGD infrastructure should provide data which describe the dataset.
		9. The OGD infrastructure should provide data about the context in which the dataset has been created.
	Data interpretation support	10. It should be clear for which purpose the data have been collected.
		11. It should provide examples of the context in which the data might be used.
		12. Domain knowledge about how to interpret and use the data should be provided.
	Data heterogeneity	13. The OGD infrastructure should allow for the publication of datasets in different formats.
		Confirmed 4: use interoperable standards to describe OGD.
	Data analysis support	14. The OGD infrastructure should offer tools that make it possible to analyse OGD.
		15. The OGD infrastructure should provide insight in the conditions for reusing the data.
Visualising OGD	Data visualisation support	16. The OGD infrastructure should provide and integrate visualisation tools.
		17. The OGD infrastructure should allow for visualising data on maps.
OGD interaction	Lack of interaction	18. The OGD infrastructure should support interaction between OGD providers, policy makers and OGD users in OGD use processes.
		19. The OGD infrastructure should allow for conversations and discussions about released governmental data.
		20. The OGD infrastructure should allow for viewing who used a dataset and in which way.

Table 4-10: Overview of the functional requirements for the OGD infrastructure derived from the case studies.

OGD use category	Clusters of factors influencing OGD use (RQ1)	Requirements for the OGD infrastructure (RQ2)
OGD inter-action	Interaction support	21. The OGD infrastructure should provide tools for interactive communications between OGD providers, policy makers, and OGD users (e.g. data request mechanisms and social media).
		22. The OGD infrastructure should provide tools for interactive communications between OGD users (e.g. discussion forums and social media).
		23. The OGD infrastructure should provide tools to keep track of amended datasets so that users know how datasets have been changed.
OGD quality analysis	Dependence on the quality of open data	24. The OGD infrastructure should provide insight in quality dimensions of OGD.
		25. It should be possible for OGD users, OGD providers and policy makers to discuss the quality of a dataset.
	Poor data quality	26. The OGD infrastructure should provide information on the context in which a person reused a particular dataset.
	Quality variation and changes	27. The OGD infrastructure should provide quality dimensions of datasets that are comparable with other datasets and with different versions of the same dataset.
		28. It should be possible to compare the quality of datasets over different data sources, over time and over data reuse on the data infrastructure.

Table 4-10 (continued): Overview of the functional requirements for the OGD infrastructure derived from the case studies.

One should note that the requirements are derived from only two cases. The cases focus on a specific type of open data in a particular context. The studied cases focused on the use of structured research OGD from the domains of social sciences and humanities, and they incorporated the use of these data by researchers outside the government through OGD infrastructures. The elicited infrastructure requirements are important in the context of these cases, yet the findings from the case studies may not be generalizable beyond this context. This needs to be examined in evaluations.

Moreover, the number of elicited requirements is based on how we described and prioritised them based on the case study findings. While we expect that other researchers would elicit the same requirements, they might describe them in a different way. For example, other scholars might integrate or split some of our requirements, which might lead to a different number of elicited

requirements. In this study we decided to separately describe each requirement that was seen as important by the interviewed case study participants and the other case study information sources, instead of integrating the requirements on a higher aggregation level. As such we could also have described the requirements on a higher level of detail. In the next chapter, we use the functional requirements to generate the design of the OGD infrastructure.

5. Design of the OGD infrastructure

Building on the functional requirements identified in the previous chapter, this chapter incorporates the third research phase and provides the design of the OGD infrastructure. It aims at answering the third research question: *Which functional elements make up an infrastructure that enhances the coordination of OGD use?* Requirement analysis, literature review and design are the research instruments used to answer this question. This chapter starts with outlining the design approach. Thereafter, based on an analysis of the functional infrastructure requirements from the previous chapter and a literature review, design propositions are defined. The design propositions are formulated on a high level of abstraction and are further developed by identifying more detailed design principles through an extended literature review. The design propositions and the design principles provide input for the OGD infrastructure design, which is described thereafter. The description of the infrastructure consists of the system design, the coordination patterns, and the function design. Finally, this chapter concludes with a summary and the answer to the third research question. Parts of this chapter have been published in Zuiderwijk, Jeffery, et al. (2012b), Zuiderwijk, Jeffery, and Janssen (2012a), Zuiderwijk, Janssen, and Jeffery (2013), Zuiderwijk and Janssen (2013a), Zuiderwijk, Janssen, and Parnia (2013) and Zuiderwijk, Janssen, and Susha (forthcoming).

5.1 Design approach

Figure 5-1 depicts the design approach that is used in this chapter. The figure shows that the design approach has been divided into three key steps, namely 1) the development of design propositions, 2) the development of design principles, 3) the development of the artefact, including the system design, the coordination patterns and the function design of the OGD infrastructure. The first two steps provide input for the design of the OGD infrastructure, whereas the last step provides the actual OGD infrastructure design. The three steps of the design approach will be explained below.

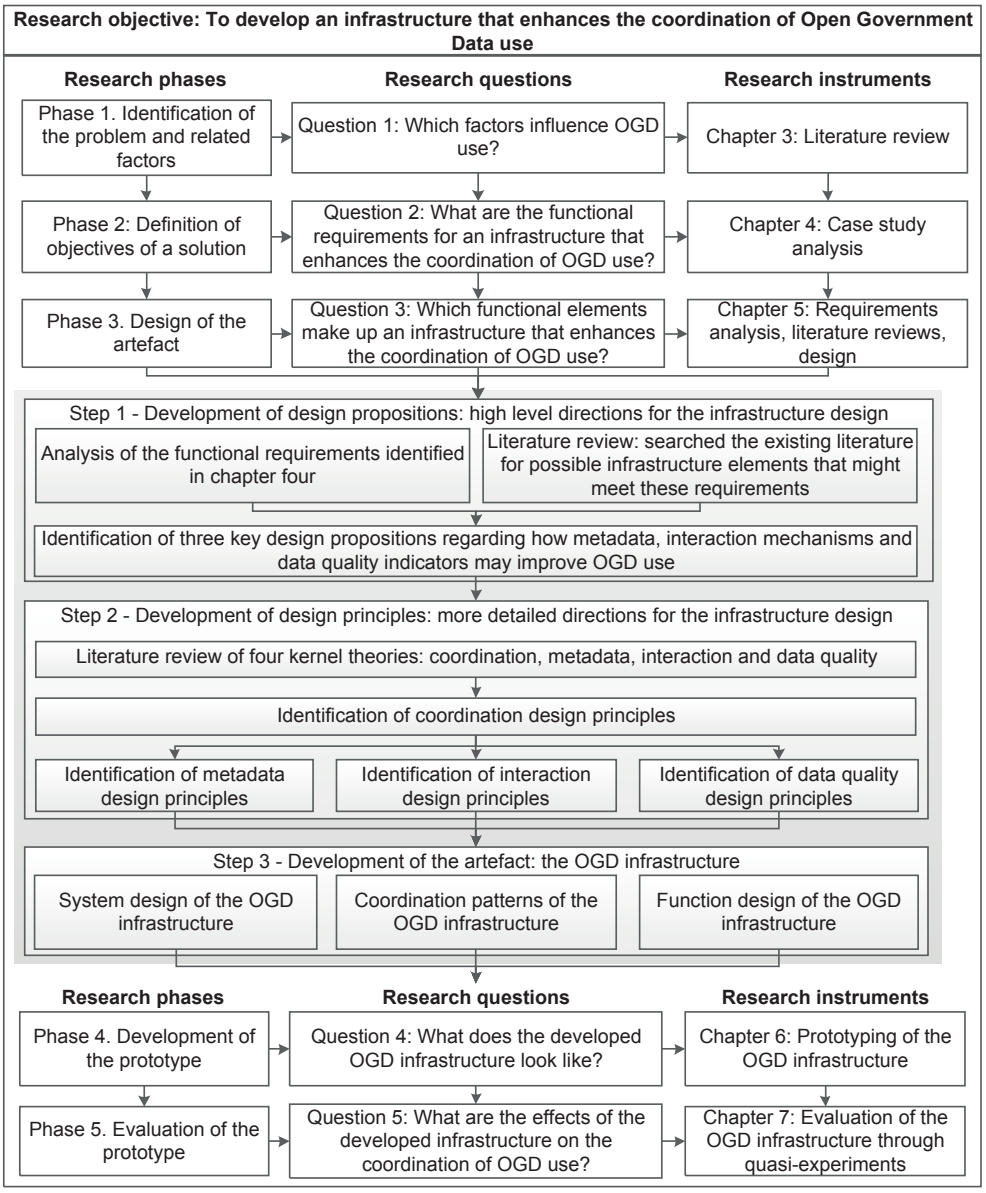


Figure 5-1: Research design including the approach for identifying and designing the functional elements of the OGD infrastructure.

5.1.1 Step 1: Development of design propositions

As a first step of the design approach we generate propositions for the design of the OGD infrastructure. A design proposition can be defined as “a general template for the creation of solutions for a particular class of field problems” (Denyer et al.,

2008, p. 395). While design propositions are not complete solutions for any business problem, on a high level of abstraction they offer input for the design of a particular solution (idem).

In order to create design propositions, we first analysed the functional infrastructure requirements that we identified in chapter four. Subsequently, we searched in the existing literature for possible functional infrastructure elements that might meet these requirements. The literature review that we performed is an extension of the literature review that was described in chapter three. We do not describe the complete approach here again, but we refer to section 3.1 for the details of the literature review. Here we only describe the differences with the literature review approach described in section 3.1. First, the motivation of the literature review was different, namely to identify propositions that provide high level input for the design of the OGD infrastructure. Second, criteria for selecting articles in the literature review were different, since these criteria included that they needed to describe information relevant to the topic of this study (i.e. functional infrastructure elements), and that the context of the references appeared appropriate for citing them in this study. Third, the conclusions drawn from the literature review are different.

From the literature review we conclude that three key functional elements may enhance the coordination of OGD use: metadata, interaction mechanisms and data quality indicators. The three functional elements were proposed based on two key criteria:

1. The functional elements needed to cover as many of our functional requirements as possible. Since time constraints would not allow us to implement a very large number of elements, the implementation of a limited number of elements covering as many functional requirements as possible would be more feasible.
2. At least basic research regarding the functional elements (in other research domains than open data) already had to be available, since it needed to be feasible to implement the elements in the OGD infrastructure in a relatively short time.

For each of the three elements, we show which functional requirements we expect to be met by them and we develop propositions. The design propositions defined in section 5.2 suggest on a high level which functional elements may be used to enhance the coordination of OGD use, and they guide the identification of design principles in section 5.3.

5.1.2 Step 2: Development of design principles

As a second step of the design approach, design principles are derived to guide the design efforts. Whereas the design propositions are defined on a relatively high level of abstraction, the design principles further develop the design input on a more detailed level. Gilb (1997, p. 165) defines principles as "rules of thumb that guide the choices and actions of engineers". Richardson, Jackson, and Dickson (1990, p. 389) state that "principles are an organisation's basic philosophies that guide the development of the architecture". Although different principle definitions exist, they have in common that design principles are normative and prescriptive, and that they give direction to the design of IS (Bharosa, 2011). We follow Bharosa (2011, p. 153), who defines design principles as "'normative' and 'directive' guidelines, formulated towards taking action by the information system architects". The design principles will provide directions for the system design and the patterns of the OGD infrastructure.

For the identification of the design principles, we extended the literature review that we started for identifying design propositions. We searched through the same databases (i.e. Scopus, JSTOR, ACM Digital Library, and Google Scholar), and again enriched the literature base by examining articles cited in the identified articles. The goal of this literature review was to identify principles that provide more detailed input for the design of the OGD infrastructure. For this goal, we added three keywords to our literature search: metadata, participation (as a commonly used term to refer to interaction in the domain of OGD), and data quality. Although the number of keywords that we added related to potential functional infrastructure elements was limited, the functional requirements elicited through the second design science research phase (see chapter 4) had already directed us towards these infrastructure elements. Our analysis of the functional requirements already suggested on a high level that metadata, interaction and data

quality might enhance the coordination of OGD use (see section 5.1.1), and therefore the focus of the extended literature review was on these keywords. The keywords also reflect our social science focus. An infrastructure is built to meet the defined requirements, but offers opportunities for new (and previously unconsidered) requirements.

Design principles were identified from four types of literature. Firstly, since we aimed to improve OGD use by enhancing coordination, we derived coordination design principles that prescribe how coordination can be enhanced from the literature on coordination. Coordination design principles are overarching and can be used for all three functional infrastructure elements (i.e. metadata, interaction mechanisms and data quality indicators). Secondly, metadata design principles, interaction design principles and data quality design principles are derived from the literature on metadata, interaction and data quality. The combination of design principles from different types of literature intends to enhance the coordination of OGD use activities. The defined design principles provide input for the development of the artefact, i.e. the OGD infrastructure design.

5.1.3 Step 3: Development of the OGD infrastructure design: the system design, coordination patterns and function design

In the third step of our design approach we develop the design of the OGD infrastructure. The infrastructure was designed in the context of our functional requirements that were derived from particular cases. The functional requirements related to the use of structured research OGD from the domains of social sciences and humanities, and they incorporate the use of these data by researchers outside the government through OGD infrastructures. The design of the infrastructure is an iterative process, and various iterations took place between the design of the OGD infrastructure and the design principles. Moreover, several case study participants were asked several times to provide feedback on different versions of the OGD infrastructure, and several iterations between the requirement analysis and the infrastructure design took place. Finally, there were iterations between the prototype creation and the design of the OGD infrastructure through various tests with potential end-users (see section 6.5).

The OGD infrastructure comprises the system design of the OGD infrastructure, the coordination patterns of the OGD infrastructure, and the function

design. The system design, coordination patterns and function design build on the design principles developed in step two of the design approach. They will be outlined for the metadata model, the interaction mechanisms and the data quality indicators. Since this study focuses on fulfilling functional requirements for an OGD infrastructure, it also incorporates the user interface. While the system design, the coordination patterns and the function design are at the core of the OGD infrastructure, the design and the development of the user interface are covered in the step that follows thereafter. As the user interface design can best be understood using visuals, the user interface is described as part of the developed prototype in chapter six (see section 6.4.2).

The system design can be defined as a "structure or behaviour-related description of a system, commonly using some formalism […] and possibly text" (Offermann et al., 2010, p. 83). Offermann et al. (2010) define a pattern as a "definition of reusable elements of design with its benefits and application context". The patterns provide insight in how the different parts of the system design are related. For instance, this encompasses the patterns of metadata standards that can be integrated or patterns of data quality indicators that need to be combined to enhance coordination of open data use. The function design refers to the design of functions. Functions have been defined as "the things that the designed object must do in order to be successful" (Dym & Little, 2004, p. 79). The function design translates the system design and patterns to concrete functionalities that can be implemented in the OGD infrastructure. The function design of the OGD infrastructure will be used for the development of the prototype in chapter six.

5.2 Design propositions

In this section the design propositions for the OGD infrastructure are developed by analysing the functional requirements for the OGD infrastructure as provided in chapter four, complemented with a literature review. A design proposition can be defined as "a general template for the creation of solutions for a particular class of field problems" (Denyer et al., 2008, p. 395). Table 5-1 summarises the functional requirements for our OGD infrastructure as identified in chapter four, and maps these to potential functional elements of the OGD infrastructure. Three key functional elements are proposed to meet the OGD infrastructure requirements,

namely metadata, interaction mechanisms and data quality indicators. The table below shows that the three proposed functional elements potentially meet all the requirements, and these elements have already been investigated in other domains than open data. While other functional elements may also meet some of these requirements, metadata, interaction mechanisms and data quality indicators are the key elements which together cover all the twenty-eight functional requirements.

OGD use category	Functional requirements for the OGD infrastructure	Functional elements for the OGD infrastructure		
		Meta-data	Inter-action mecha-nisms	Data quality indica-tors
Sear-ching for and finding OGD	1. The OGD infrastructure should be a one-stop shop for datasets and metadata from a variety of other OGD infrastructures.	X		
	2. The OGD infrastructure should allow OGD users to integrate and refer to datasets from various other OGD sources.	X		
	3. Use controlled vocabularies to describe OGD.	X		
	4. Use interoperable standards to describe OGD.	X		
	5. The OGD infrastructure should support data search through keywords, data category browsing and data querying.	X		
	6. The OGD infrastructure should support OGD use by the ability to search for data and metadata in multiple languages.	X		
	7. The OGD infrastructure should facilitate filtering, sorting, structuring and ordering relevant search results.	X		
OGD analysis	8. The OGD infrastructure should provide data which describe the dataset.	X		
	9. The OGD infrastructure should provide data about the context in which the dataset has been created.	X		
	10. It should be clear for which purpose the data have been collected.	X		
	11. It should provide examples of the context in which the data might be used.	X		
	12. Domain knowledge about how to interpret and use the data should be provided.	X		
	13. The OGD infrastructure should allow for the publication of datasets in different formats.	X		

Table 5-1: Mapping of the functional requirements from chapter four to the proposed functional OGD infrastructure elements.

OGD use category	Functional requirements for the OGD infrastructure	Functional elements for the OGD infrastructure		
		Meta-data	**Inter-action mecha-nisms**	**Data quality indica-tors**
OGD analysis	14. The OGD infrastructure should offer tools that make it possible to analyse OGD.	X		
	15. The OGD infrastructure should provide insight in the conditions for reusing the data.	X		
OGD visuali-sation	16. The OGD infrastructure should provide and integrate visualisation tools.	X		
	17. The OGD infrastructure should allow for visualising data on maps.	X		
OGD inter-action	18. The OGD infrastructure should support interaction between OGD providers, policy makers and OGD users in OGD use processes.	X	X	
	19. The OGD infrastructure should allow for conversations and discussions about released governmental data.	X	X	
	20. The OGD infrastructure should allow for viewing who used a dataset and in which way.	X	X	
	21. The OGD infrastructure should provide tools for interactive communications between OGD providers, policy makers, and OGD users (e.g. data request mechanisms and social media).	X	X	
	22. The OGD infrastructure should provide tools for interactive communications between OGD users (e.g. discussion forums and social media).	X	X	
	23. The OGD infrastructure should provide tools to keep track of amended datasets so that users know how datasets have been changed.	X	X	
OGD quality analysis	24. The OGD infrastructure should provide insight in quality dimensions of OGD.	X		X
	25. It should be possible for OGD users, OGD providers and policy makers to discuss the quality of a dataset.		X	X
	26. The OGD infrastructure should provide information on the context in which a person reused a particular dataset.	X		X
	27. The OGD infrastructure should provide quality dimensions of datasets that are comparable with other datasets and with different versions of the same dataset.	X		X
	28. It should be possible to compare the quality of datasets over different data sources, over time and over data reuse on the data infrastructure.	X		X

Table 5-1 (continued): Mapping of the functional requirements from chapter four to the proposed functional OGD infrastructure elements.

Based on the assumption that metadata, interaction mechanisms and data quality indicators can improve OGD use, three key propositions were developed. The following sections introduce these propositions.

5.2.1 Proposition 1: Metadata

Metadata are generally defined as 'data about data' (e.g., Dempsey & Heery, 1998; National Information Standards Organization, 2004; Sheth, 1999; Vardaki, Papageorgiou, & Pentaris, 2009). "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (National Information Standards Organization, 2004, p. 1). Metadata also need to describe the relationship between resources, and they characterise attributes of people, services, software components and data (Dempsey & Heery, 1998). For example, metadata can be information about the sampling subjects of a conducted survey, the method used for this survey, and the population studied (Vardaki et al., 2009).

Although scholars generally agree about this definition, there are different views on what type of data can be metadata. Jeffery (2013) and Gilliland (2008) argue that data and metadata should be defined according to the way they are used. For instance, for a library user, catalogue cards may be viewed as *metadata*, since they describe a book (e.g. the book's author) and navigate the reader towards its place (i.e. where the book can be found on the shelf). However, a librarian may use the same catalogue cards as *data*, for example, to report how many books the library contains on a certain topic. Thus, according to this definition any type of data can be metadata, yet it depends on the purpose of data use whether we actually call it metadata or data (Jeffery, 2013). Metadata can then be defined as a type of data that somehow describes internet resources for the end-user (idem).

Others have pointed at other ways to distinguish metadata and data, for instance by making a difference between which types of data can be metadata and which cannot. From this perspective, the distinction between data and metadata does not depend on the purpose for which the data are used. Proponents of the second view argue that metadata are always different from the data themselves, since metadata refer to the information that can be derived from a dataset. The

idea is that metadata differ from data in the sense that metadata provide new information that is not explicitly described in the dataset (Choenni, 2015).

In this dissertation we adopt the view that, generally, metadata can be defined as data about the data. Although for open datasets there is often a clear distinction between metadata and data, we do acknowledge that sometimes there can be an overlap. In that case, defining data as either data or metadata may depend on the purpose for which the data are used. While some types of metadata often cannot be found within the dataset itself (e.g. the creator of the dataset and the project that funded its creation), other types of metadata may be mentioned within the dataset (e.g. the variables examined in the study, the types of organisations studied, and the methods used to obtain the data). From the dataset one may then extract the needed metadata. Thus, some metadata can be derived from an open dataset, whereas other data cannot be derived in this way and needs to be collected and offered by the data provider.

Metadata has considerable potential benefits. For example, metadata facilitate the integration of data and information from heterogeneous sources (Jeffery, 2000) and they assist in the discovery of relevant data (Jeffery et al., 2014; National Information Standards Organization, 2004). Despite the many potential benefits of metadata, metadata provision for open data is often cumbersome (Martin, 2014). While the literature postulates that it is essential for the correct interpretation and use of open data to offer sufficient metadata simultaneously to data (Braunschweig et al., 2012b; Jeffery, 2000), open government initiatives in general have been criticised for providing inadequate metadata (Jurisch et al., 2015). The provision of inadequate metadata was also found for OGD initiatives in particular (Dawes, 2010; Dawes & Helbig, 2010). It has been argued that there is a "lack of consistency of metadata" and this "reduces the reusability of data" (Martin et al., 2013, p. 244). We propose metadata as a mechanism to support researchers in searching for and finding OGD, analysing OGD, visualising OGD, interacting about OGD, and analysing the quality of OGD. This leads to the following proposition.

Metadata positively influence the ease and speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD, and OGD quality analysis (P1).

Our first proposition suggests that metadata can be used to enhance the coordination of all five OGD use activities, i.e. searching for and finding OGD, analysing OGD, visualising OGD, interaction about OGD, and assessing the quality of OGD. Moreover, it suggests that metadata can improve the ease and speed of OGD use. Although successful OGD use can be measured through various aspects (e.g. satisfaction, efficiency, or effectiveness), this study focuses on the measurement of the ease and speed of OGD use, since we endorse the idea that the ease and speed of OGD use are important indicators for the successful coordination of OGD use. Moreover, the ease and speed of OGD use are interrelated with other use aspects. For instance, we expect that easier and faster use of OGD is correlated to higher user satisfaction. If OGD use would be very difficult, or if it would take considerable time to use open data, we believe that the satisfaction of OGD users will not be high. Likewise, the efficiency, effectiveness and other user aspects of OGD use are not expected to be high if ease and speed of OGD use are insufficient. Moreover, in general previous research has shown that the perceived ease of use is important. With regard to e-government services, it was found that the perceived ease of use had a positive effect on the intention to use an e-government service (Carter & Bélanger, 2005). Previous research has also shown that the ease of open data use influences the intention to use open government in general (Jurisch et al., 2015). The following sub propositions about the ease and speed of OGD use were formulated:

- *Metadata positively influence the ease of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis (P1a).*
- *Metadata positively influence the speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD, and OGD quality analysis (P1b).*

5.2.2 Proposition 2: Interaction mechanisms

The interaction between OGD providers, OGD users (i.e. researchers) and policy makers in OGD use processes may be stimulated through various mechanisms. To define the scope of the type of interaction mechanisms that we focus on in this study, we use the distinction of type of democratic engagement of Davies (2010). Davies (2010) makes a distinction between three types of democratic engagement, namely 1) formal participation in political institutions (e.g. voting, petitions), 2) participatory collaborative and/or community based action (e.g. collaboration amongst citizens and between the public and governmental agencies), and 3) individual choice (i.e. identifying citizen preferences). The first and the third type of democratic engagement are outside the scope of this study, yet the second type may be used to enhance the coordination of OGD use. This type of interaction refers to engaging citizens in the democratic processes of state forming (Parycek & Sachs, 2010; Veljković et al., 2014). Participation may then propose directions for the development of state services and guidelines (Parycek & Sachs, 2010). Electronic participation, commonly referred to as 'eParticipation' comprises citizens' participation in public service provision processes at multiple stages, including planning, decision making, implementation and evaluation (Susha & Grönlund, 2012).

Alani et al. (2008) point out that access to OGD is not enough to enhance active participation. We assume that this is also the case for interaction. Most existing OGD infrastructures do not offer interaction mechanisms. For instance, in a case on parcel data, Dawes and Helbig (2010) found that almost no feedback mechanisms exist between data users and data providers, and that "investments that users make in data improvements are not fed back into improvements in the original data sources" (p. 56). The interaction between open data providers and users in OGD processes may be stimulated through various functionalities. For example, Dawes and Helbig (2010) and Bertot et al. (2012) propose the development of formal feedback mechanisms. Garbett et al. (2011) suggest that existing social media can be used to engage people in open datasets. We propose the implementation of interaction mechanisms on the infrastructure to improve interaction about OGD. This leads to the following proposition.

Interaction mechanisms positively influence the ease and speed of interaction about OGD (P2).

This proposition again focuses on the improvement of the ease and speed of OGD use, for the reasons outlined in section 5.2.1. Ease and speed appear to be important indicators for the successful coordination of OGD use. This leads to the two following sub propositions:

- *Interaction mechanisms positively influence the ease of interaction about OGD (P2a).*
- *Interaction mechanisms positively influence the speed of interaction about OGD (P2b).*

5.2.3 Proposition 3: Data quality indicators

Open data success strongly depends on the quality of released datasets (Behkamal et al., 2014). Many different perspectives exist on the analysis and comparison of data quality (Batini et al., 2009). Data quality can refer to multiple dimensions, such as the scope (e.g. the quality of systems or processes), the system architecture (e.g. data warehousing system), and the level of abstraction (e.g. data quality metrics and meta models) (Berti-Équille, 2007). In this study the data quality scope concerns dimensions for the quality of open governmental research datasets. Open data quality dimensions include, among others, accuracy, completeness, consistency and timeliness (Batini et al., 2009). OGD reuse requires that potential data users can trust that datasets that they want to use are of sufficient quality (O'Hara, 2012). However, the quality of open data can easily be affected because of the reuse of the data (Oviedo et al., 2013). At the same time the quality of data varies widely (Auer et al., 2013; Kuk & Davies, 2011; Petychakis et al., 2014), and also depends on the purpose that one has for the reuse of an open dataset (the “fitness for use”) (Dawes, 2010). The quality of OGD may be low and open data users may be concerned about the quality of the data (Martin, 2014).

The user determines whether the quality of the data is good enough for his or her specific purposes (Van Loenen, 2006). It is therefore important that researchers using OGD can obtain more insight in the quality of OGD that they want to use. Strategies for determing the quality of data published on the Web need to be developed (Auer et al., 2013). We propose the implementation of data quality indicators on the OGD infrastructure to improve OGD quality analysis. This leads to the following proposition.

Data quality indicators positively influence the ease and speed of OGD quality analysis (P3).

The third proposition also focuses on the improvement of the ease and speed of OGD use, for the same reasons as those that apply to the first and second proposition. Ease and speed appear to be important indicators for the successful coordination of OGD use, which leads to the following sub propositions:

- *Data quality indicators positively influence the ease of OGD quality analysis (P3a).*
- *Data quality indicators positively influence the speed of OGD quality analysis (P3b).*

5.2.4 Overview of design propositions

In the previous sections we argued that metadata, interaction mechanisms and data quality indicators may enhance the coordination of OGD use by researchers. Figure 5-2 shows the three propositions that were developed. In our model, we argue that metadata support all the five OGD use activities, interaction mechanisms can be used to assist collaboration of the stakeholders involved in the open data process, and OGD quality indicators can support the generation of OGD users' trust in the dataset and in the data provider. Nevertheless, we are aware that several variables in our model may also indirectly influence each other. For instance, the ease of OGD quality analysis may also influence the ease of OGD analysis and OGD visualisation. Although this is not directly taken into account by

the model since this would not allow us to operationalise and test the propositions, the evaluations will consider these complexities by examining intermediating variables (e.g. through observations and surveys, see chapter 7).

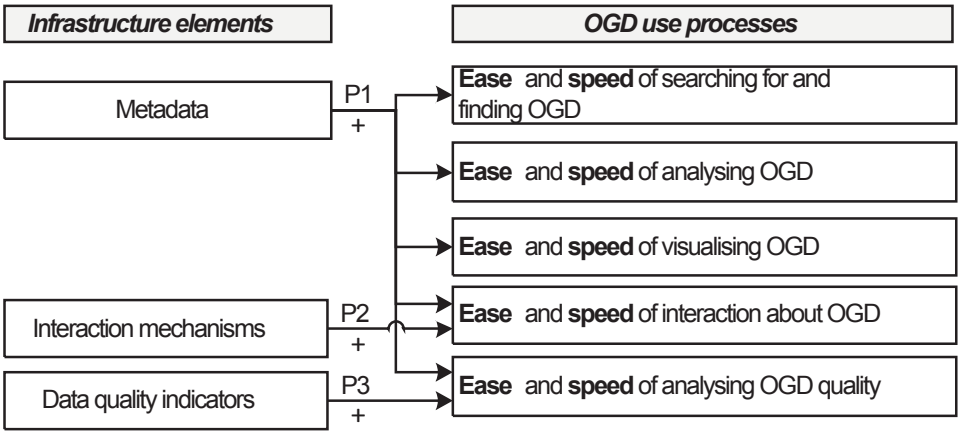


Figure 5-2: Propositions for the design of the OGD infrastructure.

In the following section we elaborate on the design propositions by identifying design principles for each design proposition.

5.3 Design principles

Section 5.2 provided the design propositions. These propositions already describe design assumptions on a high level of abstraction and can be used to identify more specific principles for the design of the OGD infrastructure. This section provides the more detailed principles that can be used for the design of the OGD infrastructure. Design principles can be defined as "'normative' and 'directive' guidelines, formulated towards taking action by the information system architects" (Bharosa, 2011, p. 153). We focus on four types of design principles derived from four kernel theories. Kernel theories are "the underlying knowledge or theory from the natural or social or design sciences that gives a basis and explanation for the design" (Gregor & Jones, 2007, p. 322). Since this study aims to improve OGD use by enhancing coordination, coordination literature is used to identify coordination design principles. Coordination design principles are overarching and can be used to enhance coordination for all three functional infrastructure elements (i.e. metadata, interaction mechanisms and data quality indicators). Moreover, metadata design principles are derived from the metadata literature. In addition,

interaction literature is used to collect interaction design principles, and finally, data quality literature is used to obtain data quality design principles. The following sections elaborate on these four types of design principles.

5.3.1 Coordination design principles

Various functional requirements identified in chapter four relate to a lack of management of dependencies between OGD use activities. For instance, in the requirements clusters of 'data heterogeneity' and 'search support', it is shown that OGD users depend on the format in which data are provided, and on the tools to find these data. When heterogeneous data are provided by governmental agencies, and when tools for using the data are not provided, this may hinder the reuse of the opened datasets by researchers.

Building on the coordination definition of Malone and Crowston (1990), we define coordination of OGD use as *the act of managing dependencies between and among activities performed to use OGD* (see section 3.2.4) Coordination theory proposes a number of mechanisms to improve the management of dependencies. These coordination mechanisms can be seen as principles for the design of our OGD infrastructure. On the basis of the work of March and Simon (1958), Thompson (1967) expounds three types of coordination mechanisms. First, *coordination by standardisation* refers to the development of routines or rules, which constrain action of each organisational part or position. This type of coordination requires an internally consistent set of rules and a stable and repetitive situation to be coordinated (Thompson, 1967). Second, *coordination by plan* requires a lower degree of stability and routines than coordination by standardisation and refers to the creation of schedules for interdependent organisational parts. These schedules may govern their actions and they are appropriate for dynamic situations, such as changing tasks (March & Simon, 1958; Thompson, 1967). Third, *coordination by mutual adjustment* is suitable for reciprocal interdependence. This type of coordination needs most communication and decisions, as it "involves the transmission of new information during the process of action" (Thompson, 1967, p. 56). Coordination by mutual adjustment is possible for variable and unpredictable situations (Thompson, 1967). March and Simon (1958) refer to this as *coordination by feedback.*

Several other scholars have described coordination mechanisms which relate to the ones mentioned by March and Simon (1958) and Thompson (1967). Galbraith (1973) mentions that certain behaviour can be determined by rules created beforehand and communicated to stakeholders, which relates to standardisation and planning. Daft and Lengel (1986) proposed seven structural mechanisms, including rules and regulations and planning. Mintzberg (1983) describes coordination mechanisms related to mutual adjustment, and standardisation of work, outputs and skills. Gittell (2002) refers to *relational coordination* with the pre-specification of tasks to be performed and the sequence in which to perform them. Another coordination mechanism identified by Gittell (2002) is the *boundary spanners* mechanism, which is the integration of work of different people. Applied to open data, the boundary spanner mechanism can be applied to requests for datasets and discussions about data use where data providers and data users can interact and 'span the boundaries' of their usual working fields. Malone et al. (1999) describe the *pull* mechanism (make to order) and the *push* mechanism (make to inventory). For instance, requests for certain datasets from OGD users to OGD providers might be seen as pull mechanisms, while the publication and integration of datasets from different infrastructures might be seen as push mechanisms. Such push and pull mechanisms may involve the use of standards or communication between individual users.

Also based on March and Simon's (1958) work, Gosain et al. (2004) argue that in an inter-enterprise setting, coordination can be attained by combining *advanced structuring* and a *dynamic adjustment* approach. Advanced structuring refers to structuring information flows and interconnected processes that exist between organisations before they take place. The advantage of this approach is that the effort related to adjusting to changing environments is reduced. Advanced structuring makes use of 'loose coupling', which means that certain elements of systems are linked (i.e. "coupled") to attain some degree of structuring, while spontaneous change may occur, leading to a certain degree of independence (i.e. "looseness"). Gosain et al. (2004) identified three aspects that advance the 'coupling' and 'looseness' in the advanced structuring approach. First, *standardisation of process and content interfaces* concerns "explicit or implicit agreement on common specifications for information exchange formats, data

repositories, and processing tasks at the interfaces between interacting supply chain partners" (Gosain et al., 2004, p. 14). Second, *modular interconnected processes*, which means "the breaking up of complex processes into sub processes (activities) that are performed by different organisations independently (such that sub processes occur through overlapping phases, or better still, fully simultaneously) with clearly specified interlinked outputs" (Gosain et al., 2004, p. 16). Third, *structured data connectivity* refers to "the ability to exchange structured transaction data and content with another enterprise in electronic form" (Gosain et al., 2004, p. 17).

The dynamic adjustment approach of Gosain et al. (2004) refers to effectively and quickly reconfiguring inter-organisational processes, so that these processes become appropriate for changed organisational environments. The reconfiguration is supported through (IT) learning and adaptation (Gosain et al., 2004). Aspects that advance the dynamic adjustment approach are 1) the *breadth of information shared* with supply chain partners, 2) the *quality of information shared* with supply chain partners and 3) *deep coordination-related knowledge*. Breadth of shared information is required to react to unexpected change, while information of high quality is needed to make effective and efficient inferences. Deep coordination-related knowledge consists of knowledge of partner competencies, process and content, organisation memory of past change episodes and understanding of causal linkages (Gosain et al., 2004).

The management of dependencies through the above-mentioned coordination mechanisms may enhance the coordination of OGD use. Various coordination mechanisms can be used as design principles for the OGD infrastructure (see Table 5-2). For instance, standardisation may be used to reduce heterogeneity of dataset terminology, and pull mechanisms may be used by OGD users to request datasets from public agencies. The table shows that coordination mechanisms cover all five types of OGD use activities. The coordination mechanisms are overarching and do not directly influence the coordination of OGD use, but aim to influence the coordination of OGD use via functional infrastructure elements (metadata model, interaction mechanisms and data quality indicators).

OGD use	Clusters of influencing factors	Coordination mechanisms	Source	Derived coordination design principles
Searching for and finding OGD data	Data fragmentation	Advanced structuring	Gosain et al. (2004)	1) Provide homogeneous access to heterogeneous data from various sources
		Planning	March and Simon (1958), Thompson (1967)	2) Plan OGD use activities
		Boundary spanners	Gittell (2002)	3) Allow for communication and interaction between OGD users, OGD providers, and policy makers
		Pull mechanisms, push mechanisms	Malone et al. (1999), Malone, Crowston, and Herman (2003)	4) Allow OGD users to request datasets from public agencies, and public agencies to push the publication of datasets
	Terminology heterogeneity	Standardisation, routines, rules	March and Simon (1958), Thompson (1967), Gosain, Lee, and Kim (2005), Daft and Lengel (1986), Mintzberg (1983) Galbraith (1973)	5) Use standards and controlled vocabularies to describe datasets
	Search support	Advanced structuring	Gosain et al. (2004)	6) Structure interfaces necessary for searching OGD
		Relational coordination	Gittell (2002), Gittell (2011)	7) Pre-define the tasks that need to be conducted to search for OGD
	Information overload	Advanced structuring	Gosain et al. (2004)	*Confirmed 6*: Structure interfaces necessary for searching OGD
		Dynamic adjustment	Gosain et al. (2004)	8) Adjust search results towards the needs of OGD users, sharing broad and high quality information
OGD analysis	Data context	Standardisation, routines, rules	March and Simon (1958), Thompson (1967), Gosain et al. (2005), Daft and Lengel (1986), Mintzberg (1983), Galbraith (1973)	9) Use standards to describe the dataset and its contextual information

Table 5-2: Coordination design principles for the design of the OGD infrastructure.

OGD use	Clusters of influencing factors	Coordination mechanisms	Source	Derived coordination design principles
OGD analysis	Data interpretation support	Standardisation, routines, rules	March and Simon (1958), Thompson (1967), Gosain et al. (2005), Daft and Lengel (1986), Mintzberg (1983), Galbraith (1973)	10) Use standards to describe the domain of the dataset
		Advanced structuring and dynamic adjustment	Gosain et al. (2004)	11) Describe the purposes for which data have been used, which can be adapted after new data use
	Data heterogeneity	Standardisation, routines, rules	March and Simon (1958), Thompson (1967), Gosain et al. (2005), Daft and Lengel (1986), Mintzberg (1983), Galbraith (1973)	12) Use interoperable standards and vocabularies to publish OGD
	Data analysis support	Advanced structuring and dynamic adjustment	Gosain et al. (2004)	13) Structure interfaces necessary for using OGD analysis tools and for viewing licenses
		Relational coordination	Gittell (2002), Gittell (2011)	14) Pre-define the tasks that need to be conducted to analyse OGD
Visualising OGD	Data visualisation support	Advanced structuring and dynamic adjustment	Gosain et al. (2004)	15) Structure the interfaces necessary for using OGD visualisation tools
		Standardisation, routines, rules	March and Simon (1958), Thompson (1967), Gosain et al. (2005), Daft and Lengel (1986), Mintzberg (1983), Galbraith (1973)	16) Use interoperable standards to integrate visualisation tools

Table 5-2 (continued): Coordination design principles for the design of the OGD infrastructure.

OGD use	Clusters of influencing factors	Coordination mechanisms	Source	Derived coordination design principles
Interaction about OGD	Lack of interaction	Advanced structuring and dynamic adjustment	Gosain et al. (2004)	17) Structure the interfaces necessary for conversations about released data and for viewing who used a dataset and in which way
		Boundary spanners	Gittell (2002)	*Confirmed 3:* Allow for communication and interaction between OGD users, OGD providers, and policy makers
	Interaction support	Advanced structuring and dynamic adjustment	Gosain et al. (2004)	18) Structure interfaces necessary for integrating social media, feedback mechanisms, and viewing dataset adjustments
		Feedback, mutual adjustment	March and Simon (1958), Thompson (1967), Mintzberg (1983)	19) Allow for discussions on data use among OGD users and for improving datasets by public agencies based on feedback from OGD users
		Boundary spanners	Gittell (2002)	*Confirmed 3:* Allow for communication and interaction between OGD users, OGD providers, and policy makers
OGD quality analysis	Dependence on the quality of open data	Advanced structuring	Gosain et al. (2004)	20) Provide functions to discuss the quality of OGD
	Poor data quality	Advanced structuring	Gosain et al. (2004)	21) Provide functions to discuss the purposes for which OGD can be used
	Quality variation and changes	Standardisation, routines, rules	March and Simon (1958), Thompson (1967), Gosain et al. (2005), Daft and Lengel (1986), Mintzberg (1983), Galbraith (1973)	22) Use standards to provide comparable quality indicators for all datasets and for data about the context of data use

Table 5-2 (continued): Coordination design principles for the design of the OGD infrastructure.

5.3.2 Metadata design principles

Table 5-3 outlines the 40 metadata design principles as identified from the literature. The metadata design principles are numbered, and the numbering continues from the numbering of the coordination design principles. The table shows considerable metadata design principles related to all the five OGD use categories that we identified in chapter three (searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis), and to each of the clusters of factors which influence OGD use within these OGD use categories. The principles show how metadata can be used in the design of the OGD infrastructure.

OGD use	Clusters of influencing factors	Derived metadata design principles	Source
Searching for and finding OGD	Data fragmentation	23) Metadata can be used to create a 'one-stop-shop' experience for data users by collecting and integrating data from various portals	Marienfeld, Schieferdecker, Tcholtchev, and Lapi (2013)
		24) Metadata facilitate the integration of data and information from heterogeneous sources	Jeffery (2000)
		25) Metadata support interoperability and legacy resource integration	National Information Standards Organization (2004)
		26) Metadata assist organising and creating order in diverse data sources	Duval et al. (2002)
	Terminology heterogeneity	27) Metadata support combining data, vocabularies and other building blocks in a syntactically and semantically interoperable way. Controlled vocabularies help to increase the precision of a description	Duval, Hodgins, Sutton, and Weibel (2002)
		28) Certain metadata standards allow for the use of controlled vocabularies	National Information Standards Organization (2004)
	Search support	29) Metadata assist in the discovery of relevant data	Schuurman, Deshpande, and Allen (2008), Qin, Ball, and Greenberg (2012), National Information Standards Organization (2004), Jeffery et al. (2014)

Table 5-3: Metadata design principles for the design of the OGD infrastructure.

OGD use	Clusters of influencing factors	Derived metadata design principles	Source
Searching for and finding OGD	Search support	30) Metadata is essential for understanding the relevance of data	Jeffery (2000)
		31) Metadata support multilinguality	Jeffery (2000)
	Information overload	32) Metadata can be used to refine queries so that they select that which the user searches for	Jeffery (2000)
		33) Metadata can help to structure the properties of unstructured data resources and can assist in automatic classification	Joorabchi and Mahdi (2011)
OGD analysis	Data context	34) Metadata can assist in describing and obtaining resources efficiently	Joorabchi and Mahdi (2011)
		35) Metadata provide insight in and understanding of its context (including quality and relevance)	Jeffery et al. (2014)
		36) Metadata are essential for data preservation	Qin et al. (2012), National Information Standards Organization (2004)
		37) High-quality and complete metadata that incorporate the context and the reference frame support the understandability of datasets for future reuse	Dawes et al. (2004)
	Data interpretation support	38) Metadata facilitate the organisation of electronic resources	National Information Standards Organization (2004)
		39) Metadata can provide end-users with additional semantics required to reconstruct the context of stored data	Vardaki and Papageorgiou (2007), Vardaki et al. (2009)
		40) Metadata are required for the user to know how the database semantics should be interpreted	Schuurman et al. (2008)
		41) Metadata facilitate distilling knowledge from information and data	Jeffery (2000)
		42) Metadata can be processed by computers using domain specific formats	Jeffery et al. (2014)
		43) Metadata can be used to identify the data properties and data quality problems	Rahm and Hai Do (2000)
		44) Data and metadata standards contribute to the ability to use data for different purposes	Dawes (2010) (p. 380)
	Data heterogeneity	45) Metadata can be described in various standards	Bailo and Jeffery (2014), Bunakov and Jeffery (2013)

Table 5-3 (continued): Metadata design principles for the design of the OGD infrastructure.

OGD use	Clusters of influencing factors	Derived metadata design principles	Source
OGD analysis	Data heterogeneity	46) Certain metadata standards allow for the use of controlled vocabularies	National Information Standards Organization (2004)
		47) Data and metadata standards contribute to data quality	Dawes (2010)
		48) Metadata allow for data of different formats to be integrated	Sen (2004)
	Data analysis support	49) Metadata can be used to integrate and establish communications between various tools	Sen (2004)
		50) Metadata can provide insight in the licenses for reusing datasets and documents	Bunakov and Jeffery (2013)
Visualising OGD	Data visualisation support	51) Metadata can support multimedia representations	Jeffery (2000)
		52) Metadata facilitate data reuse	Schuurman et al. (2008), Qin et al. (2012)
		53) Metadata pertains to be used to support data representation or visualisation	Jeffery (2000)
		54) Metadata can be used to integrate and establish communications between various tools	Sen (2004)
Interaction about OGD	Lack of interaction	55) A metadata approach can be used to facilitate access to data sources by citizens, organisations and others for secondary and further reuse	Bertot, Jaeger, Shuler, Simmons, and Grimes (2009)
		56) Provenance metadata help to derive the history of data resources from their origin	Simmhan, Plale, and Gannon (2005)
	Interaction support	57) Metadata can be used to integrate and establish communications between various tools	Sen (2004)
		58) Provenance metadata help to derive the history of data resources from their origin	Simmhan et al. (2005)
OGD quality analysis	Dependence on the quality of open data	59) Metadata assist in understanding the quality of data	Jeffery (2000)
		60) High-quality metadata is needed to assess the data quality	Dawes (2010)
	Poor data quality	61) High-quality metadata is needed to understand the nature of the data, so that they can identify the factors that determine its fitness for an intended use	Dawes (2010)
	Quality variation and changes	62) Metadata can provide insight in various quality dimensions (e.g. about completeness, accuracy, currency, explicitness and availability)	Dawes et al. (2004)

Table 5-3 (continued): Metadata design principles for the design of the OGD infrastructure.

5.3.3 Interaction design principles

In this section the principles that guide the design of interaction mechanisms in the OGD infrastructure are described. Table 5-4 depicts the fifteen interaction design principles that can be used to support interaction about OGD. The interaction design principles are numbered, and the numbering continues from the numbering of the metadata design principles.

OGD use	Clusters of influencing factors	Derived interaction design principles	Source
Interaction about OGD	Lack of interaction	63) Data access assists in citizen participation and collaboration	Parycek and Sachs (2010)
		64) A culture of participation through collaborative citizen and government networks may lead to participation in public agenda-setting and decision-making	Maier-Rabler and Huber (2011)
		65) Information-based strategies which incorporate Web 2.0 tools may facilitate online public dialogs to collect feedback, questions, and recommendations for improvements	Dawes and Helbig (2010)
		66) Information-based strategies actively encourage businesses, civic organisations, and individuals to use government information for their own purposes	Dawes and Helbig (2010)
		67) Web 2.0 facilitated feedback mechanisms allow for measuring the performance of projects related to OGD	Linders (2013)
	Interaction support	68) The integration of existing social media may facilitate the engagement of people with open data	Garbett et al. (2011)
		69) Social media and interactive communications can be used to connect people, support idea sharing, receive different types of feedback and involve people in policy-making and decision-making processes	Lee and Kwak (2012), Veljković et al. (2014), Bertot et al. (2012)
		70) Collaboration with citizens may help to improve and make administrations more efficient	Parycek and Sachs (2010)
		71) Governments can profit from the knowledge and involvement of citizens when citizens collaborate among each other	Parycek and Sachs (2010)
		72) Governments can profit from the knowledge and involvement of citizens when the government collaborates with citizens	Parycek and Sachs (2010)

Table 5-4: Interaction design principles for the design of the OGD infrastructure.

OGD use	Clusters of influencing factors	Derived interaction design principles	Source
Interaction about OGD	Interaction support	73) Collaboration amongst citizens and between citizens and governments can be beneficial to citizens since it allows for reaching common goals by working together	Parycek and Sachs (2010)
		74) The analysis of feedback from OGD users can help in improving the procedures for newly publishing or updating datasets	Kucera and Chlapek (2014), Bertot, McDermott, and Smith (2012)
		75) Users may discover and correct errors in datasets and communicate such errors and improvements to the data provider and other data users	Dawes and Helbig (2010), Bertot, McDermott, and Smith (2012)
		76) Feedback regarding errors in datasets may lead to continuous improvements to datasets of benefit to all future users of the dataset and to public agencies	Dawes and Helbig (2010)
		77) Feedback assists in engaging the public in agency operations	Bertot, McDermott, and Smith (2012)

Table 5-4 (continued): Interaction design principles for the design of the OGD infrastructure.

5.3.4 Data quality design principles

Table 5-5 shows the four data quality design principles that guide the design of the OGD infrastructure. The data quality design principles are numbered, and the numbering continues from the numbering of the interaction design principles. The data quality design principles focus on dimensions and metrics for assessing data quality levels. Batini et al. (2009) state that most information quality literature puts four quality dimensions central: 1) accuracy, 2) completeness, 3) consistency and 4) timeliness. Various other scholars have also provided frameworks for investigating the quality of information (e.g., Bharosa, 2011; Naumann & Rolker, 2000; Strong, Lee, & Wang, 1997; Zhu & Gauch, 2000). However, apart from the sources mentioned in Table 5-5 there is barely any literature on data quality dimensions and metrics for OGD in particular.

Compared to the other design principles, the data quality design principles are relatively general. The reason for this is that the quality of open datasets differs per data use purpose and per domain. For example, to identify on a high level of aggregation how many crimes occur in a certain country, it may not be problematic

that a dataset lacks detailed information of suspected offenders or of the exact types of crimes. However, in order to identify the number of suspected offenders per crime per neighbourhood it does matter whether this data is complete and accurate, and as much information as possible needs to be available. Since there are considerable differences in the data quality requirements (fitness for use), the design principles need to meet this large variety of data quality requirements and have therefore been created on a high level of abstraction.

OGD use	Clusters of influencing factors	Derived data quality design principles	Source
OGD quality analysis	Dependence on the quality of open data	78) Information about the nature of datasets and about factors that determine data quality support the assessment of data quality	Dawes (2010)
		79) Crowdsourcing can help to assess the quality of open data. The public can critically analyse datasets and assess their quality, so that the quality of open datasets can be improved	O'Hara (2012)
	Poor data quality	80) Information about fitness for an intended data use supports the assessment of data quality	Dawes (2010)
	Quality variation and changes	81) Comments from data users and providers on data quality will help to benchmark datasets against each other and to improve data quality	O'Hara (2012)

Table 5-5: Data quality design principles for the design of the OGD infrastructure.

5.4 The OGD infrastructure

Building on the design principles identified in section 5.3, this section describes the design of the OGD infrastructure. All of the 81 design principles described in the previous section of this chapter are incorporated in the infrastructure and we did not make a selection of design principles. Figure 5-3 depicts the functional design elements of the OGD infrastructure. The design of the OGD infrastructure incorporates the system design, the patterns, and the function design.

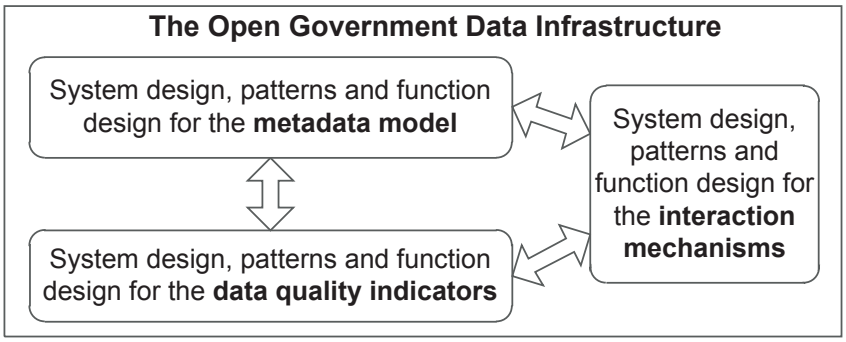


Figure 5-3: The functional design elements of the OGD infrastructure.

The OGD infrastructure was designed in collaboration with partners from the consortium of the ENGAGE-project. The ENGAGE-project was a combination of a Collaborative Project and Coordination and Support Action (CCP-CSA) funded by the European Commission under the Seventh Framework Programme. This project was led by the National Technical University of Athens (NTUA, Greece). Other partners from the ENGAGE consortium that contributed to the design of the OGD infrastructure besides NTUA and Delft University of Technology were the University of the Aegean (Greece), IBM (Israel), Intrasoft International (Luxembourgh), the Science and Technology Facilities Council (United Kingdom), Fraunhofer FOKUS (Germany), euroCRIS Current Research Information Systems (the Netherlands) and the Microsoft Innovation Center (Greece). Each partner was involved because of its particular expertise in a certain area related to OGD (e.g. technical expertise, expertise regarding requirement collection or expertise concerning system and project evaluations). In the following sections we discuss for each of the infrastructure elements which elements were created uniquely by the author of this dissertation and which were created in collaboration with partners from the ENGAGE consortium.

5.4.1 System design

The system design describes the structure and the behaviour of the system (Offermann et al., 2010). In the following sub sections, the system design will be described for the three functional infrastructure elements: the metadata model, the interaction mechanisms and the data quality indicators. The coordination design principles will be applied to these three elements and to the OGD infrastructure in general, since they cannot be used as elements themselves. For example, the

standardisation mechanisms need to be applied to other elements ('something' needs to be standardised).

Metadata model system design

To integrate the metadata design principles that were described in section 5.3.2, a metadata model was developed. The metadata model system design has been created in collaboration with all the ENGAGE-project partners, but in particular with metadata experts from euroCRIS and the Science and Technology Facilities Council. Building on research conducted by Jeffery et al. (2013), Bailo and Jeffery (2014) and Jeffery et al. (2014), the metadata model uses a three-layer structure, and includes discovery metadata, contextual metadata and domain specific metadata (see Figure 5-4). Each of the three layers will be explained in the following sections.

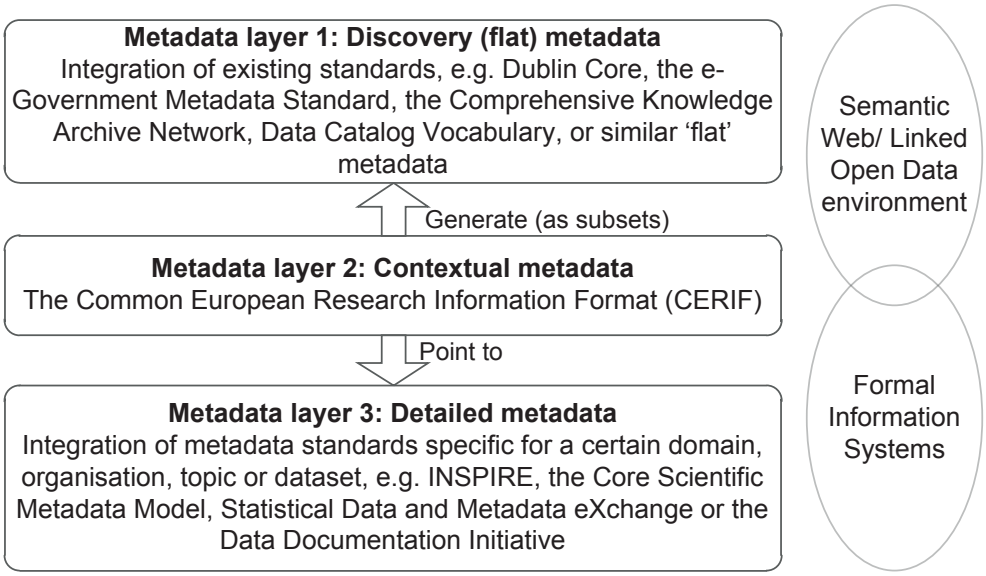


Figure 5-4: The three-layer metadata architecture (adapted from Zuiderwijk, Jeffery, et al., 2012b).

Discovery (flat) metadata layer

Discovery metadata assists in discovering relevant datasets by browsing or querying. Examples of discovery metadata include data about the identifier, title, creator, publisher, country, source, type, format, and language of the dataset. Existing standards for describing discovery metadata can be used, including Dublin

Core (DC) (Dublin Core Metadata Initiative, 2010), the e-Government Metadata Standard (e-GMS) (ESD Standards, 2004), the Comprehensive Knowledge Archive Network (CKAN) (Open Knowledge Foundation, 2007), Data Catalog Vocabulary (DCAT) (World Wide Web Consortium, 2014), or similar 'flat' metadata. The integration of existing metadata standards allows for providing homogenous access to heterogeneous datasets. The data can be offered in standards that are already typically used by many OGD providers.

The 'flat' discovery metadata standards used in the discovery layer are relatively simple, and they allow for the easy linkage of open datasets. The use of a Semantic Web or Linked Open Data (LOD) environment allows simple linkages between datasets and flexibility, and is useful for quick bottom-up linking of data sources. Any data over the web can be linked using a LOD environment (Van den Brink, Janssen, Quak, & Stoter, 2014). In this way the proposed metadata model supports the linkage and integration of open datasets from different infrastructures, and allows for the creation of a one-stop-shop to datasets and metadata.

However, only using the layer of discovery metadata has several disadvantages for researchers using OGD. First, discovery metadata insufficiently describe the relationships between data, persons, projects and other contextual aspects. The discovery metadata that are typically provided on OGD infrastructures are usually insufficient to offer OGD users adequate knowledge of the context in which a dataset has been created and the context in which it can be reused. Second, only using a Semantic Web or LOD environment comes at the cost of not being able to add constrains and therefore lacks integrity. The semantics provided with the flat metadata standards are often rudimentary, insufficiently formal, and they do not handle well multilinguality and temporal relationships. This leaves a void in data necessary for the understanding of the quality, the content and other detailed information, which are essential for OGD use. Third, Semantic Web and LOD environments use the Resource Description Framework (RDF) to link datasets, which suffers from a number of problems. RDF uses simple triples <subject><link><object> to link data, but the simple <link> cannot adequately represent temporal or geospatial relationships (Perry, Sheth, & Jain, 2011). Temporal or geospatial relationships then require multiple linked RDF statements, which makes the use of only RDF complicated for open datasets.

The Common European Research Information Format (CERIF) (EuroCRIS, 2010) facilitates the utilisation of relational database technology and it interconverts between two types of representations of contextual metadata, namely a relational representation and an RDF-representation. This makes it possible to use OGD in an integrated semantic web context and a typical information system context. For these reasons a Semantic Web / LOD environment was not used directly, and a second layer of contextual metadata using CERIF was added to the discovery metadata layer to generate the Semantic Web / LOD environment.

Contextual metadata layer

Contextual metadata allows for the provision of rich information on the context in which a dataset has been created and the context in which it can be reused. Contextual metadata includes persons, organisations, projects, publications, conditions of use (e.g. licenses), provenance, quality dimensions and many other aspects associated with the dataset. These metadata clarify for which purpose the data have been collected, by whom and various other contextual aspects, which allows for identifying for which purposes the data may be used in the future. The contextual metadata layer supports interoperability among common metadata formats used to describe OGD. This means that existing standards that are used to describe OGD can be integrated and a single point of access to the datasets described with these standards can be offered. Contextual metadata can mitigate information overload by reducing the set of relevant search results and by better determining the most relevant ones for the user. Moreover, the contextual metadata points to the detailed metadata (the third metadata layer in our architecture).

The contextual metadata layer uses the CERIF standard (EuroCRIS, 2010), which is an EU recommendation to member states. “CERIF […] is the most complete contextual metadata ‘standard’ with formal syntax and declared semantics” (Jeffery et al., 2014, p. 4) and is commonly used for data derived from research. It provides a superset exchange mechanism for common metadata formats, and offers highly structured relationships. This allows for temporally defined role-based relationships between instances of entities. CERIF also allows for the use of various controlled vocabularies, multiple languages and for

structuring data properties. By using CERIF, searching for OGD can be structured, since it facilitates the filtering and ordering of data search results. CERIF also allows for the integration of various tools, such as data analysis and visualisation tools. The contextual metadata about persons facilitates identifying OGD users, OGD providers and policy makers, as well as the purposes for which certain datasets have been used. CERIF is already adopted by various governments (e.g. the United Kingdom, Norway, Denmark, Sweden, Slovakia, Slovenia, Ireland and the Netherlands) and by European institutions including European Research Council (ERC) and European Science Foundation (ESF), which ensures a large user-base of this standard. Moreover, CERIF is maintained by euroCRIS[6], an independent organisation that ensures continuity and adoption to changing needs.

Detailed metadata layer

A third layer of detailed metadata is added to the discovery and contextual metadata, because many datasets can be described by metadata that are specific to a certain domain, organisation, topic or even to a certain dataset. Detailed metadata are metadata that are specific to a (research) domain, such as healthcare or crime, an organisation, or even to a specific dataset, and are described in a formalised way. Examples include the INSPIRE metadata for European environmental and geospatial data (European Commission, 2008; The European Parliament and the Council of the European Union, 2007), the Core Scientific Metadata Model (CSMD) for scientific research data (Matthews et al., 2010), Statistical Data and Metadata eXchange (SDMX) for statistical data (Statistical Data and Metadata eXchange, 2011), and the Data Documentation Initiative (DDI) for social science data (DDI Alliance, 2009). The layer of detailed metadata allows for the provision of such domain or dataset specific metadata in a formalised way, and provides the OGD user with additional information that can assist in the adequate interpretation of the dataset.

Interaction mechanisms system design

The interaction mechanisms system design has been created in collaboration with all the ENGAGE-project partners, but in particular in collaboration with the National Technical University of Athens, the University of the Aegean, and euroCRIS.

[6] www.eurocris.org

Section 5.3.3 provided the interaction design principles. From these design principles we identified two key interaction mechanisms: 1) feedback and 2) collaboration and discussions (see Figure 5-5). The mechanisms can be used to stimulate interaction of actors in OGD use. The interaction mechanisms developed in this study fall in Davies' (2010) democratic engagement category of participatory collaborative and/or community based action. The system design of the two interaction mechanisms is explained below.

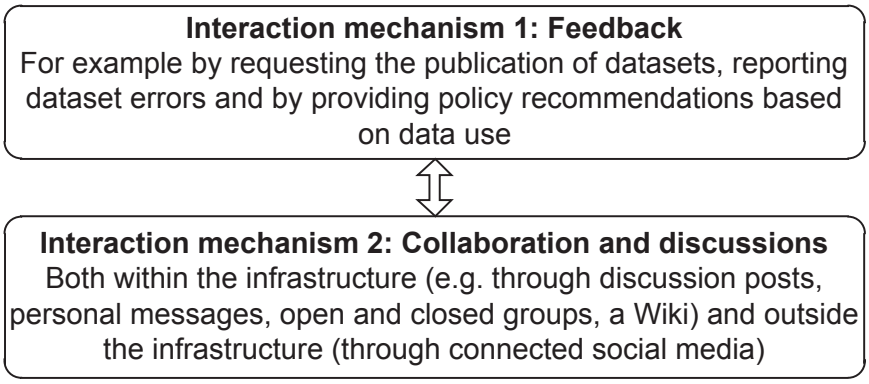


Figure 5-5: The two main interaction mechanisms of the OGD infrastructure.

Feedback

Feedback can be given about the provision of datasets by a particular governmental organisation. Researchers using OGD may request the publication of other datasets than the ones that are already available, or they may ask other OGD users for help in finding certain datasets which they suspect to be already available on the internet. Feedback may also be provided after OGD use. A researcher using OGD may find errors in the dataset and discuss these with the OGD provider through public or private messages on the infrastructure, so that the provision of useful OGD can be improved. Moreover, feedback after OGD use can be supplied in the form of policy recommendations or other feedback that is interesting for the public agency. Researchers using OGD can provide public agencies with recommendations derived from the use of datasets.

Collaboration and discussions

The infrastructure allows for collaborations and discussions between OGD users, OGD providers and policy makers, as well as among OGD users. Discussions and

collaborations can take place both within and outside the infrastructure environment. Within the infrastructure environment, collaboration between OGD users, OGD providers and policy makers can take place in public and in private groups. The private group function is offered because researchers using OGD may be reluctant to interact about OGD use when they know that everything they write is visible to anyone. The OGD users, OGD providers and policy makers who want to collaborate in groups can find each other through an overview of data provider and data user profiles (subject to trust, security and privacy). Historical information on co-operations between actors on the infrastructure is also available.

Moreover, collaboration can take many other forms. The infrastructure implements a cooperative working environment (e.g. a Wiki) to support discussions and learning from each other. Furthermore, conversations about released datasets are facilitated through open discussion forums. The infrastructure also enables the provision of converted datasets to other users of the infrastructure (e.g. data users may not know how to convert a dataset and request other users to do this for them), or the addition of metadata to a dataset (e.g. after data analysis a data user notices that certain metadata fields are incomplete and adds some of the missing metadata based on his analysis of the dataset). Researchers using OGD may also discuss with each other through public or personal messages.

Discussions do not necessarily need to take place on the OGD infrastructure itself, but can also take place outside the infrastructure. The integration of social media tools, such as Twitter, Facebook and LinkedIn, allow for interactive communications between OGD users, OGD providers and policy makers, and among OGD users. Through these means, researchers can see who used a dataset and in which way, and they can help each other by discussing experiences with specific datasets.

Data quality indicators system design

Within the ENGAGE-project, the author of this dissertation has been responsible for the creation of the data quality indicators system design, yet she collaborated with the ENGAGE-project partners for the creation of the design (particularly with the National Technical University of Athens). Section 5.3.4 described the data quality design principles. From these principles three key data quality indicators

were identified: structured OGD quality rating, free text reviews of OGD quality, and evaluator information (see Figure 5-6). These OGD quality indicators will be explained in the remainder of this section.

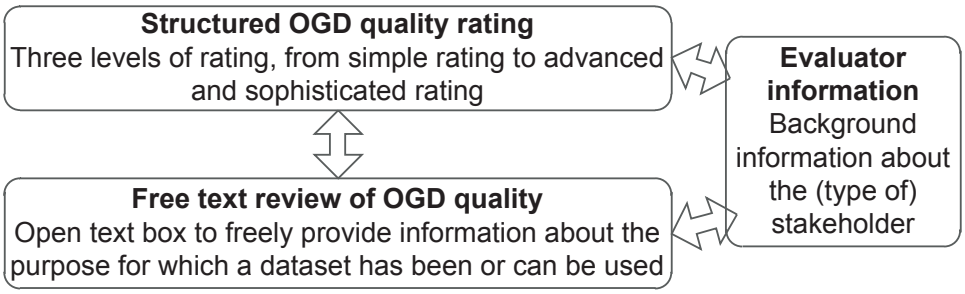


Figure 5-6: The three key data quality indicators of the OGD infrastructure.

Structured OGD quality rating

A rating system is a system in which OGD providers and users can (subjectively) assess the quality of an open dataset. Three levels of rating were designed for the OGD infrastructure.

1. First, a simple rating system is in place in which OGD users and providers can provide an overall score for the entire dataset. For instance, they can choose a rate between one and five stars. Such a simple rating system does not allow for nuances for various quality dimensions, yet it does allow OGD users and providers to quickly and easily communicate their overall assessment to other users. However, only having a simple system in place to rate a complete dataset does not allow for indicating differences in quality dimensions for the dataset. The overall score is refined by a more advanced rating system.
2. In the more advanced rating system, the OGD provider or user is asked to assess different quality dimensions of the dataset. These quality dimensions can be diverse, yet data providers and users may be more motivated to assess a limited set of dimensions. We decided to let OGD providers and users rate the four most common information quality dimensions (Batini et al., 2009):
    - Accuracy, the extent to which information represents the underlying reality;

- Completeness, the extent to which information is not missing and is of sufficient breadth and depth for task execution;
- Consistency, the extent to which information is presented in the same format;
- Timeliness, the extent to which information is sufficiently up to date for task execution.

A definition of each of the quality dimensions needs to be provided to the OGD user. Assessment of the quality dimensions is done on a scale from 0 to 10, where 0 refers to a very negative score (e.g. the dataset is totally inaccurate) and 10 refers to a very positive score (e.g. the dataset is totally accurate). A risk of this type of rating is that persons who rate the dataset interpret the quality dimensions differently. The rating can therefore be advanced by using scales that have been examined in previous research.

3. On the third level of sophisticated rating, quality evaluators are asked to assess a number of data quality dimensions. Following Lee, Strong, Kahn, and Wang (2002): the following statements need to be assessed on a scale from 0 to 10 (0= totally disagree and 10=totally agree):
   - "Free of error / Accuracy
     - This information is correct.
     - This information is incorrect (reversely coded).
     - This information is accurate.
     - This information is reliable.
   - Completeness
     - This information includes all necessary values.
     - This information is incomplete (reversely coded).
     - This information is complete.
     - This information is sufficiently complete for our needs.
     - This information covers the needs of our tasks.
     - This information has sufficient breadth and depth for our task.
   - Consistency
     - This information is consistently presented in the same format.

    - This information is not presented consistently (reversely coded).
    - This information is presented consistently.
    - This information is represented in a consistent format.
- Timeliness
    - This information is sufficiently current for our work.
    - This information is not sufficiently timely (reversely coded).
    - This information is not sufficiently current for our work (reversely coded).
    - This information is sufficiently timely.
    - This information is sufficiently up-to-date for our work" (Lee et al., 2002, pp. 143-144).

The combination of the three rating systems makes it possible for the quality evaluator to choose the system that suits him or her. A person who wants to view the rates for a certain dataset cannot only see how many people provided information about the above-mentioned data quality indicators, but also in which way these people rated the quality. The distribution of the ratings including the standard deviation as well as the means of rates for different user groups (e.g. separated for researchers as data users and for data providers) needs to be provided. Nevertheless, only having this three-level rating system in place does not allow OGD users to relate the rated quality dimensions to the purpose for which the dataset has been used, or to the background of the person who has provided the quality assessment. While the quality of a dataset may be adequate for certain reuse purposes, it may be insufficient for other purposes. Two other types of data quality indicators need to be added to the OGD infrastructure to address this.

Free text review of OGD quality

The second layer of data quality indicators comprises free text reviews of OGD quality. In this quality review, OGD users and providers can freely provide information about the purpose for which a dataset has been or can be used. Any other information relevant to the quality of a dataset that cannot be provided in the first layer of quality rating can be provided through the free text review.

Furthermore, more information about the context of the datasets can be provided in the free text review.

Evaluator information

Additionally, OGD users who rate and/or review a dataset (irrespective of the simple advanced or sophisticated rating) are asked to provide information about themselves (contextual metadata) through the infrastructure. This background information might make it easier for other OGD users to determine in which context and from which perspective an evaluator has assessed a dataset. It may also show how the dataset has been reused in the past and how it can be reused in the future. For instance, a researcher may conclude from this information that another researcher has found the quality of a particular dataset sufficient to write a scientific paper. Although this information on itself is not very useful, combined with the rating information and the free text review this may help the researcher to assess whether the quality of the dataset is adequate for his data use purposes. This type of mechanism relates to so-called recommender systems (Resnick & Varian, 1997; Ricci, Rokach, Shapira, & Kantor, 2011) and reputation systems (Resnick, Kuwabara, Zeckhauser, & Friedman, 2000; Resnick & Zeckhauser, 2002) that are used in the domain of Artificial Intelligence and Multi-Agent Systems. The following information is requested from the evaluator.

- Whether the person is a data user or a data provider.
- To which user group(s) the evaluator belongs (e.g. researcher, journalist, civil servant, entrepreneur, archivist/librarian, citizen, developer).

Although this should be done with caution, the dataset rating levels, the reviews and the evaluator information can be compared with other datasets and over time. The comparability of quality information depends on quality assessment of many individuals. The design of the data quality assessment should therefore be as user-friendly as possible, so that a large user-base can be attracted.

The three types of quality assessment mechanisms need to be combined to obtain a useful overview of the quality of open datasets. Obtaining comprehensive quality information may lower the threshold for making use of OGD, and may enhance the coordination of OGD use. Merely providing one of these

quality assessment mechanisms is not sufficient to meet the functional requirements that we described in chapter four.

5.4.2 Coordination patterns

This section discusses the coordination patterns of the designed OGD infrastructure. Whereas the author of this dissertation collaborated with the ENGAGE-project consortium to create the coordination patterns, the description of the patterns has been created by the author. The patterns define the reusable parts of the design with their benefits and an explanation of how they can be applied, and with the relation between them (Offermann et al., 2010). Figure 5-7 depicts the coordination patterns. The figure extends Figure 1-1, Figure 1-2 and Figure 3-2 that we developed throughout the previous chapters. Just like we mentioned in section 5.2.4, several variables in this model may indirectly influence each other. Although this is not directly taken into account by the model since this would not allow us to operationalise and test the propositions, the evaluations will consider these complexities by examining intermediating variables (see chapter 7). The patterns depicted in this figure are explained in the following sections.

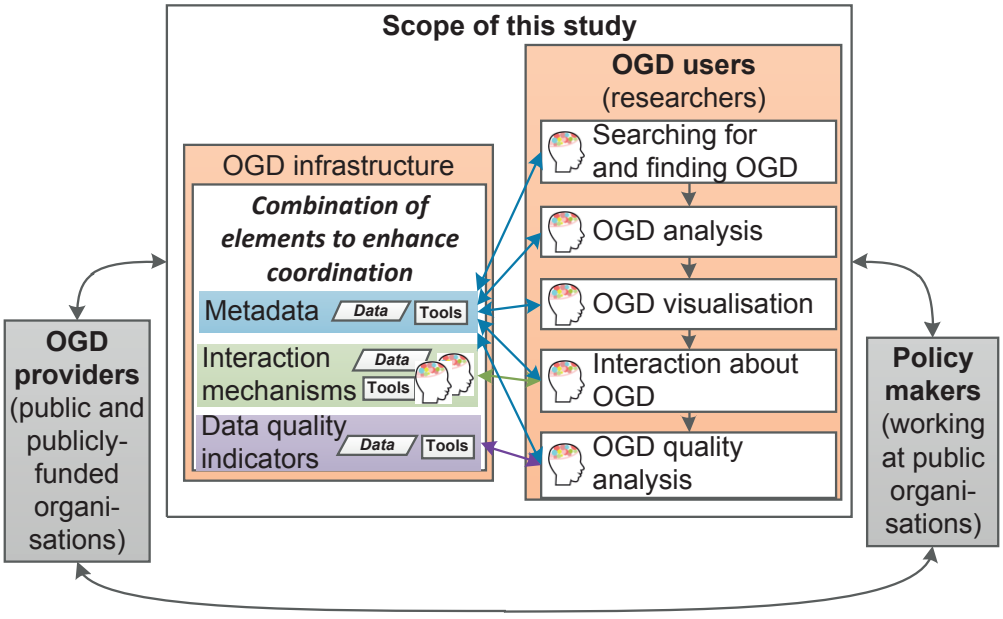


Figure 5-7: Coordination patterns: metadata, interaction and data quality elements to enhance the coordination of OGD use.

Metadata patterns

The blue dashed arrows in Figure 5-7 show that metadata can be used to improve all the five key activities conducted by researchers using OGD. Metadata allow for improving searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis by OGD use. For instance, our metadata model offers tools to link datasets to each other, so that it becomes easier to search for related datasets, and they describe the context in which the data were created, which makes it easier to analyse the data and its quality. The three-layer metadata model supports adding data relevant for discovery and about the context and details in standard formats. CERIF functions as the superset exchange mechanism for common metadata formats. From the CERIF contextual metadata layer the discovery metadata can be generated, since the flat metadata standards are subsets of CERIF or can be subsumed by CERIF. This facilitates a semantic web representation of the formalised metadata. The contextual metadata layer points at the detailed metadata layer. In this layer conventional information systems capability with structured query are maintained. Thus, the three layer architecture combines formalised metadata with easy to use RDF. The three metadata layers need to be combined to lower the threshold for using OGD by researchers. Merely providing one of these metadata layers is not sufficient to meet the functional requirements that we described in chapter four.

Interaction patterns

The green arrow in Figure 5-7 suggests that interaction mechanisms can be used to improve interaction about OGD. Interaction about OGD is facilitated through collaborations and discussions. This mechanism makes it possible for researchers using OGD to work together to attain a common goal. The infrastructure provides a location for OGD stakeholders to virtually meet each other through the infrastructure and to exchange ideas. Moreover, through the feedback mechanisms data users can provide feedback to data providers, which can also be useful to other data users when data publication processes are improved. The feedback can be used by governmental agencies to improve datasets, data publication processes, and policy-making. The two created interaction mechanisms need to be integrated to enhance the benefits that can be obtained. The integration of the

interaction mechanisms intends to support OGD users, OGD providers and policy makers to interact about OGD use and to learn from each other.

Data quality patterns

The purple arrow in Figure 5-7 points at the coordination of OGD quality analysis. The combination of different levels of structured OGD quality rating makes it possible for quality evaluators to choose the system that suits them. When the evaluator has little time, the simple rating system can be used, while rating at the more advanced and sophisticated levels can be used for more in-depth data quality analysis. The structured rating facilitates comparing the quality of datasets over different data sources, time and reuse. Moreover, whereas the structured rating supports the provision of a structured review of different OGD quality dimensions, the free text review allows for the provision of more contextual information. With the free-text evaluation, evaluators are free to describe any aspects that they believe are relevant for the assessment of the quality of particular datasets. Furthermore, the third indicator that can be helpful to assess the quality of OGD is information about the evaluator. Evaluator information intends to show in which context someone has assessed the dataset, and to make it easier for OGD users to determine in which context a dataset has been reused in the past and how it can be reused in the future. These three data quality indicators need to be combined to meet the requirements from chapter four. The structured OGD quality rating can be integrated in the CERIF metadata standard, recording the quality indicators (e.g. completeness and consistency) and the measure entities (e.g. “this information is presented consistently” assessed on a scale from 0 to 10). In addition, the OGD quality rating can be related to the information and context about the evaluator of the dataset as well as to the context of the dataset being evaluated (e.g. the free text review).

5.4.3 Function design

This section describes the function design of the OGD infrastructure. While the author of this dissertation collaborated with the ENGAGE-project consortium in describing the function design, and particularly with the National Technical University of Athens, the overview of functions below has been created by the author of this dissertation. Functions have been defined as “the things that the

designed object must do in order to be successful" (Dym & Little, 2004, p. 79). The function design translates the system design and coordination patterns to concrete functions that can be implemented in the OGD infrastructure. The functions must be determined "in order to ensure that our final design does what it is supposed to do" (Dym & Little, 2004, p. 80). Dym and Little (2004) suggest several methods to identify functions. The method used here was simply to enumerate all the functions that we identified based on the design principles. Although this method has disadvantages, the function design was refined based on the tests that we performed with OGD users (see section 6.5). Making use of the prototype in which the function design was incorporated, OGD users provided feedback on the available functions, which allowed us to complement the functions of the OGD infrastructure. In the following sections we list the identified functions related to metadata, interaction and data quality.

Metadata function design

The metadata design principles described in section 5.3.2 and the tests conducted with OGD users (see section 6.5) led to the thirty metadata infrastructure functions as depicted in Table 5-6.

OGD use	Function	Function description
Searching for and finding OGD	1) Upload dataset	Anyone can upload a dataset.
	2) Enhance metadata	Anyone can add metadata.
	3) Acquire datasets	Users can use a single point of access to acquire datasets from various OGD infrastructures. The infrastructure harvests datasets from different governmental OGD infrastructures.
	4) Acquire metadata	Users can acquire metadata. The infrastructure can harvest metadata from different governmental OGD infrastructures
	5) Retrieve data by query	Datasets can be queried through the SPARQL Protocol and RDF Query Language (SPARQL) and through the Structured Query Language (SQL).
	6) Retrieve data by facets	Facetted search is possible so that datasets can be ordered in multiple ways through filters desired by the user, e.g. they can be filtered or ordered by geospatial and temporal coordinates, the country where the data comes from, data categories (e.g. environment, finance or education), the data publisher and the dataset license. Controlled vocabularies are integrated.

Table 5-6: Function design of the metadata model.

OGD use	Function	Function description
Searching for and finding OGD	7) Retrieve data by keywords	Users can enter a simple keyword search to find datasets.
	8) Search multilingually	The infrastructure translates the keywords from the original language to various other languages, resulting in multilingual search results.
	9) Request data	Data users can request governmental agencies or other OGD users to open a certain dataset that they cannot find through the infrastructure.
OGD analysis	10) Download data	Datasets can be downloaded to the personal computers of users.
	11) Obtain a structured metadata overview	An overview of discovery, contextual and detailed metadata is visible to the user (e.g. the dataset maintainer, date of last update, dataset release date). The metadata are described following the existing standards as described in section 5.4.1.
	12) Display data services	For each dataset it is shown which processing services are available.
	13) Obtain a multilingual dataset overview	All the available information about the dataset is automatically translated to the language entered by the user.
	14) Viewing the dataset online without downloading	Datasets can be viewed and explored online without the need to download the data. Interactive views, such as the Excel Online Web Application can be used for this.
	15) Create an extension graph and manage different versions of datasets	Users can see hierarchically how an extended or derived dataset relates to the original dataset and how the original dataset was reused. When a dataset has been extended (e.g. when metadata are added to it or when additional formats of the same dataset are added), users can see a graph of the extensions, as well as the type of extension.
	16) Cleanse data	For each dataset it will be shown which services are available to cleanse datasets (e.g. using Open Refine).
	17) Convert data format	Data conversion facilities are provided to enable the creation of different file formats.
	18) Enhance metadata	After data analysis users are encouraged to supply additional metadata.
	19) Refer to data	The infrastructure refers to related or comparable datasets using enhanced metadata.
	20) Obtain license information	Metadata are provided about the license for reusing a dataset.
OGD visualisation	21) Visualise data in a table	For each dataset it will be shown which services are available to visualise datasets in tables (e.g. through the Excel Online Web Application).
	22) Visualise data in a chart	For each dataset it will be shown which services are available to visualise datasets in charts (e.g. through the Excel Online Web Application).

Table 5-6 (continued): Function design of the metadata model.

OGD use	Function	Function description
OGD visualisation	23) Visualise data on a map	For each dataset it will be shown which services are available to visualise datasets with geographical variables on maps.
Interaction about OGD	24) Register a user and create a profile	Users can register (e.g. with one of their social media accounts) and create a profile.
	25) Search through user profiles	CERIF provides the feature to provide metadata describing users (as persons with various roles, e.g. responsibilities, authorities and usage history). This allows users to find other users not just by searching for people who they already know 'in real life', but also by searching for users with a certain background. For example, a researcher may search for a developer or for a civil servant.
	26) Link data manually	Users can indicate that there is a linkage between two datasets, which can be recorded in the CERIF metadata.
	27) Follow user	Users may subscribe for following the activities conducted by another user.
	28) Follow dataset	Users may subscribe for following datasets so that they receive a notification when the dataset had been changed or updated.
	29) Obtain overview of interaction tools	For each dataset it will be shown which tools are available to provide feedback on the dataset, to discuss the dataset, and to collaborate in data use.
OGD quality analysis	30) Obtain data quality metadata	Contextual metadata is provided about the dataset, the person who created it and other contextual aspects. This allows OGD users to evaluate their confidence in the data quality and in the data provider.

Table 5-6 (continued): Function design of the metadata model.

Interaction mechanisms function design

The interaction design principles described in section 5.3.3 and the tests conducted with OGD users (see section 6.5) led to the design of the eleven interaction mechanism functions as shown in Table 5-7. Some functions overlap with the metadata functions, such as the function to request datasets.

OGD use	Function	Function description
Interaction about OGD	1) Request data	OGD users can request datasets from governmental organisations and from other OGD users
	2) Provide feedback to data providers	OGD users can provide feedback to governmental organisations and to other OGD users (e.g. concerning errors in the dataset).
	3) Provide feedback to policy makers	OGD users can provide feedback derived from the use of the dataset (e.g. policy recommendations and contributions to decision making) to other OGD users and to governmental organisations.
	4) Submit related items	Users can submit an item related to the original dataset (e.g. a publication that was written based on the dataset, a report about the data collection method or a visualisation or application of the dataset).
	5) Write a message to discuss data or data use	Users can post a message to discuss a dataset or to discuss conclusions based on data use (e.g. users can describe how they used a dataset and what they learned from this). For each message it is visible who posted it.
	6) Write a personal message	Users of the infrastructure can send each other personal messages that are delivered in the form of e-mails.
	7) Obtain community overview	Users of the infrastructure can obtain an overview of all the users registered on the OGD infrastructure. The profiles of OGD providers, OGD users and policy makers can be searched, e.g. by keyword, pre-defined organisations or user group.
	8) Enter an open collaboration group	Open collaboration groups are accessible to anyone. A user of the infrastructure can enter an open collaboration group to collaborate with other users on the analysis of a specific dataset or a project.
	9) Enter a closed collaboration group	Closed collaboration groups can be accessed only by a selection of infrastructure users. The closed groups can be used to collaborate with other users.
	10) Post Wiki articles	Users can post articles about open data use in general (so not related to particular dataset) on a Wiki. For example, the Wiki contains documentation and tutorials about how the infrastructure can be used to visualise and curate datasets.
	11) Share data or data use findings on social media	Users can share a dataset or findings from data use via social media (e.g. Twitter, Facebook, LinkedIn). Social media are integrated in the OGD infrastructure to allow for building online networks of OGD providers, OGD users and policy makers.

Table 5-7: Function design of the interaction mechanisms.

Data quality indicators function design

The data quality design principles described in section 5.3.4 and the tests conducted with OGD users (see section 6.5) resulted in the five functions as described in Table 5-8.

OGD use	Function	Function description
OGD quality analysis	1) Assess or examine structured data quality ratings	Users can assess or examine the quality of a dataset on pre-defined quality dimensions, using the three level structure from section 5.4.1.
	2) Obtain an overview of the distribution of ratings	Users can obtain information about how the quality ratings of the dataset are distributed.
	3) Write a free text review of the data quality	Users can discuss or they can view a discussing on the quality of a dataset. Users can write a review and describe the purpose for which the dataset was used.
	4) Analyse the data quality	Users can compare different quality ratings and reviews.
	5) Obtain quality evaluator information	A selection of background information about the evaluator of the data quality was visible to all users of the infrastructure.

Table 5-8: Function design of the data quality indicators.

5.5 Summary: overview of functional infrastructure elements and answer to the third research question

In this chapter we answered the third research question: *which functional elements make up an infrastructure that enhances the coordination of OGD use?* The design of the functional elements builds on the functional requirements elicited through the case studies (see chapter 4) that focused on the use of structured research OGD from the domains of social sciences and humanities, incorporating the use of these data by researchers outside the government through OGD infrastructures.

In this chapter we first analysed the functional infrastructure requirements from chapter four. Assisted by a literature review of potential infrastructure elements of the OGD infrastructure, we mapped the functional requirements to the functional elements. The analysis of the requirements and the literature suggested that metadata, interaction mechanisms and data quality indicators can enhance the coordination of OGD use by researchers. Moreover, it was found that to-date metadata, interaction mechanisms and quality indicators are only limitedly provided

by OGD infrastructures, and there is no infrastructure that combines these three functional elements. The findings led to the development of three propositions:

- Metadata positively influence the ease and speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis.
- Interaction mechanisms positively influence the ease and speed of interaction about OGD.
- Data quality indicators positively influence the ease and speed of OGD quality analysis.

The design propositions provided high-level guidance regarding what the OGD infrastructure should look like. As a second step of our design approach, we identified design principles from the three propositions, which provided more detailed guidance for the design of the OGD infrastructure. Since we aimed to improve OGD use by enhancing coordination, design principles were both derived from the literature regarding coordination theory and from the literature concerning metadata, interaction mechanisms and data quality indicators.

Subsequently, in the third step of our design approach, the OGD infrastructure was described, which included the system design, coordination patterns and the function design. The OGD infrastructure was developed through several iterations between the case study analysis and the design principle analysis. With regard to the system design, a three-tier metadata model was developed incorporating discovery metadata, contextual metadata and detailed metadata. Two types of interaction mechanisms were designed, namely feedback mechanisms and collaboration and discussion mechanisms. Then a data quality indicator model was developed which incorporated structured data quality rating, free text quality reviews and evaluator information. With regard to the coordination patterns, it was explained how the functional elements of the OGD infrastructure together enhance coordination of OGD use. Finally, the function design was outlined. Table 5-9 summarises the 46 identified functions of the OGD infrastructure. Most of these functions are selected for the development of the prototype in chapter six.

	Metadata functions	Interaction functions	Data quality functions
Searching for and finding OGD	1) Upload dataset		
	2) Enhance metadata		
	3) Acquire datasets		
	4) Acquire metadata		
	5) Retrieve data by query		
	6) Retrieve data by facets		
	7) Retrieve data by keywords		
	8) Search multilingually		
	9) Request data		
OGD analysis	10) Download data		
	11) Obtain a structured metadata overview		
	12) Display data services		
	13) Obtain a multilingual dataset overview		
	14) Viewing the dataset online without downloading		
	15) Create an extension graph and manage different versions of datasets		
	16) Cleanse data		
	17) Convert data format		
	18) Enhance metadata		
	19) Refer to data		
	20) Obtain license information		
OGD visualisa-tion	21) Visualise data in a table		
	22) Visualise data in a chart		
	23) Visualise data on a map		
OGD inter-action	24) Register a user and create a profile	1) Request data	
	25) Search through user profiles	2) Provide feedback to data providers	
	26) Link data manually	3) Provide feedback to policy makers	
	27) Follow user	4) Submit related items	
	28) Follow dataset	5) Write a message to discuss data or data use	

Table 5-9: Function design of the OGD infrastructure.

	Metadata functions	Interaction functions	Data quality functions
Inter-action about OGD	29) Obtain overview of interaction tools	6) Write a personal message	
		7) Obtain community overview	
		8) Enter an open collaboration group	
		9) Enter a closed collaboration group	
		10) Post Wiki articles	
		11) Share data or data use findings on social media	
OGD quality analysis	30) Obtain data quality metadata		1) Assess or examine structured data quality ratings
			2) Obtain an overview of the distribution of ratings
			3) Write a free text review of the data quality
			4) Analyse the data quality
			5) Obtain quality evaluator information

Table 5-9 (continued): Function design of the OGD infrastructure.

6.Prototype of the OGD infrastructure

This chapter addresses the fourth research phase of our study, namely the development of the prototype. It aims to answer the fourth research question: *What does the developed OGD infrastructure look like?* Prototyping is used as the major research instrument in this chapter. The first section of this chapter discusses the approach used for the development of the prototype. Thereafter, the prototyping objectives are described, followed by an overview of the selected prototype functions. Subsequently, the construction of the prototype and the prototype testing are discussed. This final part of the chapter also discusses the various iterations that took place in the development of the prototype. The chapter concludes with a summary of the prototyping phases and the answer to the fourth research question. Parts of this chapter have been published in Alexopoulos, Loukis, Charalabidis, and Zuiderwijk (2013), Zuiderwijk, Janssen, and Parnia (2013), Zuiderwijk, Janssen, and Jeffery (2013), Alexopoulos, Zuiderwijk, Charalabidis, Loukis, and Janssen (2014), and Zuiderwijk et al. (forthcoming).

6.1 Prototyping approach

This section provides the prototyping approach that is used in this chapter. Prototyping refers to building a working version of various aspects of a system (Bernstein, 1996). While prototyping may have disadvantages (Mason & Carey, 1983; Pliskin & Shoval, 1987), consensus is growing that in certain situations prototypes can be effective for application development methodologies (Mason & Carey, 1983). As stated by Martin (2003, p. 573) “prototyping […] can add significant value to the application system development process” and can be beneficial for the development of systems for various reasons. First, prototypes allow for understanding aspects, risks and costs of the system which might have remained unknown without prototype models (Martin, 2003). Second, prototypes express feasibilities, weaknesses and system requirements in a way that is understandable for end-users, and prototypes enable clear communication between the developers and the users of a system (Bernstein, 1996; Martin, 2003).

The provision of a realistic view of a system makes it possible for users to relate what they see to their needs (Mason & Carey, 1983). Furthermore, prototypes make it possible to involve users early in the development of the system and to detect errors and correct the system before the initial system delivery (Martin, 2003). Finally, software prototypes can be used as validation tools (Ince & Hekmatpour, 1987, p. 8), as they allow for continuously testing and validating and for the evolvement of user requirements (Bernstein, 1996; Martin, 2003).

The approach used to develop the prototype is depicted in Figure 6-1. In line with our OGD use focus (see section 1.2), the developed prototype focused on the use of structured research OGD from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures. Following Ince and Hekmatpour (1987), the development of the prototype in this study encompassed 1) defining the objectives of the prototype, 2) selecting the functions of the prototype, 3) constructing the prototype and, 4) testing the prototype. Although these phases are described separately and linearly in this chapter, much iteration between the prototyping phases took place.

First, the prototyping objectives were defined. The prototype needed to allow for refining and detailing the user requirements, as well as for measuring the effects of the designed OGD infrastructure. Since the prototype needed to be modified multiple times to refine the functional user requirements, we used an evolutionary prototyping approach, which is further explained in section 6.2. Second, the functions implemented in the prototype were defined. Functions were selected from the functions designed in chapter five and described in section 6.3. Third, the construction of the prototype took place and this is described in section 6.4. The programmes used to develop the prototype and the user interface are outlined. We refer to the prototype with the term ‘ENGAGE’. Finally, in the fourth prototyping phase the prototype was tested through alpha and beta tests. The tests clarify the weaknesses and strengths of the prototype, and show to which extent the functional user requirements are met. Multiple versions of the prototype were developed and released. The evolutions are described in section 6.5. The final version of the prototype was used for the evaluation that is described in chapter seven. In the following sub sections of this chapter we describe each of the prototyping phases and their results.

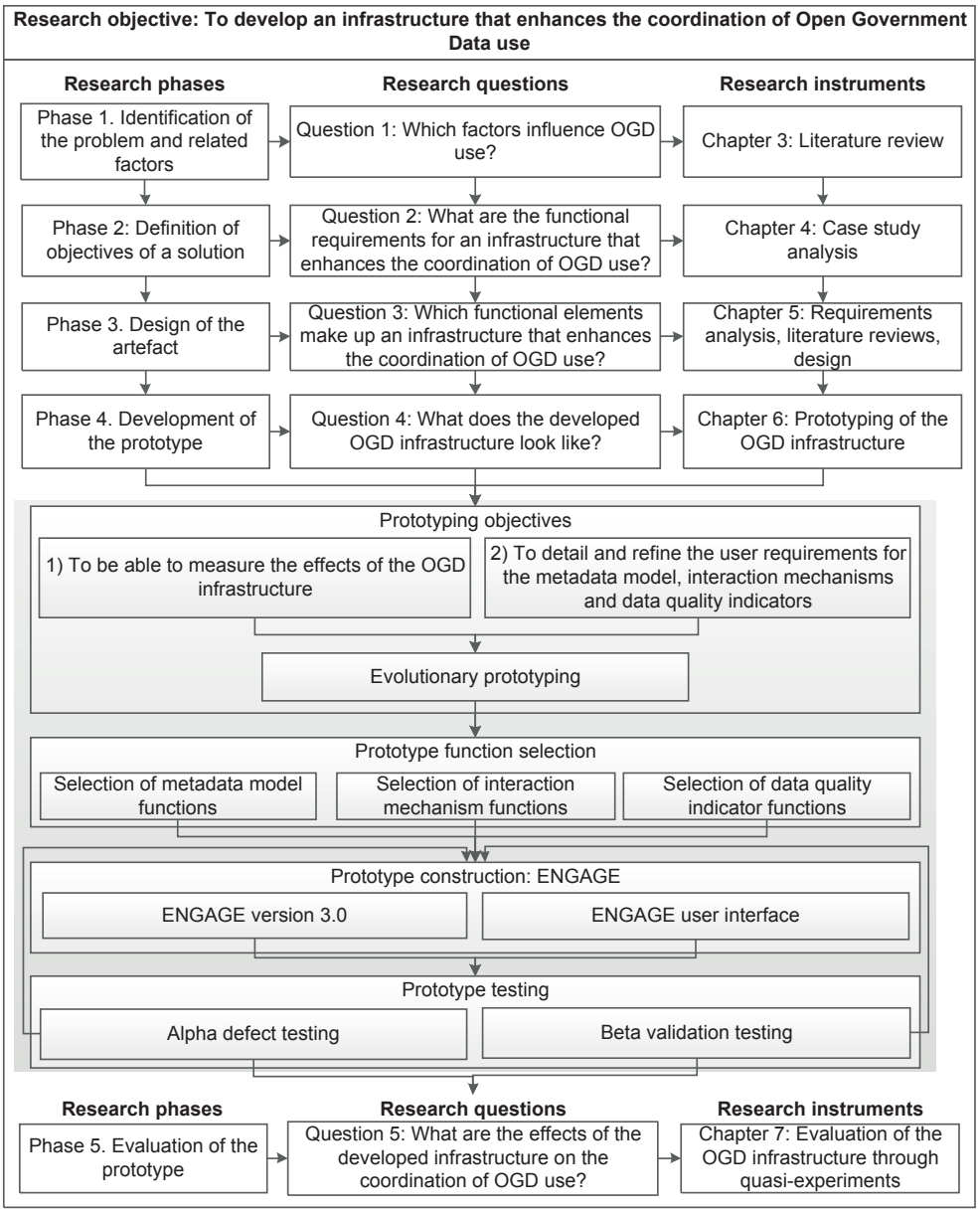


Figure 6-1: Research design including the prototyping approach.

6.2 Prototyping objectives

The first prototyping stage comprises identifying what exactly the prototype aims to achieve (Ince & Hekmatpour, 1987). Ince and Hekmatpour (1987) make a distinction between three types of software prototyping: throw-it-away prototyping,

incremental prototyping and evolutionary prototyping. Throw-it-away prototyping, also referred to as non-operational prototyping, refers to creating an early version of the software system while requirements are still being gathered, and can be used to further specify requirements (Ince & Hekmatpour, 1987; Martin, 2003). Incremental prototyping involves gradually developing a prototype and designing different parts of the system in different phases. Evolutionary prototyping is equal to incremental prototyping in the sense that both forms are constructed gradually. The main difference is that evolutionary prototyping allows for evolving the design of the system throughout its use. Thus, the system is modified during its use (Ince & Hekmatpour, 1987). Martin and Carey (1991) and Martin (2003) refer to incremental and evolutionary prototyping as iterative prototyping, "a series of evolutionary changes based upon user feedback" (Martin, 2003, p. 567), where the user feedback provides input for changes that lead to the final system (Martin & Carey, 1991). The type of prototype determines which development method is appropriate and how many resources are used to develop the prototype.

The type of prototype to be created depends on the prototyping objectives. The objective of the prototype development in this study was twofold. First, the prototype was used to refine and detail the functional user requirements regarding the metadata model, the interaction mechanisms and the data quality indicators. In the testing phase many iterations took place in consultation with potential end-users of the prototype (see section 6.5). Second, since we wanted to evaluate the effects of the OGD infrastructure in a realistic setting, and in practice there were no examples of OGD infrastructures which contained a combination of the designed metadata model, interaction mechanisms and data quality indicators, the prototype was developed to be able to measure the effects of the designed OGD infrastructure. Without the development of the prototype it would not be possible to evaluate the effects of the designed OGD infrastructure in a realistic setting.

Since the prototype needed to allow for refining and detailing the user requirements, as well as for improving the design of the prototype throughout its use, evolutionary prototyping was selected as the appropriate prototyping approach. Evolutionary prototyping allowed for gradually developing the prototype and for the evolvement of the OGD infrastructure design throughout its use. The gradual development facilitated the detection of errors and their correction early in

the development process. Applying evolutionary prototyping intended to provide a working version of the prototype that contained as few errors as possible and could be used widely by OGD users in the evaluations. The evolutions that took place are described in section 6.5.

6.3 Prototype function selection

The second phase of prototype development involves the selection of functions that needed to be prototyped (Ince & Hekmatpour, 1987). Prototype functions were selected based on a number of criteria. The main selection criterion was that the functions needed to be measurable – the functions needed to allow for measuring the key effects of metadata, interaction mechanisms and data quality indicators on the coordination of OGD use. Moreover, they needed to be measurable in the limited time frame of the evaluations. Another key criterion was that the functions needed to allow for refining the user requirements for the three functional infrastructure elements, which means that the functions needed to be testable in the limited time frame of the alpha and beta tests that we wanted to conduct (see section 6.5).

Out of the 46 functions described in chapter five, 40 were selected to be implemented. Six functions were not implemented because it would be too time-consuming to use them within the limited time frame of the evaluations. These 6 functions were not central to the five OGD use activities of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. Without these functions, the five OGD use activities could still be evaluated. In this section the selected functions will be described for the metadata model, the interaction mechanisms and the data quality indicators.

6.3.1 Metadata model functions

Table 6-1 demonstrates the metadata functions that were selected for implementation in the prototype. Almost all functions described in chapter five were selected, except for ‘convert data format’, ‘refer to data’ and ‘link data manually’. These three functions were not selected for implementation because using them in the evaluations would be too time-consuming and these three functions are not central to the five OGD use activities of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis.

Without the four functions the key tasks of OGD use by researchers could still be evaluated.

OGD use	Infrastructure function	Function description
Searching for and finding OGD	Upload dataset	Anyone can upload a dataset.
	Enhance metadata	Anyone can add metadata.
	Acquire datasets	Users can use a single point of access to acquire datasets from various OGD infrastructures. The infrastructure harvests datasets from different governmental OGD infrastructures.
	Acquire metadata	Users can acquire metadata. The infrastructure can harvest metadata from different governmental OGD infrastructures.
	Retrieve data by query	Datasets can be queried through the SPARQL Protocol and RDF Query Language (SPARQL) and through the Structured Query Language (SQL).
	Retrieve data by facets	Facetted search is possible so that datasets can be ordered in multiple ways through filters desired by the user, e.g. they can be filtered or ordered by geospatial and temporal coordinates, the country where the data comes from, data categories (e.g. environment, finance or education), the data publisher and the dataset license. Controlled vocabularies are integrated.
	Retrieve data by keywords	Users can enter a simple keyword to find datasets.
	Search multilingually	The infrastructure translates the keywords from the original language to various other languages, resulting in multilingual search results.
	Request data	Data users can request governmental agencies or other OGD users to open a certain dataset that they cannot find through the infrastructure.
OGD analysis	Download data	Datasets can be downloaded to the personal computers of users.
	Obtain a structured metadata overview	An overview of discovery, contextual and detailed metadata is visible to the user (e.g. the dataset maintainer, date of last update, dataset release date). The metadata are described following the existing standards as described in section 5.4.1.
	Display data services	For each dataset it is shown which processing services are available.
	Obtain a multilingual dataset overview	All the available information about the dataset is automatically translated to the language entered by the user.
	Viewing the dataset online without downloading	Datasets can be viewed and explored online without the need to download the data. Interactive views, such as the Excel Online Web Application can be used for this.

Table 6-1: Metadata model functions selected for implementation in the prototype.

OGD use	Infrastructure function	Function description
OGD analysis	Create an extension graph and manage different versions of datasets	Users can see hierarchically how an extended or derived dataset relates to the original dataset and how the original dataset was reused. When a dataset has been extended (e.g. when metadata are added to it or when additional formats of the same dataset are added), users can see a graph of the extensions, as well as the type of extension.
	Cleanse data	For each dataset it will be shown which services are available to cleanse datasets (e.g. using Open Refine).
	Enhance metadata	After data analysis users are encouraged to supply additional metadata.
	Obtain license information	Metadata are provided about the license for reusing a dataset.
OGD visualisation	Visualise data in a table	For each dataset it will be shown which services are available to visualise datasets in tables (e.g. through the Excel Online Web Application).
	Visualise data in a chart	For each dataset it will be shown which services are available to visualise datasets in charts (e.g. through the Excel Online Web Application).
	Visualise data on a map	For each dataset it will be shown which services are available to visualise datasets with geographical variables on maps.
interaction about OGD	Register a user and create a profile	Users can register (e.g. with one of their social media accounts) and create a profile.
	Search through user profiles	CERIF provides the feature to provide metadata describing users.
	Follow user	Users may subscribe for following the activities conducted by another user.
	Follow dataset	Users may subscribe for following datasets so that they receive a notification when the dataset had been changed or updated.
	Obtain overview of interaction tools	For each dataset it will be shown which tools are available to provide feedback on the dataset, to discuss the dataset, and to collaborate in data use.
OGD quality analysis	Obtain data quality metadata	Contextual metadata is provided about the dataset, the person who created it and other contextual aspects. This allows OGD users to evaluate their confidence in the data quality and in the data provider.

Table 6-1 (continued): Metadata model functions selected for implementation in the prototype.

6.3.2 Interaction mechanism functions

Table 6-2 shows which interaction mechanism functions were selected for implementation in the prototype. Two functions described in chapter five were not selected, namely ‘enter an open collaboration group’ and ‘enter a closed collaboration group’, because it would be too time-consuming to develop these functions and to use them within the limited time frame of the evaluations.

OGD use	Infrastructure functions	Function description
Interaction mechanisms	Request data	OGD users can request datasets from governmental organisations and from other OGD users
	Provide feedback to data providers	OGD users can provide feedback to governmental organisations and to other OGD users (e.g. concerning errors in the dataset).
	Provide feedback to policy makers	OGD users can provide feedback derived from the use of the dataset (e.g. policy recommendations and contributions to decision making) to other OGD users and to governmental organisations.
	Submit related items	Users can submit an item related to the original dataset (e.g. a publication that was written based on the dataset, a report about the data collection method or a visualisation or application of the dataset).
	Write a message to discuss data or data use	Users can post a message to discuss a dataset or to discuss conclusions based on data use (e.g. users can describe how they used a dataset and what they learned from this). For each message it is visible who posted it.
	Write a personal message	Users of the infrastructure can send each other personal messages that are delivered in the form of e-mails.
	Obtain community overview	Users of the infrastructure can obtain an overview of all the users registered on the OGD infrastructure. The profiles of OGD providers, OGD users and policy makers can be searched, e.g. by keyword, pre-defined organisations or user group.
	Post Wiki articles	Users can post articles about open data use in general (so not related to particular dataset) on a Wiki. For example, the Wiki contains documentation and tutorials about how the infrastructure can be used to visualise and curate datasets.
	Share data or data use findings on social media	Users can share a dataset or findings from data use via social media (e.g. Twitter, Facebook, LinkedIn). Social media are integrated in the OGD infrastructure to allow for building online networks of OGD providers, OGD users and policy makers.

Table 6-2: Interaction mechanism functions selected for implementation in the prototype.

6.3.3 Data quality indicator functions

Out of the data quality indicator functions described in chapter five, the comparison of different quality ratings and reviews was the only function that was not implemented in the prototype. The use of this function would require users to compare different quality ratings and reviews, yet this activity depends on a thorough analysis of datasets and would be too time consuming to test in our evaluations. Moreover, the other functions still made it possible to analyse the quality of OGD. Table 6-3 depicts the implemented data quality indicator functions.

OGD use	Infrastructure function	Function description
OGD quality indicators	Assess or examine structured data quality ratings	Users can assess or examine the quality of a dataset on pre-defined quality dimensions, using the three level structure from section 5.4.1.
	Obtain an overview of the distribution of ratings	Users can obtain information about how the quality ratings of the dataset are distributed.
	Write a free text review of the data quality	Users can discuss or they can view a discussing on the quality of a dataset. Users can write a review and describe the purpose for which the dataset was used.
	Obtain quality evaluator information	A selection of background information about the evaluator of the data quality was visible to all users of the infrastructure.

Table 6-3: Data quality indicator functions selected for implementation in the prototype.

6.4 Prototype construction: ENGAGE

The third prototyping phase concerns the development required to produce the prototype (Ince & Hekmatpour, 1987). First, a description of the final version of the constructed prototype is provided, and, second, the user interface is discussed.

6.4.1 ENGAGE version 3.0

As it was decided to apply evolutionary prototyping, considerable effort was put in the development of the prototype. The prototype was constructed as part of the ENGAGE-project, which was a combination of a Collaborative Project and Coordination and Support Action (CCP-CSA) funded by the European Commission under the Seventh Framework Programme. In this project The National Technical University of Athens was the main responsible partner for constructing the prototype, based on the requirements that have been described in this study. The prototype was called 'ENGAGE', which refers to its functions related to engaging end-users. Since the ENGAGE prototype has been developed evolutionary, several versions of the prototype have been created. The development period of the prototype was approximately two and a half years. Initial requirements for the prototype were collected, and these requirements were refined in the prototype testing phase (see section 6.5). The final version of the prototype was referred to as ENGAGE 3.0.

In ENGAGE 3.0, a four-tier architecture was applied, including a user interface layer, a presentation logic layer, a business logic layer and a data access

layer. The user interface layer contained the user interface components for the external interfaces, and was used for the communication between end-users and the rest of the system. The presentation logic layer supported workflows for user activities on the ENGAGE prototype and the provision of meaningful information to users of the prototype. In the business logic layer business logic decisions, data processing and process scheduling were enabled, while the data access layer provided access to stored data underlying the user activities.

Moreover, the three-tier metadata scheme as described in chapter five was implemented (see Figure 6-2). The first level provided discovery metadata by providing a superset of Dublin Core (DC) and the Comprehensive Knowledge Archive Network (CKAN), which assisted basic searches for datasets using a limited and easy to learn vocabulary. The advantage of using these standards is that they are used by most existing open data infrastructures and therefore contribute to the interoperability of these infrastructures. Yet, since discovery metadata do not very well capture semantic interrelations among entities (Argyzoudis, Mouzakitis, Yaeli, & Glikman, 2014), contextual metadata were provided at the second level. These CERIF metadata captured the semantic relationships of datasets with other datasets, of datasets with other entities (e.g. persons, organisations and documents), and of dataset classifications. The contextual metadata allowed for obtaining information about the context in which a dataset was created and its provenance, purpose and coverage. The third level comprised domain metadata, which refer to metadata standards for data from certain types or domains. Domain metadata can be used for services tied to domain specific activity and tools. The superset of DC and CKAN metadata was enriched with catalogue metadata fields and vocabularies from the DDI metadata standard.

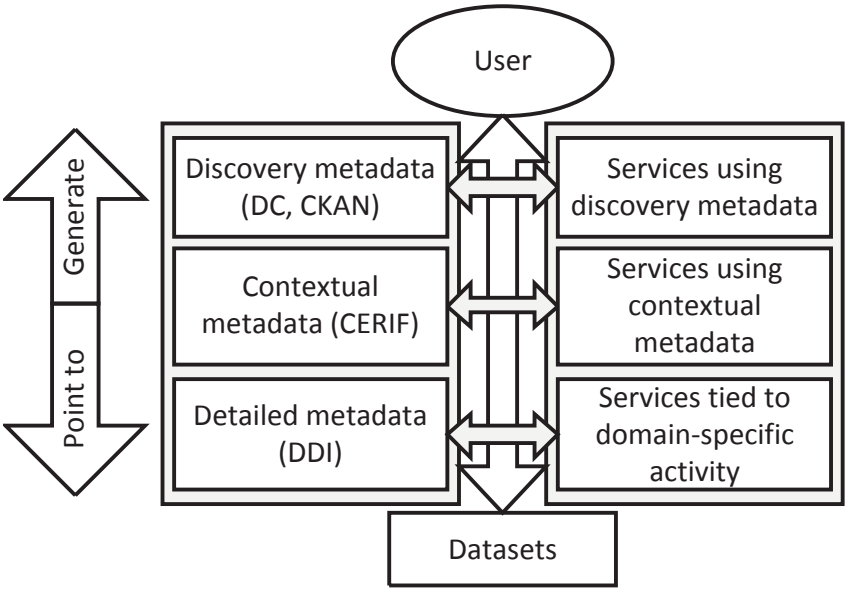


Figure 6-2: The ENGAGE multilevel metadata approach (based on Argyzoudis et al., 2014; Zuiderwijk, Jeffery, et al., 2012b)

At first a relatively limited amount of discovery, contextual and domain metadata was implemented in the prototype. Iterative steps of prototype testing with potential users of the system (see section 6.5) and prototype construction took place, which allowed for further refinement of the prototype. Based on the iterative testing and construction phases, more metadata were added to the initial selection of metadata. Figure 6-3 shows the functional metadata elements that were incorporated in the final version of the prototype.

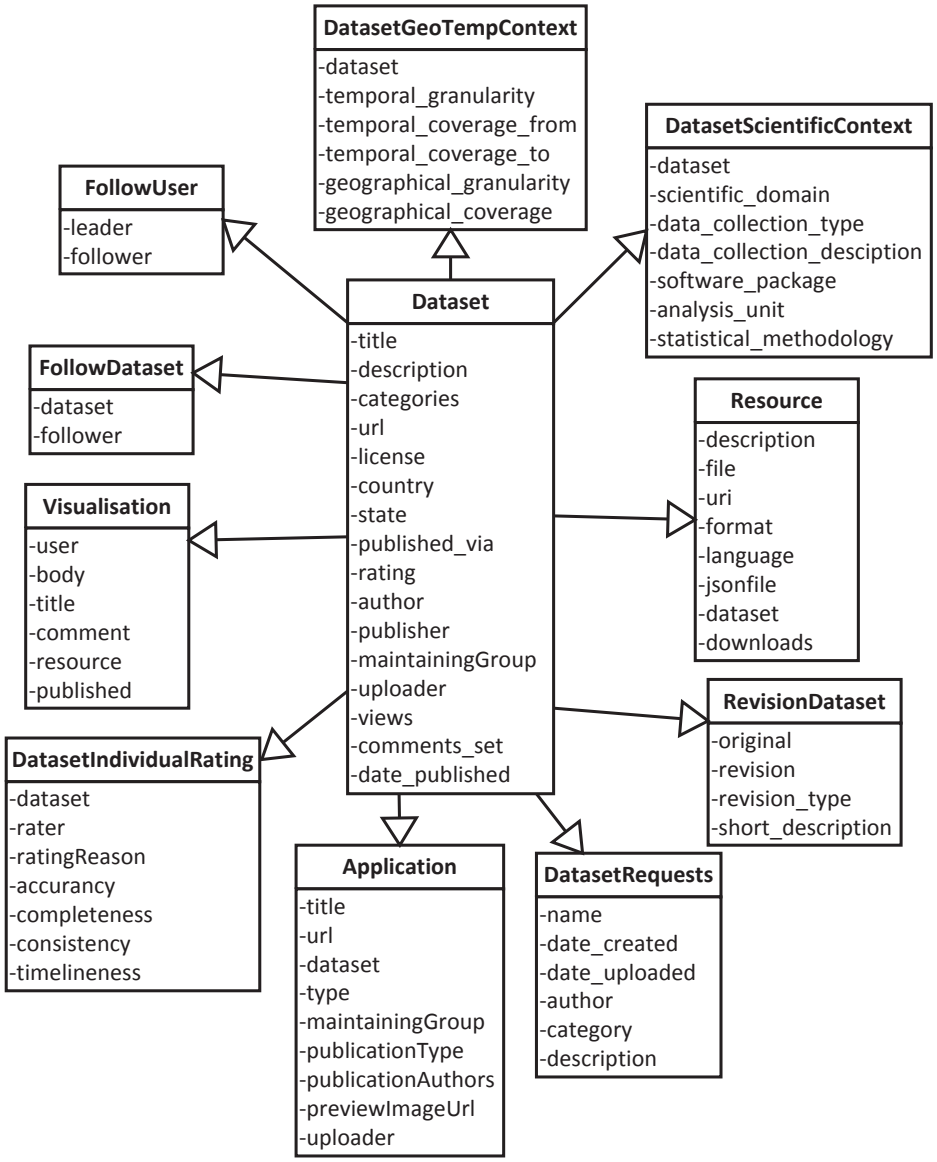


Figure 6-3: Overview of the metadata fields incorporated in the prototype.

The Integrated Development Environment (IDE) selected to develop the prototype was Aptana Studio 3. This IDE was chosen because it is free and open source, it is a well-accepted IDE, it works on multiple operating systems and it supports various other interdependent work courses (e.g. a team programming environment). The HTTP 1.1 protocol was used for communication, since this is a “generic protocol for distributed, collaborative, hypermedia information systems” (Fielding et al., 1999, p.

6). HTTP 1.1 allows for access to resources available from diverse applications (idem). Moreover, URI's were used to refer to specific resources. Furthermore, several programming languages were supported, including Java, Phyton, PHP and several other languages. The ENGAGE infrastructure was JavaScript enabled, since this is a relatively easy scripting language and it is relatively fast in its use.

The prototype enabled an interface that adheres to the linked data technology stack. This means that datasets and other resources must be identifiable by a Unique Resource Identifier (URI), all information needed to be accessible using the HTTP protocol, dataset metadata must be available in RDF and dataset metadata can be accessed by SPARQL queries (Bizer, Heath, & Berners-Lee, 2009). Furthermore, interoperability was encouraged by seamless integration and federation of content with other open data portals, which was enabled by a public API. Following these interoperability requirements makes metadata more useful and allows for interlinking datasets stored in other data hubs (Argyzoudis et al., 2014). In addition, the ENGAGE infrastructure could run in any hardware and operating system combination, including operating systems such as Windows, Mac OS X, Linux and Solaris.

The design and implementation of the interface may require considerable effort (Ince & Hekmatpour, 1987). Regarding the interface, the prototype allowed for access in two ways. First, easy access to the prototype was enabled via any modern web browser (e.g. Google Chrome, Mozilla Firefox, Internet Explorer). No plug-ins or other software were required. Second, developers could use the Application Programming Interface (API) to automatically retrieve datasets and metadata stored in the prototype. Furthermore, the API made it possible for data publishers to integrate the publishing workflow in their own applications. The API could be accessed by any third-party software over an HTTP connection. More details about the programmes that were used for the development of the prototype can be found in Argyzoudis et al. (2014).

6.4.2 ENGAGE User Interface

In this section the key characteristics of the ENGAGE user interface are described. The ENGAGE prototype was accessible for the public via a website. Figure 6-4 depicts the home page of the website. On the home page, the users could first

select the language that they preferred to work in. Then the prototype allowed for searching for open datasets in different ways. Data could simply be entered in a search bar and then be filtered, ordered and analysed more in depth. The multilingual search allowed for entering a search term in one language, and obtaining results in other languages. Moreover, the prototype assisted in searching for data without entering a specific search term but by searching for certain topics. An initial examination of the search results could take place by sorting data, filtering and categorising them based on country, data provider, data topic and licence. The search results could be translated to many different languages automatically by Microsoft Translator.

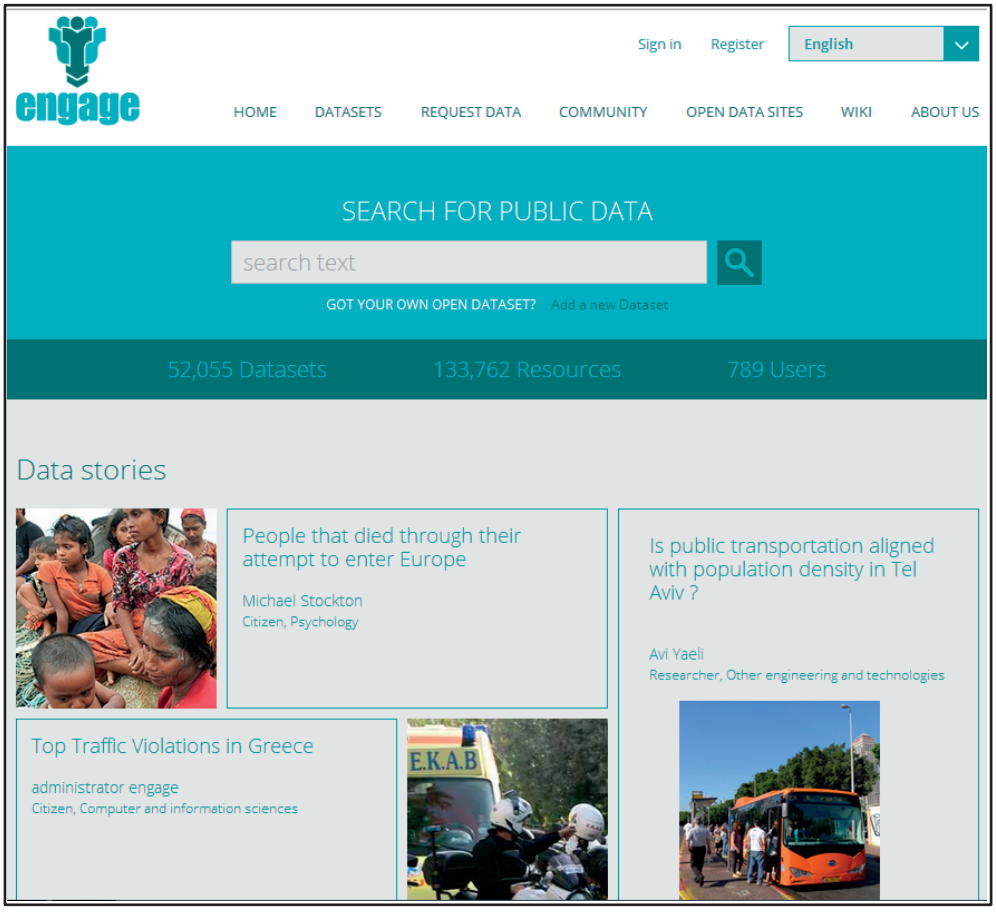


Figure 6-4: Screenshot - homepage of the prototype.

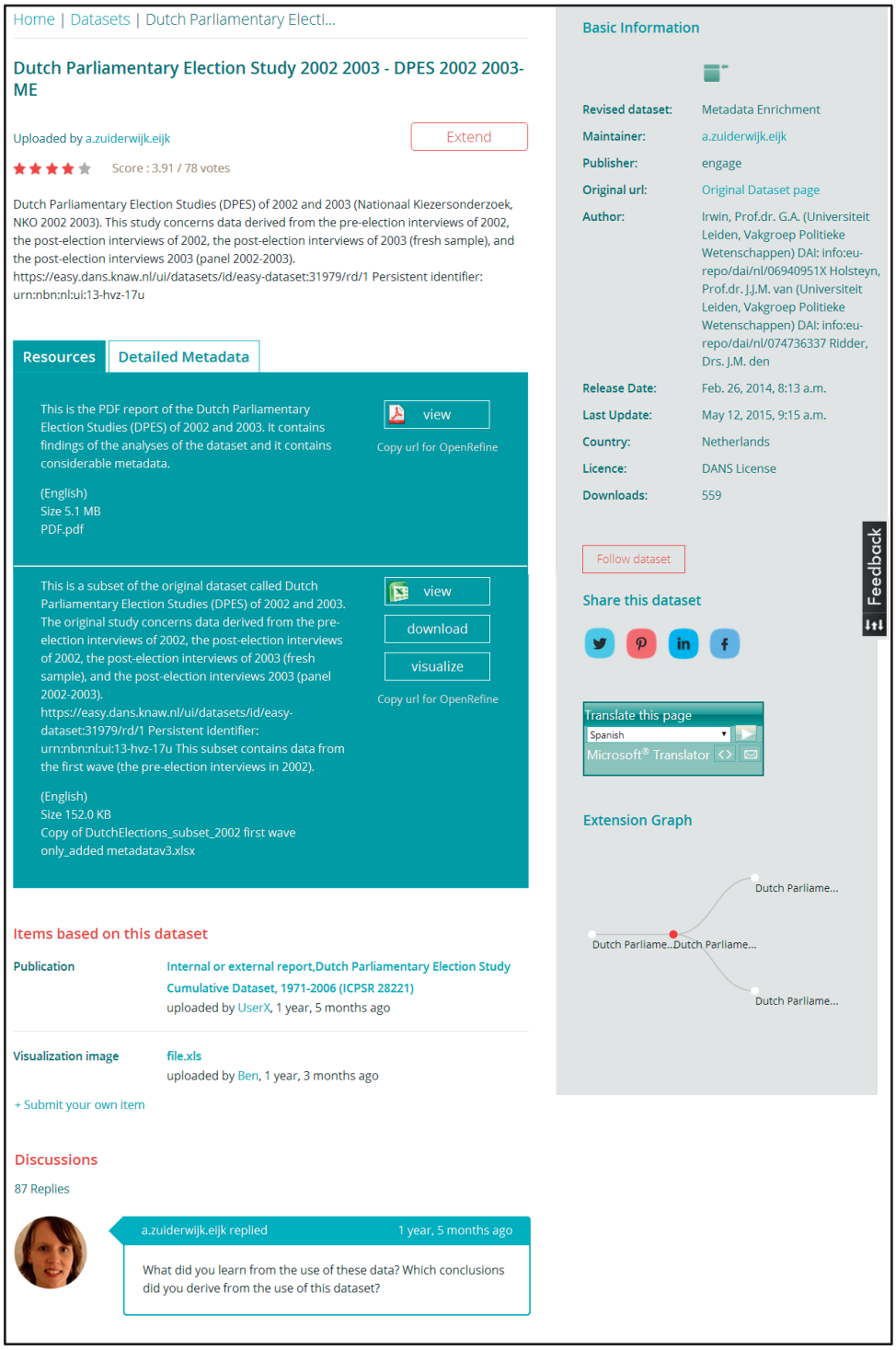


Figure 6-5: Screenshot - dataset overview of the prototype.

When a dataset had been found, it could be selected, and this then led the user to a dataset overview as illustrated in Figure 6-5. The dataset overview provided basic information about the dataset. A data quality assessment score was provided on top of the screen. Below the quality score, a description of the content of the dataset was given, the resources (e.g. a CSV-file, a PDF-file and other files) were provided, the options for viewing, downloading and visualising data and an overview of items based on the dataset were presented (e.g. publications and applications), comments and remarks on the dataset or data use were provided, 'detailed metadata' could be viewed about the context of the data and several other possibilities were provided.

Researchers using the prototype could analyse datasets by exploring the various options provided in the dataset overview. For instance, they could view the metadata to obtain insight in the context in which the dataset was created, they could read a description of the dataset and they could view the dataset without downloading it by clicking on 'view' in the resources overview. They could also see which other users had extended or amended the dataset in the extension graph.

Users could register on the infrastructure or sign in at any moment they wanted to, although it was required to do this at least before visualising a dataset, before providing feedback and discussing data and before assessing its quality, since this supported identifying which (types of) users had used a dataset in a certain way. The prototype allowed for creating tables and charts in two ways, namely in the ENGAGE visualisation tool and in the Excel Online Viewer. In this way, users were flexible and could make use of the tool that they preferred to work with. Visualising data on maps could only be done with the ENGAGE visualisation tool, since the Excel Online Viewer did not support this activity.

Figure 6-6 illustrates a part of the data interaction section. Researchers using the prototype could use this section to give feedback on datasets and processes related to data provision and use, and they could discuss what could be learned from the use of the data. The provided comments in this section have been anonymised for their publication in this dissertation. In the prototype the users could see who had provided which comment.

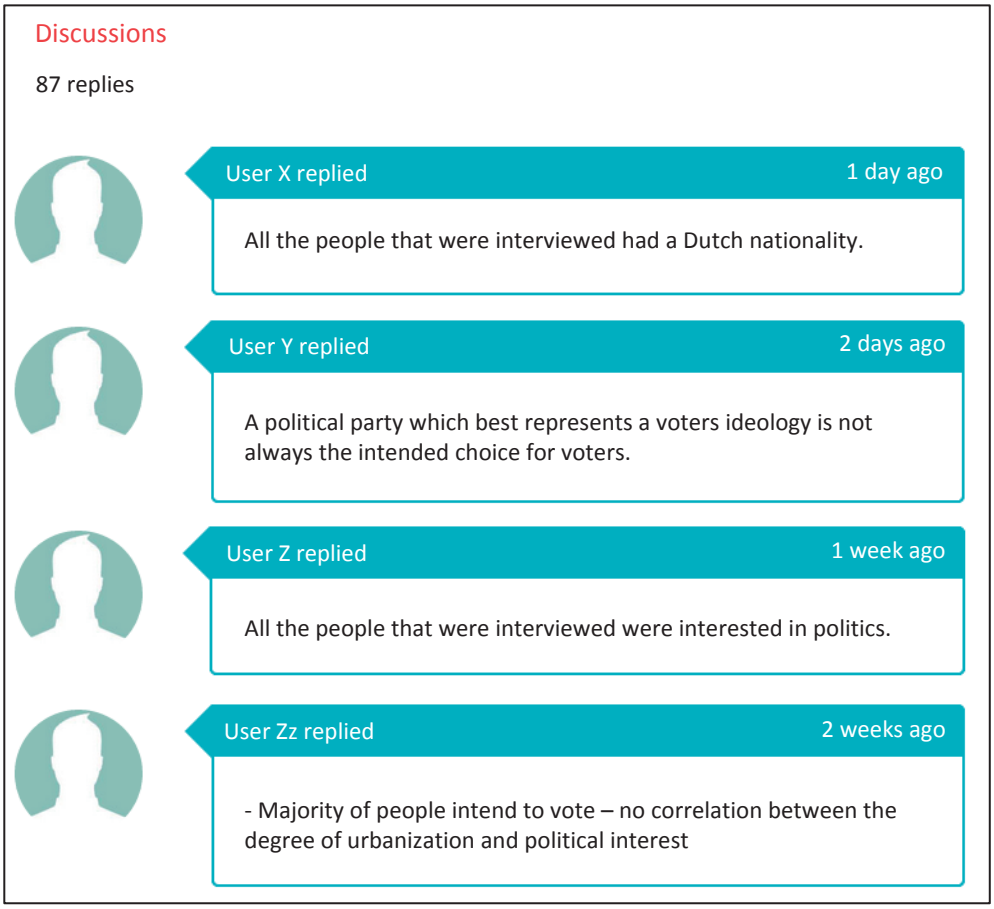


Figure 6-6: Screenshot - discussion and feedback section of the prototype.

Another part of the prototype focused on the implemented quality indicators. Users of the prototype could first rate the quality of datasets by simply selecting stars. The selection of one star represents in general a low quality, whereas the selection of five stars represents the highest possible quality. Since datasets can be used for different purposes, and the quality of datasets is related to this purpose, more advanced quality rating was also enabled. After the simple rating the user was asked to provide a more detailed assessment of various quality aspects of the dataset, including accuracy, completeness, consistency and timeliness. Thereafter users were asked to write down a review of the dataset in an open text box. In this box they could elaborate on how they had used the dataset and they could explain their assessment of the data quality indicators. Information about the data

evaluator was also available. An example of the more advanced data quality rating and a quality review are shown in Figure 6-7.

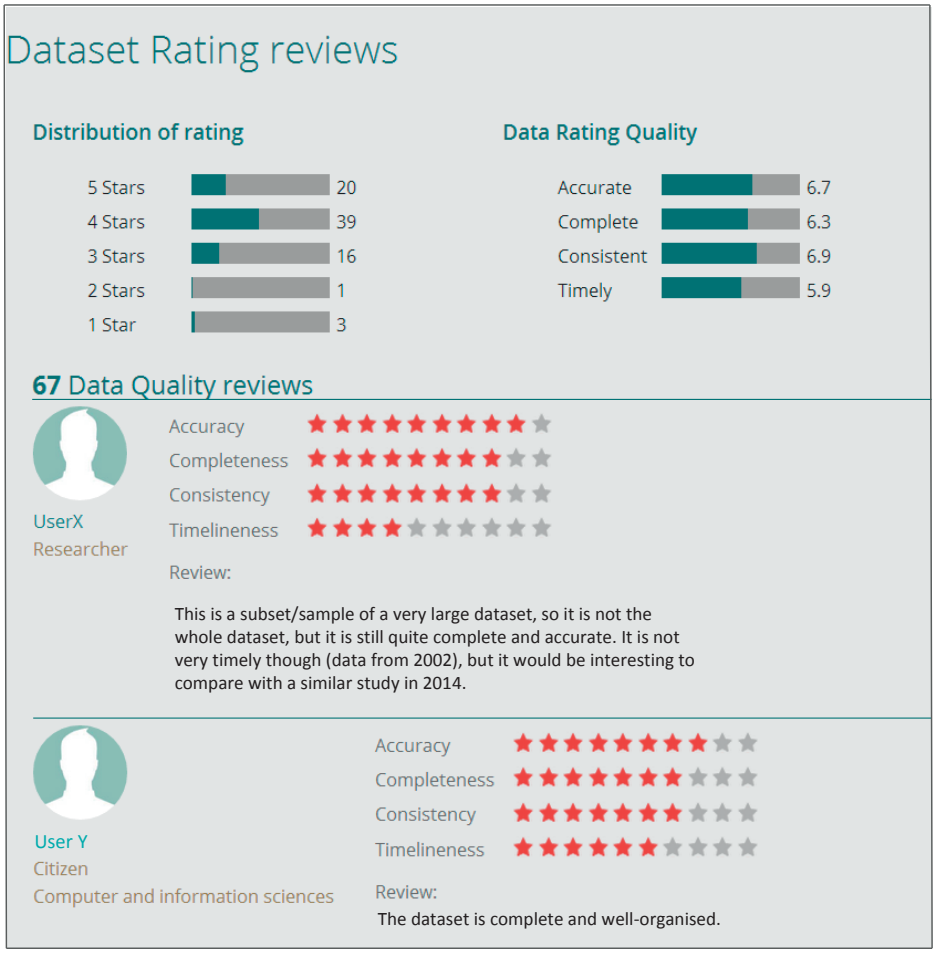


Figure 6-7: Screenshot - data quality assessment in the prototype.

6.5 Prototype testing

The fourth prototyping phase consisted of testing the prototype. Testing can be used to identify incorrect and undesirable behaviour of software, referred to as defect testing, and to demonstrate that software meets its requirements, referred to as validation testing (Sommerville, 2011). Defect testing is characterised by searching for undesirable system behaviour, including system crashes, unwanted interactions with other systems, incorrect computations and data corruption (idem).

Validation testing concerns the testing of all system features and the combinations of these features that the final design comprises (idem). Various validation and defect tests were conducted.

Before the release of each new version of the prototype to potential end-users, a number of defect tests were carried out at the developer's site. This type of testing is commonly referred to as alpha testing (Sommerville, 2011). Alpha testing is useful to discover apparent problems and issues before a first version of the software is released (Sommerville, 2011). The alpha tests helped to search for undesirable system behaviour. The key errors and omissions that were found through alpha testing were elicited. The results from the alpha tests will not be discussed in this dissertation, because there were too many to give a correct impression of the main errors and omissions of the prototype, and they only reflect testing within the internal development environment.

Secondly, potential end-users of the prototype conducted validation tests with the prototype outside the developer's environment. Sommerville (2011) refers to these kind of tests as beta tests. The key strength of testing with potential end-users is that it takes into account the user's working environment, which may influence the reliability, performance, usability and robustness of the system (idem). Beta testing is useful for discovering interaction problems between the developed software and the environment in which it is used (idem).

Since the development of the prototype was evolutionary and new features were continuously implemented, multiple beta tests were organised. In total five beta tests were incorporated to test the different releases of the prototype (see Table 6-4). The first four beta tests involved students who followed open data courses at Delft University of Technology and at The Hague University of Applied Sciences, and two researchers who were involved in open data research and publication at the Dutch Ministry of Security and Justice, and who were also involved in our case studies. In these four beta tests the users had to conduct a number of OGD use tasks that represented the expected use of the prototype. Participants were asked to indicate on a five point Likert scale to which extent it was difficult or easy to conduct each of these tasks. Subsequently, in all the first four beta tests the participants completed an online survey regarding various usability aspects, such as the user satisfaction and the speed and ease of

conducting the scenario tasks. Finally, these four beta tests ended with a discussion among the participants and the facilitator to identify the positive and negative aspects of the prototype, and to define recommendations for improvements. In addition to the first four beta tests, a fifth beta test was incorporated through an online feedback tool that was integrated in the prototype and that allowed for giving continuous feedback by anyone who accessed the prototype[7]. On the website, the online feedback tool was constantly visible to prototype users on the right side of their screen. When the prototype users clicked on the tool, they could easily report problems and bugs, suggestions, questions and data license issues.

	Beta test 1	**Beta test 2**	**Beta test 3**	**Beta test 4**	**Beta test 5**
Number and type of participants	21 Masters students	15 Bachelors students, 2 governmental researchers from the cases	19 Bachelors students	20 Masters students	31 persons (online)
Scenario tasks (usability test)	X	X	X	X	-
Post-test survey	X	X	X	X	-
Plenary discussion	X	X	X	X	-
Reports written by participants	X	-	-	X	-
ThinkTank discussion	-	X	-	-	-
Online feedback tool	-	-	-	-	X

Table 6-4: Characteristics of the conducted beta tests.

In the first, second and fourth beta test, additional tools were used to collect feedback on the prototype. In the first and the fourth beta tests the participants additionally developed open data scenarios as part of a course, tested these on the prototype, and wrote a report in which they described the scenarios, the possibilities and impossibilities of the prototype and several recommendations for improving the prototype. Moreover, in the second beta test a so-called 'ThinkTank' was used to

[7] The WebEngage tool was used, see http://webengage.com/

stimulate discussion about the prototype among the participants (www.thinktank.tudelft.nl). ThinkTank is a tool which shows its users a number of questions which they need to answer online. All participants were divided into different groups. Each of the groups started answering the following questions in ThinkTank in a different order.

- Which problems did you face during the use of the ENGAGE open data infrastructure?
- Which functions of the ENGAGE open data infrastructure did you find useful and why?
- Which functions are lacking in the ENGAGE open data infrastructure at this moment that could be useful?

The answers were provided by typing them in the online ThinkTank tool. While answering the questions, the participants could see the answers of their fellow students appearing on their computer screen and they could respond to these answers. In this way, they could easily indicate whether they agreed with the answers of other students and they could complement them. The ThinkTank tool appeared to stimulate the discussion about the prototype considerably. All answers to the above-mentioned questions were logged.

The feedback that was obtained from the beta tests led to the refinement of the functional requirements and subsequently to improvements of the prototype. Many iterative processes of the prototype construction and prototype testing took place, which ensured the evolution of the prototype. Table 6-5 depicts the main changes that were implemented in the prototype after each of the beta tests. The scenario tasks that were conducted in the beta tests became more complex after each test. For instance, in the first beta test the participants conducted relatively simple scenario tasks, such as ‘search for a dataset’ and ‘analyse the dataset’, which allowed for identifying gross system errors and omissions. In the second beta test more complex scenario tasks were tested, such as ‘search for a dataset about crime and use the advanced search functionality’ and ‘analyse the dataset using the online Excel Online Data Viewer’. The third and the fourth beta test prescribed the execution of advanced scenario tasks. The number of errors and omissions of the prototype was reduced after each beta test.

Beta test	Tasks	Key errors and omissions found in the beta tests	Prototype adjustments after the beta tests
1	Simple scenario tasks by Masters students	- Difficulties when using the prototype with many people simultaneously - Long loading time of pages - Non-advanced search functions - Limited dataset formats - Lack of data visualisation tools - Lack of understandable metadata - Difficult to find the metadata - Difficult to assess data quality - Errors while uploading data	- More capacity, possible to use the prototype with more people at the same time - Faster response, less time to load pages - Improved user interface - More metadata - Simple visualisation tools - Simple quality rating tool
2	More complex scenario tasks by Bachelors students and govern-mental resear-chers	- Too few (useful) datasets - Difficulties with data visualisation - Request for more contextual metadata, especially on the data source, terms of use and dataset license - Difficulties with viewing datasets - Unclear use of terms and vocabularies - Lack of Internet Explorer compatibility - Prototype is not intuitive; user interface should be improved - Long loading time of pages - Downloading data does not always work - More advanced rating system for datasets would be useful to assess data quality and usability	- Elastic and filtered search - More datasets integrated - Improved simple integrated visualisation tools - More metadata about licenses - Microsoft Excel Web App for viewing, visualising and analysing data online without the need to download them - Open Refine for more advanced data analysis - Integration of controlled metadata vocabularies - Improved user interface - More advanced dataset quality rating - Sign up with social media accounts - User profiles/groups - User messages and notifications - Dataset extension graph - Dataset requests - Discussion tools - Simple Wiki
3, 4	Advanced scenario tasks by Bachelors and Masters students	- Search problems (e.g. de-selection of filters) - No search function for the Wiki, the overview of users and the data requests - Limited information on the Wiki - Visualisation difficulties - Suggestion to include e-mail notifications alongside website notifications when someone replies to your dataset/ comment/ request	- Multilingual search through automated data translations - Added search functions to the Wiki, user profiles and data requests - Improved Wiki, allowed for posting articles and comments - Visualisation tool improvements - Possible to link items (e.g. visualisations, apps) to original datasets (e.g. integrated and external visualisations

Table 6-5: Outcomes of the beta tests and adjustments made to the prototype after the tests.

Beta test	Tasks	Key errors and omissions found in the beta tests	Prototype adjustments after the beta tests
3, 4	Advanced scenario tasks by Bachelors and Masters students	- Long loading time of pages - No tools for converting data formats - More metadata about the correctness and validity of the data needed	- Integration of data use stories (blog style feature) - Mechanisms to follow datasets and users - Social media sharing and discussions - Improved Internet Explorer compatibility - More datasets integrated
5	No pre-scribed tasks	- Suggestions for search functions (e.g. add more data categories to the filters) - Problems with links to datasets which did not lead to any dataset (i.e. 'broken links') - Visualisation problems for specific datasets - Problems with user interaction (e.g. not every user can be added to a user group) - Suggestions to improve user interface	- Possible to unselect filters - Many filters could be selected - Solving issues with specific datasets - User activity is rewarded by showing the latest user activity on the home page

Table 6-5 (continued): Outcomes of the beta tests and adjustments made to the prototype after the tests.

A third type of testing described by Sommerville (2011) is acceptance testing, where customers test a system in the customer environment. Acceptance testing is performed after the release of the final software product has been tested (ibid). The evaluations that we describe in chapter seven can be seen as a form of acceptance testing, since they aim to find out to which extent the prototype can be accepted to enhance the coordination of OGD use. We will report on the acceptance tests that were conducted in the following chapter.

6.6 Summary: overview of prototype and answer to the fourth research question

This chapter addressed the fourth research phase of our study, namely the development of the prototype. It aimed to answer the fourth research question: *what does the developed OGD infrastructure look like?* A prototype of the OGD infrastructure was developed. The creation of the prototype in this chapter has to be viewed from the perspective of the cases that we studied (see chapter 4), which focused on the use of structured research OGD from the domains of social

sciences and humanities by researchers outside the government through OGD infrastructures. Following Ince and Hekmatpour (1987), the development of the prototype in this study encompassed 1) defining the objectives of the prototype, 2) selecting the functions of the prototype, 3) constructing the prototype and, 4) testing the prototype. Although these phases were described linearly in this chapter, much iteration between the prototyping phases and between the prototyping phases and the design phases (chapter 5) took place.

In the first phase two prototyping objectives were identified. First, the prototype needed to allow for refining and detailing the functional user requirements regarding the metadata model, the interaction mechanisms and the data quality indicators, and second, the prototype was developed to be able to measure the effects of the designed OGD infrastructure. To meet these objectives, evolutionary prototyping was used since this allowed for gradually developing the prototype and for the evolvement of the OGD infrastructure design throughout its use. The second phase of prototype development involved the selection of functions that needed to be prototyped (Ince & Hekmatpour, 1987). The functions that needed to be implemented in the prototype were defined, including the implemented metadata model, interaction mechanisms and data quality indicators. The third prototyping phase concerned the development required to produce the prototype (Ince & Hekmatpour, 1987). A description of the final version of the constructed prototype as well as its user interface was provided. Finally, the fourth prototyping phase consisted of the prototype testing. Various alpha and beta tests were organised to test the developed prototype and to obtain feedback for further improvements. Feedback derived from these tests was used to gradually improve the prototype, and this ensured its evolution. In the following chapter we describe how the final version of the developed prototype (ENGAGE 3.0) was used to evaluate the effects of the OGD infrastructure.

7. Evaluation of the prototype

This chapter addresses the final phase of our research, namely the evaluation of the prototype. It aims to answer the fifth research question of this dissertation: *What are the effects of the developed infrastructure on the coordination of OGD use?* Quasi-experiments are used as the key research instrument for the infrastructure evaluation. This chapter starts with a description of the approach and structure of this chapter, followed by the evaluation methodology. Next, the preparation of the research data for analysis is described, and the evaluation results are presented. The evaluation results are followed by a summary of this chapter and the answer to the fourth research question. Parts of this chapter have been published in Zuiderwijk and Janssen (2015), Zuiderwijk, Janssen, and Davis (2014) and Zuiderwijk et al. (forthcoming).

7.1 Approach and structure of this chapter

Figure 7-1 depicts the approach and structure of this chapter. The figure shows that this chapter first describes the evaluation methodology (section 7.2), which incorporates the argumentation for taking a quasi-experimental approach and an overview of the pre-test and post-test conditions, as well as the treatment group and control group conditions. Furthermore, the roles of the different actors involved in the quasi-experiments are described, and it is explained which measures were taken to enhance the evaluation validity. As a final part of the evaluation methodology, the structure of the quasi-experiments is outlined to show which measures were used to obtain the required research data and how the quasi-experiments were organised.

After the evaluation methodology is explained, section 7.3 describes how the research data that we collected through this methodology were prepared for analysis. Thereafter, the results from the data analysis are provided. Section 7.4 describes the main characteristics of the participants of the quasi-experiments, including their gender, age, nationality, and experience. Based on these characteristics, we argue that the research data from the first and second quasi-experiment can be combined. Subsequently, section 7.5 and 7.6 provide the

outcomes of the proposition testing. They discuss the key results regarding the ease and speed of OGD use, as well as the enhancement of the coordination of OGD use. Finally, the findings from this chapter are summarised and the answer to the fifth research question is provided in section 7.7.

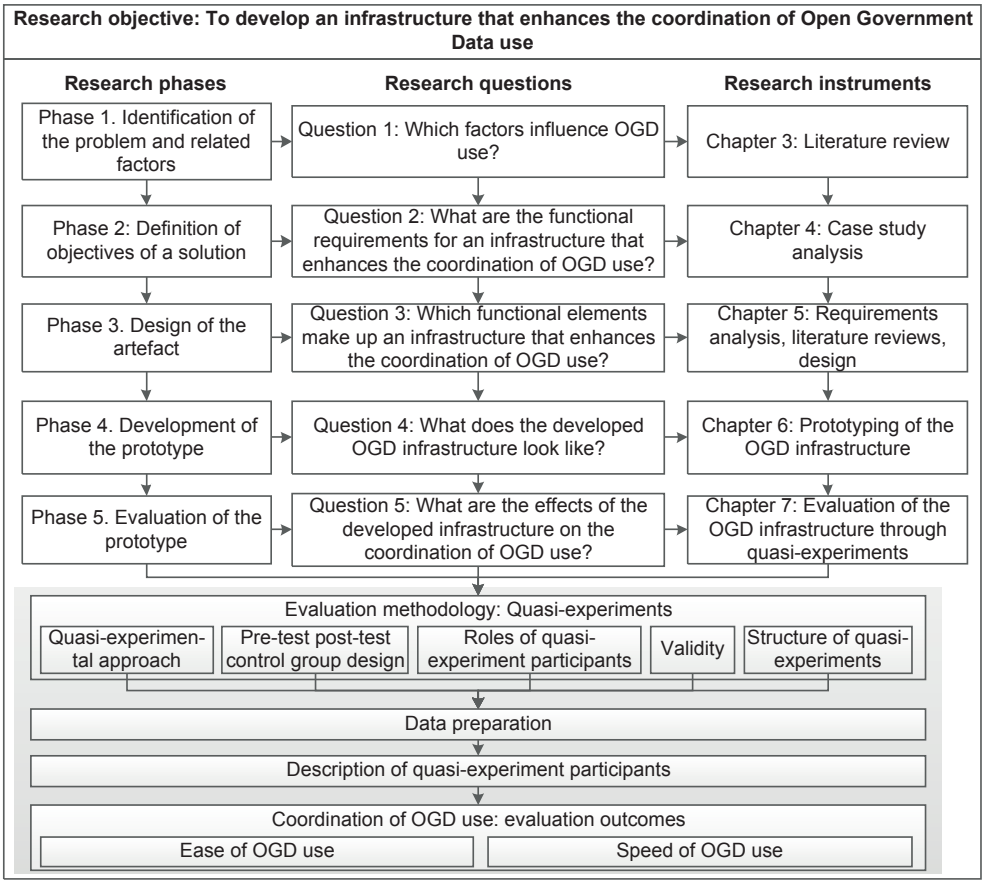


Figure 7-1: Research design including the evaluation approach.

7.2 Evaluation methodology

Evaluation can be defined as "the systematic determination of merit, worth, and significance of something […] or someone" (Hevner & Chatterjee, 2010, p. 109). The evaluation aimed at examining to which extent the functional OGD infrastructure elements (i.e. metadata, interaction mechanisms and data quality indicators) can enhance the coordination of OGD use. In line with our definition of

OGD use in section 1.2, the evaluations focus on evaluating the use of structured research OGD from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures. In this section the evaluation methodology is described, incorporating a description of the quasi-experimental approach, followed by a description of the pre-test post-test control group design of the quasi-experiments. Then the roles of the actors involved in the quasi-experiments are defined, as well as an outline of how the evaluations deal with validity. Finally, the structure of the quasi-experiments is discussed.

7.2.1 Quasi-experimental approach

In chapter five three propositions were developed:

- P1: Metadata positively influence the ease and speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis.
- P2: Interaction mechanisms positively influence the ease and speed of interaction about OGD.
- P3: Data quality indicators positively influence the ease and speed of OGD quality analysis.

This chapter aims at evaluating the ease and speed of the different types of OGD use as indicators for the enhancement of OGD use coordination. This means that the three key variables in our evaluation, i.e. metadata, interaction mechanisms and data quality indicators, as well as the OGD use processes needed to be controlled to ensure as much as possible that the effects of the OGD infrastructure can be attributed to these variables. Furthermore, the evaluation of the prototype needed to take place in a realistic setting in which the prototype had to be operated to measure the effects of the infrastructure. At the time of the quasi-experiments it was not possible to find examples of functioning OGD infrastructures in practice which contained the three functional infrastructure elements. Therefore, merely using surveys or interviews to ask people for their experiences with such OGD infrastructures would not result in the desired type of outcomes. This led us to an experimental approach.

Experiments can be conducted to manipulate variables and observe their effects upon other variables (Campbell & Stanley, 1969, p. 2). An experiment can

be defined as "a study in which an intervention is deliberately introduced to observe its effects" (Shadish, Cook, & Campbell, 2002, p. 12), and can be either a *true experiment* or a *quasi-experiment*. True or natural experiments have more than one purposively created group, common measured outcome(s) and random assignment (Gribbons & Herman, 1997). Quasi-experiments differ from true experiments in the sense that the experimental subjects in a quasi-experiment are not randomly assigned to conditions. Quasi-experiments encompass 1) a treatment and a control condition, 2) a pre-test and a post-test, and 3) a model that reveals the treatment and the control group effects over time, given no treatment effects (Kenny, 1975). In quasi-experiments, researchers can have control over selecting and scheduling measures, how the participants are assigned non-randomly, over the type of control group with which the treatment group is compared, and over how the treatment is organised (Shadish et al., 2002). Since it was not possible for our evaluations to randomly draw a sample of evaluation participants from the entire population of OGD users, we cannot refer to the evaluations as true experiments. Therefore, we conducted quasi-experiments.

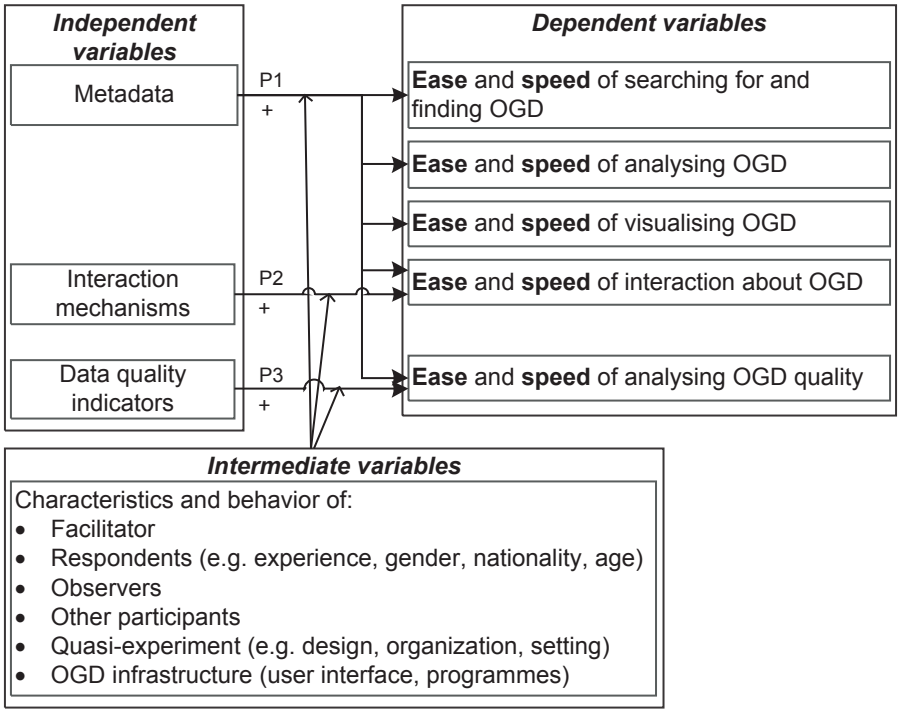


Figure 7-2: Variables in the quasi-experiments.

Figure 7-2 shows the variables involved in the quasi-experiments, including the three propositions. The figure shows that metadata, interaction mechanisms and data quality indicators are the three independent variables. Since these three functional elements aim to enhance the coordination of OGD use, the five types of OGD use are the main dependent variables in the evaluation model (i.e. searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality assessment). For each of these independent variables the effect on the dependent variables will be determined. Moreover, for each of the categories of OGD use activities, the ease and speed of the activity will be investigated to indicate the coordination of OGD use. As explained in chapter five, this will be done because these use aspects are believed to be most important for OGD use.

Since the literature does not clearly reveal the nature of the relationship between the functional elements of the OGD infrastructure on the one hand (the independent variables), and OGD use on the other hand (the dependent variables), we also evaluate whether intermediate variables influence this relationship. The potential intermediating influence on the relation between the independent and the dependent variables will be examined for the following six variables.

1. *The role of the facilitator.* This variable was investigated since the facilitator of the quasi-experiments may have introduced the evaluation session in such a way that participants were directed towards certain responses. For instance, the facilitator may have shown a positive instead of a neutral attitude towards the OGD infrastructure, which may have resulted in more positive evaluation results.
2. *Characteristics of the respondents.* Characteristics of the respondents may have influenced the results. As data use requires certain skills and techniques (Puron-Cid, Gil-Garcia, & Luna-Reyes, 2012), the levels of skills and techniques that the participants have acquired may vary. Moreover, some individuals may already have access to open data infrastructures, hardware, software, financial and educational resources and skills (Gurstein, 2011) and may be more experienced with OGD use. These individuals may be able to make better use of OGD than others. Experience is therefore one of the investigated intermediate variables in the evaluation model. The skills, techniques and experience may vary

across countries, age and gender of the participants, and these variables are therefore also included in the evaluation model.

3. *Observers.* One of the measures used in the evaluation concerned observations. Since observers were present in the room where respondents participated in the evaluation sessions, we need to examine whether the observers influenced the behaviour of the participants or the way that they answered the questions.
4. *Other participants*. Participants may have influenced each other's behaviour in the evaluation session, for example by adapting their answers towards the answers of the other participants. Although various arrangements were made to avoid this type of influence (e.g. emphasising that the participants should not speak to each other and using partitions between their computers), it is important to test the influence of participants on each other's behaviour.
5. *The design, organisation and setting of the quasi-experiments*. The influence of the design, organisation and setting will be investigated as intermediate variables, since they may have influenced OGD use. Examples of such intermediate variables are the room in which the evaluation took place, the lights and the sounds heard during the evaluation sessions.
6. *The OGD infrastructure*. Characteristics of the infrastructure may have influenced the user experience of the infrastructure, such as the user interface, the tools and programmes available on the infrastructure and the number of offered datasets.

Even though various arrangements were made to keep the influence of the above-mentioned variables as limited as possible, their potential influence on OGD needs to be investigated. The inclusion of intermediate variables in the evaluation model offers insight in which variables influence the coordination of OGD use on the infrastructure and allows for investigating rival explanations for certain OGD use outcomes. Examining intermediate variables strengthens our conclusions about the relationship between the independent and dependent variables.

7.2.2 Pre-test post-test control group design

An important characteristic of evaluations concerns the selection of participants. Ideally, evaluation participants are randomly selected from a representative population to reduce the risk that evaluation findings are caused by a selection bias (Campbell & Stanley, 1963). For this study it was not possible to randomly select a sample of participants, since there was no central overview of people who belong to the population of OGD users from which we could randomly draw such a sample. The quasi-experiments therefore contained non-randomly designed groups, to which Campbell and Stanley (1969) refer to as non-equivalent groups. Students and open data professionals were involved in the evaluation. The following inclusion criteria, the predefined characteristics that qualify potential participants for including them in the study (Salkind, 2010), were defined:

- Participants had to have the skills to work with computers;
- Participants had to be at least 20 years old;
- Participants had to have attended presentations concerning the basics of open data;
- Participants had to live in the Netherlands;
- Participants had to be available for the quasi-experiments;
- Participants had to be willing to participate in the evaluations.

The non-randomly selected evaluation participants were split into a control and a treatment group. While the participants of the treatment group operated the developed OGD infrastructure (i.e. the treatment infrastructure), the participants of the control group operated a control infrastructure. Participants were free to choose their seat, which determined whether they would participate in the treatment or control group. The participants did not know this in advance. It was not possible to divide the group of professionals into a treatment and control group, since this would not allow for having sample sizes of at least 30 participants per treatment group and 30 participants per control group. Thus, dividing the group of professionals into a treatment and control group would not allow for conducting statistical tests. Therefore, the professionals only participated in the treatment group. Participants from the control group were not matched to a participant from the treatment group before the quasi-experiments took place (see Reichardt, 1979 for more information about matching). However, by inviting a particular group of

persons to participate in the quasi-experiments, we tried to select participants (students and professionals) with a similar background (e.g. with regard to their experience in open data use and with regard to the focus on open data derived from research). Moreover, the characteristics of the participants in the control and treatment group were compared through a non-pair wise matching process. The characteristics of the different groups of participants were analysed and compared. Finally, in addition to dividing the evaluation participants in a treatment and control group, it is advised to conduct pre-tests and post-tests for both of these groups (Reichardt, 1979). Pre-tests measure or observe participants prior to the treatment, while post-tests measure or observe them after the treatment (idem). The quasi-experiments in this study incorporated a pre-test and a post-test in the form of surveys.

In sum, a pre-test post-test control group design was used to conduct the quasi-experiments, and participants were selected non-randomly. Three quasi-experiments were conducted in March and April 2014 (see Table 7-1).

Characteristics	**Quasi-experiment 1**	**Quasi-experiment 2**	**Quasi-experiment 3**
Date	March 3, 2014	March 5, 2014	April 23, 2014
Type of participants	3rd year Bachelors students	1st year Masters students	Professionals (researchers, policy-makers, citizens, entrepreneurs and others)
Duration	100 minutes	100 minutes	95 minutes
Location	Delft University of Technology	Delft University of Technology	Delft University of Technology
Involved groups	Treatment and control group	Treatment and control group	Treatment group only
Motivation for participation	Mandatory part of a course (Policy, Economy and Law) that they had to follow.	Mandatory part of a course (Business Process Management and Technology) that they had to follow.	Part of a 4-hour workshop for which the participants had registered. The participants volunteered to participate in the workshop.

Table 7-1: Characteristics of the three quasi-experiments.

All evaluations lasted between 95 and 100 minutes and took place in the same computer room at Delft University of Technology. The computers were separated by partitions, so that the participants had their own work place and it was difficult for them to see what other participants did on their computers. This stimulated

working individually. Whereas participants of the first and second quasi-experiment were obliged to participate in the quasi-experiments as part of their education, participants of the third quasi-experiment participated voluntarily.

7.2.3 Roles in the quasi-experiments

Table 7-2 depicts the roles that were assigned to various actors involved in the quasi-experiments. First, the *session manager* designed the quasi-experiments, managed the time during the quasi-experiments, answered questions of the participants, led the plenary discussion at the end of the quasi-experiment and ended the quasi-experiment. The author of this dissertation fulfilled the role of session manager throughout all the quasi-experiments. All questions were answered by this one session manager to ensure that similar answers were provided to all participants. Moreover, the answers were provided in such a way that the session manager expected to have the least possible influence on the results provided by the participants, for example by referring to the neutral explanations of the quasi-experiment described on the hand-out or in the surveys that the participants had already received.

Roles in the quasi-experiment	Tasks
Session manager	- Designed the quasi-experiment - Managed the time during the quasi-experiment - Answered questions - Led the plenary discussion - Ended the quasi-experiment
Facilitator	- Introduced the quasi-experiment - Presented the instructions using PowerPoint slides and a hand-out - Explained the activities performed in the quasi-experiment
Observers	Observed the behaviour of the participants during the quasi-experiment
Participants: users of the prototype and the control OGD infrastructure	- Completed questionnaires - Conducted scenario tasks - Participated in the plenary discussion

Table 7-2: Roles and tasks in the quasi-experiments.

The *facilitator* was responsible for providing an introduction to the evaluation session in which the activities that would be performed during the session were introduced and explained to the participants. As the facilitator might influence the results of the evaluation, a facilitator was chosen who had not been involved in the research before, and who did not aim to use the results of the study for any

research or other purposes. Moreover, the facilitator was not involved in the development or use of the developed prototype. Neutrality of the facilitator was further stimulated by selecting a facilitator who was not involved in the design of the quasi-experiments. The same person fulfilled the role of facilitator throughout all the quasi-experiments. This kept the facilitator's influence on the outcomes as limited as possible. The influence from the facilitator was also evaluated by the observers and the participants of the quasi-experiments and will be reported upon in this chapter.

The role of *observer* was fulfilled 18 times by 11 persons in the three quasi-experiments, which means that some persons observed in multiple quasi-experiments. Webb, Campbell, Schwartz, and Sechrest (1973) note that one has to be very much aware of the role of patently visible observers in quasi-experiments, as they may cause changes to the behaviour of the participants, which could decrease the validity of the study. Even if the observers are very well-integrated, they can still bias the production of the data: "the bias may be a selective one to jeopardize internal validity, or, perhaps more plausibly, it may cripple the ability of the social scientist to generalize his findings very far beyond the sample" (Webb et al., 1973, p. 113). For this reason, we only involved a relatively small number of observers. We involve observers because there may be certain behaviour noticed by these observers that may not be found through the other measures (i.e. in a questionnaire or by time measures). The roles of the observers were explained by providing instructions in training sessions and on paper and maps. The observers observed on average seven persons per observation, with a maximum of twelve participants observed.

Finally, the role of *participant* was fulfilled by students and professional open data users. They completed various questionnaires, performed scenario tasks and participated in the plenary discussion. Table 7-3 shows how many participants were involved in the quasi-experiments. As a rule of thumb, one can say that to be able to measure medium to large effects, each sample should have a size of at least 30 participants per cell (Cohen, 1988; Van Voorhis & Morgan, 2001). The smaller the sample size, the less likely it is that the data are normally distributed. The first quasi-experiment did not contain samples of 30 or more persons. However, as the first and the second quasi-experiment consisted of

participants with background characteristics that were relatively similar, the results of the first and second quasi-experiment were combined (see section 7.4.5). In total 127 persons participated in the quasi-experiments, including 41 control group participants and 86 treatment group participants.

Quasi-experiment (QE)	Treatment group participants	Control group participants	Total number of participants
QE1	10	9	19
QE2	40	32	72
QE3	36	0	36
Total	86	41	127

Table 7-3: Number of participants involved in the quasi-experiments.

7.2.4 Validity

Evaluations should pay explicit attention to four types of validity, namely internal validity, construct validity, external validity and statistical conclusion validity. These types of validity should be established as much as possible (Cook & Campbell, 1979). Construct validity refers to the establishment of correct operational measures for the constructs that are investigated (Cronbach & Meehl, 1955). Table 7-4 lists the key characteristics of the quasi-experiments related to construct validity.

Tactic to enhance construct validity	Implementation in the quasi-experiments
Use multiple sources of evidence (Campbell & Fiske, 1959; Jick, 1979; Webb et al., 1973; Yin, 2003)	- Conducted multiple quasi-experiments to see whether replicating the evaluations would provide the same results - Quantitative surveys and time measures were combined with qualitative semi-structured observations
Establish a chain of evidence (Yin, 2003)	- Focused on data collection that allowed for investigating the propositions developed in chapter 5 - Developed procedures and protocols for the evaluations - Tested the quasi-experimental design and organisation in advance (e.g. to examine whether the survey questions were clear and whether scenarios could be completed within the set time).
Have key informants review draft reports of the findings (Yin, 2003)	- Findings were presented to and discussed with open data experts (the developers of the ENGAGE prototype, persons who maintained the control OGD infrastructure and other open data experts) - Evaluation participants interested in the results of the evaluations received a summary of the findings on which they could comment - Evaluation results were peer-reviewed and published in scientific articles (e.g. see Zuiderwijk et al., forthcoming)

Table 7-4: Tactics to enhance construct validity in the quasi-experiments.

Internal validity is the establishment of a causal relationship, showing that certain conditions lead to other conditions (Yin, 2003). Table 7-5 shows how internal validity was enhanced in this study.

Tactic to enhance internal validity	**Implementation in the quasi-experiments**
Match patterns (Yin, 2003)	- Examined pre-test and post-test results - Investigated various characteristics of the treatment/control group and the student/professionals group (e.g. the distribution of participants with OGD experience in these groups)
Build explanations (Yin, 2003)	- Searched for explanations in the survey, observation and time measure results - Considered and discussed the explanations given by the respondents and the observers of the evaluations additionally to own explanations
Search for rival explanations (Campbell, 1969; Campbell & Stanley, 1963; Yin, 2003)	- Investigated rival explanations by conducting a non-pair wise comparison of the treatment and control group - Conducted a pre-test and post-test: measured just before and just after the artefact was used - Examined intermediate variables - Examined the rival explanations listed by Campbell (1969) (e.g. examined whether changes in observers, and biases in the recruitment of comparison groups might have influenced the outcomes)
Use logic models (Yin, 2003)	Developed a logic model (see Figure 7-2)

Table 7-5: Tactics to enhance internal validity in the quasi-experiments.

External validity refers to the establishment of the domain to which the findings of the research can be generalised (Yin, 2003). Table 7-6 depicts the tactics used to enhance external validity in the quasi-experiments.

Tactic to enhance external validity	Implementation in the quasi-experiments
Use theories and existing research (Yin, 2003)	- Incorporated theories and existing research in the evaluations. E.g. the survey questions were based on a model developed of Venkatesh, Thong, Chan, Hu, and Brown (2011) that integrates the Unified Theory of Acceptance and Use of Technology (UTAUT) and the two-stage expectation confirmation theory of Information Systems (IS) continuance
Replication logic (Yin, 2003)	- Generated participant inclusion criteria (see section 7.2.2) - Opted for evaluating the OGD infrastructure in a realistic setting by using scenarios: detailed descriptions of interactions between the OGD users and the infrastructure - Developed protocols and instructions (e.g. an observation protocol) to allow for repetition of the operations of the study (e.g. data collection procedures) - Provided training and detailed instructions to the evaluation facilitator and the observers - Developed and used the same pre-test and post-test surveys in all evaluations for both the treatment and the control group - Used similar scenarios and instructions for all the evaluations -Developed detailed surveys to obtain the same type of results from the observations - Participants randomly chose where they were going to sit in the room, which determined whether they would participate in the treatment or control group (this was unknown to the participants)

Table 7-6: Tactics to enhance external validity in the quasi-experiments.

Finally, statistical conclusion validity concerns the statistical power of a study, reasonable evidence of co-variation of the presumed cause and effect, and the strength of the co-variation (Cook & Campbell, 1979). Table 7-7 shows how statistical conclusion validity was addressed in the evaluations.

Tactic to enhance statistical conclusion validity	Implementation in the quasi-experiments
Analysing the amount of power one has to detect the effect of a magnitude given the variances and sample sizes (Cook & Campbell, 1979)	- Used the required sample size for the evaluations - Used both a 95 per cent and 99,9 per cent confidence interval for the analysis

Table 7-7: Tactics to enhance statistical conclusion validity in the quasi-experiments.

7.2.5 Structure of the quasi-experiments

Figure 7-3 shows the structure of the quasi-experiments. In the following sections we elaborate on the different boxes of the figure.

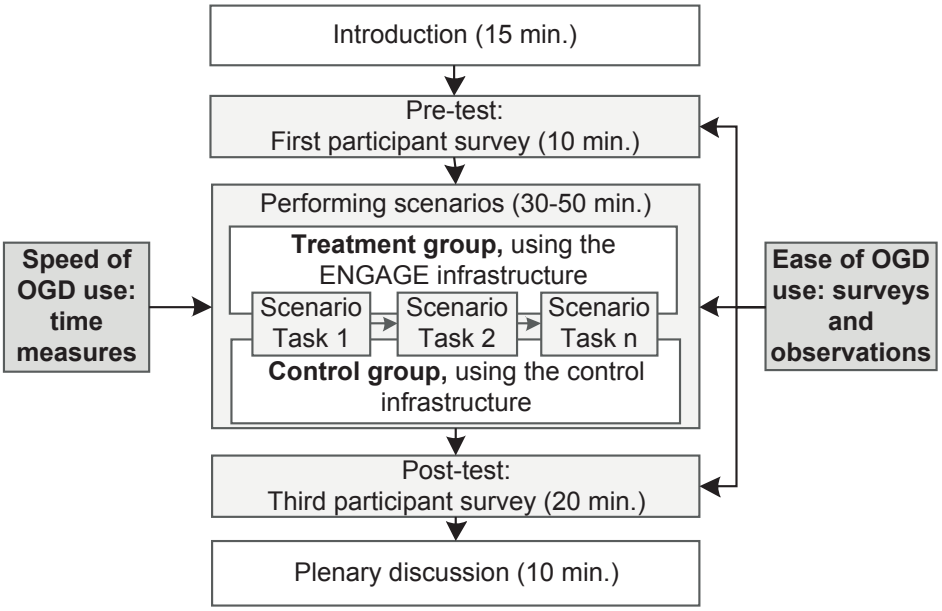


Figure 7-3: Structure of the quasi-experiments.

Introduction

In the introduction the participants were informed about the objectives of the quasi-experiments and they received instructions. All participants received a participant code that they needed to write down in the three participant surveys, so that the results from the different surveys could be linked to one individual participant. The participant code was also visible for the observers, so that they could relate particular behaviour to a participant. The participants received a hand-out with general instructions, a time plan, links to the surveys and an overview of the five main parts of the quasi-experiments on paper. Participants were ensured that the information that they would provide in the quasi-experiments would be treated confidentially. Students were ensured that the results would not be used to assess their course performance. It was emphasised that the term “open data” in the surveys and scenarios referred to data opened by governments.

Pre-test: first participant survey

The first participant survey (the pre-test) was completed online before the treatment. It consisted of 19 questions (see Table 7-8 and Appendix C), and most questions were mandatory, i.e. it was not possible to skip these questions. It took approximately 8 to 12 minutes for the respondents to complete the first survey.

Parts of the first participant survey	Description
1. General questions	12 questions on the demographics of the respondents (e.g. gender, age, daily occupation, experience with data publication and data use)
2. Open data metadata	1 question with several sub questions on the participants' experience with metadata
3. Open data interaction mechanisms and data quality indicators	1 question with several sub questions on the participants' experience with interaction mechanisms and data quality indicators
4. Expectations of the open data infrastructure	4 questions with several sub questions about the participants' expectations of the open data infrastructure that they would investigate in the scenario tasks
5. Suggestions and comments	1 open question about suggestions and comments

Table 7-8: Structure of the first participant survey.

Scenarios

After the completion of the pre-test, the quasi-experiments proceeded with scenarios. The scenario-based design, the scenario tasks, the control infrastructure and the dataset used for the quasi-experiments are explained below.

Scenario-based design

A scenario-based design was used for the treatment. Scenarios are narrative descriptions of interactions between users and proposed systems (Potts, 1995). More specifically, "scenarios highlight goals suggested by the appearance and behaviour of the system, what people try to do with the system, what procedures are adopted, not adopted, carried out successfully or erroneously, and what interpretations people make of what happens to them" (Carroll, 1999, p. 2). Scenarios can be used for various purposes in the interactive systems development processes, and can be written from many perspectives and at multiple levels (Carroll, 1999; Lim & Sato, 2003). The power of scenarios comes from the their ability to provide a view of the whole of a situation in a way that allows people to reason from (Alexander & Maiden, 2004). Scenarios can evoke reflection about design issues, as they provide descriptions of end-user experiences. Furthermore, scenarios can be abstracted and categorised to create knowledge to discuss problems (Carroll, 1999). The above-mentioned characteristics make scenario-based evaluations an appropriate approach for the treatment in our quasi-experiments.

The design of the scenarios was guided by the criterion that it should cover the range of OGD use activities that were identified in chapter three, and that it should examine the coordination of OGD use by including the management of dependencies between and among activities performed to use OGD. To identify the exact tasks that needed to be covered by the scenarios, we used open datasets from the case studies to conduct activities, such as analysing, visualising, discussing, providing feedback on data and reviewing the quality of these open datasets. From this test case we identified typical tasks that needed to be conducted to complete the scenarios on both the treatment and control infrastructure. Five scenarios were developed which comprised eighteen scenario tasks in total. The scenario tasks were described as part of the second participant survey.

Scenario tasks: second participant survey

The scenarios were presented to the participants in the form of a survey. This second participant survey included scenario tasks, instructions and questions. Table 7-9 provides an overview of the main content of the second participant survey (see Appendix D).

The evaluation focused on what the OGD infrastructure did and how the quasi-experiment participants could use it. The participants completed scenarios that prescribed them to use various tools, to interact with other OGD users and to use tools that allowed for interaction with OGD providers and policy makers. This means that they used OGD in a way that corresponds to our definition of coordination (see section 3.2.4). The scenarios did not explicitly focus on the evaluation of the user interface. Since we were constrained by time limitations of the quasi-experiments, we decided to focus on the evaluation of the system, its coordination patterns and its functions. Although the user interface was not evaluated explicitly in the quasi-experiments, it was examined as an intermediate variable.

Parts of the second participant survey	Description
Introduction	A question to write down the participant code
Time measure 1	A question to write down the time at that moment
Scenario 1. Searching for and finding OGD	- Task descriptions (task 1-4) - A question to assess the difficulty or ease of task 1-4 on a five-point Likert scale (very difficult – very easy) - A question to rank the difficulty of tasks 1-4 - A question on whether tasks 1-4 could be completed
Time measure 2	A question to write down the time at that moment
Scenario 2. OGD analysis	- Task descriptions (task 5-7) - A question to assess the difficulty or ease of task 5-7 on a five-point Likert scale (very difficult – very easy) - A question to rank the difficulty of tasks 5-7 - A question on whether tasks 5-7 could be completed
Time measure 3	A question to write down the time at that moment
Scenario 3. OGD visualisation	- Task descriptions (task 8-10) - A question to assess the difficulty or ease of task 8-10 on a five-point Likert scale (very difficult – very easy) - A question to rank the difficulty of tasks 8-10 - A question on whether tasks 8-10 could be completed
Time measure 4	A question to write down the time at that moment
Scenario 4. Interaction about OGD	- Task descriptions (task 11-14) - A question to assess the difficulty or ease of task 11-14 on a five-point Likert scale (very difficult – very easy) - A question to rank the difficulty of tasks 11-14 - A question on whether tasks 11-14 could be completed
Time measure 5	A question to write down the time at that moment
Scenario 5. OGD quality analysis	- Task descriptions (task 15-18) - A question to assess the difficulty or ease of task 15-18 on a five-point Likert scale (very difficult – very easy) - A question to rank the difficulty of tasks 15-18 - A question on whether tasks 15-18 could be completed
Time measure 6	A question to write down the time at that moment
6. General questions	Questions about the influence of the user interface, number of datasets, available programmes and other factors on the difficulty or ease of conduct the scenario tasks

Table 7-9: Structure of the second participant survey.

The participants received the scenario tasks and instructions on paper. It was decided not to provide the scenario tasks in an online form, as this would require the participants to continuously switch between two tabs on their computer. Moreover, each of the six sections in the scenario tasks was displayed on a different page and with a different colour to make it easier for the observers to identify on which tasks the participants were working and whether they found it difficult or easy to complete the tasks. In the scenario instructions the participants were asked to indicate to which extent they found it difficult or easy to conduct the

scenario tasks, to rank the tasks based on their ease or difficulty and to indicate whether they were able to complete each individual task. Participants had been told that if they could not conduct a certain task, they had to move on to the next task. In that case they had been told to write down in the questionnaire that it was 'very difficult' to conduct the task and to state that they were not able to conduct the task. Completing the scenario tasks took the participants between 30 and 50 minutes.

Control infrastructure

During the treatment, participants of the treatment group used the ENGAGE infrastructure to complete scenarios, whereas participants of the control group used the DANS infrastructure (see www.dans.knaw.nl/en) to conduct the same scenario tasks. We refer to the DANS infrastructure as the control infrastructure. The control infrastructure has not been created as part of this study, but already existed before our research started. It was selected as the control infrastructure for the following reasons.

1) This infrastructure provided English translations of all the functions that were required for conducting the quasi-experiments. Because of the participation of non-Dutch participants in the quasi-experiments, the availability of the infrastructure in English was a pre-condition.
2) This infrastructure gave a realistic impression of open data provision by public agencies in The Netherlands. Twenty Dutch public organisations are already disclosing their data via the control infrastructure, including agencies of the Ministry of the Interior and Kingdom Relations, the Ministry of Health, Welfare and Sport, Statistics Netherlands, the Netherlands Organization for Scientific Research, the Research and Documentation Centre and The Netherlands Institute for Social Research. The control infrastructure is well-accepted for publishing research data from social sciences, humanities, behavioural sciences, geospatial sciences and other sciences, and Dutch ministries are nowadays obliged to store their data resulting from policy-oriented research at DANS (Data Archiving and Networked Services, 2014a).
3) This infrastructure was already used to store the crime and social data that we studied in our cases. A comparison of ENGAGE with the control infrastructure

allowed for examining whether the use of the examined crime and social data can be improved through metadata, interaction mechanisms and data quality indicators.

4) Compared to other existing open data infrastructures, the control infrastructure allowed for executing relatively many tasks related to open data use.

The selected control infrastructure was the most comparable existing open data infrastructure that we could find compared to the ENGAGE prototype. A comparison of ENGAGE and the control infrastructure showed that ENGAGE provided more metadata than the control infrastructure. Moreover, contextual and detailed metadata were hardly available on the control infrastructure. The control infrastructure barely provided any of the interaction mechanisms that were available on ENGAGE. The control infrastructure enabled viewing and adding data quality rating and data quality reviews, yet only in a limited way and it did not contain a wiki, social media sharing options or functionalities for the submission of related items.

Dataset selected for scenarios

All participants in both the treatment and the control group used the same dataset to ensure that differences in the findings could not be attributed to differences in the datasets that the participants used. The used dataset concerned Dutch parliamentary elections in 2002 and 2003 (referred to as the Parliamentary Elections Dataset). This dataset was chosen for the following reasons.

1) The dataset was completely in English which was required since non-Dutch participants were involved in the quasi-experiments;
2) The dataset could be analysed and visualised on the control infrastructure through an additional tool that could not be used for most other datasets;
3) Several data quality reviews were available for this dataset, while this was not the case for many other datasets in the control infrastructure;
4) The dataset was available through open access, whereas various other datasets on the control infrastructure first required obtaining permission for its use;

5) The dataset was well-connected to the social data case study that was described in chapter four (the SCP had been a partner in the creation of this dataset);
6) The dataset was relatively straightforward and well-documented, so that it allowed for drawing conclusions based on the data analysis (this was required for several scenario tasks).

Apart from this dataset, no other dataset was available that met all the above-mentioned criteria. However, the selected dataset did not contain any geographical coordinates, and therefore it did not allow for visualisations on a map, while this was a task in the visualisation scenario. For this reason another dataset that was stored at the control infrastructure was used only for the scenario task in which participants had to visualise the dataset on a map. This dataset concerned barrow groups in the southern part of The Netherlands in the Bronze Age (referred to as the Barrow Groups Dataset).

Time measures

Time measures were used to test whether the speed of OGD use by the treatment group was different from the speed of OGD use by the control group. Participants registered the time before and after they conducted each of the five scenarios. This enabled us to investigate whether the OGD use speed was improved by the metadata model, the interaction mechanisms and the data quality indicators. Time duration measures can be used to find out how much attention a person paid to an object (Webb et al., 1973, p. 134). In this study we assume that the more time is spent on a task, the more attention a person needs to perform the task and the more difficult the OGD use described in this task is. Yet, it should be noted that other factors may also influence how much time a person spends on a task, such as a person's character and perseverance, and the feeling of pressure from other participants and the facilitator to complete the tasks. It was therefore emphasised in the instructions that time measures were not used to assess the performance of the participants and that the participants should use as much time as they needed to conduct the scenario tasks.

Post-test: third participant survey

A third participant survey was used to test whether the ease of OGD use by the treatment group was different from the ease of OGD use by the control group. Table 7-10 shows the questions that were included in the third participant survey. The third participant survey consisted of 26 questions and was provided online (see Appendix G). It took the participants 15 to 20 minutes to conduct this final participant survey.

Parts of the third participant survey	Description
1. Evaluation of the session and the scenarios	1 question with 3 sub questions concerning the quasi-experiment and the scenarios
2. Evaluation of open data metadata	1 question with several sub questions on the assessment of metadata provided by the investigated open data infrastructure
3. Evaluation of open data interaction mechanisms and data quality indicators	1 question with several sub questions on the assessment of interaction mechanisms and data quality indicators provided by the investigated open data infrastructure
4. Evaluation of the prototype or control open data infrastructure	22 questions on the evaluation of the investigated open data infrastructure
5. Suggestions and comments	1 open question about suggestions and comments

Table 7-10: Structure of the third participant survey.

Observations

Observations were used to observe whether the ease of OGD use by the treatment group was different from the ease of OGD use by the control group. Observations were carried out during the pre-test, the scenarios and the post-test. Out of the eighteen observations that were conducted for the three quasi-experiments, twelve were conducted by persons who had significant experience with open data. All the involved observers had experience with Information and Communication Technologies.

Riley (1963) refers to two types of errors that participant-observation studies can be subject to. First, Riley refers to the 'control effect', which is the situation in which the measure process itself becomes an agent working for change and is unsystematic. Second, the biased-viewpoint effect refers to the situation in which the human observer "may selectively expose himself to the data, or selectively perceive them, and, worse yet, shift over time the calibration of his observation measures" (Webb et al., 1973, p. 114). By using an observation protocol and a semi-structured observer survey (see Table 7-11 and Appendices E

and F), we reduced the risk on having these two types of errors in the quasi-experiments. The observers were provided with the protocol, the observer survey and the scenario tasks before the quasi-experiments took place, and they were asked to read these documents carefully in advance. The protocol, the semi-structured observer survey and the scenario tasks were also explained in dedicated training sessions.

Parts of the observer survey	Description
Part 1: Observations per scenario task	- Time. For each of the five scenarios the observers were asked to write down approximately how long it took for the participants to complete the scenarios. - Difficulty. The observers were asked to which extent they thought that the participants found it difficult or easy to conduct each of the five scenarios (rated on a 5-point Likert scale from very difficult to very easy) and to explain their answer in an open text field.
Part 2: Observations of the scenarios in general	- Difficulty. The observers were asked to which extent they thought that the participants found it difficult or easy to conduct the scenarios in general (rated on a five-point Likert scale from very difficult to very easy) and to explain their answer in an open text field. - Influence from other factors. The observers were asked to which extent they thought that the user interface, the facilitator, the observers, other participants, the participant's gender, the participant's age, the participant's nationality, the participant's experience, the setting, the organisation of the quasi-experiment and any other aspects influenced the way that the observed participants conducted the scenarios
Part 3: Observations of collaboration and questioning	- Influence from other factors. The observers were asked to which extent the observed participants worked individually all the time while conducting the scenarios without discussing with anyone else, how many participants discussed at least once with (one of) their neighbour(s), how many participants intensively discussed more than once with (one of) their neighbour(s), how many participants asked questions to the facilitator of the quasi-experiment, and if any of the participants (and if so how many) complained about the difficulty of conducting the scenarios. - Other remarks. An open text field was provided for other comments.

Table 7-11: Structure of the observer survey.

Observers were responsible for observing a particular group of participants. It was defined in advance which observer would stand where exactly and which participants would be observed (see Figure 7-4). A participant code was assigned to each of the work places and was visible to the observers. The observers had

also received a map which allowed for knowing which participant had which code. As a consequence, the behaviour of participants could be related to the codes, which could then be related to the findings from the surveys and time measures.

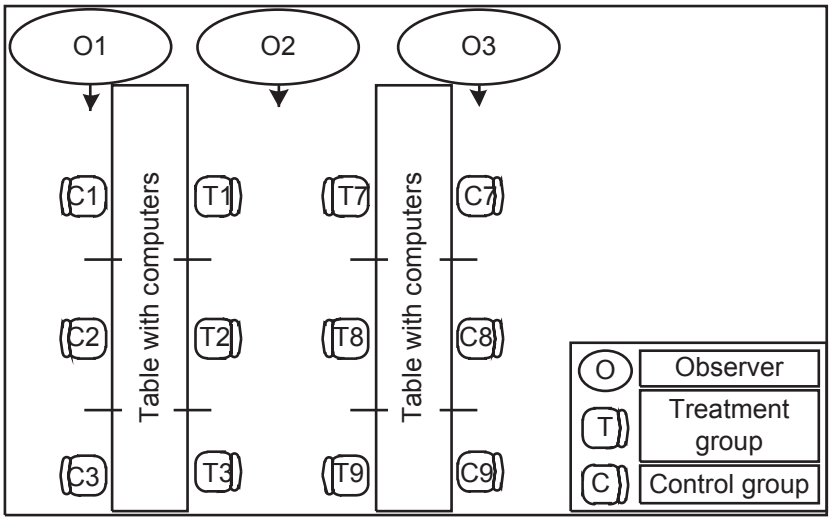


Figure 7-4: Organisation of the quasi-experiments.

Plenary discussion

Finally, in the plenary discussion the participants were asked which tasks they found most difficult, which tasks they found easiest and whether they had any suggestions to improve the investigated infrastructure. These plenary discussions were mainly used to obtain more information about limitations of the study, recommendations for further development of the prototype, and the influence from intermediate variables.

7.3 Data preparation

IBM SPSS Statistics version 20 was used to prepare and analyse the collected data. As part of the data preparation a reliability analysis was conducted to measure the consistency of the constructs of the model, which was required since the different types of OGD use were measured through a number of statements. Reliability can be defined as “the degree to which measures are free from error and therefore yield consistent results” (Peter, 1979, p. 6). Cronbach’s Alpha, which is also known as the reliability coefficient, was calculated to obtain information about the reliability of the constructs. Values of 0.7-0.8 are acceptable values for

Cronbach's alpha (Field, 2009, p. 675). Murphy and Davidshofer (1988) state that alpha values below 0.6 are unacceptable, values of 0.7 are low, values between 0.8 and 0.9 are moderate to high and values around 0.9 are high. Others (e.g., Davis, 1964; Nunnally, 1967) have recommended a lower acceptance boundary and believe that Alpha values between 0.5 and 0.6 can still be acceptable. Table 7-12 shows Cronbach's alpha values for the five constructs that are used in our model for both the pre-test and post-test. Except for the open data analysis construct in the pre-test, all Cronbach's Alpha values were moderate (.726) to high (.921). Cronbach's Alpha value for the open data analysis construct in the pre-test is lower (.633), yet not unacceptable.

	OGD use constructs	Number of items	Cronbach's Alpha
Pre-test	Searching and finding OGD	4	.772
	OGD analysis	4	.633
	OGD visualisation	3	.817
	Interaction about OGD	5	.899
	OGD quality analysis	4	.921
Post-test	Searching and finding OGD	4	.855
	OGD analysis	4	.795
	OGD visualisation	3	.726
	Interaction about OGD	5	.921
	OGD quality analysis	4	.917

Table 7-12: Reliability analysis of the constructs included in the pre-test and post-test (n=120, 7=missing).

7.4 Description of the quasi-experiment participants

This section elaborates on the characteristics of and the differences between the quasi-experiment participants, including their gender and age, nationality, open data experience, OGD use, and a discussion regarding the combination of data from different quasi-experiments. The participant information was derived from the first survey. Out of the 127 persons who participated in the quasi-experiments, 116 completed the first, second and third participant survey. Eleven persons completed only one or two of these surveys.

7.4.1 Gender and age

In all three quasi-experiments and in both the control and the treatment groups, the majority of the participants were male. The percentage of males per condition (control or treatment) in the three quasi-experiments ranged from 61 to 90. The

percentage of males in the treatment and control groups of the first quasi-experiment and of the control group in the second quasi-experiment were almost equal. The treatment group of professionals and the treatment group of students in the second quasi-experiment contained relatively more females (31% and 30% respectively) than the other groups.

The average age of all 120 participants who provided age information was 27,9 years. The youngest participant was 20 years old, while the oldest was 65. Participants of the third quasi-experiment were relatively older (μ=38,7, σ=12,4) than participants of the first quasi-experiment (μ=21,8, σ=1,7) and second quasi-experiment (μ=24,4, σ=2,0). The differences in age between the control and treatment groups within the first and second quasi-experiment were relatively small (see Figure 7-5).

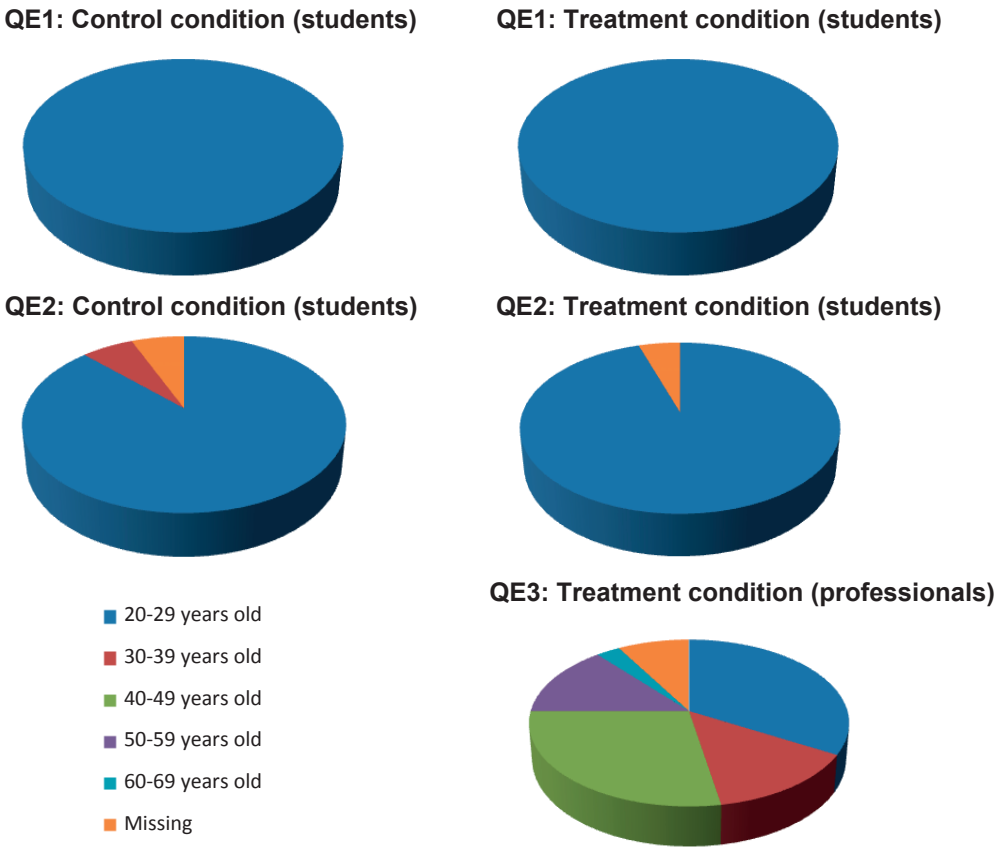


Figure 7-5: Age of the quasi-experiment participants.

7.4.2 Nationality

Out of the nine global country clusters identified by Northouse (2010), eight clusters were represented in the quasi-experiments (see Figure 7-6). In the first quasi-experiment all participants were Dutch (both in the control and in the treatment group), while in the second and third quasi-experiment more nationalities were represented. In the second quasi-experiment half of the treatment group as well as half of the control group consisted of participants from the Germanic European cluster. Other nationalities represented in the second quasi-experiment differed for the treatment and the control group. The control group of the second quasi-experiment contained more participants from the Southern Asian cluster (34,4% versus 5,0%), whereas the treatment group of the second quasi-experiment contained more participants from the Eastern European cluster (0,0% versus 15,0%) and from the Latin European cluster (3,1% versus 15,0%). These differences need to be taken into account in interpreting the results of this study.

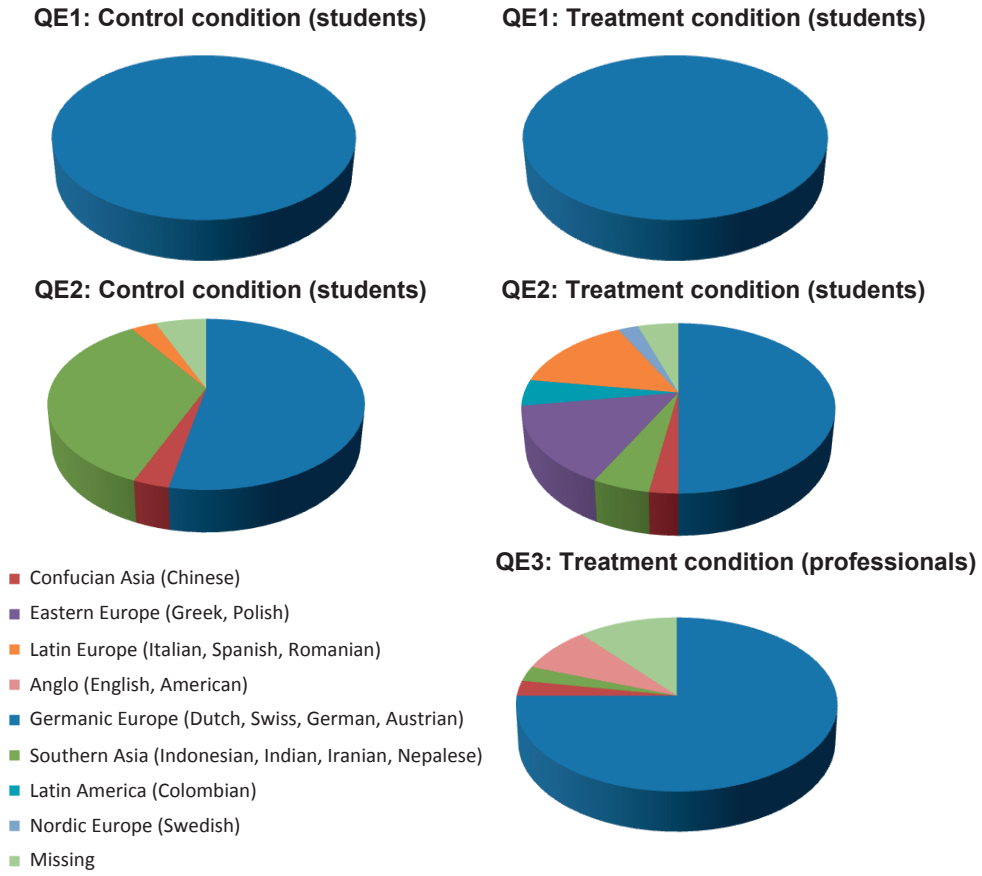


Figure 7-6: Nationality of the quasi-experiment participants.

7.4.3 Experience

In all quasi-experiments the minority of participants indicated that they had never used open data, varying from 10 to 41 per cent of the participants in the different groups of the three quasi-experiments. Compared to the treatment group participants, more control group participants of the second quasi-experiment had never used open data before (40,6% versus 27,5%). Yet, the control group of the second quasi-experiment contained more people who had used open data daily, and it also contained people who had used open data monthly or weekly (see Figure 7-7). Even though we involved professionals interested in and concerned with open data use in the third quasi-experiment, still ten per cent of the professionals stated that they had never used open data, yet they were professionally involved in open data.

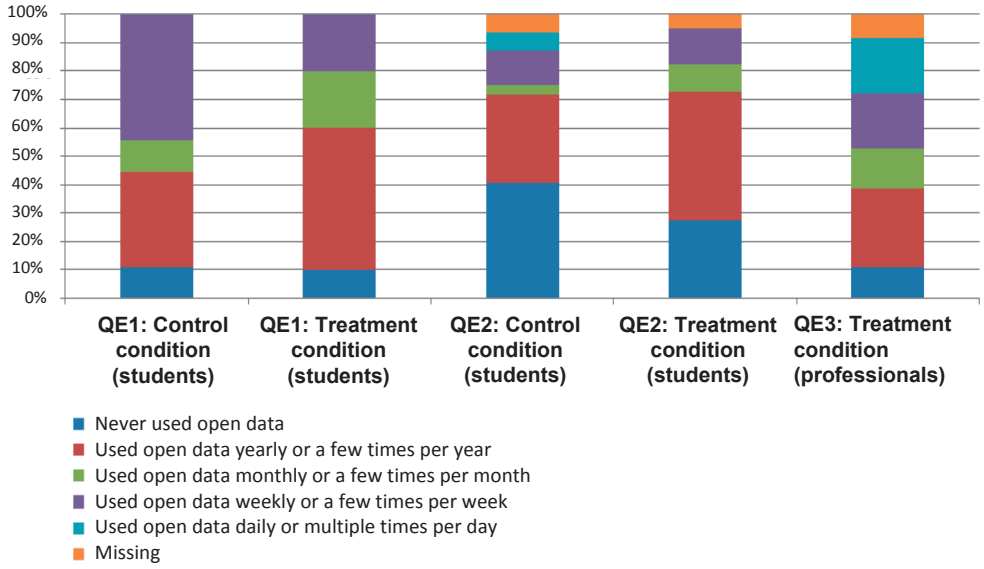


Figure 7-7: Open data use experience of the quasi-experiment participants.

Furthermore, in both the first and second quasi-experiment the self-reported level of experience of control group participants appeared to be slightly higher than the experience of the treatment group participants. On a scale from 1 to 10, the average self-reported experience varied between 4,0 and 6,3 (see Figure 7-8). Of those participants who had been involved in open data use, most had been involved in using open data for 2 to 5 years (33 participants) or for 5 to 10 years (24 participants).

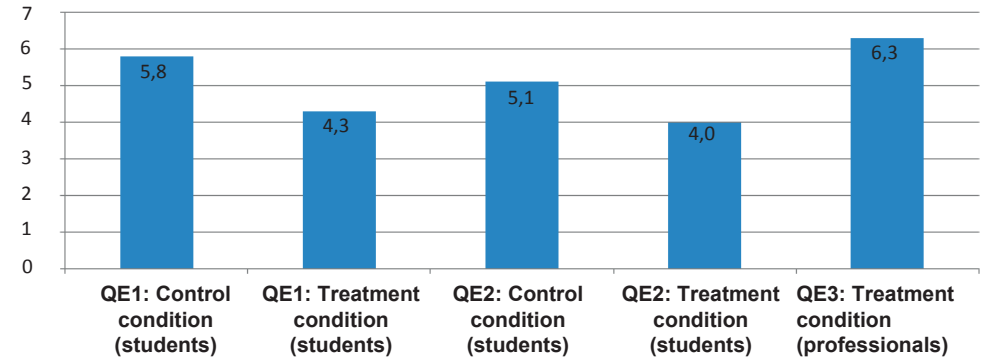


Figure 7-8: Average self-reported experience with open data use.

As far as the participants' involvement with data publication was concerned, the majority of the participants had never published open data. In the first and second quasi-experiment out of the 91 participants only 16 had published open data yearly or a few times per year, and only one person had done this monthly or a few times per month. As expected, the professionals in the third quasi-experiment had published open data more often. About one-third of the participants had published open data yearly or a few times per year, one-third had done this monthly or more often, and one-third had never done this.

7.4.4 OGD use

In line with our definition of OGD use in section 1.2, the evaluations focused on the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government. The evaluations involved a dataset that focused on the domain of social sciences (see section 7.2.5). Out of the 119 participants of the quasi-experiments who provided information about their daily occupation, 84 per cent stated that the role of researcher or student/citizen described them best. Since even within these roles the purposes for open data use may differ, in the quasi-experiments all participants had to conduct similar task from a similar role (see section 7.2.5). The tasks that the evaluation participants had to conduct were focused on studying a dataset through research to obtain useful information for policy making by governments. The data use was operational (to obtain insights for a study), and detailed data were needed to complete the scenario tasks. Just like in the case studies, the evaluations incorporated mainly Dutch open data users, yet also people with other nationalities.

7.4.5 Combining data from the first and second quasi-experiment

The first quasi-experiment involved only 10 treatment group participants and 9 control group participants. These two groups were too small to statistically analyse their results, since such an analysis would require each group to have at least 30 participants. It was therefore researched whether the results from 9 control group participants of the first quasi-experiments could be combined with the results of the 32 control group participants of the second quasi-experiment. These two groups both contained students with the same type of education, of approximately the same age (most were 20-29 years old) and of the same gender (more than 80%

were male). In both groups most students had some experience with open data use. In both groups most students were Dutch, although the treatment group of the first quasi-experiment contained only Dutch participants, while the treatment group of the second quasi-experiment contained also participants with other nationalities.

Moreover, it was examined whether the results from the 10 treatment group participants in the first quasi-experiment could be combined with the 40 treatment group participants of the second quasi-experiment. These two treatment groups both consisted of students with the same type of education, and of about the same age (>95% was 20-29 years old). In both groups most students had some experience with open data use. Although both groups contained mainly males, the treatment group of the first quasi-experiment contained more males than the treatment group of the second quasi-experiment (90% and 65% respectively). In both groups most participants were Dutch, although the percentage of Dutch participants also varied (100% versus 45%).

Although there were several differences between the participants on some measured characteristics, for most measured characteristics the participants from the treatment groups and the control groups were relatively similar. Since the treatment group participants of the first quasi-experiment appeared to be relatively comparable with the treatment group participants of the second quasi-experiment, the results of these two groups were combined. The same was done for the control group participants of the first and second quasi-experiment. We acknowledge that the respondents from the treatment groups and from the control groups may still be different with regard to certain characteristics that we did not measure. A limitation of this study is that we do not have insight in this.

7.5 Ease of OGD use

This section aims to present the evaluation results regarding the ease of OGD use (as an indicator for levels of coordination) utilising the developed OGD infrastructure. The findings of our evaluations have to be seen in the context of this study's focus on the use of structured research OGD from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures. The following three sub propositions are investigated.

- P1a: Metadata positively influence the ease of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis.
- P2a: Interaction mechanisms positively influence the ease of interaction about OGD.
- P3a: Data quality indicators positively influence the ease of OGD quality analysis.

Insight in the ease of OGD use was obtained through the participant surveys (see section 7.5.1) and the observations (see section 7.5.2).

7.5.1 Surveys

This section first discusses the results from the three participant surveys. The results from the pre-test and post-test are compared, and intermediating variables are described. Finally, we report on the overall ease of OGD use for each of the examined participant groups.

First survey: pre-test results

"To get a 'reasonable' estimate of the treatment effect, the analysis must properly […] control for the effects of […] initial differences [between groups]" (Reichardt, 1979, p. 149). For this reason pre-tests were conducted. The pre-tests were used to find out to which extent existing OGD infrastructures that the participants had already worked with allowed for conducting the tasks from our OGD use scenarios. With regard to the OGD use tasks of the first three scenarios (searching for and finding OGD, OGD analysis and OGD visualisation), the participants indicated that on average they neither agreed nor disagreed that at least one of the open data infrastructures they knew enabled them to conduct the activities related to searching for and finding open data, analysing open data and visualising open data. The only exception concerned one function related to searching for and finding OGD, as they on average slightly agreed that they knew an open data infrastructure in which they could use various options to search for data through keywords, categorisations, filters and translations. It was found that the means of the control and treatment groups of students and professionals were relatively similar for the first three scenarios.

As far as the OGD use tasks of the fourth and the fifth scenario were concerned (interaction about OGD and OGD quality analysis), it was found that the results of the control and treatment groups were not as equal as for the first, second and third scenario tasks. The control group participants indicated that on average they neither agreed nor disagreed that at least one of the open data infrastructures they knew enabled them to conduct activities related to interaction about OGD and OGD data quality analysis. The treatment group of students on average slightly disagreed that they knew an open data infrastructure which enabled four out of the nine functions, whereas they were neutral about the availability of the other five functions. For all functions the treatment group of professionals slightly disagreed or disagreed that they knew an open data infrastructure which enabled functions related to interaction mechanisms and data quality indicators.

It is unclear why the control group was slightly more positive in the pre-test than the treatment groups of students and professionals. We speculate that the difference might be caused by differences in the nationality of the participants. For example, the nationality of the control group participants may have led to answering the questions in a way that the participants thought was more socially desirable. However, the differences in nationalities seem not to have influenced the results of the pre-test for the first, second and third scenario. Moreover, the participants of all the involved participant groups had indicated that they disagreed with the statement that the facilitator of the quasi-experiments had influenced their behaviour ($\mu = 3.21$ out of 7 for the control group, $\mu = 3.21$ for the students treatment group and $\mu = 2.61$ for the professionals treatment group). On average the participants agreed that the facilitator of the quasi-experiments had a neutral attitude ($\mu=5.74$ out of 7 for the control group, $\mu=5.83$ for the students treatment group and $\mu=5.70$ for the professionals treatment group).

Second survey: functionalities enabled by the infrastructure

In the second participant survey, the participants were asked for each scenario task whether they were able to complete it. It was found that only a small percentage of the participants was not able to complete the tasks related to the first scenario (searching for and finding OGD). About 84 per cent of the control group

participants could complete the tasks related to scenario one, while this was possible for 96 per cent of the students treatment group and for 79 per cent of the professionals treatment group. This means that the participants of the professional treatment group were slightly less often able to conduct these scenario tasks than the control group participants, although the difference was small. For the second scenario (OGD analysis) these numbers were 64 per cent, 93 per cent, and 79 per cent respectively, meaning that the control group participants had more difficulties with conducting the tasks of the second scenario. On average 38 per cent of the control group participants could complete the tasks of the third scenario (OGD visualisation), while this was possible for 77 per cent of the students treatment group and 58 per cent of the professional treatment group.

For the fourth and fifth scenario the differences between the control group and treatment group were larger than for the first, second and third scenario. Moreover, participants of the treatment group of students were able to complete scenario tasks related to scenario four and five more often than the treatment group of professionals. The tasks of the fourth scenario (interaction about OGD) could be completed by 20 per cent of the control group participants, while 79 per cent of the treatment group of students and 61 per cent of the treatment group of professionals could do this. On average, the tasks of the fifth scenario (OGD quality analysis) could be completed by 84 per cent of the students' treatment group and by 57 per cent of the professionals' treatment group. Only 8 per cent of the control group participants indicated that they could complete the tasks of the fifth scenario. In general, the second participant survey showed that the treatment group participants were more often able to conduct the scenario tasks except for the first scenario.

Third survey: post-test results

After the participants had completed the five scenarios, they were asked to complete a third survey. With this third survey we measured the post-test results (also see Table 7-13). It was found that the control group participants were neutral or slightly disagreed that the open data infrastructure that they had used enabled the tasks of the first, second and third scenario (the scenarios related to searching for and finding OGD, OGD analysis and OGD visualisation). The professional open

data users in the treatment group were relatively more positive than the control group. The students who had participated in the treatment group were most positive. On average, the students of the treatment group slightly agreed that the infrastructure that they had used enabled the scenario tasks of the first three scenarios.

With regard to scenario four (interaction about OGD) and scenario five (OGD quality analysis) the findings were slightly different. Even though the control group was more positive in the pre-test, it was found that the post-test values for the treatment group were all higher than for the control group, meaning that the participants agreed that the OGD infrastructure had enabled them to complete the tasks related to interaction about OGD and OGD quality analysis. The differences between the control and treatment group were slightly larger for scenarios four and five than they were for scenarios one, two and three.

The tasks of the first three scenarios that were easiest to conduct for the control group concerned understanding what the dataset that participants found is about (μ=4.79) and viewing datasets without downloading them (μ=4.54). All mean values of the control group were between 1.85 and 3.18 for the fourth and fifth scenario, which indicates that the control group was on average relatively negative about the use of the open data infrastructure for these tasks. The highest values for the fourth and fifth scenario concerned the tasks of viewing quality ratings of the dataset (μ=3.18) and rating different quality aspects of the dataset (μ=2.33). The most difficult tasks for the control group in the first three scenarios concerned visualising data on a map (μ=2.23), visualising data in a chart (μ=3.54), and drawing conclusions based on the data that participants found (μ=3.54). The most difficult tasks in the fourth and fifth scenario according to the control group were discussing what can be learned from data use by leaving a discussion post (μ=1.85) and discussing what can be learned from the data use on a wiki or forum (μ=2.03).

The treatment groups were most positive about viewing datasets without downloading them (μ=6.38 for the student and μ=5.76 for the professionals) and visualising data in a table (μ=6.37 for the students and μ=5.73 for the professionals). In the post-test of scenarios four and five, the treatment group was most positive about viewing quality ratings of datasets (μ=6.13 for the student

treatment group and μ=4.82 for the professional treatment group) and discussing what can be learned from data use by leaving a discussion post (μ=5.75 for the student treatment group and μ=4.82 for the professional treatment group). The most difficult task found for the treatment group concerned discussing what can be learned from the data use on a wiki or forum (μ=4.25 for the student treatment group and μ=3.85 for the professional treatment group). On average the participants in the student treatment group were relatively positive about the use of the infrastructure for interaction about OGD and OGD quality analysis. The treatment group of professionals was slightly less positive than the students, yet they were still more positive than the control group and more positive than they were in the pre-test.

Comparison of pre-test and post-test results

Table 7-13 provides the means and standard deviations of the pre-tests and post-tests for each of the five OGD use scenarios. The scenario tasks were rated on a Likert scale from 1 to 7. In the pre-test the participants responded to statements that started with “at least one of the open data infrastructures that I know enables me to…”, which were followed by an OGD use activity. In the post-test the statements were formulated in the form of: “the open data infrastructure enabled me to...”, followed by the same OGD use tasks. A mean score of 1 means that respondents strongly disagreed with the statement, indicating a negative response. A mean score of 7 means that respondents strongly agreed with the statement, indicating a positive response. Mean values around 4 point at a neutral respondent’s attitude.

	Ease of OGD use					
	Control group (students, *n=39, 2 missing*)		Treatment group (students, *n=48, 2 missing*)		Treatment group (professionals, *n=33, 3 missing*)	
	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test
Scenario 1: Searching for and finding OGD	μ: 4.89 σ: 0.77	μ: 4.25 σ: 1.37	μ: 4.83 σ: 0.88	μ: 5.45 σ: 0.81	μ: 4.70 σ: 1.16	μ: 4.93 σ: 1.23
Scenario 2: OGD analysis	μ: 4.72 σ: 0.76	μ: 4.14 σ: 1.41	μ: 4.74 σ: 0.86	μ: 5.67 σ: 0.69	μ: 4.48 σ: 1.15	μ: 4.86 σ: 1.06
Scenario 3: OGD visualisation	μ: 4.52 σ: 1.16	μ: 3.21 σ: 1.45	μ: 4.50 σ: 1.07	μ: 5.22 σ: 1.19	μ: 4.36 σ: 1.56	μ: 4.43 σ: 1.51
Scenario 4: Interaction about OGD	μ: 4.39 σ: 0.84	μ: 2.16 σ: 1.10	μ: 3.94 σ: 1.22	μ: 4.95 σ: 0.89	μ: 3.22 σ: 1.59	μ: 4.45 σ: 1.42
Scenario 5: OGD quality analysis	μ: 4.31 σ: 1.16	μ: 2.45 σ: 1.23	μ: 3.87 σ: 1.29	μ: 5.63 σ: 1.02	μ: 2.90 σ: 1.60	μ: 4.48 σ: 1.85

Table 7-13: Means and standard deviations of the open data use scenarios on a Likert scale from 1 to 7 (n=127).

When the pre-test results for the five scenarios are compared to the post-test results for these scenarios, it can be found that the post-test results of the control group are all more negative than the pre-test results of the control group. This suggests that the control OGD infrastructure functioned worse than the participants had expected based on their experience with other existing OGD infrastructures.

On the contrary, the post-test results of the students treatment group for the first three scenarios were more positive than the pre-test results of this group, except for one function (i.e. to use various options to search for data), as the participant were already relatively positive about this function in the pre-test. When we compare the mean values from the pre-test and the post-test for scenarios four and five, the post-test values of the treatment groups are all higher than the pre-test values. This suggests that the OGD infrastructure used by the student treatment group in general performed better than other OGD infrastructures that the participants had experience with.

For the treatment group of professionals, eight out of the eleven post-test results were assessed more positively than the pre-test results of this group. This was not the case for the findings related to three functions, namely 1) to draw conclusions based on the data that they found, 2) to visualise data in a chart and 3) to visualise data on a map. On these three aspects the infrastructure functioned slightly worse than the participants had expected based on their experience with

other OGD infrastructures. The problems with data visualisations were illustrated by quotes of the participants. For example, professional open data users stated that they *"didn't find the visualization tools easy to use"*, *"the visualising (chart, graph, map) was a bit difficult to use"*, and *"the icons for table, graphs and map in the visualisation part seem redundant"*. This shows that there is potential to further improve the developed OGD infrastructure, yet the results from the treatment groups are still more positive than the results from the control group.

To be able to measure whether the level of difficulty of conducting the scenario tasks was significantly different for the control and treatment groups, the Mann-Whitney Test was conducted (Mann & Whitney, 1947). This test is appropriate since the quasi-experiments produced one independent categorical outcome variable with two categories (whether a person participated in the control or treatment group) and one continuous dependent variable (the mean value of the level of difficulty of conducting scenario tasks). The Mann-Whitney test is the non-parametric equivalent of the independent t-test (Field, 2009, p. 540), which had to be used since the sample did not meet the assumptions for parametric tests (the data were not normally distributed).

Table 7-14 provides the results of the Mann-Whitney test and the medians of the compared groups. The Mann-Whitney test shows that the level of difficulty of scenario tasks related to all five open data scenarios of the student treatment group differed significantly from the level of difficulty of these tasks of the student control group. On average the students in the treatment group found it easier to conduct scenario tasks related to searching for and finding OGD (scenario 1), OGD analysis (scenario 2), OGD visualisation (scenario 3), interaction about OGD (scenario 4) and OGD quality analysis (scenario 5) than the students in the control group.

It was not possible to compare a control group of professionals to a treatment group of professionals. The control group of students can be compared to the treatment group of professionals. However, since we found that there are certain differences between the professionals and the students which may have influenced the outcomes (e.g. differences between age, daily occupation and experience with open data use, see section 7.4), these findings need to be interpreted with caution.

	Median of control group (students, n=39, 2 missing)	Median of treatment group (students, n=48, 2 missing)	Mann-Whitney U
Scenario 1: Searching for and finding OGD	4.50	5.50	410.50**
Scenario 2: OGD analysis	4.25	5.75	318.00**
Scenario 3: OGD visualisation	3.67	5.33	259.50**
Scenario 4: Interaction about OGD	2.00	5.00	68.50**
Scenario 5: OGD quality analysis	2.00	6.00	63.50**

* p <.05, ** p <.001

Table 7-14: Comparison of the level of difficulty of scenario tasks for the student control group and the student treatment group.

The Mann-Whitney test revealed that the level of difficulty of the scenario tasks related to all five open data scenarios of the professionals treatment group differed significantly from the level of difficulty of these tasks of the student control group (see Table 7-15). The findings suggest that on average the professionals in the treatment group found it significantly easier to conduct scenario tasks related to searching for and finding OGD (scenario 1), OGD analysis (scenario 2), OGD visualisation (scenario 3), interaction about OGD (scenario 4) and OGD quality analysis (scenario 5) than the students of the control group.

	Median of control group (students, n=39, 2 missing)	Median of treatment group (professionals, n=33, 3 missing)	Mann-Whitney U
Scenario 1: Searching for and finding OGD	4.50	5.25	450.50*
Scenario 2: OGD analysis	4.25	5.00	439.50*
Scenario 3: OGD visualisation	3.67	4.33	372.50*
Scenario 4: Interaction about OGD	2.00	4.80	143.50**
Scenario 5: OGD quality analysis	2.00	4.50	241.50**

* p <.05

** p <.001

Table 7-15: Comparison of the level of difficulty of scenario tasks for the student control group and the professionals treatment group.

Intermediate variables reported in the participant surveys

Although we found that the developed infrastructure enhances the coordination of OGD use, we cannot claim that these effects have only been caused by the implemented metadata model, interaction mechanisms and data quality indicators. The participant surveys provided insight in several intermediate variables which may have influenced the OGD use on the developed infrastructure. Table 7-16 provides the means and standard deviations for six statements regarding intermediate variables concerning characteristics of the facilitator of the quasi-experiment and the quasi-experiment itself. A mean of one indicates that the respondent strongly disagrees with the statement, while a mean of seven refers to strong agreement. The table shows that the participants of all quasi-experiments on average agreed that the quasi-experiment was well-organised and well-structured. The participants of all control and treatment groups disagreed that the facilitator of the quasi-experiment had influenced their behaviour and they agreed that the facilitator had a neutral attitude. On average, the participants indicated that they were neutral about or slightly agreed with the statement that the scenarios reflected open data use in a realistic way. The participants of the treatment group stated that they learned something by participating in the quasi-experiment, while the control group participants were more neutral about this.

	Control group (students, n=39, 2=missing)	**Treatment group (students, n=48, 2=missing)**	**Treatment group (professio-nals, n=33, 3=missing)**
The practical session on open data use was well-organised.	μ: 5.08 σ: 1.60	μ: 6.19 σ: 1.02	μ: 5.15 σ: 1.18
The session was well-structured (clear sequence).	μ: 5.51 σ: 1.62	μ: 6.38 σ: 0.67	μ: 5.82 σ: 0.73
The facilitator of this session influenced my behaviour during the session.	μ: 3.21 σ: 1.56	μ: 3.21 σ: 1.88	μ: 2.61 σ: 1.46
The facilitator of this session had a neutral attitude.	μ: 5.74 σ: 1.09	μ: 5.83 σ: 1.12	μ: 5.70 σ: 1.02
The scenarios reflected the use of open data in a realistic way.	μ: 4.51 σ: 1.37	μ: 4.83 σ: 1.48	μ: 4.64 σ: 1.54
I learned something by participating in the session.	μ: 4.31 σ: 1.64	μ: 5.27 σ: 1.33	μ: 5.48 σ: 1.15

Table 7-16: Means and standard deviations for questions about intermediate variables in the quasi-experiment derived from the participant surveys.

In the surveys, the quasi-experiment participants were also asked which other intermediate variables may have influenced their performance. The participants of the quasi-experiments pointed at the considerable influence of the user interface. In the control group 63 per cent of the participants indicated that the user interface had negatively influenced the difficulty to conduct the scenario tasks, while 20 per cent said that the user interface had had a positive influence and 10 per cent said it had not influenced their performance. In contrast, 58 per cent of the participants of the student treatment group and 41 per cent of the participants in the professionals treatment group stated that the user interface of the infrastructure had positively influenced their performance.

Participants in the treatment condition also mentioned that their performance on the open data infrastructure was positively influenced by the use of clear and big buttons. For instance, a participant from the treatment group stated that *"using big buttons at places where you expect them […] makes using the infrastructure very user-friendly".* In addition, the clarity of the buttons, headings and logo's, and the organisation of the interface were seen as positive aspects of the open data infrastructure. Participants stated that that "*the clear buttons, logical symbols and clear setup of the page*" and "*putting the information under clear headings and using logos*" had positively influenced their performance. Another participant pointed at the importance of contrast and colours on the infrastructure (*"important buttons have a different contrast or colour"*). Several participants wrote that the user interface had positively influenced their performance by clearly presenting the possibilities of the infrastructure and the results of data analysis. It was stated that: "*the simple design of the website makes it easier to find data*", the "*clear friendly interface makes users easily find useful information*", the "*clear structure by the different sections makes things easy to find*" and "*it was easy to view the data in an efficient way*".

On the other hand, the user interface had also negatively influenced the performance of control group participants. They mentioned that the use of small symbols and the non-intuitive user interface had a negative influence on the way that the control open data infrastructure could be used. One control group participant reported: *"the UI [(User Interface)] is not intuitive, a lot of tasks couldn't be completed"*, which was confirmed by other participants: *"it's not clear where to*

click to conduct the tasks", "it was not clear what functions could be used and where data could be found", "*not a very friendly user interface*", and *"DANS really needs to improve its interface (make it more intuitive)".* Especially the font size seemed to have a negative influence (*"the size of the words was very small. It was not very clear what to do where".* Other control group participants wrote down: *"very small symbols", and "the pictograms being very small, it is more difficult to know they are to be used"*).

An additional hindering intermediate variable concerned a lack of tools to search for and filter data. Participants of the control group said that *"it was not clear […] where data could be found"* and "*too much data is 'duttered' and the navigation menu [is] also confusing"*). Regarding difficulties resulting from a lack of experience, participants stated that there was a negative influence from *"having no experience at all with this kind of programs", "having never worked with ENGAGE"* and *"general IT-skills".* Another participant stated: *"this was the first time I looked really into it so it took me some time to search".* Moreover, regarding the number of datasets, it appeared that the number of datasets provided may have a negative influence. A control group participant mentioned that *"the index gives too many different datasets to be a good overview",* and a treatment group participant wrote that the "*huge number of datasets makes searching become hard*". Required registration and problems with signing in on the infrastructure were other hindering factors. The control group participants mentioned, for example: *"registration is always required when you want to analyse something. And you need more than one account"*, *"constantly logged out (against will), 'relog' wasn't always possible"* and *"login was not possible, frustration occurred"*.

The programmes used appeared to be another intermediating variable, that both hindered and enabled OGD use. For instance, a participant of the treatment group stated *"Excel online is a little difficult to use"*, while another participant disagreed with this and pointed at the positive influence of *"programs that help to visualise the data (Excel)"*. One participant wrote that *"the programmes are very useful and easy to understand"* and another pointed at the *"easy to use graphs and filters"*, whereas yet another participant stated that "*the way to create a chart was not easy to use"*. It is important to keep the above-mentioned intermediating variables in mind when interpreting the results of this study and in the development

of OGD infrastructures, since these factors appeared to either positively or negatively influence the coordination of OGD use.

Overall ease of OGD use in the entire infrastructure

Apart from investigating the ease of OGD use per scenario task, the participants of the quasi-experiments were also asked to which extent they believed that the complete infrastructure improved OGD use (see Table 7-17). A mean of one in the table means that respondents strongly disagreed with the statement (e.g. that they were strongly dissatisfied), indicating a very negative response. A mean of seven means that respondents strongly agreed with the statement (e.g. that they were strongly satisfied), indicating a very positive response. Mean values around four indicate a neutral attitude of the respondent. The table shows that participants of the treatment group stated that the infrastructure had enabled them to use open data more easily than the control group participants.

	Control group (students, n=39, 2=missing)	**Treatment group (students, n=48, 2=missing)**	**Treatment group (professio-nals, n=33, 3=missing)**
Using the infrastructure makes it easier to use open data.	μ: 4.03 σ: 1.69	μ: 5.67 σ: 0.91	μ: 5.03 σ: 1.36
I find it easy to use the open data infrastructure.	μ: 3.05 σ: 1.69	μ: 5.77 σ: 0.88	μ: 5.06 σ: 1.37
Learning to use the infrastructure is easy for me.	μ: 3.56 σ: 1.54	μ: 6.02 σ: 0.76	μ: 5.27 σ: 1.38
It is easy for me to become skilful at using the infrastructure.	μ: 3.62 σ: 1.53	μ: 5.81 σ: 0.89	μ: 5.12 σ: 1.39

Table 7-17: Means and standard deviations for the ease of OGD use.

7.5.2 Observations

The ease of OGD use was not only measured through surveys, but also through observations. This section describes the findings from the observations. Table 7-18 gives an overview of the assessment of the quasi-experiments by the observers. Out of the eighteen observers that were involved, five persons observed the control group, six observed the treatment group of students and seven observed the treatment group of professionals. In the following sections the findings from the observations are discussed for each of the five OGD use scenarios, and for the

OGD use scenarios as a whole. Subsequently, the observation results regarding potential intermediate variables are provided.

		Very difficult or difficult	Not difficult, nor easy	Easy or very easy	Total number of observations
Scenario 1: Searching for and finding OGD	Control group (students)	3	1	1	5
	Treatment group (students)		1	5	6
	Treatment group (professionals)		1	6	7
Scenario 2: OGD analysis	Control group (students)	3	2		5
	Treatment group (students)	1	3	2	6
	Treatment group (professionals)	1	3	3	7
Scenario 3: OGD visualisation	Control group (students)	4	1		5
	Treatment group (students)		2	4	6
	Treatment group (professionals)	5	2		7
Scenario 4: Interaction about OGD	Control group (students)	4			5 (1 missing)
	Treatment group (students)		3	3	6
	Treatment group (professionals)			7	7
Scenario 5: OGD quality analysis	Control group (students)	4	1		5
	Treatment group (students)	1		5	6
	Treatment group (professionals)			6	7 (1 missing)
All scenarios	Control group (students)	5			5
	Treatment group (students)		3	3	6
	Treatment group (professionals)		3	4	7

Table 7-18: Number of observers and their assessment of the difficulty of the OGD use scenarios.

Observations of OGD searching and finding scenario

In general the observers found that the first scenario on searching for and finding OGD was easier to conduct by the treatment group than by the control group. The control group observers wrote that most students can find the right file, but that "*in all cases simple keyword search gives too many results*", that there was a "*lack of an appropriate search mechanism*" and that they found it difficult to sort and filter the data. Two observers of the treatment groups stated that it was not that easy to conduct the first scenario, although none of them indicated that it was difficult. In contrast, other observers wrote that the students "*almost instantly found the data*" and "*they found the relevant dataset quickly*".

Observations of OGD analysis scenario

In the second scenario, the participants analysed open data, for instance, by looking at the metadata and by deriving conclusions regarding patterns in the data. The second scenario was not assessed as easy by any of the observers of the control group. One observer postulated that the participants became frustrated, as they "*spent much time looking for useful information*". To quote from another observer, *"a considerable large percentage of the participants had quite a lot of issues to complete the tasks"*.

The second scenario was observed also not to be so easy for the participants of the treatment group, although in general the observers of the treatment groups were more positive than the observers of the control group. Two observers of the student treatment groups stated that the participants "*spent quite some time to find the solutions"* and that *"this part obviously takes more time than others*", while another observer stated that the participants "*found the data and tools quickly but also liked to explore other options*". Exploring other options might have resulted in spending more time on this scenario. This also seemed to be the case in the treatment group of professionals, since an observer of the professionals stated: *"there are a lot of functionalities (buttons), they spent a lot of time to play with the tool"*. Another observer wrote that "*it's easy to view the dataset and metadata, but participants found it difficult to draw a conclusion*".

Observations of OGD visualisation scenario

The third scenario concerned data visualisations, including visualisations in tables and charts and on maps. This scenario was observed to be easiest for the treatment group of students. The control group and the treatment group of professionals had more difficulties with this scenario. An observer of the control group stated: "*I could not see many of them being able to visualise data*." Two observers of the professional treatment group stated that "*most of the participants had some small issues with the visualisation*" and "*some people find it easy to view the dataset in table format, but it's not easy to view the dataset in chart or map format*". The treatment group of students was found to have fewer difficulties with the third scenario: "*quite an easy task, for all found the solutions relatively easy*".

Observations of interaction about OGD scenario

The interaction about OGD scenario encompassed giving feedback on and discussing data use. Participants of the quasi-experiments could post messages, discuss data use on social media and a Wiki and submit an item related to the original dataset (e.g. an improved or extended dataset). The observers of the control group observed that it was difficult or very difficult for the participants to use the control infrastructure for these purposes. The observers of the control group wrote that "*participants found it quite difficult […] and they gave up too quickly*". One observer mentioned that he or she "*had not seen them writing comments or giving feedback*".

The observers of the treatment group of students took the view that it was easier for the students to conduct the interaction scenario. Most of these participants could find the functions that they needed to participate in OGD discussions, although it still took them some time: "*they are trained to form opinions but taking the task seriously they took time to formulate their opinion well*". One observer mentioned the difficulty for the students in the treatment group to find the Wiki: "*the students can easily find the discussion under the selected dataset, but it is difficult for them to find a forum or wiki related to the dataset*". The observers of the treatment group of professionals all stated that the participants found it easy or very easy to conduct the interaction scenario ("*it's really easy for them to look at other people's comments under the specific dataset*"). The professional OGD users may be more trained to interact about OGD use and to formulate their opinion regarding various aspects of the dataset.

Observations of OGD quality analysis scenario

The fifth scenario concerned the analysis of OGD quality, for example through ratings and reviews. Participants could view how other users assessed the dataset, they could discuss the quality of the dataset by leaving a message or review in the data quality rating system, and they could view information about the person who had evaluated the quality of the data. The observers indicated that it was relatively difficult for the control group participants to conduct this scenario, as illustrated by one control group observer who said that "*all of them could not finish the tasks*". Another control group observer stated that "*only some of them saw the rating indexes. No one has been seen to leave a message*". It was observed that it was

difficult to use the control OGD infrastructure to rate and review the quality of open datasets.

On the other hand, the observations showed that the students and professionals of the treatment groups found this scenario less difficult to complete. The observers of the treatment group wrote that this was an "*easy task for all the participants*" and that "*even the slowest participant did it in a few minutes*". Another observer mentioned that "*students find it easy to rate the dataset*". It was observed by all six observers of the professional treatment group that it was easy or very easy to use the developed OGD infrastructure for data quality analysis purposes. While five other observers also mentioned that students in the treatment group found it easy or very easy to use the developed OGD for data quality analysis purposes, one observer indicated that the observed students found this difficult or very difficult. A motivation for this statement was not provided.

Observations in general

Finally, the observers were asked to assess the difficulty and ease of all scenarios together. The observers of the control group all said that it was difficult or very difficult to complete the total of five scenarios, whereas the observers of the treatment group postulated that it was neither difficult nor very difficult. Opinions of the observers of the treatment groups were divided; some stated that it was neither difficult nor easy to conduct all five scenarios, while others stated that it was observed to be easy. In general, most scenarios were observed to be less difficult to conduct for the treatment groups than for the control group. Even though the observers did not always agree about the difficulty of conducting scenarios, the observers generally believed that the five scenarios were more difficult to conduct for the control group than for the treatment groups.

Observations of intermediate variables

Various intermediate variables were investigated through the observations. Intermediate variables can be defined as variables which might influence the effect of the independent on the dependent variables (Pearl, 2001), yet it is not clear whether this influence actually exists and what its nature is. An example of such an intermediate variable is the setting (e.g. location, noise and light) in which the quasi-experiments took place, which might have influenced the results. The

observers were asked to which extent they believed that various intermediate variables besides the functional elements of the infrastructure may have influenced the difficulty or ease of performing the scenario tasks by the participants of the quasi-experiments. We first discuss the intermediate variables observed by the observers of the control group, followed by those observed in the student treatment group and the professional treatment group.

Out of the five observers of the control group, four wrote that the performance of the control group had negatively been influenced by the user interface. It was mentioned that the user interface of the control infrastructure was not user-friendly and could be improved. Furthermore, two observers wrote that some of the observed participants might have been influenced slightly by the observers, since it was visible to the participants that they were observed. Four observers wrote that some participants may to some extent have influenced other participants, and one wrote that previous experience had influenced their performance ("*their previous experience was important*"). Moreover, one observer wrote that the setting may have influenced the performance of the participants, because it was difficult for some participants to see the screen with instructions, although these participants could still hear the instructions as presented by the facilitator of the quasi-experiments, and they had also received a summary of the instructions on paper. Finally, one observer wrote that the temporal unavailability of the server of the control group had frustrated some control group participants. No influence was observed from the facilitator, the nationality of the participants or the organisation of the quasi-experiment ("*it was clear for all the participants what to do*").

The observers also investigated intermediate variables for the student treatment group. Out of the six observers of the student treatment group, six wrote that the difficulty or ease of performing the scenarios had been influenced by the user interface ("*I believe that the ease and the simplicity of the user interface played a major role*"). This finding indicates that the usability of the OGD infrastructure may not only be the result of the functional infrastructure elements, but that the user interface also plays an important role. Since the observers of the control group also found that the user interface could negatively influence the

usability of the infrastructure, it is important for OGD infrastructures to devote sufficient attention to the development of the user interface.

Two of the six observers of the student treatment group wrote that some participants might have been influenced slightly by the observers, since the participants could see that they were observed. Two observers wrote that some participants may to some extent have influenced other participants, as a few participants asked each other questions. Two observers wrote that the nationality of two participants may have made it slightly more difficult to conduct the tasks for non-Dutch participants, which may have been caused by the fact that the participants had to reuse a dataset (in English) that concerned Dutch elections. The participants themselves did not complain about this. One observer stated that some noise of participants who had completed their tasks in the quasi-experiment may have influenced the performance of other participants who were still working on the tasks. It was observed that the organisation of the quasi-experiment was clear to almost all participants. One observer mentioned that "*one person didn't understand some instructions*", while other observers stated that there was no influence from the organisation of the quasi-experiment at all ("*everything was well-prepared, participants were not distracted by unclear aspects*"). The observers of the student treatment group did not observe any influence from the facilitator of the quasi-experiment. It was noted that the experience of the participants in the student treatment group mattered ("*their e-skills are so high, they immediately got used to the interface*").

The third group that was observed in the quasi-experiments included the treatment group of professional OGD users. Out of the seven observers of the professional treatment group, five wrote that the difficulty or ease of performing the scenarios had been influenced by the user interface. This observation confirms the observations of the participants of the control and the student treatment groups, and it shows the importance of paying considerable attention to the user interface to improve OGD use. Moreover, three observers wrote that the participants' experience had influenced their performance ("*more experienced participants performed better and faster*"). This finding suggests that OGD use can be improved by training potential users, so that they obtain experience with the use of OGD infrastructures.

Some of the observers of the professional treatment group noticed that several participants rushed through the fifth scenario and some also rushed through the fourth scenario when they found out that only little time was left to complete the scenarios. Some observers stated that these participants indicated that the scenario tasks of scenario four and five were very difficult to complete, even though they had not even tried to complete them. This may have resulted in a slightly more negative mean for the professionals.

Moreover, some observers said that older participants worked more slowly than younger participants. There were no significant differences observed between males and females who participated in the quasi-experiments. In addition, no influence was observed from the nationality of the participants or the setting of the quasi-experiment. Almost no influence was observed from the organisation of the quasi-experiments ("*Well-structured organisation. No guidance questions from the participants*"). No influence was observed from the observers in the treatment group of professionals, and only one participant was found to be influenced by other participants. In all quasi-experiments the computers were separated by partitions, so that all the participants had their own work place and it was difficult for them to see what other participants did on their computers. This stimulated working individually.

7.6 Speed of OGD use

This section reports on the results concerning the speed of OGD use (as an indicator for levels of coordination) utilising the developed OGD infrastructure. It examines the following sub propositions.

- P1b: Metadata positively influence the speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis.
- P2b: Interaction mechanisms positively influence the speed of interaction about OGD.
- P3b: Data quality indicators positively influence the speed of OGD quality analysis.

The speed of OGD use was examined through time measures. The following sections describe the results from the time measures as well as the intermediate variables that were found.

7.6.1 Time measures

Results regarding time measures were obtained through the scenario tasks. Table 7-19 depicts the average number of minutes that the participants spent on conducting each of the scenarios, as well as the standard deviations. The table shows that participants of the control group needed more time to conduct all the five scenarios than the participants of the students and professionals treatment groups. On average the professional open data users in the treatment group conducted the scenarios slightly faster than the students in the treatment group. The participants of the control group needed on average 42 minutes to complete all the five scenarios, while the students of the treatment group needed 29 minutes and the professionals of the treatment group 27 minutes.

Time spent on each scenario	**Number of respondents (N), average number of minutes spent (μ) on scenario and standard deviation (σ)**		
	Control group (students)	**Treatment group (students)**	**Treatment group (professionals)**
Duration scenario 1 (searching for and finding OGD)	N: 40 μ: 6 minutes σ: 3 minutes	N: 50 μ: 4 minutes σ: 1 minutes	N: 32 μ: 5 minutes σ: 3 minutes
Duration scenario 2 (OGD analysis)	N: 38 μ: 11 minutes σ: 4 minutes	N: 50 μ: 10 minutes σ: 5 minutes	N: 31 μ: 9 minutes σ: 3 minutes
Duration scenario 3 (OGD visualisation)	N: 37 μ: 9 minutes σ: 5 minutes	N: 50 μ: 6 minutes σ: 2 minutes	N: 31 μ: 8 minutes σ: 3 minutes
Duration scenario 4 (Interaction about OGD)	N: 37 μ: 9 minutes σ: 4 minutes	N: 49 μ: 5 minutes σ: 2 minutes	N: 27 μ: 5 minutes σ: 3 minutes
Duration scenario 5 (OGD quality analysis)	N: 36 μ: 4 minutes σ: 2 minutes	N: 47 μ: 3 minutes σ: 1 minutes	N: 24 μ: 2 minutes σ: 0 minutes
Total duration scenarios 1-5	N: 36 μ: 42 minutes σ: 9 minutes	N: 47 μ: 29 minutes σ: 6 minutes	N: 24 μ: 27 minutes σ: 6 minutes

Table 7-19: Speed of OGD use for the five scenarios.

A Mann-Whitney test was conducted to test whether the average number of minutes spent on the scenarios was significantly different for the control and treatment group. This test was appropriate, since there was one independent categorical variable (whether a person participated in the control or treatment group) and there was one continuous dependent variable (the number of minutes that a participant spent on a scenario). The Mann-Whitney test is the non-parametric equivalent of the independent t-test (Field, 2009, p. 540), which had to be used since the sample did not meet the assumptions for parametric tests (the data was not normally distributed).

The Mann-Whitney test showed that the number of minutes that the students of the treatment group used to conduct the five open data use scenarios (*Mdn* = 29) differed significantly from the number of minutes that the students of the control group used to conduct these scenarios (*Mdn* = 45), $U = 215.00$, $p < .001$. Moreover, the number of minutes that the professionals of the treatment group used to conduct the five open data use scenarios (*Mdn* = 27) differed significantly from the number of minutes that the students of the control group used to conduct these scenarios (*Mdn* = 45), $U = 81.50$, $p < .001$.

7.6.2 Intermediate variables

In interpreting the above-mentioned findings from the time measures, one should take into account a number of intermediate variables that influence the coordination of OGD use. First, the control infrastructure server was temporarily not available for some of the control group participants during the second quasi-experiment. This may have resulted in longer time durations to conduct the tasks for some participants of the control group. Some participants had no problems, while others had to wait for some minutes before they could use the control open data infrastructure again to conduct the tasks. For most participants the server problem occurred when they were conducting the tasks of the third scenario.

Second, due to time limitations it was not possible to conduct scenario tasks longer than 50 minutes. A few participants were not able to complete the scenarios in this time frame. This seemed to be caused mainly by having a problem with one of the preceding tasks, which resulted in spending much time on

this particular task. For some functions the participants desired to spend more time on that task. For instance, data quality rating required a thorough analysis of a dataset, and it can be difficult to assess the quality of an open dataset in a short time frame. The time limitation may have resulted in finding a lower number of minutes spent on the scenarios for both the control and treatments groups than the number of minutes that was actually required to complete the tasks.

7.7 Summary: overview of evaluation outcomes and answer to the fifth research question

This chapter described the results of the evaluations that took place to answer the fourth research question: *what are the effects of the developed infrastructure on the coordination of OGD use?* The findings of our evaluations have to be seen in the context of this study's focus on the use of structured research OGD from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures. Using the prototype described in chapter six, three quasi-experiments were conducted with 127 participants in total (students and professional open data users). The propositions developed in chapter five guided the design of the quasi-experiments, since they showed which functional elements of the OGD infrastructure needed to be investigated. We used the ease and speed of OGD use as indicators for the coordination of OGD use.

The quasi-experiments used a pre-test post-test control group design. The control and the treatment groups conducted the same scenario tasks concerning the use of research OGD. The participants completed scenarios that prescribed them to use various tools, to interact with other OGD users and to use tools that allowed for interaction with OGD providers and policy makers. This means that they used OGD in a way that corresponds to our definition of coordination (see section 3.2.4). In the quasi-experiments we examined to which extent the ease and the speed of OGD use was improved by the developed OGD infrastructure, and we examined the coordination of OGD use by including the management of dependencies between and among activities performed to use OGD in the evaluation scenarios.

Three types of measures were used to investigate whether the OGD infrastructure improved the ease and speed of OGD use: surveys and observations

were used to examine the ease of OGD use, while time measures were used to research the speed of OGD use. The three types of measures complemented each other. For instance, the second survey showed that quite a large part of the professional treatment group found it difficult to conduct the fourth and the fifth scenario, while the observations showed that a few participants had rushed through these scenarios and had stated that they were difficult to complete while they had not tried to complete the tasks. Without the observations we could not have made the necessary differentiations. In the following sections we discuss the findings form the survey, the observations and the time measures, as well as the theoretical contributions of the quasi-experiments, the limitations of the quasi-experiments and the findings, and finally the areas for improvement.

7.7.1 Ease of OGD use: survey and observation results

In our quasi-experiments, the ease of OGD use was measured through three surveys and through semi-structured observations. The first survey was used as a pre-test to measure to which extent participants knew at least one open data infrastructure that enabled them to conduct tasks related to metadata, interaction mechanisms and data quality indicators. This survey showed that the control and treatment groups of students and professionals were relatively equal with regard to knowing an open data infrastructure that enabled activities related to the first three scenarios (searching for and finding OGD, OGD analysis and OGD visualisation), while differences regarding experiences with interaction about OGD and OGD quality analysis were slightly larger. With regard to the latter, the control group of students was already more positive in the pre-test than the treatment groups of students and professionals, and the treatment group of students was already more positive in the pre-test than the treatment group of professionals.

The second survey showed that the assessed difficulty of scenario tasks related to all five open data scenarios of the student treatment group differed significantly from the assessed difficulty of these tasks of the student control group. On average the students and the professionals in the treatment group found it significantly easier to conduct scenario tasks related to searching for and finding OGD (scenario 1), OGD analysis (scenario 2), OGD visualisation (scenario 3), interaction about OGD (scenario 4) and OGD quality analysis (scenario 5) than the

students in the control group. More specifically, on average the students in the treatment group (*Mdn* = 5.5) found it easier to conduct scenario tasks related to searching for and finding OGD than the students of the control group (*Mdn* = 4.5), $U = 410.50$, $p < 0.01$. The students of the treatment group (*Mdn* = 5.75) also found it easier to conduct scenario tasks related to OGD analysis than the students of the control group (*Mdn* = 4.25), $U = 318.00$, $p < 0.01$. Students of the treatment group who visualised OGD (*Mdn* = 5.33) found this activity easier than students of the control group (*Mdn* = 3.67), $U = 259.50$, $p < 0.01$. Students of the treatment group that completed the OGD interaction scenario (*Mdn* = 5.0) found this significantly easier than students of the control group (*Mdn* = 2.0), $U = 68.5$, $p < .001$. Finally, the students of the treatment group who conducted a data quality analysis (*Mdn* = 6.0) found this significantly easier than the students of the control group (*Mdn* = 2.0), $U = 63.5$, $p < 0.01$.

The third survey showed that the open data infrastructure used by the control group performed worse than the participants would have expected based on their previous experiences with open data infrastructures (measured through the pre-test). The treatment infrastructure performed better than other open data infrastructures that the participants had experience with. The participants of the control group were neutral or slightly disagreed that the open data infrastructure that they had used enabled the tasks of the first three scenarios. The professional open data users in the treatment group were relatively more positive than the control group, although most means were between 'neutral' and 'slight agreement'. The students of the treatment group slightly agreed that the infrastructure that they had used enabled the scenario tasks of the three scenarios. With regard to scenarios four and five related to interaction about OGD and OGD quality analysis, it was found in the third survey that even though the control group was more positive in the pre-test, in the post-test the treatment group was more positive. The differences between the control and treatment group were slightly larger for scenarios four and five than they were for scenarios one, two and three. Participants of the treatment group were generally more satisfied with the ease-of-use of the open data infrastructure, and they stated that the infrastructure had enabled them to use open data relatively easily.

The observations confirmed the findings from the three surveys. The observations showed that in general the scenarios were easier to conduct for the treatment groups than for the control group. The observers of the control group all said that it was difficult or very difficult to complete the five scenarios, whereas the observers of the treatment group postulated that it was neither difficult nor easy or that it was easy. In sum, the three participants surveys and the observations showed that the OGD infrastructure positively influenced the ease of all five types of OGD use, i.e. the ease of searching for and finding OGD, the ease of OGD analysis, the ease of OGD visualisation, the ease of interaction about OGD and the ease of OGD quality analysis.

7.7.2 Speed of OGD use: time measure results

To investigate the speed of OGD use on the developed infrastructure, we measured how much time the participants needed to complete the scenario tasks. The time measures showed that participants of the control group needed more time to conduct all the five scenarios than the participants of the treatment groups. On average the professional treatment group conducted the scenarios slightly faster than the student treatment group. The Mann-Whitney test showed that the number of minutes that the students of the treatment group spent to conduct the five open data use scenarios (*Mdn* = 29) differed significantly from the number of minutes that the students of the control group used to conduct these scenarios (*Mdn* = 45), $U = 215.00$, $p < .001$. Moreover, the number of minutes that the professionals of the treatment group used to conduct the five open data use scenarios (*Mdn* = 27) differed significantly from the number of minutes that the students of the control group spent to conduct these scenarios (*Mdn* = 45), $U = 81.50$, $p < .001$.

7.7.3 Theoretical contributions of the quasi-experiments

The evaluations contributed to our four kernel theories concerning coordination, metadata, interaction and data quality. We contributed to coordination theory by showing that coordination of OGD use does not merely require a focus on processes, but additionally requires the integration of social aspects and technology into these processes, as well as the interaction between the social and technical perspective. The quasi-experiments emphasised the importance of this finding. Regarding the contributions to the metadata kernel theory, the quasi-

experiments confirmed several recent studies that different types of metadata (discovery, contextual and detailed metadata) need to be combined to enhance the coordination of OGD use, while OGD infrastructures traditionally mainly provide discovery metadata. In addition, whereas kernel theories concerning coordination, metadata, interaction and data quality are often studied separately, this study revealed that combining metadata, interaction mechanisms and data quality indicators in one OGD infrastructure is an essential condition for managing the dependencies of researchers using OGD on different tools, on other researchers, and on other actors.

In our quasi-experiments, coordination theory influenced the other three kernel theories, while metadata, interaction and data quality more directly influenced OGD use through the developed infrastructure. The influence of the metadata kernel theory was mainly technical (although it also had a social impact), the influence of the interaction kernel theory was mainly social (although it also had a technical impact) and the influence of the data quality kernel theory was both social and technical.

7.7.4 Limitations of the quasi-experiments and the findings

The quasi-experiments suffered from a number of limitations. First, although the treatment groups in our quasi-experiments reported higher levels of ease of OGD use than the control group, one should note that the differences between the treatment and control groups were sometimes small. Moreover, while the control group sometimes disagreed with statements, the treatment groups often provided a neutral response or showed slight agreement with a statement, rather than strong agreement. This means that even though the treatment groups found OGD use easier than the control group, their level of ease can still be improved further.

Second, in the quasi-experiments the developed OGD infrastructure was compared to a control infrastructure. The chosen control infrastructure may have influenced the findings of this study. The statistical tests showed that the control group results differed significantly from the treatment group results. Using another infrastructure as the control infrastructure might have shown other differences between the control and treatment group results. Yet, there were sound reasons to choose the control infrastructure as the control infrastructure, since this control

infrastructure could be used in English, provided more OGD use functions than other infrastructures, and was already widely used by Dutch governmental agencies (see section 7.2.5). Further research might include several other OGD infrastructures to compare their effects.

Third, in the second quasi-experiment, the control infrastructure was temporarily not available for some of the control group participants. Some participants had no problems, while others had to wait for several minutes before they could use the control open data infrastructure again to conduct the tasks. For most participants the server problem occurred when they were conducting the tasks of the third scenario. This may have influenced the speed of OGD use, and may have resulted in extended time durations for conducting several scenario tasks by some participants.

It was tried to keep the variables between the control and the treatment groups in the quasi-experiments as equal as possible. Nevertheless, in our quasi-experiments the participants were not matched pair-wise in advance, but they were matched non-pair wise afterwards, based on their characteristics that they described in the surveys. Ideally, we would have matched the participants pair-wise before the quasi-experiments took place. Moreover, ideally the treatment and the control infrastructure would only vary regarding the metadata model, the interaction mechanisms and the data quality indicators. However, there may have been differences between the control and treatment groups that might have influenced the outcomes of the evaluations. The surveys, observations and time measures showed that various intermediate variables played a role in the coordination of OGD use.

First, intermediate variables were found with regard to the infrastructure. Both the surveys and the observations pointed at the user interface as an important intermediating variable. While a user-friendly interface can support OGD use, a non-user-friendly interface was found to hinder OGD use. Other intermediate variables mentioned in the surveys included the variety of tools to search for and filter data, the number of provided datasets, required registration and signing in on the infrastructure, and the types of programmes and tools used on the OGD infrastructure. An intermediate variable that influenced the ease of OGD use mentioned by the observers concerned the temporal unavailability of the

server for the control group. This intermediate variable also influenced the speed of OGD use on the infrastructure. Temporal unavailability of the control group infrastructure may have resulted in longer time durations to conduct the tasks for some participants of the control group. It is important to keep the above-mentioned intermediating variables in mind for the development of OGD infrastructures.

Moreover, intermediate variables were found for the quasi-experimental design. Both the surveys and the observations indicated that the participants' experience with OGD use was important for the ease of OGD use. Other intermediate variables that were mentioned by the observers concerned the potential influence from the observers, from other participants, from the setting, from the nationality of the participants, and from noise that was heard at the end of the second quasi-experiment. An intermediate variable that was also found to have influenced the speed of OGD use concerned the time limitations for conducting the scenario tasks. Time limitations may have resulted in finding a lower number of minutes spent on the scenarios for both the control and treatment groups than the number of minutes that was actually required to complete the tasks. It is important to keep the above-mentioned intermediating variables in mind when interpreting the results of this study, since these factors appeared to either positively or negatively influence the performance of users of the OGD infrastructure.

7.7.5 Areas for improvement

The evaluations suggest various areas for improvements for the OGD infrastructure and for improvements of the quasi-experimental design. Improvements of the OGD infrastructure can focus on the functions that are provided by the infrastructure. For instance, improvements can focus on making it easier to visualise data in charts and on maps, and to draw conclusions based on open datasets. It was found that visualising OGD was not an easy task, as expressed by one quasi-experiment participant: *"the visualisation (chart, graph, map) was a bit difficult to use".* Even though metadata helped to improve OGD visualisations, additional mechanisms may be used to improve this further. The interface design appeared to be important to improve the ease and speed of OGD visualisations. Second, the section where participants could discuss data use could be improved. One participant mentioned that he did not like the fact that he *"had to*

scroll through all the comments". Another participant stated that he was *"unable to reply to a comment"* and that this *"made it a bit frustrating".* Moreover, the infrastructure users suggested to enlarge the section with items connected to the dataset (e.g. visualisation and applications), as *"the 'items based on this dataset' section is not easy to see, you can pass it easily"*.

Furthermore, improvements can involve the identified intermediating variables, such as providing a very user-friendly user interface, providing sufficient programmes and tools to use OGD on the infrastructure, and offering more options to search for and filter open data. Various participants stated that the user interface can be improved. This is illustrated by the following quotes: "*ENGAGE has high potential but needs some work in terms of interface and what users will need from the platform"* and "s*ome of the interface is intuitive but some things […] made it a bit frustrating"*.

The evaluations also point at the importance of examining non-functional requirements, such as people's experience with OGD use and the long-term availability of the infrastructure. For instance, the evaluations show that sustainability is important for the use of OGD. One quasi-experiment participant expressed this by saying: *"I need to see transparency of the business model that makes ENGAGE sustainable over a longer period of time"*. Additionally, trust in the OGD infrastructure and privacy aspects influence OGD use. Some users of the infrastructure indicated that there was a lack of information about what ENGAGE did with their account information, such as the information about their social media account. Especially since it was possible for users to login with their social media accounts, and there was no information about how their login details would be used by the infrastructure, there were some concerns about trust and privacy.

8. Conclusions

This study focuses on the operational use of structured Open Government Data (OGD) by researchers outside the government. Moreover, it focuses on a particular type of OGD, namely research OGD from the domains of social sciences and humanities. In the envisioned situation, the use of OGD will support OGD publication and governmental policy making, since OGD providers and governmental policy makers can learn from the insights obtained through the use of OGD data. This study focuses on a specific type of OGD use (also see section 1.2), and OGD providers and governmental policy makers are outside the scope of this study.

The use of OGD requires several actors, activities and tools, however, these are fragmented and depending on each other. In addition, our literature review showed that governments and scholars mainly focus on the publication of OGD, whereas the actual use of the data resulting in benefits is often neglected. An OGD infrastructure can enhance the coordination of OGD use by researchers, as shown in our quasi-experiments. The objective of this study is *to develop an infrastructure that enhances the coordination of OGD use.* Core components of the OGD infrastructure developed in this study are an advanced and interoperable three-tier *metadata model* to find, analyse, visualise, interact about and assess OGD, *interaction mechanisms* to stimulate interaction between OGD users and other actors, and *data quality indicators* to assess the data's fitness for use.

This study is among the first to describe the design of an OGD infrastructure. This dissertation contributes to science by providing a comprehensive overview of barriers and functional requirements for OGD use from the perspective of the OGD user, by defining functional building blocks for the design of the OGD infrastructure, and by developing and evaluating a prototype of the OGD infrastructure. Furthermore, this study is the first to apply coordination theory in the field of OGD and shows that coordination of OGD use does not merely require a focus on processes, but additionally requires a technical perspective including the integration of tools, a social perspective including interaction between researchers, OGD providers and policy makers, and

interaction between the social and technical perspective. Moreover, while OGD infrastructures traditionally mainly provide discovery metadata, this study confirms several recent studies that different types of metadata (discovery, contextual and detailed metadata) need to be combined to improve OGD use. In addition, whereas kernel theories concerning coordination, metadata, interaction and data quality are often studied separately, this study reveals that it is essential for the development of OGD infrastructures to combine these four kernel theories.

This chapter provides the key findings from our study including the scientific and practical contributions (section 8.1), followed by the design theory for OGD infrastructures developed in our research (section 8.2), a description of how the kernel theories can be combined (section 8.3), and this study's limitations (section 8.4). The findings in this chapter apply to the specific type of OGD use and users that we studied.

8.1 Findings from this study

This section first discusses the findings from our study for each of the five research questions (sections 8.1.1- 8.1.5), followed by an overall reflection on the realisation of the research objective (section 8.1.6).

8.1.1 Research question 1: factors influencing OGD use

The first research question aimed to identify factors influencing OGD use (see chapter 3). This question helped to scrutinise the problem that this study addresses and the factors related to it. Through a literature, we searched for factors influencing five types of OGD use: searching for and finding OGD, analysing OGD, visualising OGD, interaction about OGD and OGD quality analysis (see section 3.3). The following factors were identified (also see Appendix A).

- Factors influencing searching for and finding OGD
    - Data fragmentation: data users find it difficult to locate the datasets that they want to use, since the data are offered at many different places and it may be unclear to users where they should search for the data that they need;
    - Terminology heterogeneity: different terminologies are used to describe datasets, so that users often do not know which terms they should use to search for the data that they need;

  - Search support: most OGD portals provide simple search functions and there is a lack of more advanced and multilingual data query functions;
  - Information overload: an information overload may be faced when the amounts of OGD become overwhelming to OGD users.
- Factors influencing OGD analysis
  - Data context: OGD providers often do not provide extensive contextual data about the data, which complicates the analysis and interpretation of the data;
  - Data interpretation support: because of a lack of data interpretation support, there is potential for the misuse, misunderstanding and misinterpretation of open data;
  - Data heterogeneity: heterogeneous data formats and heterogeneous semantics;
  - Data analysis support: there is a lack of integrated tools provided by OGD infrastructures to support data analysis.
- Factors influencing OGD visualisation
  - Data visualisation support: there is a need for visualisation tools to make sense of OGD.
- Factors influencing interaction about OGD
  - Lack of interaction: the cumbersome involvement of stakeholders in OGD use, for instance, because of a lack of conversations about released data;
  - Interaction support and tools: a lack of interaction support and tools at OGD infrastructures.
- Factors influencing OGD quality analysis
  - Dependence on the quality of open data: OGD use depends on the quality of the data;
  - Poor data quality: many open datasets are expected to suffer from poor data quality;
  - Quality variation and changes: the quality of datasets may vary (e.g. per source and after reuse over time).

The answer to the first research question of this study contributed to the literature by providing a comprehensive overview of factors that influence OGD use, including the barriers that traditionally hindered OGD use. Before this study, there was no comprehensive overview of factors that influence OGD use, nor was there a comprehensive overview of barriers that hinder OGD use. The studies on factors influencing OGD use that had been conducted before this study had often defined influencing factors on a high level of abstraction or only focused on a specific type of OGD use. They often did not focus on the barriers related to the dependence of OGD users on different tools, on each other, and on other actors. This study is among the first to provide a comprehensive overview of the factors and the barriers that need to be taken into account when one wants to enhance the coordination of OGD use.

8.1.2 Research question 2: functional requirements for the OGD infrastructure

After we identified clusters of factors influencing OGD use in general, we searched for functional infrastructure requirements within each of the identified clusters. We answered the second research question: *what are the functional requirements for an infrastructure that enhances the coordination of OGD use?* (see chapter 4) Functional requirements were sought for one specific target group of OGD users, namely researchers outside the government, using structured research OGD from the domains of social sciences and humanities. The clusters of factors derived from the literature review (see section 3.4) were used as a framework to elicit functional infrastructure requirements from two case studies. The two selected cases concerned the use of judicial data provided by the Research and Documentation Centre, and the use of social data provided by The Netherlands Institute for Social Research. In these cases, governmental research data had already been made available to the public for several years. Documents, archival records, notes from open and semi-structured interviews, direct and participant observations and datasets were studied for both cases. The cases allowed for eliciting functional requirements for an OGD infrastructure that aims at enhancing the coordination of OGD use for our particular target group of OGD users. The following Functional Requirements (FR) were elicited.

- Functional infrastructure requirements related to searching for and finding OGD
    - Data fragmentation: FR1) the OGD infrastructure should be a one-stop shop for datasets and metadata from a variety of other OGD infrastructures, FR2) the OGD infrastructure should allow OGD users to integrate and refer to datasets from various other OGD sources;
    - Terminology heterogeneity: FR3) use controlled vocabularies to describe OGD, FR4) use interoperable standards to describe OGD;
    - Search support: FR5) the OGD infrastructure should support data search through keywords, data category browsing and data querying, FR6) the OGD infrastructure should support OGD use by the ability to search for data and metadata in multiple languages;
    - Information overload: FR7) the OGD infrastructure should facilitate filtering, sorting, structuring and ordering relevant search results.
- Functional infrastructure requirements regarding OGD analysis
    - Data context: FR8) the OGD infrastructure should provide data which describe the dataset, FR9) the OGD infrastructure should provide data about the context in which the dataset has been created;
    - Data interpretation support: FR10) it should be clear for which purpose the data have been collected, FR11) it should provide examples of the context in which the data might be used, FR12) domain knowledge about how to interpret and use the data should be provided;
    - Data heterogeneity: FR13) the OGD infrastructure should allow for the publication of datasets in different formats;
    - Data analysis support: FR14) the OGD infrastructure should offer tools that make it possible to analyse OGD, FR15) the OGD infrastructure should provide insight in the conditions for reusing the data.
- Functional infrastructure requirements regarding OGD visualisation
    - Data visualisation support: FR16) the OGD infrastructure should provide and integrate visualisation tools, FR17) the OGD infrastructure should allow for visualising data on maps.
- Functional infrastructure requirements regarding interaction about OGD

  - Lack of interaction: FR18) the OGD infrastructure should support interaction between OGD providers, policy makers and OGD users in OGD use processes, FR19) the OGD infrastructure should allow for conversations and discussions about released governmental data, FR20) the OGD infrastructure should allow for viewing who used a dataset and in which way;
  - Interaction support: FR21) the OGD infrastructure should provide tools for interactive communications between OGD providers, policy makers, and OGD users (e.g. data request mechanisms and social media), FR22) the OGD infrastructure should provide tools for interactive communications between OGD users (e.g. discussion forums and social media), FR23) the OGD infrastructure should provide tools to keep track of amended datasets so that users know how datasets have been changed.
- Functional infrastructure requirements regarding OGD quality analysis
  - Dependence on the quality of open data: FR24) the OGD infrastructure should provide insight in quality dimensions of OGD, FR25) it should be possible for OGD users, OGD providers and policy makers to discuss the quality of a dataset;
  - Poor data quality: FR26) the OGD infrastructure should provide information on the context in which a person reused a particular dataset;
  - Quality variation and changes: FR27) the OGD infrastructure should provide quality dimensions of datasets that are comparable with other datasets and with different versions of the same dataset, FR28) it should be possible to compare the quality of datasets over different data sources, over time and over data reuse on the data infrastructure.

The answer to the second research question contributed to the literature by offering a comprehensive overview of user requirements for enhancing the coordination of OGD use through infrastructures based on practical case studies. Although the need for taking a user perspective had already been acknowledged in the literature before we started this study, there was a lack of insight in the user requirements for an infrastructure that enhances the coordination of OGD use. In total we identified

twenty-eight functional requirements (see section 4.3). All these requirements were found in both cases, and the requirements identified in the first case were confirmed by the second case. The comprehensive user requirement overview directed us towards objectives of a solution, namely the design of the OGD infrastructure.

8.1.3 Research question 3: functional elements of the OGD infrastructure

The third research question, *which functional elements make up an infrastructure that enhances the coordination of OGD use?,* was answered in chapter five. The functional requirements that were identified through the second research question were mapped to potential functional elements of the OGD infrastructure, and we searched the literature for potential functional infrastructure elements that would meet these requirements. Based on two criteria (i.e. 1. the functional elements needed to cover as many of our functional requirements as possible and 2. at least basic research regarding the functional elements (in other research domains than open data) already had to be available), three functional elements were proposed to meet the OGD infrastructure requirements, namely 'metadata', 'interaction mechanisms' and 'data quality indicators'. When we conducted this study, there was no comprehensive body of literature that showed that these functional elements could be used in the context of open data. The case studies described in chapter four were used to identify functional requirements for the OGD infrastructure, which subsequently pointed at potential functional infrastructure elements that had already been used in other contexts outside the field of open data.

Based on the assumption that metadata, interaction mechanisms and data quality indicators can improve the coordination of OGD use, three design propositions were developed:

1. Metadata positively influence the ease and speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis.
2. Interaction mechanisms positively influence the ease and speed of interaction about OGD.

3. Data quality indicators positively influence the ease and speed of OGD quality analysis.

While other functional elements than metadata, interaction mechanisms and data quality indicators may also meet some of the functional requirements, metadata, interaction mechanisms and data quality indicators were the key elements which together covered all the twenty-eight functional requirements. While metadata technically support different OGD use activities, interaction mechanisms can be used to support the collaboration of the stakeholders involved in open data use, and OGD quality indicators are useful to generate OGD users' trust in the dataset and in the data provider, and to see for which purposes a dataset can be used. Together the three functional elements intend to improve the management of dependencies (i.e. coordination) of OGD use by researchers.

The design propositions (see section 5.2) suggest on a high level which functional elements may be used to enhance the coordination of research OGD use. The propositions are abstractions, and various mechanisms underlie these abstractions. Building on the design propositions, we developed 81 design principles, which provide more detailed directions for the design of the OGD infrastructure (see section 5.3). Kernel theories were used to aid the development of design principles. Following Gregor and Jones (2007, p. 322), we endorse the idea that a kernel theory is "the underlying knowledge or theory from the natural or social or design sciences that gives a basis and explanation for the design". In our study different types of kernel theories were used to develop design principles. Coordination theory and literature regarding metadata, interaction, and data quality underlay the design of the OGD infrastructure.

Building on the design principles, the design of the OGD infrastructure was described (see section 5.4). The OGD infrastructure was designed in collaboration with partners from the consortium of the ENGAGE-project. The ENGAGE-project was a combination of a Collaborative Project and Coordination and Support Action (CCP-CSA) funded by the European Commission under the Seventh Framework Programme. The OGD infrastructure design incorporated the system design, the coordination patterns and the function design. The system design described the structure and the behaviour of the system. A three-tier metadata model was developed incorporating discovery metadata, contextual metadata and detailed

metadata. Two types of interaction mechanisms were designed, namely feedback mechanisms and collaboration and discussion mechanisms. Then a data quality indicator model was developed which incorporated structured data quality rating, free text quality reviews and evaluator information. The patterns defined the reusable parts of the design with their benefits and an explanation of how they can be applied, and the relation between them. With regard to the coordination patterns, it was explained how the functional elements of the OGD infrastructure can together enhance the coordination of OGD use by researchers. The OGD infrastructure design was developed through iterative phases. The design was also presented to the case study participants, and the case study participants were asked to provide feedback.

Finally, the function design translated the system design and the coordination patterns to concrete functions that could be implemented in the OGD infrastructure, and showed what the final design was supposed to do. Functions were related to each of the five OGD use activities (i.e. searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis) and to each of the three functional infrastructure elements (i.e. the metadata model, interaction mechanisms and data quality indicators). In total 46 functions were defined. Examples of functions are dataset upload, multilingual search, data cleansing, following users, personal messages, social media sharing, and distribution of ratings.

By answering the third research question, this study contributed to the literature on open data infrastructures. Existing research often described functional elements on a high level of abstraction and did not describe them in such a way that they could be implemented directly in OGD infrastructures. This study provided detailed functional elements that can be implemented in the design of existing or new OGD infrastructures. The coordination patterns explain how the functional elements can be used in and applied to practice, and the function design shows which functions can be implemented by developers of existing and future OGD infrastructures. This study is among the first to describe the design of an OGD infrastructure, including the elements it encompasses.

In addition, this study contributes to the literature concerning the four kernel theories regarding metadata, interaction, data quality and coordination.

Kernel theories about coordination, metadata, interaction and data quality are often studied separately (e.g., Malone & Crowston, 1990), and metadata, interaction mechanisms and data quality indicators have never been combined in one OGD infrastructure. This study revealed that combining metadata, interaction mechanisms and data quality indicators in one OGD infrastructure is an essential condition for managing the dependencies of OGD users (i.e. researchers) on different tools, on each other, and on other actors (see section 8.3). Coordination theory provided coordination mechanisms that were applied to the metadata model, the interaction mechanisms and the data quality indicators to enhance the coordination of OGD use by researchers. The literature regarding metadata showed how the use of research OGD can be supported technically, the literature regarding interaction was used to develop mechanisms that support the interaction between OGD users (i.e. researchers) and other actors, and the literature concerning data quality helped to design quality indicators that provided insight in the purposes for which a dataset can be used. The combination of the three functional elements and the four kernel theories is needed to enhance the coordination of OGD use by researchers.

Furthermore, this study contributes to the literature regarding coordination. Traditionally many studies on coordination have been conducted (e.g., Crowston et al., 2004; Malone & Crowston, 1990), and insights from these studies can be used to enhance coordination of open data use activities. However, this study found that the literature on coordination is mainly focused on improving processes (e.g., Malone & Crowston, 1990). This study builds on the coordination literature (Crowston et al., 2004; Malone & Crowston, 1990) and shows that the coordination of OGD use through infrastructures does not merely require a focus on processes, but additionally requires a technical perspective including the integration of tools, a social perspective including interaction between involved actors, and the interaction between the social and technical perspective.

Traditionally, it has been argued that OGD infrastructures should focus on providing discovery metadata, using standards such as CKAN (e.g. see Marienfeld et al., 2013). Recently several studies have shown that different types of metadata (discovery, contextual and detailed metadata) need to be combined to improve OGD use (Bailo & Jeffery, 2014; Jeffery et al., 2014; Jeffery et al., 2013). This

study builds on the existing literature regarding metadata (e.g., Gilliland, 2008; Jeffery et al., 2013; Vardigan et al., 2008) and confirms that different types of metadata (discovery, contextual and detailed metadata) and metadata standards need to be integrated to enhance the coordination of OGD use by researchers. This study is among the first to use CERIF as a superset exchange mechanism for common metadata standards in the field of research OGD.

8.1.4 Research question 4: development of the OGD infrastructure

The development of the prototype was described in chapter six. It aimed to answer the fourth research question: *what does the developed OGD infrastructure look like?* The prototyping approach encompassed four steps, namely 1) defining the objectives of the prototype, 2) selecting the functions of the prototype, 3) constructing the prototype and, 4) testing the prototype. Although these phases were described separately, much iteration between the prototyping phases as well as between the prototyping phases on the one hand and the design of the OGD infrastructure on the other hand took place. The prototype was implemented within the ENGAGE-project funded by the European Commission under the Seventh Framework Programme.

The development of the prototype started with defining the prototyping objectives (see section 6.2), which were twofold. First, the prototype was developed to be able to refine and detail the user requirements regarding the metadata model, the interaction mechanisms and the data quality indicators. Second, since we wanted to evaluate the effects of the OGD infrastructure in a realistic setting, and in practice there were no examples of OGD infrastructures which contained a combination of the designed metadata model, interaction mechanisms and data quality indicators, the prototype was developed to be able to measure the effects of the designed OGD infrastructure.

In the second prototyping step, a selection of prototype functions was made (see section 6.3) from the function design described in chapter five. The main selection criterion was that the functions needed to allow for measuring the key effects of metadata, interaction mechanisms and data quality indicators, and for refining the user requirements for these three functional infrastructure elements. Almost all functions described in chapter five were selected, except for three

metadata-enabled functions (convert data format, refer to data and link data manually), two interaction mechanism functions (enter open and private collaboration groups) and one data quality indicator function (compare different quality ratings and reviews) (see section 5.4.3). These functions were not implemented because it would be too time-consuming to use them within the limited time frame of the evaluations, or because they were not central to the five OGD use activities of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. Without these functions, the five OGD use activities could still be evaluated.

In the third prototyping step the prototype was constructed (see section 6.4). The prototype was called 'ENGAGE'. The ENGAGE prototype was accessible for the public via a website (www.engagedata.eu). The prototype allowed for searching for open datasets in different ways (e.g. entering data in a search bar, filtering, sorting, ordering, categorisation, multilingual search). For each dataset an overview of basic information was provided (e.g. contextual metadata, general data quality assessment score, main content and resources, the options for viewing, downloading and visualising data, comments and remarks on the dataset) as well as more detailed information (e.g. detailed metadata). Users could analyse datasets by exploring the various options provided in the dataset overview (e.g. viewing a dataset without downloading it, viewing which other users had extended or amended the dataset). The prototype allowed for using different tools to create tables, charts and maps of open datasets. Data interaction mechanisms could be used to give feedback on datasets and processes related to data provision and use, and they can discuss what could be learned from the use of the data. Various quality indicators were available, including rating the quality of datasets by assessing the accuracy, completeness, consistency and timeliness of a datasets, by writing a review of the dataset in an open text box (e.g. to elaborate on the purpose of data use), and by viewing information about the data evaluator. These elements and functions together comprised the prototype.

Finally, in the fourth prototyping phase the prototype was tested (see section 6.5). Whereas the initial requirements for the prototype were collected through the case studies, these requirements were further refined through various iterations in the prototype testing phase. A number of defect alpha tests were

conducted at the developers' site before the release of each new version of the prototype. These alpha defect tests were used to identify incorrect and undesirable behaviour of the infrastructure. Moreover, beta validation tests were carried out to examine whether the infrastructure met its requirements. The feedback that was obtained from the beta tests led to the refinement of the functional requirements and subsequently to improvements of the prototype. In addition to students, two case study participants were also involved in one of the beta tests.

Our answer to the fourth research question contributed to the literature by showing how the designed OGD infrastructure can be developed. Whereas some studies had already described functional elements for the development of OGD infrastructures at the time that we started this study (e.g., Charalabidis et al., 2011), traditionally, research had not described what an infrastructure that combines metadata, interaction mechanisms and data quality indicators may look like and how it can be developed. This research contributes to the literature by providing a description of what the designed OGD infrastructure including its three functional elements may look like and how it can be developed. Developers can use these findings for the development of OGD infrastructures.

8.1.5 Research question 5: effects of the OGD infrastructure

Chapter seven of this study aimed to evaluate the developed OGD infrastructure. In the previous research phases we assumed that the design of the OGD infrastructure could enhance the coordination of OGD use, which was evaluated in this final research phase. The fifth research question was answered: *what are the effects of the developed infrastructure on the coordination of OGD use?* The final version of the OGD infrastructure (ENGAGE 3.0) was evaluated through quasi-experiments. Metadata, interaction mechanisms and data quality indicators were the three independent variables of the evaluation model (see section 7.2.1). For each of the independent variables the effects on the dependent variables was determined. The dependent variables were the five types of OGD use that were identified in chapter three (i.e. searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality assessment). Since the literature did not clearly reveal the nature of the relationship between the functional elements of the OGD infrastructure on the one hand, and OGD use on the other

hand, we also evaluated whether intermediate variables, such as the characteristics of the respondents, the observers or the design of the quasi-experiments, influenced this relationship.

Three quasi-experiments were conducted with 127 participants (students and professional open data users). The participants were split into control and treatment groups. While the participants of the treatment group operated the developed OGD infrastructure (i.e. the treatment infrastructure), the participants of the control group operated a control infrastructure. The control and the treatment groups conducted the same scenario tasks concerning the use of research OGD. The participants completed scenarios that prescribed them to use various tools, to interact with other OGD users and to use tools that allowed for interaction with OGD providers and policy makers. This means that they used OGD in a way that corresponds to our definition of coordination (see section 3.2.4). In the quasi-experiments we examined to which extent the ease and the speed of OGD use was improved by the developed OGD infrastructure, and we examined the coordination of OGD use by including the management of dependencies between and among activities performed to use OGD in the evaluation scenarios.

The ease of OGD use was measured through three surveys and through observations. With regard to the surveys, the Mann-Whitney test showed that on average the students and the professionals in the treatment group found it significantly easier to conduct scenario tasks related to searching for and finding OGD (scenario 1), OGD analysis (scenario 2), OGD visualisation (scenario 3), interaction about OGD (scenario 4) and OGD quality analysis (scenario 5) than the students in the control group (see section 7.5.1). On average the students in the treatment group (Mdn = 5.5) found it significantly easier to conduct scenario tasks related to searching for and finding OGD than the students of the control group (Mdn = 4.5), U = 410.50, $p < 0.01$. The students of the treatment group (Mdn = 5.75) also found it significantly easier to conduct scenario tasks related to OGD analysis than the students of the control group (Mdn = 4.25), U = 318.00, $p < 0.01$. Students of the treatment group who visualised OGD (Mdn = 5.33) found this activity significantly easier than students of the control group (Mdn = 3.67), U = 259.50, $p < 0.01$. Students of the treatment group who completed the OGD interaction scenario (Mdn = 5.0) found this significantly easier than students of the

control group (*Mdn* = 2.0), *U* = 68.5, *p* < .001. Finally, the students of the treatment group who conducted a data quality analysis (*Mdn* = 6.0) found this significantly easier than the students of the control group (*Mdn* = 2.0), *U* = 63.5, *p* < 0.01.

The observations confirmed the surveys with regard to the OGD use findings (see section 7.5.2). They showed that most scenarios were easier to conduct for the treatment groups than for the control group. The survey and observation measures showed that the ease of OGD use was higher in the treatment groups than in the control group, indicating higher levels of coordination of OGD use in the treatment groups.

Subsequently, we examined to which extent metadata, interaction mechanisms and data quality indicators affected the speed of OGD use (see section 7.6). In the quasi-experiments the speed of OGD use was recorded through time measures, which showed that participants of the control group needed more time to conduct all the scenarios than the participants of the student and professional treatment groups. The Mann-Whitney test showed that the number of minutes that the students of the treatment group used to conduct the five open data use scenarios (*Mdn* = 29) differed significantly from the number of minutes that the students of the control group needed to conduct these scenarios (*Mdn* = 45), *U* = 215.00, *p* < .001. Moreover, the number of minutes that the professionals of the treatment group used to conduct the five open data use scenarios (*Mdn* = 27) differed significantly from the number of minutes that the students of the control group used to conduct these scenarios (*Mdn* = 45), *U* = 81.50, *p* < .001. Again these findings suggest higher levels of coordination of OGD use in the treatment groups.

Although the treatment groups in our quasi-experiments reported higher levels of ease of OGD use than the control group, one should note that the differences between the treatment and control groups were sometimes small. Moreover, while the control group sometimes disagreed with statements, the treatment groups often provided a neutral response or showed slight agreement with a statement, rather than strong agreement. This means that even though the treatment groups found OGD use easier than the control group, their level of ease can still be improved further. Various areas for the improvement of the OGD infrastructure and for the quasi-experimental design were identified. Improvements

of the OGD infrastructure may focus mainly on the functions that are provided by the infrastructure. For instance, functions related to the visualisation of data in charts and on maps and drawing conclusions based on open datasets can be improved. Furthermore, improvements can involve the identified intermediating variables, such as providing a user-friendly interface, providing sufficient programmes and tools to use OGD on the infrastructure, and offering more options to search for and filter open data. This study also points at the importance of examining non-functional requirements, such as people's experience with OGD use and the long-term availability of the infrastructure.

This study contributed to the literature by providing insight in the strengths and weaknesses of the developed OGD infrastructure. At the start of this study, there was no overview of the effects of functional OGD infrastructure elements on the coordination of OGD use. Research regarding these effects is needed to determine to which extent functional OGD infrastructure elements can be used to enhance coordination, and subsequently to improve OGD use for governmental policy making. The effects of functional infrastructure elements on the coordination of OGD use had barely been investigated. This study contributes to the literature by using quasi-experiments to investigate the effects of the developed OGD infrastructure in a systematic way. The findings from the quasi-experiments confirmed that the developed OGD infrastructure can be used to enhance coordination of OGD use, and that the infrastructure improved the ease and speed of OGD use tasks.

With regard to the practical contributions, this study provided insight in the effects of the infrastructure and its limitations. The outcomes of the evaluation offer practical recommendations about which functional elements are important in the design of OGD infrastructures and which further work the designed OGD infrastructure needs. Moreover, the evaluations allow for drawing conclusions about how and to which extent different OGD use tasks can be coordinated through the developed infrastructure. The insights obtained through the evaluations can be used by practitioners to further improve OGD infrastructures for researchers. Moreover, policy makers can use the infrastructure evaluation findings to derive useful information from OGD use by researchers, and may consider this in the development of governmental policy making.

8.1.6 Research objective: does the developed infrastructure enhance the coordination of OGD use?

We started this study with the objective to develop an infrastructure that enhances the coordination of OGD use. We focused on the operational use of structured research OGD from the domains of social sciences and humanities. In addition, the focus was on the use of these data by researchers outside the government through OGD infrastructures. Building on the functional requirements identified in two cases concerning the use of judicial and social research data, an infrastructure was developed incorporating the functional elements of metadata, interaction mechanisms and data quality indicators. We proposed the functional elements as coordination mechanisms: mechanisms that enhance the coordination of OGD use.

To assess to which extent the developed infrastructure and its functional elements enhanced the coordination of OGD use, the infrastructure was evaluated through quasi-experiments. In these quasi-experiments, participants completed scenario tasks concerning the use of research OGD. The tasks prescribed them to use tools for activities such as finding, analysing and visualising OGD, interacting with other OGD users and interacting with OGD providers and policy makers. This means that the quasi-experiment participants used OGD in a way that corresponds to our definition of coordination (i.e. *the act of managing dependencies between and among activities performed to use OGD*, see section 3.2.4).

In the quasi-experiments, the ease and the speed of OGD use were used as indicators of the coordination of OGD use. The findings showed that the five scenarios concerning the use of research OGD were significantly easier to conduct using the developed OGD infrastructure than using the control infrastructure. Moreover, the scenarios were conducted significantly faster using the developed OGD infrastructure compared to the control infrastructure. Participants found it easier to use data analysis and visualisation tools, to use tools to interact with other OGD users and to use tools that allowed for interaction with OGD providers and policy makers. In general, our study showed that the developed OGD infrastructure enhances the coordination of OGD use, although there are also several areas for improvement, such as improving some of the functions of the infrastructure (e.g. making it easier to visualise data in charts and on maps, and to draw conclusions based on open datasets), involving the identified intermediating variables (e.g.

providing a very user-friendly user interface and offering more tools to use OGD), and examining non-functional requirements (e.g. the long-term availability of the infrastructure).

This study builds on two particular cases, namely cases concerning open judicial data use and open social data use. The functional requirements were elicited from these cases, and subsequently the design, the prototyping and the evaluation of the OGD infrastructure were developed based on these cases. It is important to consider the context of the cases in interpreting the findings from this study. Both cases focused on the disclosure of research data. The data were collected by organisations that operate as part of a Ministry, yet they are both to a large extent independent of this ministry. Furthermore, in these cases most data were collected on a micro level, and both data providing organisations maintained a one or two year embargo period for the disclosure of the data. The key differences between the cases concerned the type of data that they focused on (judicial and social) and the sensitivity of the data. The judicial data were more privacy sensitive, while the social data were mainly non-sensitive. Nevertheless, in this study the differences between the cases did not lead to differences in functional requirements.

It is important to be aware of the characteristics of the cases, since they have influenced the elicited functional requirements, the designed functional elements, the developed prototype and the obtained evaluation outcomes of our study. Key characteristics of the cases that influenced the outcomes concern the type of data that the cases concentrated on, the type of users, and the way that the data were used. The focus of the cases on judicial and social data led to functional requirements related to data fragmentation and terminology heterogeneity. One stop shops where OGD from different sources are integrated in OGD infrastructures as well as controlled vocabularies and standards already exist for OGD from other domains. For instance, in the domain of environmental and geographical OGD all INSPIRE datasets and services have to be published centrally through the European INSPIRE geoportal (Van Loenen & Grothe, 2014), and standards support the collection of metadata for this type of data. The functional requirements related to data fragmentation and terminology heterogeneity may not apply to other types of OGD.

Furthermore, our cases focused on the supply of research data with researchers as envisioned data users. The focus on research data led to functional requirements related to data context, data interpretation support, data analysis support and data visualisation support. It was found that the use of OGD by researchers is complex and requires considerable support through tools. This support may not be needed for other types of OGD users. Additionally, to be able to use OGD for research, the quality is important. For other types of uses, this OGD quality may be less relevant. The functional requirements that we identified concerning the 'dependence on the quality of open data', 'poor data quality' and 'quality variation and changes' are related to the purpose of OGD use, and to our focus on using research OGD by researchers. Focusing on other types of OGD users, such as developers and entrepreneurs, might have led to other functional requirements. For example, developers might have wanted other types of search support (e.g. accessing data through an API) and entrepreneurs may not have requirements related to data fragmentation (e.g. entrepreneurs can earn money by combining, integrating and selling fragmented open datasets or services based on this fragmentation).

Finally, the cases focused on the use of OGD outside the government, and we considered OGD use that may be helpful for the formulation of governmental policies. This means that at least three key actors are involved, namely data providers, data users and policy makers. However, in other cases, data users may also obtain datasets directly from the data providers without considering how the data can be used to support governmental policy making. For those cases, interaction support may not be needed. The aforementioned characteristics of the cases have influenced the functional requirements that were elicited, and subsequently the infrastructure design, the prototype and the evaluation outcomes. Other types of cases may lead to other functional requirements, infrastructures, prototypes and evaluation outcomes.

8.2 A design theory for OGD infrastructures

A design science research approach was used for the development of the OGD infrastructure. A theory for design and action was developed, which prescribes how an artefact can be created, including the methods, techniques and principles for

the development of the artefact (Gregor, 2006). Design science was used because we aimed to develop an artefact (i.e. an OGD infrastructure) that did not yet exist, and since we aimed to contribute to scientific and practical developments for the design of OGD infrastructures. Design science research can be used to create something new that does not exist in nature (Vaishnavi & Kuechler Jr, 2008).

Although some research on the design of OGD infrastructures has been conducted, none of the existing studies have developed a design theory for the development of OGD infrastructures. This research is the first to develop a design theory that prescribes how an OGD infrastructure for open government research data can be created and evaluated. Gregor and Jones (2007) argue that an IS design theory comprises eight key components. Table 8-1 depicts these components and explains how this thesis provided design theory contributions for each of the components.

IS design theory components (Gregor & Jones, 2007, p. 322)	Design theory contributions
Purpose and scope of the theory; "what the system is for", the type of artefact to which the theory applies, and the scope or boundaries of the theory.	The aim is to develop an Open Government Data (OGD) infrastructure that enhances the coordination of OGD use. Governments and researchers usually focus on the disclosure of OGD, whereas the actual use of the data resulting in benefits is often neglected. This design theory is focused on a specific type of OGD use through infrastructures, namely the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government (see section 1.2). Outside the scope of this study are the data providers and the policy makers, and a premise is that improved OGD use will support policy making.
Constructs; representations of the entities of interest in the theory.	The constructs in the design theory are OGD, OGD infrastructures, OGD use and the coordination of OGD use (see section 3.2). OGD are *structured, machine-readable and machine-actionable data which governments and publicly-funded research organisations actively publish on the internet for public reuse and which can be accessed without restrictions and used without payment.* An OGD infrastructure is defined as *a shared, (quasi-)public, evolving system, consisting of a collection of interconnected social elements (e.g. user operations) and technical elements (e.g. open data analysis tools and technologies, open data services) which jointly allow for OGD use.* The theory focuses on the coordination of five types of OGD use activities, namely searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. Building on the coordination definition of Malone and Crowston (1990), we define coordination of OGD use as *the act of managing dependencies between and among activities performed to use OGD.*
Principles of form and function; principles that define the structure, organisation and functioning of the design product or method.	Twenty-eight functional requirements for the OGD infrastructure were elicited from two case studies concerning the use of judicial and social open government research data (see section 4.3). Three functional elements were proposed to meet the OGD infrastructure requirements, namely 'metadata', 'interaction mechanisms' and 'data quality indicators' (see section 5.2). Coordination design principles, metadata design principles, interaction design principles and data quality design principles guided the design of the OGD infrastructure (see section 5.3). The OGD infrastructure is the artefact created in this study, and consists of the system design, the coordination patterns and the function design (see section 5.4).
Artefact mutability; the changes in state of the artefact anticipated in the theory, that is, what degree of artefact change is encompassed by the theory.	OGD users often employed single tools that were not working in concert. The infrastructure should be viewed as part of a wider open data ecosystem in which each tool can add value. Each of the elements of the OGD infrastructure can be reused for the development of existing or future OGD infrastructures, and it can be tested to which extent they also apply in other contexts (e.g. for other types of data). The reuse and evaluation of the infrastructure's elements supports the evolvement of open data infrastructures.

Table 8-1: A design theory for OGD.

IS design theory components (Gregor & Jones, 2007, p. 322)	Design theory contributions
Testable propositions; truth statements about the design theory.	In the context of this study, three design propositions were elicited: ➢ Proposition 1: Metadata positively influence the ease and speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. ➢ Proposition 2: Interaction mechanisms positively influence the ease and speed of interaction about OGD. ➢ Proposition 3: Data quality indicators positively influence the ease and speed of OGD quality analysis. Moreover, our research showed that the metadata model, the interaction mechanisms and the data quality indicators need to be combined to support the five OGD use activities (i.e. searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis). The propositions are testable. Our evaluations of the OGD infrastructure provided support for the three propositions.
Justificatory knowledge; the underlying knowledge or theory from the natural or social or design sciences that gives a basis and explanation for the design (kernel theories).	OGD users conduct various activities for which they depend on different tools, on each other, and on other actors. A broad view on theories and underlying knowledge was adopted to refer to design-type knowledge to take into account the dependencies. Justificatory knowledge and kernel theories were used regarding coordination, metadata, interaction and data quality. Coordination theory and the literature underlying metadata, interaction and data quality assisted in identifying principles to guide the design of the OGD infrastructure (see section 5.3). ➢ Coordination theory provided coordination mechanisms that were applied to the metadata model, the interaction mechanisms and the data quality indicators to enhance the coordination of OGD use. This study shows that coordination of OGD use by researchers does not merely require a focus on processes, but additionally requires a technical perspective (e.g. the integration of tools), a social perspective (e.g. the use and the interaction between involved actors), and the interaction between the social and technical perspective. ➢ The literature regarding metadata showed how the use of OGD can be supported technically. This study builds on the existing literature regarding metadata and confirms that different types of metadata (discovery, contextual and detailed metadata) and metadata standards need to be integrated to enhance the coordination of OGD use by researchers. ➢ The literature regarding interaction was used to develop mechanisms that support the interaction between OGD users (i.e. researchers) and other actors, and the literature concerning data quality helped to design quality indicators that provided insight in the purposes for which a dataset can be used. The combination of the four kernel theories is needed to enhance the coordination of OGD use by researchers (see section 8.3).

Table 8-1 (continued): A design theory for OGD.

IS design theory components (Gregor & Jones, 2007, p. 322)	Design theory contributions
Principles of implementation; a description of processes for implementing the theory (either product or method) in specific contexts.	In total, 81 design principles have been developed (see section 5.3), encompassing 22 coordination design principles, 40 metadata design principles, 15 interaction design principles and 4 data quality design principles. Building on these design principles, the system design, the coordination patterns and the function design of the OGD infrastructure were described (see section 5.4). The system design, the coordination patterns and the function design support interoperation and may also be applied to the design of other OGD infrastructures.
Expository instantiation; a physical implementation of the artefact that can assist in representing the theory both as an expository device and for purposes of testing.	In this study a prototype of the OGD infrastructure (the artefact) was developed and evaluated (see chapter 6). Through quasi-experiments we examined to which extent the ease and the speed of OGD use was improved by the developed OGD infrastructure, and we examined the coordination of OGD use by including the management of dependencies between and among activities performed to use OGD in the evaluations. Studies that describe the development of OGD infrastructure prototypes are scarce. Moreover, while several studies have tried to evaluate open data use, this study is among the first to evaluate OGD use through quasi-experiments. The developed evaluation methodology can be used and further extended by other scholars. The prototype and the evaluations illustrated how the designed OGD infrastructure can be used to improve the ease and speed of OGD use and to enhance the coordination of OGD use by researchers.

Table 8-1 (continued): A design theory for OGD.

8.3 Combining the kernel theories

Kernel theories concerning coordination, metadata, interaction and data quality are often studied separately, whereas this study reveals that it is essential for the development of OGD infrastructures to combine these four kernel theories. This section discusses how the kernel theories can be combined, their social and/or technical impact, how each of the kernel theories contributed to enhancing the coordination of OGD use, and how the kernel theories influenced each other.

First, we contributed to the metadata kernel theory by confirming several recent studies that different types of metadata (discovery, contextual and detailed metadata) need to be combined to enhance the coordination of OGD use rather than only providing discovery metadata (see section 8.1.3). The theory and underlying knowledge related to metadata (i.e. the metadata kernel theory) supported the technical aspects of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis by OGD

users. Yet, this kernel theory also had a social impact, since it also provided technical elements that allowed users to interact. Metadata technically support OGD use coordination by managing the dependence of OGD users on different tools, that on their turn also depend on other tools and that need to be interoperable. Metadata also provide social support for managing dependencies by providing tools that researchers can use to interact with each other and with other actors (e.g. OGD providers). In this way the metadata kernel theory influenced the coordination of OGD use. The metadata kernel theory influenced the interaction mechanisms by providing design principles such as 'metadata can be used to integrate and establish communications between various tools', and it influenced the data quality kernel theory by providing design principles such as 'high-quality metadata is needed to assess the data quality'.

Second, the theory and underlying knowledge regarding interaction mechanisms (i.e. the interaction kernel theory) was used to improve interaction related to OGD use. This kernel theory mainly supported the social aspects of OGD use, such as the discussions of researchers with OGD providers and with other researchers. At the same time, this kernel theory had a technical impact by supporting users in their interaction that might be needed to use the OGD infrastructure tools. Interaction mechanisms socially support OGD use by managing the dependence of OGD users on each other and on other actors.

Third, the theory and underlying knowledge regarding data quality indicators (i.e. the data quality kernel theory) was used to enhance the analysis of OGD quality. This kernel theory was used to both address social and technical aspects of OGD use. For instance, the data quality indicators refer to the technical parameters of data quality, yet data quality also influences the use of the data and may enhance or complicate its interpretation, which can be considered a social aspect. Data quality indicators socially support OGD use by managing the dependence of OGD users on each other and on other actors (e.g. through free text data quality assessment), and they technically support OGD use by managing the dependence of OGD users on different tools (e.g. through tools that facilitate structured rating of data quality dimensions).

Fourth, this study was the first to apply coordination theory in the field of OGD and showed that coordination of OGD use does not merely require a focus on

processes, but additionally requires a technical perspective including the integration of tools, a social perspective including interaction between researchers, OGD providers and policy makers, and the interaction between the social and technical perspective (see section 8.1.3). Coordination theory both had a social and a technical impact on the use of OGD. The kernel theory of coordination was used to increase the management of dependencies. Coordination theory was applied to the metadata kernel theory, interaction mechanism kernel theory and data quality kernel theory. For instance, the coordination mechanism of 'standardisation, routines and rules' influenced the metadata model design by suggesting 'the use of standards and controlled vocabularies to describe datasets'. The coordination mechanism of 'boundary spanners' influenced the interaction mechanisms by suggesting the design principle 'to allow for communication and interaction between OGD users, OGD providers, and policy makers'. And the coordination mechanism of 'advanced structuring' influenced the design of the data quality indicators by suggesting the design principle 'to provide functions to discuss the quality of OGD'.

In sum, four kernel theories concerning coordination, metadata, interaction and data quality were combined in this study. Coordination theory influenced the other three kernel theories, while metadata, interaction and data quality more directly influenced OGD use through the developed infrastructure.

8.4 Research limitations

This section describes the limitations of our study. Research limitations are discussed with regard to taking an interpretivistic and open data proponent perspective, not considering non-functional requirements for the OGD infrastructure, the generalisation of the findings from the selected cases, the evaluation of the prototype instead of the complete infrastructure, and the generalisation of the findings from the quasi-experiments.

8.4.1 Taking an interpretivistic and an open data proponent perspective

This study has been conducted from the interpretivistic paradigm, which advocates that multiple realities exist, and that realities are socially constructed by human actors (Vaishnavi & Kuechler Jr, 2008; Walsham, 2001). Moreover, this study started from the perspective that OGD can be used to obtain benefits. An optimistic

view was adopted of an open data proponent. However, interpretive research has been criticised for not having objective evaluation criteria (Chen & Hirschheim, 2004), and previous research has also shown that OGD may also have negative side effects. For instance, releasing OGD may conflict with an individual's right to information privacy (Kulk & Van Loenen, 2012), the conventional wisdom that opening data results in transparency has been challenged (Bannister & Connolly, 2011), and open data might mainly empower those who are already 'empowered' (Gurstein, 2011). Although our quasi-experiments showed that the developed OGD infrastructure can enhance the coordination of OGD use, the infrastructure may also have negative effects. Although we discussed some of these effects in chapter seven, not all of them were evaluated in this study. Yet, metadata, interaction mechanisms and data quality indicators cannot assure that OGD will not be misinterpreted and misused.

To avoid giving a one-sided, biased representation of this study's findings and to deal with the criticisms on interpretivist research, various measures were taken by the researcher. First, different perspectives were examined. For instance, we did not only focus our literature review on the identification of open data benefits, but also on the barriers for OGD use, as well as on the potential negative effects of OGD use. In addition, we spoke to different case study participants, and we used multiple evaluation measures.

Second, we tried to make the process that led from data and experiences to findings as transparent as possible, and we took various measures to allow for replicating this study, so that generalisations become possible. For example, a protocol was developed for the elicitation of infrastructure requirements in the two case studies. The protocol described how the case studies optimised different types of validity. To discuss different aspects of the cases, we did not only consider the proponent perspective in the case study, but we also discussed negative aspects of OGD use. In the interviews both arguments for and against OGD publication and use were identified and described. A protocol was also developed for the quasi-experimental evaluation. We tried to reduce the researcher bias by involving a facilitator and observers in the quasi-experiments who had not been involved in the research before. The role of the actors (e.g. the facilitator, the observers, other participants) involved in the quasi-experiments was also evaluated

through the survey and showed that their influence on the quasi-experiment participants was limited.

Finally, both the positive and the negative effects of the OGD infrastructure were described. Sections 7.7 and 8.1.5 did not only discuss the infrastructure's positive effects as identified through the surveys, observations and time measures, but they also discussed its negative effects, weaknesses, and areas for improvement.

8.4.2 Non-functional requirements are not considered

This study focused on functional requirements for an OGD infrastructure that enhances the coordination of OGD use. Nevertheless, our definition of OGD infrastructures showed that various non-functional requirements of OGD infrastructures are also important. For instance, the evolvement of the infrastructure over time requires flexibility, sustainability and maintainability of the infrastructure. This study made the assumption that all the relevant non-functional requirements were met. A limitation of this research is that the non-functional requirements have not been investigated, and non-functional requirements could be critical success factors for implementing OGD infrastructures. Non-functional requirements, such as costs, performance, security and privacy may limit the capabilities or relevance of the functional requirements, and may in this way also affect user satisfaction concerning an OGD infrastructure.

8.4.3 Limitations regarding the generalisation of the findings from the cases

Two cases were studied to identify infrastructure requirements. The two cases were selected based on our focus on the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government. This focus led to the study of a very specific target group of OGD users in the case studies, namely only those people who are interested in using and who can use research data. This type of OGD use is different from other types of OGD use, such as the use of OGD by citizens to collect information about the quality of schools for their children or by entrepreneurs to develop an application. Additionally, other types of OGD, such as data derived from sensors, are different and might have led to different functional requirements for the OGD infrastructure.

Furthermore, both case studies were conducted in The Netherlands, and they both involved Dutch governmental organisations. The use of OGD may differ per country, since it may be influenced by, for example, national open data policies and the budget that a public agency makes available for the development of an OGD infrastructure. Moreover, two types of OGD were central in the cases, namely judicial and social data. The type of data may affect OGD disclosure and use. Furthermore, the cases involved two data providing organisations that both operated on a high level of the Dutch governmental hierarchy. Both organisations functioned as part of ministries, and this study did not involve cases regarding the use of municipal or regional OGD. The support for OGD provision and use may differ on different government levels.

In this study, we selected cases that fit in the scope of this study and that differed as little as possible, to be able to investigate whether the findings from the cases confirmed each other. At the same time, we varied with the type of data (judicial and social research data with diverse levels of sensitivity). Although the cases involved different types of data, they confirmed each other's findings. Moreover, the OGD infrastructure that was developed was evaluated with a variety of OGD users (mainly students and researchers). The OGD users in the quasi-experiments came from eighteen different countries, and many of them had experience with using various types of OGD. Therefore, the findings from the quasi-experiments suggest that generalisation of the case study findings is possible.

8.4.4 Evaluation of prototype instead of completely designed OGD infrastructure

The findings from the evaluations showed that the infrastructure can be used to improve the ease and speed of OGD use and to enhance the coordination of OGD use. Nevertheless, it is important to note that the infrastructure that was evaluated in the three quasi-experiments did not encompass the complete infrastructure design that was described in chapter five (see section 6.3). Although the evaluated prototype did contain the three functional elements of the OGD infrastructure, i.e. metadata, interaction mechanisms and data quality indicators, its function design was slightly restricted in comparison with the function design described in section 5.4.3:

- Out of the 30 metadata model functions (see section 6.3.1), 27 functions were selected for implementation in the prototype. The functions 'convert data format', 'refer to data' and 'link data manually' were not implemented.
- Out of the eleven interaction mechanism functions (see section 6.3.2), nine functions were implemented in the prototype. The two functions 'enter an open collaboration group' and 'enter a closed collaboration group' were not implemented.
- Out of the five data quality indicator functions (see section 6.3.3), the comparison of different quality ratings and reviews was the only function that was not implemented in the prototype.

The selection of functions took place based on the criterion that the functions needed to allow for measuring the key effects of metadata, interaction mechanisms and data quality indicators, and for refining the user requirements for these three functional infrastructure elements. The six infrastructure functions were not implemented in the prototype because it would be too time-consuming to use within the limited time frame of the evaluations, or because they were not central to the five OGD use activities of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. Without these functions, the five OGD use activities could still be evaluated. Our conclusions regarding the enhancement of the coordination of OGD use with the OGD infrastructure are based on the prototype of the developed OGD infrastructure and they do not concern OGD use that is related to the non-implemented functions.

8.4.5 Limitations regarding the generalisation of the findings from the quasi-experiments

A first limitation regarding the generalisation of the quasi-experiment results concerns the tasks conducted in the quasi-experiments. The scenarios executed in the quasi-experiments focused on a number of specific tasks related to metadata, interaction mechanisms and data quality indicators. The quasi-experiments did not evaluate the use of OGD in daily life, but only in a controlled setting. The difficulty or ease of using OGD also depends on the complexity of the tasks that the quasi-experiment participants needed to conduct. This means that if we would have chosen more complex scenario tasks, the assessed ease and speed of OGD use may have been different. In addition, since time was limited, the scenario could

include only a certain number of tasks, which raises the question to which extent these tasks represent daily OGD use by researchers. Moreover, the scenario tasks needed to be conducted within a limited time frame. This may have influenced the measured speed of OGD use.

A second limitation regarding the generalisation of the quasi-experiment results is that the quasi-experiment participants were not matched pair-wise. In our study, it was not possible to match a control group of professional open data users to a treatment group of professional open data users, since the number of participants in the treatment group of professionals was too small to divide the group into both a control and treatment group to still be able to conduct statistical tests (which required at least thirty participants per group). However, by inviting a particular group of persons to participate in the quasi-experiments, we tried to select participants (students and professionals) with a similar background (e.g. with regard to their experience in open data use and with regard to the focus on open data derived from research). Moreover, the characteristics of the participants in the control and treatment group were compared through a non-pair wise matching process. The characteristics of the different groups of participants were analysed and compared.

The first participant survey collected information about various characteristics of the quasi-experiment participants. Although the participants from the treatment groups and the control groups were relatively similar, they also differed with regard to some of the measured characteristics. For instance, the treatment group of professionals and the treatment group of students in the second quasi-experiment contained relatively more females than the other groups, and the professional open data users of the third quasi-experiment were relatively older. The first quasi-experiment also only contained Dutch participants (both in the control and in the treatment group), while more nationalities were represented in the second and third quasi-experiment. The professionals' level of experience with OGD use was higher than the student's level of experience with OGD use. Besides these measured differences, there may also have been differences between the control and treatment groups of the three quasi-experiments that we did not measure. A limitation of this study is that we do not have insight in these characteristics and their potential influence.

A third limitation for the generalisation of the quasi-experiments is that various intermediate variables we not investigated in-depth. For instance, the user interface was investigated, but not in detail. A variety of intermediate variables may have influenced the outcomes of the quasi-experiments, such as the user interface, experience with open data use and search options. More insight in the intermediate variables needs to be obtained.

9. Epilogue

This chapter provides an epilogue of our research. We first reflect on this study and then we provide recommendations for future research.

9.1 Reflection on this study

The reflection has been divided into six categories: using OGD for improving policy making, deciding to open or not to open governmental data, the stimulation of interaction regarding open data use, making money with open data, the role of the context in which datasets are published and used, and the evolvement of open data infrastructures.

9.1.1 Can we use open data for policy making?

We started this study by stating that OGD can be used to improve governmental policy making. The quasi-experiments conducted for this study evaluated to which extent the developed infrastructure improved the ease and speed of OGD use. The combination of the functional infrastructure elements, namely metadata, interaction mechanisms and data quality indicators, enhanced coordination by facilitating the use of various tools, the interaction with other OGD users and the interaction with OGD providers and governmental policy makers. We assume that policy makers can use the infrastructure to obtain insight in open data use, and that they can analyse the feedback provided by open data users to improve policy making. They may also use log data from the infrastructure to examine the use of OGD. Nevertheless, one could say that policy makers then use a form of 'spying' on OGD users, which could also have negative effects. For example, governments can then collect considerable data about the behaviour of citizens, which could violate data protection legislation, and governments might also use the gathered data for other purposes than policy making, such as watching citizens who use datasets about certain topics. Since policy making was outside the scope of this study, it is important that future research investigates the conditions under which OGD can be used to improve policy making.

The infrastructure provided a number of functions that are aimed at supporting policy makers. Policy makers could access the infrastructure and see

how people had used governmental datasets. For example, policy makers could see messages of OGD users in which they had discussed data or data use, they could use social media to see which findings from OGD use had been shared with others, and they could see which types of datasets users had requested to be released. The analysis of this type of feedback could be used to improve governmental policy making. Policy makers could also see what kind of people had used the datasets and in which ways, and they could contact them. For instance, they could see whether a dataset had been used by an entrepreneur, researcher, journalist, civil servant, librarian or other type of OGD user, and they could contact them with a personal message, in a collaboration group, in a discussion or via social media. It is premised that these functions indirectly improve policy making.

Nevertheless, apart from the infrastructure functions that can help policy makers to use open data, policy makers can face various other challenges. For example, the insights that OGD users obtain with OGD use may be based on wrongful data use. If policy makers would use these incorrect insights, this may not lead to the improvement of governmental policies, but to their deterioration. Furthermore, there are various institutional challenges for governments to use open data, such as the traditional top-down decision-making culture (also see section 9.1.2) (Meijer & Thaens, 2013; Mergel, 2012). Thus, although policy makers can in theory use the open data infrastructure for governmental policy making, future research should also consider the institutional barriers that hinder the use of open data for policy making.

9.1.2 To open or not to open?

To be able to use OGD on an infrastructure, the availability of governmental data is a precondition. However, our research showed that various institutional and legal barriers hinder the publication of governmental data, as we discussed in various articles (Zuiderwijk, Gascó, Parycek, & Janssen, 2014; Zuiderwijk & Janssen, 2014b; Zuiderwijk, Janssen, Choenni, et al., 2014; Zuiderwijk, Janssen, Meijer, et al., 2012). First, datasets can be sensitive. Public agencies risk violating the law when they release personal data. At the same time, the combination of different datasets could also result in the identification of persons. Governments can never be sure that the datasets that they release will not be combined with other datasets

in the future and that this will not lead to the identification of personal information. Since it is impossible to guarantee a priori people's privacy, governments are reluctant to publish their data (Zuiderwijk & Janssen, 2014b). Moreover, datasets can be policy-sensitive, for example, when they concern a certain topic that is under discussion by politicians. Publishing sensitive data may harm the reputation of politicians and organisations that publish the data. Additionally, for certain types of data the law prohibits their publication, even if a governmental organisation would want to disclose the data to the public.

Second, datasets can be created by multiple organisations which have different levels of security, different policies and which have to comply with different laws. When datasets are owned by different organisations, these organisations all need to give permission for the disclosure of the data, which makes it difficult to publish the data. Third, the data that are published can be biased or of low quality. Governmental organisations fear the misinterpretation and misuse of opened datasets, since this could damage the reputation of the data provider. The publication of the data may then have negative consequences for the government. Finally, governmental organisations may first want to reuse the datasets themselves, before they disclose the data, which could prohibit the publication of timely data. In sum, the extent to which the OGD infrastructure developed in this study can be used strongly depends on the availability of governmental research data, and the publication of these data is hindered by various institutional and legal barriers. For each dataset, governmental organisations need to balance the arguments for and against its publication.

9.1.3 How to stimulate interaction?

This study showed that OGD providers, OGD users and policy makers can interact to utilise open data for improving governmental policy making. OGD users, OGD providers and policy makers may not be motivated intrinsically to participate in OGD use activities. The use of OGD infrastructures offers a new way of working for policy makers, since open data are traditionally not used to improve governmental policy making. Furthermore, it offers a different way of working to OGD providers, for instance, through discussions about open data use or by responding to requests for disclosing certain datasets. Moreover, most functions that we tested in

the quasi-experiments require a critical mass of users. For example, we do not expect discussion messages and data rating reviews to be successful when only few people provide them. Therefore, open data infrastructures using metadata, interaction mechanisms and quality indicators can profit from a large user base consisting of OGD providers, OGD users and policy makers.

However, potential OGD providers and OGD users may not be motivated to participate in discussions about OGD use or to share the findings from open data use. The designed metadata, interaction mechanisms and data quality indicators to a large extent rely on sharing findings from OGD use. For instance, connecting a visualisation of a dataset to the data themselves can rely on sharing metadata about the visualisation, and assessing the quality of data can rely on explaining for what purposes the data were used. This raises the question what incentives open data users have to share this information with other persons.

In discussions that we had with OGD users and developers of OGD infrastructures, it was suggested to acknowledge the activities of OGD providers and OGD users by awarding them with so-called 'kudos'. Each activity on the OGD infrastructure would yield the OGD providers and users a certain number of kudos, and the OGD providers and users with most kudos would be mentioned on the home page of the infrastructure as 'top OGD providers/users'. This might motivate them to be active on the OGD infrastructure. Various other functionalities to stimulate interaction on OGD infrastructures have also been proposed during discussions with OGD infrastructure developers. One option is to better connect OGD providers to the infrastructure by making a dataset request be delivered to a data provider directly, rather than requiring a data provider to go to the OGD infrastructure to search for data requests that concern their data. Another option is to add a functionality that allows OGD users to uprate or down rate open dataset requests and comments on datasets. For example, if many users uprate a request for a certain dataset, the OGD provider can see that opening this particular dataset would be interesting to many users, and if many users down rate a certain comment on a dataset, this could show that these users do not agree with each other. Further research is recommended to pay attention to how the interaction between OGD providers, OGD users and policy makers can be stimulated.

Interaction also has a number of other limitations. One of them includes the subjectivity of feedback. For instance, data quality assessment is subjective and depends to a large extent on the user's purpose for the data use and the user's frame of reference. Having a larger user base to assess datasets may contribute to reducing this problem, although data quality assessment will always remain subjective. It is therefore important that quality assessment of OGD takes into account different types of data use, different data quality indicators and that users provide a nuanced view on the assessment by describing the context of the way that they used the data. It should be explained for which types of use a certain dataset was useful or for which purposes it was not useful. These aspects cannot be controlled strictly, although quality checks can already be performed before datasets actually appear online. The maintainer of the dataset may conduct an initial data quality check before the data are published on the infrastructure.

9.1.4 Making money with open data

While this study focused on researchers as OGD users, OGD can also be used by other types of users. Beyond the scope of this study, open data changes the way that organisations and individuals can make money. The philosophy behind open data is that data need to be provided to data consumers without payment for accessing and using the data. Various organisations that used to make money by selling their data in the past may now change their revenue models and think of other ways to earn money. These organisations may now provide a basic set of data for free, while requiring the data user to pay for additional services, such as quality checks, regular updates or real time data. They may expect that researchers and citizens would use the basic dataset, while companies that make money with their data would be willing to pay for the additional services. Other organisations ask the federal government for subsidies that would help them opening data for free, or they participate in projects funded by the European Commission or national governments that allow them to run a pilot to release datasets.

While governmental organisations think of new business models, data users consider how the opened datasets can be used. For open datasets that are accompanied by a license, the conditions for open data use may differ per license.

Some licenses allow data users to use open data commercially, for instance by developing services and products with which they can make money based on open data. Our research showed that a whole range of new business models for infomediary OGD users is emerging (Janssen & Zuiderwijk, 2014). Key infomediary business models that are emerging include:

- Apps that provide real-time services, such as data about weather, quality of restrooms, vehicles, houses, or pollution. The app processes the data and presents it visually to the user. The infomediary company may earn money from selling the app or from selling advertisement through the app.
- Information aggregators and comparisons that combine open data sources and process and sell them for subsequent presentation to the users (e.g. a transportation planner that combines open transport data from various organisations or a company that combines information about the quality of schools). The company may earn money from selling the aggregated or combined open data sources.
- Service platforms that offer features for searching, importing, cleansing, processing, and visualizing information (e.g. www.junar.com). The company can earn money from selling the services and features to other companies using open data or to governments who may implement the services and features in their open data platforms (Janssen & Zuiderwijk, 2014).

Other licenses may not allow for commercial open data usage. Yet, our research showed that even when commercial open data use is allowed, the creation of competitive advantage with open data requires companies to have in-house capabilities and resources for open data use (Zuiderwijk, Janssen, Poulis, & Vandekaa, 2015). Thus, both the release and the use of open data require governmental organisations, companies and entrepreneurs to change their mind-set and to consider new revenue models.

9.1.5 Can we use the OGD infrastructure outside the context of this study?

This study focused on a specific type of open data, namely structured judicial and social research data. The developed OGD infrastructure focuses on meeting the functional requirements for OGD use for these particular types of data. Moreover,

this study focused on a variety of barriers for OGD use, yet not on all the barriers that exist for OGD use. Several barriers are not solved with the OGD infrastructure developed in this study, such as the barrier that 'considerable different terminology is used to describe datasets' and the barrier that 'datasets are released in numerous different formats'. This raises the question whether the developed OGD infrastructure can be used outside the context of structured judicial and social research data. Several contextual aspects influence the usability of the infrastructure outside the context of this study. For instance, the research data that we studied was created with the aim to contribute to governmental policy making, which is different from data that have been generated in other ways (e.g. through sensors). Other datasets may be bigger and unstructured, which complicates the use of these data and which could lead to different user requirements. Moreover, a large part of the investigated judicial data and also some of the social data was sensitive, which may have influenced the functional requirements for the OGD infrastructure. For example, the sensitivity of the data affected the type of license that was used for the opened datasets and it affected the number of datasets that was published. In addition, if policy makers would have been involved in this study we might also have found other functional requirements for the development of the OGD infrastructure. For example, the involvement of policy makers in the case studies might have shown that even more metadata would be required to use OGD for policy making. Policy makers might also have pointed at non-functional requirements related to a lack of trust in using the findings derived from OGD use for governmental policy-making. Further research needs to consider the contextual aspects that influence the usability of the developed OGD infrastructure outside the context of our case studies.

9.1.6 How will open data infrastructures evolve?

While infrastructures are important, they often provide only one piece of the puzzle for open data providers and users which often deploy more than one infrastructure to publish or process open data. Even if the infrastructure is not going to be maintained by a single governmental organisation or by a company, the different elements of the infrastructure can be reused by future OGD infrastructures. As such, we plea for viewing infrastructures as part of a wider open data ecosystem in

which each instrument and tool can add value. For instance, from our list of 46 functions, developers of OGD infrastructures can select the ones that they believe need to be implemented in their infrastructure. Each of these functions can be developed further, and based on the extensive use of the functions they may be improved and needs for additional functions may be elicited. The reuse of each instrument and tool in the open data ecosystem supports the evolvement of open data infrastructures and their sustainability, as we showed in one of our papers (Zuiderwijk, Janssen, & Davis, 2014). The research described in the paper showed that "open data ecosystems develop through user adaptation, feedback loops and dynamic supplier and user interactions" (p. 23).

9.2 Towards an agenda for open data research

Chapter eight and the previous sections of this chapter highlighted what has been accomplished in this study, yet they also revealed various limitations of this study. In this section we identify a number of recommendations for an emerging open data research agenda.

- *Recommendation 1: balance the benefits and the risks of OGD publication.*

This study started from the perspective that OGD can be used to obtain benefits. An optimistic view was adopted of an open data proponent. Nevertheless, previous research has also shown that OGD may have negative side effects and there is a diversity of risks when data are published. Future research is recommended to examine both the positive and negative effects of OGD. This type of new research should help OGD providers, OGD users and policy makers to weigh the advantages and the disadvantages of OGD publication and may support the publication of OGD. This future research is expected to influence the efficiency of the OGD publication processes and could help OGD users to obtain the data that they need.

- *Recommendation 2: examine the evolving design of the OGD infrastructure*

The application of a design science approach in the field of open data is recommended for the development of OGD infrastructures. In our quasi-experiments we saw how the infrastructure evolves when it is used. However, the

further evolvement of the infrastructure was outside the scope of this study. We propose that future research examines how the design of the OGD infrastructure is adapted through the interaction of infrastructure users with the technology. This type of research may help to better understand the importance of non-functional requirements, such as sustainability and maintainability, and their relation to the functional requirements that we elicited in this study. The findings can be used to further improve the coordination of OGD use through infrastructures, so that users can contribute to obtaining the benefits of OGD including improved governmental policy-making.

- *Recommendation 3: study to which extent the identified infrastructure requirements also apply to other types of data, other types of data use, on other governmental levels and in other countries and cultures.*

The case studies conducted to identify functional requirements for the OGD infrastructure focused on a specific type of OGD use, namely the operational use of structured research OGD from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures. The judicial and social data were created and collected by governmental organisations that operated on a high level of the Dutch national government. Moreover, the number of studied cases was limited, and we focused on functional requirements only. Some of the identified requirements may be typical for judicial or social data, and may not be applicable to other types of data, or to the same type of data from other organisations. Future research is recommended to focus on studying to which extent the identified requirements also apply to other contexts, such as cases involving other types of open data (e.g. big open data and data concerning other topics), other types of OGD use, in other countries, in other cultures, or on the level of local goverments. This type of research is expected to stimulate the development of OGD infrastructures in various domains, and can support the use of OGD from multiple disciplines. Moreover, future research is recommended to investigate the non-functional requirements for OGD infrastructures, since the non-functional requirements (e.g. costs, sustainability, security and privacy) may influence the user satisfaction with an OGD infrastructure and may limit the capabilities and the relevance of the functional requirements. Investigating the

generalizability of the OGD infrastructure as well as its non-functional requirements could support the wider adoption of OGD.

- *Recommendation 4: further improve OGD use utilising the developed OGD infrastructure*

Even though the evaluations showed higher levels of the ease and speed of OGD use for the treatment group compared to the control group, OGD use on the infrastructure can still be improved further. While chapter three discussed a wide variety of barriers for OGD use, this study did not address all of these barriers. Various barriers were not considered in this study, such as the barrier that 'considerable different terminology is used to describe datasets' and the barrier that 'datasets are released in numerous different formats'. These barriers are not completely removed with the developed OGD infrastructure. Future research could focus on removing the barriers that have not been addressed in this study to further enhance the coordination of OGD use.

In addition, for various OGD use activities the quasi-experiment participants answered that they neither agreed nor disagreed with the statement, or that they slightly agreed, where higher agreement pointed at higher ease of OGD use. Especially the professional open data users did not strongly agree with statements. Future research may comprise an examination of how the ease and speed of OGD can be increased further. This may also involve studying additional kernel theories (e.g. concerning user interface design, data visualisation, business models, trust and privacy) besides the four kernel theories that contributed to the design of the infrastructure in this dissertation, as well as the development of additional functional infrastructure elements. This type of research can then further enhance the coordination of OGD use and make the use of OGD easier and faster.

Moreover, the four kernel theories that contributed to the design of the infrastructure can be improved further. For example, the kernel theory regarding the use of metadata for OGD infrastructures can be developed further by adding additional metadata fields, and by developing tools that allow governmental organisations to collect metadata about the context in which the data were created at the same time as they collect the data. The other three kernel theories used in this study can also be developed further. For instance, additional mechanisms for

participation of policy makers and OGD providers on OGD infrastructures can be generated, such as dashboards where they can obtain statistics about the use of all the datasets that they uploaded at a single glance. Future research is recommended to evaluate in more detail how other elements, such as interface design, trust and privacy influence the coordination of OGD use, and how the used kernel theories can be expanded[8]. This type of future research is expected to contribute to theory building in the area of infrastructures for OGD use, and may support the development of specific theories. These theories can help scientists and practitioners to (re)design and improve OGD infrastructures.

- *Recommendation 5: evaluate the OGD infrastructure*

The conducted quasi-experiments involved a particular group of participants which may not be representative for the entire OGD use community. We focused on the use of OGD by researchers outside the government. While the quasi-experiments involved mainly students and researchers, other types of OGD users were not so well-represented, such as archivists, entrepreneurs, civil servants, librarians and journalists. Furthermore, the control group did not include professional OGD users. We recommend future research to evaluate OGD use utilising the developed infrastructure in various contexts, such as in other countries, with other types of OGD users and with larger numbers of participants. Moreover, future evaluations of the infrastructure can involve professional OGD users both in treatment and control groups. In addition, future research may evaluate to which extent the OGD infrastructure meets the non-functional requirements. This type of research will help to obtain insight in the generalizability of the developed OGD infrastructure and in the adaptations that are needed for specific contexts, such as other countries and other types of OGD users.

[8] We will conduct further research in this area as part of the VRE4EIC project (A Europe-wide interoperable Virtual Research Environment to Empower multidisciplinary research communities and accelerate Innovation and Collaboration).

Reference list

Ackoff, R. L. (1989). From data to wisdom: Presidential address to ISGSR, June 1988. *Journal of applied systems analysis, 16*(1), 3-9.

Alani, H., Hall, W., O'Hara, K., Shadbolt, N., Chandler, P., & Szomszor, M. (2008). Building a pragmatic semantic web. *IEEE Intelligent Systems, 23*(3), 61-68.

Alexander, I. F., & Maiden, N. (2004). *Scenarios, stories, use cases through the systems development life-cycle*. West Sussex: John Wiley & Sons.

Alexopoulos, C., Loukis, E., Charalabidis, Y., & Zuiderwijk, A. (2013). *An evaluation framework for traditional and advanced open public data e-infrastructures*. Paper presented at the 13th European Conference of Electronic Government, Como, Italy.

Alexopoulos, C., Spiliotopoulou, L., & Charalabidis, Y. (2013). *Open data movement in Greece: a case study on open government data sources*. Paper presented at the 17th Panhellenic Conference on Informatics, Thessaloniki, Greece.

Alexopoulos, C., Zuiderwijk, A., Charalabidis, Y., Loukis, E., & Janssen, M. (2014). *Designing a second generation of open data platforms: integrating open data and social media* Paper presented at the 13th conference on Electronic Government, Dublin, Ireland.

Altinay, L., & Paraskevas, A. (2008). Chapter 4 - Research philosophies, approaches and strategies. In L. Altinay & A. Paraskevas (Eds.), *Planning Research in Hospitality & Tourism* (pp. 69–88). Oxford: Elsevier.

Archer, P., Dekkers, M., Goedertier, S., & Loutas, N. (2013). Study on Business Models for Linked Open Government Data (BM4LOGD). Retrieved February 23, 2015, from https://joinup.ec.europa.eu/community/semic/document/study-business-models-linked-open-government-data-bm4logd.

Argyzoudis, E., Mouzakitis, S., Yaeli, A., & Glikman, Y. (2014). Deliverable 5.5.4. ENGAGE services development report and documentation, 3rd version.

Auer, S., Lehmann, J., Ngomo, A.-C. N., & Zaveri, A. (2013). Introduction to linked data and its lifecycle on the web. In S. Rudolph, G. Gottlob, I. Horrocks & F. van Harmelen (Eds.), *Reasoning Web. Semantic Technologies for Intelligent Data Access* (pp. 1-90). Mannheim: Springer.

Babüroglu, O. N., & Ravn, I. (1992). Normative action research. *Organization Studies, 13*(1), 019-034.

Bailo, D., & Jeffery, K. (2014). *EPOS: a novel use of CERIF for data-intensive science*. Paper presented at the International Conference on Current Research Information Systems, Rome, Italy.

Bannister, F., & Connolly, R. (2011). The Trouble with Transparency: A Critical Review of Openness in e-Government. *Policy & Internet, 3*(1), Article 8.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 1-52.

Behkamal, B., Kahani, M., Bagheri, E., & Jeremic, Z. (2014). A metrics-driven approach for quality assessment of linked open data. *Journal of Theoretical and Applied Electronic Commerce Research, 9*(3), 64-79.

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of Information Systems. *MIS Quarterly, 11*(3), 369-386.

Berners-Lee, T. (2009). Linked data. Retrieved May 7, 2014, from http://www.w3.org/DesignIssues/LinkedData.html.

Bernstein, L. (1996). Foreword: importance of software prototyping. *Journal of Systems Integration, 6*(1-2), 9-14.

Berti-Équille, L. (2007). Quality awareness for managing and mining data. Retrieved July 10, 2015, from https://www.researchgate.net/profile/Laure_Berti-Equille/publication/251573080_L._Berti-quille_Quality_Awareness_for_Data_Managing_and_Mining_Habilitation__Diriger_des_Recherches_Universit_de_Rennes_1_Juin_2007/links/0046351f12d600266b000000.pdf.

Bertot, J. C., Jaeger, P. T., Shuler, J. A., Simmons, S. N., & Grimes, J. M. (2009). Reconciling government documents and e-government: Government information in policy, librarianship, and education. *Government Information Quarterly, 26*(3), 433–436.

Bertot, J. C., McDermott, P., & Smith, T. (2012). *Measurement of open government: metrics and process*. Paper presented at the 45th Hawaii International Conference on System Sciences, Hawaii, U.S.A.

Bharosa, N. (2011). *Netcentric information orchestration. Assuring information and system quality in public safety networks.* Oisterwijk: Uitgeverij BOXPress.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - The story so far. *International Journal on Semantic Web, 5*(3), 1-22.

Böhm, C., Freitag, M., Heise, A., Lehmann, C., Mascher, A., Naumann, F., . . . Schmidt, M. (2012). *GovWILD: integrating open government data for transparency*. Paper presented at the 21st international Conference Companion on World Wide Web, Lyon, France.

Bonoma, T. V. (1983). *A case study in case research: marketing implementation*. Boston, Massachusetts: Harvard University Graduate School of Business Administration.

Braa, J., Hanseth, O., Heywood, A., Mohammed, W., & Shaw, V. (2007). Developing health information systems in developing countries: the flexible standards strategy. *MIS Quarterly, 31*(2), 381-204.

Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012a). *OPEN - Enabling non-expert users to extract, integrate, and analyze open data*. Paper presented at the Datenbank-Spektrum. Zeitschrift für Datenbanktechnologien und Information Retrieval.

Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012b). *The state of open data. Limits of current open data platforms*. Paper presented at the International World Wide Web Conference, Lyon, France.

Bryman, A. (2012). *Social research methods* (fourth ed.). Oxford: Oxford University Press.

Bunakov, V., & Jeffery, K. (2013). *Licence management for public sector information*. Paper presented at the Conference for e-Democracy and Open Government, Krems an der Donau, Austria.

Burrell, G., & Morgen, G. (1979). *Sociological paradigms and organizational analysis*. Burlington: Ashgate Publishing Company.

Bygstad, B. (2010). Generative mechanisms for innovation in information infrastructures. *Information and Organization, 20*(3), 156-168.

Campbell, D. T. (1969). Reforms as experiments. *American psychologist, 24*(4), 409.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research* Boston: Houghton Mifflin.

Campbell, D. T., & Stanley, J. C. (1969). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carroll, J. M. (1999). *Five reasons for scenario-based design*. Paper presented at the 32nd Hawaii International Conference on System Sciences, Hawaii, U.S.A.

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: citizen trust, innovation and acceptance factors. *Information Systems Journal, 15*(1), 5-25.

Charalabidis, Y., Loukis, E., & Alexopoulos, C. (2014). *Evaluating second generation open government data infrastructures using value models*. Paper presented at the 47th Hawaii International Conference on System Sciences Hawaii, U.S.A.

Charalabidis, Y., Ntanos, E., & Lampathaki, F. (2011). An architectural framework for open governmental data for researchers and citizens. In M. Janssen, A. Macintosh, J. Scholl, E. Tambouris, M. Wimmer, H. d. Bruijn & Y. H. Tan (Eds.), *Electronic government and electronic participation joint proceedings of ongoing research and projects of IFIP EGOV and ePart 2011* (pp. 77-85). Delft Springer.

Checkland, P. (1981). *Systems thinking, systems practice*. Chichester: John Wiley & Sons.

Checkland, P., & Poulter, J. (2010). Soft systems methodology. In M. Reynolds & S. Holwell (Eds.), *Systems approaches to managing change: a practical guide* (pp. 191-242). London: Springer.

Chen, W., & Hirschheim, R. (2004). A paradigmatic and methodological examination of information systems research from 1991 to 2001. *Information Systems Journal, 14*(3), 197-235.

Chen, W., & Paik, I. (2013). Improving efficiency of service discovery using linked data-based service publication. *Information Systems Frontiers, 15*(4), 613-625.

Choenni, R. (2015). Metadata. In A. Zuiderwijk (Ed.).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (second ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Conradie, P., & Choenni, S. (2012). *Exploring process barriers to release public sector information in local government*. Paper presented at the 6th international conference on theory and practice of electronic governance, New York, U.S.A.

Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. *Government Information Quarterly, 31*(supplement 1), S10–S17.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Cowan, D., & McGarry, F. (2014). *Perspectives on open data: issues and opportunities*. Paper presented at the IEEE International Conference on Software Science, Technology and Engineering, Ramat Gan, Israel.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Crowston, K., Rubleske, J., & Howison, J. (2004). Coordination theory: a ten-year retrospective. In P. Zhang & D. Galletta (Eds.), *Human-Computer Interaction in Management Information Systems*. New York: M.E. Sharpe.

Dadzie, A.-S., & Rowe, M. (2011). Approaches to visualising linked data: a survey. *Semantic Web, 2*(2), 89-124.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

DANS. (2013). About DANS. Retrieved August 16, 2013, from http://dans.knaw.nl/en/content/about-dans.

Danyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan & A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 671-689). London: Sage Publications Ltd.

Data Archiving and Networked Services. (2014a). Partners. Retrieved June 11, 2014, from http://www.dans.knaw.nl/en/content/data-archive/partners.

Data Archiving and Networked Services. (2014b). Welcome to EASY. Retrieved November 27, 2014, from https://easy.dans.knaw.nl/ui/home.

Data Archiving and Networked Services, Centraal Bureau voor de Statistiek, Huygens Instituut, Internationaal Instituut voor Sociale Geschiedenis, Koninklijke Bibliotheek, & Vereniging voor Geschiedenis en Informatica. (2008). Data uit beleidsonderzoek voor rijk: verplicht archiveren. *e-data&research, 3*(1), 1.

Davies, T. (2010). Open data, democracy and public sector reform. A look at open government data use From data.gov.uk. Retrieved May 19, 2014, from http://www.opendataimpacts.net/report/.

Davies, T. G., & Bawa, Z. A. (2012). The promises and perils of Open Government Data (OGD). *The Journal of Community Informatics, 8*(2), n.p.

Davis, F. B. (1964). *Educational measurements and their interpretation*. Belmont: Wadsworth.

Dawes, S. (2010). Stewardship and usefulness: policy principles for information-based transparency. *Government Information Quarterly, 27*(4), 377–383.

Dawes, S., & Helbig, N. (2010). *Information strategies for open government: challenges and prospects for deriving public value from government transparency*. Paper presented at the 9th International Conference on e-Government, Lausanne, Switzerland.

Dawes, S., Pardo, T., & Cresswell, A. (2004). Designing electronic government information access programs: a holistic approach. *Government Information Quarterly, 21*(1), 3-23.

DDI Alliance. (2009). DDI. Data Documentation Initiative. Retrieved January 28, 2015, from http://www.ddialliance.org/.

De Vocht, L., Dimou, A., Breuer, J., Van Compernolle, M., Verborgh, R., Mannens, E., . . . Van de Walle, R. (2014). *A visual exploration workflow as enabler*

for the exploitation of linked open data. Paper presented at the International Semantic Web Conference, Trentino, Italy.
Dempsey, L., & Heery, R. (1998). Metadata: a current view of practice and issues. *Journal of Documentation, 54*(2), 145–172.
Denyer, D., Tranfield, D., & van Aken, J. E. (2008). Developing design propositions through research synthesis. *Organization Studies, 29*(3), 393-413.
Detlor, B., Hupfer, M. E., Ruhi, U., & Zhao, L. (2013). Information quality and community municipal portal use. *Government Information Quarterly, 30*(1), 23-32.
Dimou, A., de Vocht, L., van Grootel, G., van Campe, L., Latour, J., Mannens, E., . . . van de Walle, R. (2014). *Visualizing the information of a linked open data enabled research information system*. Paper presented at the Current Research Information Systems Conference, Rome, Italy.
Ding, L., Peristeras, V., & Hausenblas, M. (2012). Linked open government data. *Intelligent Systems, IEEE, 27*(3), 11-15.
Dubé, L., & Paré, G. (2003). Rigor in information systems positivist case research: current practices, trends, and recommendations. *MIS Quarterly, 27*(4), 597-635.
Dublin Core Metadata Initiative. (2010). Dublin Core metadata element set, version 1.1. Retrieved April 20, 2015, from http://dublincore.org/documents/dces/.
Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-lib magazine, 8*(4).
Dym, C. L., & Little, P. (2004). *Engineering design: a project-based introduction* (Second ed.). New York: John Wiley & Sons.
Easterby-Smith, M., Thorpe, R., & Lowe, A. (2002). *Management research: an introduction*. London: SAGE Publications.
Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532-550.
Eppler, M. J., & Mengis, J. (2004). The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society, 20*(5), 325-344.
ESD Standards. (2004). e-GMS-e-government metadata standard version 3.0. Retrieved April 22, 2015, from http://www.esd.org.uk/standards/egms/.
EuroCRIS. (2010). CERIF Releases Retrieved April 22, 2015, from http://www.eurocris.org/Index.php?page=CERIFreleases&t=1.
European Commission. (2003). Directive 2003/98/EC of the European Parliament and of the council of 17 November 2003 on the re-use of public sector information. Retrieved March 16, 2015, from http://ec.europa.eu/information_society/policy/psi/rules/eu/index_en.htm.
European Commission. (2008). Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata. Retrieved July 15, 2015, from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008R1205&from=EN.
European Commission. (2011). Digital agenda: Commission's open data strategy, questions & answers. Retrieved October 27, 2014, from http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/11/891&format=HTML&aged=1&language=EN&guiLanguage=en.

European Commission. (2013). Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the Re-use of Public Sector Information Retrieved April 7, 2015, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE.

Fielding, R., Mogul, J., Frystyk, H., Masinter, L., Leach, P., & Berners-Lee, T. (1999). Hypertext Transfer Protocol HTTP/1.1. Retrieved February 25, 2015, from www.w3.org/Protocols/rfc2616/rfc2616.txt.

Foulonneau, M., Martin, S., & Turki, S. (2014). How open data are turned into services? In M. Snene & M. Leonard (Eds.), *Exploring services science* (pp. 31-39). Geneva: Springer International Publishing.

Galbraith, J. (1973). *Designing complex organizations*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Garbett, A., Linehan, C., Kirman, B., Wardman, J., & Lawson, S. (2011). *Using social media to drive public engagement with open data*. Paper presented at the Digital Engagement, Newcastle, UK.

Geiger, C. P., & von Lucke, J. (2012). Open government and (linked) (open) (government) (data). *Journal of e-Democracy and Open Government, 4*(2), 265-278.

Gibbs, A. (2005). The paradigms, contexts, and enduring conflicts of social work research. In L. M. Stoneham (Ed.), *Advances in Sociology Research* (pp. 135-164). New York: Nova Science Publishers.

Gilb, T. (1997). Towards the engineering of requirements. *Requirements Engineering, 2*(3), 165-169.

Gilliland, A. J. (2008). Setting the stage. In M. Baca (Ed.), *Introduction to metadata 3.0* (second ed., pp. 1-19). Los Angeles: Getty Publications.

Gittell, J. H. (2002). Coordinating mechanisms in care provider groups: relational coordination as a mediator and input uncertainty as a moderator of performance effects. *Management Science, 48*(11), 1408-1426.

Gittell, J. H. (2011). New directions for relational coordination theory. In G. Spreitzer & K. Cameron (Eds.), *The Oxford handbook of positive organizational scholarship* (pp. 400-411). New York: Oxford University Press.

Goldkuhl, G. (2004). Design theories in Information Systems - a need for multi-grounding *Journal of Information Technology Theory and Application, 6*(2), 59-72.

Gonzalez, R. A. (2010). *A framework for ICT-supported coordination in crisis response*. Enschede: Ipskamp Drukkers BV.

Gosain, S., Lee, Z., & Kim, Y. (2005). The management of cross-functional inter-dependencies in ERP implementations: emergent coordination patterns. *European Journal of Information Systems, 14*(4), 371-387.

Gosain, S., Malhotra, A., & El Sawy, O. A. (2004). Coordinating for flexibility in e-business supply chains. *Journal of Management Information Systems, 21*(3), 7-45.

Gregg, D. G., Kulkarni, U. R., & Vinzé, A. S. (2001). Understanding the philosphical underpinnings of software engineering research in Information Systems. *Information Systems Frontiers, 3*(2), 169-183.

Gregor, S. (2006). The nature of theory in Information Systems. *MIS Quarterly, 30*(3), 611-642.

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems, 8*(5), 312-335.

Gribbons, B., & Herman, J. (1997). True and quasi-experimental designs. *Practical Assessment, Research & Evaluation, 5*(14), n.p.

Grootveld, M., & Egmond, J. v. (2011). Data reviews, peer-reviewed research data. *DANS studies in digital archiving.* Retrieved January 6, 2015, from http://www.dans.knaw.nl/sites/default/files/file/publicaties/DANS_SDA_5_Data_Reviews_peer_reviewed_research_data_NL_DEF.pdf.

Gurin, J. (2014). *Open data now. The secret to hot startups, Smart investing, savvy marketing, and fast innovation*. New York: Mc Graw Hill Education.

Gurstein, M. (2011). Open data: empowering the empowered or effective data use for everyone? *First Monday, 16*(2), n.p.

Hanseth, O. (2004). From systems and tools to networks and infrastructures - from design to cultivation. Towards a theory of ICT solutions and its design methodology implications. Retrieved April 22, 2015, from http://heim.ifi.uio.no/oleha/Publications/ib_ISR_3rd_resubm2.html.

Hanseth, O., & Lyytinen, K. (2010). Design theory for dynamic complexity in Information Infrastructures: the case of building internet. *Journal of Information Technology, 25*, 1-19.

Hart, C. (1998). *Doing a literature review: releasing the social science research imagination*. London: Sage Publications Ltd.

Henfridsson, O., & Bygstad, B. (2013). The generative mechanisms of digital infrastructure evolution. *MIS Quarterly, 37*(3), 907-931.

Hevner, A. R., & Chatterjee, S. (2010). *Design research in Information Systems*. New York: Springer.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in Information Systems research. *MIS Quarterly, 28*(1), 75-105.

Hirschheim, R., & Klein, H. K. (1989). Four paradigms of Information Systems development. *Communications of the ACM, 32*(10), 1199-1216.

Ho, J., & Tang, R. (2001). *Towards an optimal resolution to information overload: an infomediary approach*. Paper presented at the International ACM SIGGROUP Conference on Supporting Group Work, Boulder, Colorado, U.S.A.

Hooker, J. N. (2004). Is design theory possible? *Journal of Information Technology, Theory and Application, 6*(2), 73-83.

Hovland, C. I. (1959). Reconciling conflicting results derived from experimental and survey studies of attitude change. *American psychologist, 14*(1), 8-17.

Huijboom, N., & van den Broek, T. (2011). Open data: an international comparison of strategies. *European Journal of ePractice, 12*(1), 4-16.

Iivari, J., & Venable, J. R. (2009). *Action research and design science research - Seemingly similar but decisively dissimilar*. Paper presented at the 17th European Conference on Information Systems, Verona, Italy.

Immonen, A., Palviainen, M., & Ovaska, E. (2014). Requirements of an open data based business ecosystem. *IEEE Access, 2*, 88-103.

Ince, D. C., & Hekmatpour, S. (1987). Software prototying - progress and prospects. *Information and Software Technology, 29*(1), 8-14.

Janssen, K. (2011). The influence of the PSI directive on open government data: an overview of recent developments. *Government Information Quarterly, 28*(4), 446-456.

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.

Janssen, M., Chun, S. A., & Gil-Garcia, J. R. (2009). Building the Next Generation of Digital Government Infrastructures. *Government Information Quarterly, 26*(2), 233–237.

Janssen, M., & Zuiderwijk, A. (2014). Infomediary business models for connecting open data providers and users. *Social Science Computer Review, 32*(5), 694-711.

Jeffery, K. (2000). Metadata: the future of Information Systems. In J. Brinkkemper, E. Lindencrona & A. Sølvberg (Eds.), *Information Systems Engineering: State of the art and research themes*. London: Springer Verlag.

Jeffery, K. (2013). Metadata models. *Samos Summer School 2013* [Powerpoint presentation].

Jeffery, K., Asserson, A., Houssos, N., Brasse, V., & Jörg, B. (2014). *From open data to data-intensive science through CERIF*. Paper presented at the 12th International Conference on Current Research Information Systems, Rome, Italy.

Jeffery, K., Asserson, A., Houssos, N., & Jörg, B. (2013). *A 3-layer model for metadata*. Paper presented at the International Conference on Dublin Core and Metadata Applications, Lisbon, Portugal.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: triangulation in action. *Administrative science quarterly, 24*(4), 602-611.

Joorabchi, A., & Mahdi, A. E. (2011). An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. *Journal of Information Science, 37*(5), 499-514.

Jurisch, M. C., Kautz, M., Wolf, P., & Krcmar, H. (2015). *An international survey of the factors influencing the intention to use open government*. Paper presented at the 48th Hawaii International Conference on System Sciences, Hawaii, U.S.A.

Karr, A. F. (2008). Citizen access to government statistical information. In H. Chen, L. Brandt, V. Gregg, R. Traunmuller, S. Dawes, E. Hovy, A. Macintosh & C. A. Larson (Eds.), *Digital government: E-government research, case studies, and implementation* (pp. 503−529). New York: Springer.

Kassen, M. (2013). A promising phenomenon of open data: a case study of the Chicago open data project. *Government Information Quarterly, 30*(4), 508–513.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin, 82*(3), 345.

Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in Information Systems. *MIS Quarterly, 23*(1), 67-94.

Kroeze, J. H. (2012). Postmodernism, interpretivism, and formal ontologies. In M. Mora, O. Gelman, A. Steenkamp & M. S. Raisinghani (Eds.), *Research methodologies, innovations and philosophies in software systems*

engineering and information systems (pp. 43-62). Hershey: Information Science Reference.

Krotoski, A. K. (2012). Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise. *Insights: the UKSG journal, 5*(1), 28-32.

Kucera, J., & Chlapek, D. (2014). Benefits and risks of open government data. *Journal of Systems Integration, 5*(1), 30-41.

Kuk, G., & Davies, T. (2011). *The roles of agency and artifacts in assembling open data complementarities*. Paper presented at the Thirty Second International Conference on Information Systems, Shanghai, China.

Kulk, S., & Van Loenen, B. (2012). Brave new open data world? *International Journal of Spatial Data Infrastructures Research, 7*, 196-206.

Lee, A. (1989). A scientific methodology for MIS case studies. *MIS Quarterly, 13*(1), 33-52.

Lee, G., & Kwak, Y. H. (2012). An open government maturity model for social media-based public engagement. *Government Information Quarterly, 29*(4), 492–503.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*(2), 133-146.

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline, 9*(1), 181-212.

Lewin, K. (1947). Frontiers in group dynamics. *Human relations, 1*(2), 143-153.

Lim, Y.-K., & Sato, K. (2003). *Scenarios for usability evaluation: using design information framework and a task analysis approach*. Paper presented at the 15th Technical Congress of the International Ergonomics Association Seoul, Korea.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalist inquiry*. Beverly Hills: Sage.

Linders, D. (2013). Towards open development: leveraging open data to improve the planning and coordination of international aid. *Government Information Quarterly, 30*(4), 426-434.

Lindman, J., Kinnari, T., & Rossi, M. (2014). *Industrial open data: case studies of early open data entrepreneurs*. Paper presented at the 47th Hawaii International Conference on System Sciences, Hawaii, U.S.A.

Liu, T., Bouali, F., & Venturini, G. (2014). EXOD: A tool for building and exploring a large graph of open datasets. *Computers & Graphics, 39*, 117-130.

Lu, Y., Xiang, C., Wang, B., & Wang, X. (2011). What affects information systems development team performance? An exploratory study from the perspective of combined socio-technical theory and coordination theory. *Computers in Human Behavior, 27*(2), 811–822.

Ma, L. (2012). Meanings of information: The assumptions and research consequences of three foundational LIS theories. *Journal of the American society for information science and technology, 63*(4), 716-723.

Macionis, J. J., & Plummer, K. (2005). *Sociology: a global introduction*. Essex: Pearson Education.

Magalhaes, G., Roseira, C., & Manley, L. (2014). *Business models for open government data*. Paper presented at the International Conference on Theory and Practice of Electronic Governance, Guimarães, Portugal.

Magalhaes, G., Roseira, C., & Strover, S. (2013). *Open government data intermediaries: a terminology framework*. Paper presented at the International Conference on Theory and Practice of Electronic Governance, Seoul, Republic of Korea.

Maier-Rabler, U., & Huber, S. (2011). „Open“: the changing relations between citizens, public ddministration and political authority. *eJournal of eDemocracy & Open Government, 3*(2), 48-58.

Malone, T. W., & Crowston, K. (1990). *What is Coordination Theory and how can it Help Design Cooperative Work Systems?* Paper presented at the 3rd Conference on Computer-supported Cooperative Work, Los Angeles, U.S.A.

Malone, T. W., & Crowston, K. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys, 26*(1), 87-119.

Malone, T. W., Crowston, K., & Herman, G. A. (2003). Chapter 1: tools for inventing organizations — toward a handbook of organizational processes. In T. W. Malone, K. Crowston & G. A. Herman (Eds.), *Organizing Business Knowledge: The MIT Process Handbook*. Massachusetts: Massachusetts Institute of Technology.

Malone, T. W., Crowston, K., Lee, J., Pentland, B., Dellarocas, C., Wyner, G., . . . O'Donnell, E. (1999). Tools for inventing organizations: toward a handbook of organizational processes. *Management Science, 45*(3), 425-443.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50–60.

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: John Wiley and Sons.

March, S. T., & Smith, G. (1995). Design and natural science research on information technologies. *Decision Support Systems, 15*(4), 251-266.

March, S. T., & Storey, V. C. (2008). Design science in the information systems discipline: an introduction to the special issue on design science research. *MIS Quarterly, 32*(4), 725-730.

Marienfeld, F., Schieferdecker, I., Tcholtchev, N., & Lapi, E. (2013). *Metadata aggregation at GovData.de – an experience report*. Paper presented at the International Symposium on Wikis and Open Collaboration, Hong Kong, China.

Markus, M. L., Majchrzak, A., & Gasser, L. (2002). A design theory for systems that support emergent knowledge processes. *MIS Quarterly, 26*(3), 179-212.

Martin, C. (2014). Barriers to the open government data agenda: taking a multi-level perspective. *Policy & Internet, 6*(3), 217-240.

Martin, M. P. (2003). Prototyping. *Encyclopedia of Information Systems, 3*, 565-573.

Martin, M. P., & Carey, J. M. (1991). Converting prototypes to operational systems: evidence from preliminary surveys. *Information and Software Technology, 33*(5), 351–356.

Martin, S., Foulonneau, M., & Turki, S. (2013). 1-5 Stars: metadata on the openness level of open data sets in Europe. In E. Garoufallou & J. Greenberg (Eds.), *Metadata and Semantics Research* (pp. 234-245). Thessaloniki, Greece: Springer International Publishing.

Mason, R. E. A., & Carey, T. T. (1983). Prototyping interactive information systems. *Communications of the ACM, 26*(5), 347-354.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., & Kleese, K. (2010). Using a core scientific metadata model in large-scale facilities. *The International Journal of Digital Curation, 1*(5), 106-118.

McGibbney, L. J., & Kumar, B. (2013). An intelligent authoring model for subsidiary legislation and regulatory instrument drafting within construction and engineering industry. *Automation in Construction, 35*, 121-130.

Meijer, A., & Thaens, M. (2009). Public information strategies: making government information available to citizens. *Information Polity, 14*(1-2), 31–45.

Meijer, A., & Thaens, M. (2013). Social media strategies: Understanding the differences between North American police departments. *Government Information Quarterly, 30*(4), 343-350.

Mergel, I. (2012). The social media innovation challenge in the public sector. *Information Polity, 17*(3), 281-292.

Ministry of the Interior and Kingdom Relations. (2008). *Onderzoeksovereenkomst ARVODI-2008 geconsolideerde versie Accessibility Monitor*.

Mintzberg, H. (1983). *Structure in fives. Designing effective organizations*. Englewood-Cliffs: Prentice-Hall.

Monteiro, E., Pollock, N., Hanseth, O., & Williams, R. (2012). From artefacts to infrastructures. *Computer Supported Cooperative Work, 22*(4-6), 575-607.

Mora Segura, A., Sanchez Cuadrado, J., & De Lara, J. (2014). *ODaaS: towards the model-driven engineering of open data applications as data services.* Paper presented at the 18th International Enterprise Distributed Object Computing Conference, Workshops and Demonstrations, Ulm, Germany.

Murphy, K. R., & Davidshofer, C. O. (1988). *Psychological testing: principles and applications*. Englewood Cliffs: Prentice-Hall.

Napoli, P. M., & Karaganis, J. (2010). On making public policy with publicly available data: The case of U.S. communications policy making. *Government Information Quarterly, 27*(4), 384-391.

National Information Standards Organization. (2004). *Understanding metadata*. Bethesda: National Information Standards Organization Press.

Naumann, F., & Rolker, C. (2000). Assessment methods for information quality criteria. *Information Quality*, 148-162.

Niehaves, B. (2007). *On epistemological diversity in design science: New vistas for a design-oriented IS research?* Paper presented at the International Conference on Information Systems, Montreal, Canada.

Northouse, P. G. (2010). *Leadership: theory and practice* (5th ed.). California: SAGE.

Novais, T., Albuquerque, J. P. d., & Craveiro, G. S. (2013). *An account of research on open government data (2007-2012): A systematic literature review*. Paper presented at the 12th Electronic Government and Electronic Participation Conference, Koblenz, Germany.

Nunnally, J. C. (1967). *Psychometric theory* (1st ed.). New York: McGraw-Hill.

O'Hara, K. (2012). *Data Quality, Government Data and the Open Data Infosphere*. Paper presented at the AISB/IACAP World Congress 2012: Information Quality Symposium, Birmingham, Great Britain.

Oates, B. J. (2006). *Researching Information Systems and computing*. Los Angeles: Sage.

Offermann, P., Blom, S., Schönherr, M., & Bub, U. (2010). Artifact types in information systems design science – a literature review. In R. Winter, J. L. Zhao & S. Aier (Eds.), *Global Perspectives on Design Science Research* (Vol. 6105, pp. 77-92): Springer Berlin Heidelberg.

Open Knowledge Foundation. (2007). CKAN. Retrieved December 2, 2011, from http://ckan.org/.

Open Knowledge Foundation. (2015). Open Definition version 2.0. Retrieved February 17, 2015, from http://opendefinition.org/od/.

Orlikowski, W. J. (2004). Managing and designing. Attending to reflexiveness and enactment. In R. B. Jr & F. Collopy (Eds.), *Managing as designing* (pp. 90-...?). Stanford: Stanford University Press.

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying information technology in organizations: research approaches and assumptions. *Information Systems Research, 2*(1), 1-28.

Orlikowski, W. J., & Iacono, C. S. (2001). Research commentary: desperately seeking the "IT" in IT research. A call to theorizing the IT artifact. *Information Systems Research, 12*(2), 121–134.

Overheid.nl. (2012). Besluit houdende instelling van een Sociaal en Cultureel Planbureau. 4760, from http://wetten.overheid.nl/BWBR0002873/geldigheidsdatum_31-03-2012#Opschrift.

Oviedo, E., Mazon, J. N., & Zubcoff, J. J. (2013). *Towards a Data Quality Model for Open Data Portals*. Paper presented at the XXXIX Latin American Computing Conference, Club Puerto Azul, Venezuela.

Parycek, P., & Sachs, M. (2010). Open government – information flow in web 2.0. *European Journal of ePractice, 9*, 1-12.

Pearl, J. (2001). *Direct and indirect effects*. Paper presented at the 17th Conference on Uncertainty in Artificial Intelligence, Seattle, U.S.A.

Peffers, K., Tunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Perry, M., Sheth, A. P., & Jain, P. (2011). SPARQL-ST: Extending SPARQL to support spatiotemporal queries. In N. Ashish & A. P. Sheth (Eds.), *Geospatial semantics and the semantic web: foundations, algorithms, and applications* (pp. 61-86). New York: Springer US.

Peter, J. P. (1979). Reliability: a review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 16*(1), 6-17.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden: Blackwell Publishing.

Petychakis, M., Vasileiou, O., Georgis, C., Mouzakitis, S., & Psarras, J. (2014). A state-of-the-art analysis of the current public data landscape from a functional, semantic and technical perspective. *Journal of Theoretical and Applied Electronic Commerce Research, 9*(2), 34-47.

Pliskin, N., & Shoval, P. (1987). End-user prototyping: sophisticated users supporting system development. *ACM SIGMIS Database, 18*(4), 7-17.

Potts, C. (1995). *Using schematic scenarios to understand user needs*. Paper presented at the 1st conference on Designing Interactive Systems: Processes, Practices, Methods, & Techniques, Ann Arbor, U.S.A.

Pries-Heje, J., & Baskerville, R. (2008). The design theory nexus. *MIS Quarterly, 32*(4), 731-755.

Puron-Cid, G., Gil-Garcia, J. R., & Luna-Reyes, L. F. (2012). *IT-enabled policy analysis: new technologies, sophisticated analysis and open data for better government decisions*. Paper presented at the 13th Annual International Conference on Digital Government Research, Maryland, U.S.A.

Qin, J., Ball, A., & Greenberg, J. (2012). *Functional and architectural requirements for metadata: supporting discovery and management of scientific data*. Paper presented at the International Conference on Dublin Core and Metadata Applications, Kuching, Malaysia.

Radin, B. (2006). *Challenging the performance movement: accountability, complexity, and democratic values*. Washington: Georgetown University Press.

Rahm, E., & Hai Do, H. (2000). Data cleaning: problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4), 3-13.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation. Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *331, 11*(Challenges and opportunities of open data in Ecology), 703-705

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM, 43*(12), 45-48.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56-58.

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *The Economics of the Internet and E-commerce, 11*(2), 23-25.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender Systems Handbook*: Springer U.S.

Richardson, G. L., Jackson, B. M., & Dickson, G. W. (1990). A principles-based enterprise architecture: lessons from Texaco and Star Enterprise. *MIS Quarterly, 14*(4), 385-403.

Riley, M. W. (1963). *Sociological research: I. A case approach*. New York: Harcourt, Brace & World.

Roethlisberger, F. J. (1977). *The elusive phenomena: an autobiographical account of my work in the field of organizational behavior at the Harvard Business School*. Boston, Massachusetts: Harvard Business School.

Salkind, N. J. (2010). *Encyclopedia of research design*. Thoasand Oaks: Sage Publications.

Sayogo, D. S., Pardo, T. A., & Cook, M. (2014). *A framework for benchmarking open government data efforts*. Paper presented at the 47th Hawaii International Conference on System Sciences, Hawaii, U.S.A.

Schuurman, N., Deshpande, A., & Allen, D. (2008). Data integration across borders: a case study of the abbotsford-sumas aquifer (British Columbia/Washington State). *JAWRA Journal of the American Water Resources Association, 44*(4), 921-934.

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action Design Research. *MIS Quarterly, 35*(1), 37-56.

Sen, A. (2004). Metadata management: past, present and future. *Decision Support Systems, 152*(37), 151-173.

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., & Schraefel, M. C. (2012). Linked open government data: lessons from data.gov.uk. *IEEE Intelligent Systems, 27*(3), 16-24.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Sheth, A. P. (1999). Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In M. Goodchild, M. J. Egenhofer, R. Fegeas & C. Kottman (Eds.), *Interoperating geographic information systems* (pp. 5-29). Norwell: Kluwer Academic Publishers.

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record, 34*(3), 31-36.

Simon, H. A. (1996). *The sciences of the artificial* (3rd edition ed.). Massachusetts Massachusetts Institute of Technology.

Smith, J. K. (1983). Quantitative versus qualitative research: an attempt to clarify the issue. *Educational Researcher, 12*(3), 6–13.

Sociaal en Cultureel Planbureau. (2013a). Data. Retrieved November 23, 2013, from http://www.scp.nl/Onderzoek/Tijdsbesteding/Achtergronden/Data.

Sociaal en Cultureel Planbureau. (2013b). Wat doet het SCP? Retrieved November 23, 2013, from http://www.scp.nl/Organisatie/Wat_is_het_SCP/Wat_doet_het_SCP.

Sociaal en Cultureel Planbureau. (2013c). Wat is het SCP? Retrieved November 23, 2013, from http://www.scp.nl/Organisatie/Wat_is_het_SCP.

Sommerville, I. (2011). *Software engineering* (9th ed.). Boston: Pearson Education.

Staatscourant. (2011). Organisatieregeling Ministerie van Veiligheid en Justitie, nr. 22848. Retrieved April 22, 2015, from https://zoek.officielebekendmakingen.nl/stcrt-2011-22848.html.

Statistical Data and Metadata eXchange. (2011). SDMX Standards: Section 1. Framework for SDMX Technical Standards Version 2.1. Retrieved January 28, 2015, from http://sdmx.org/wp-content/uploads/2011/04/SDMX_2-1_SECTION_1_Framework.pdf.

Stellman, A., & Greene, J. (2005). *Applied software project management*. Sebastopol: O'Reilly Media.

Stowers, G. (2013). The use of data visualization in government. Retrieved February 23, 2015, from http://www.businessofgovernment.org/sites/default/files/The%20Use%20of%20Visualization%20in%20Government.pdf.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103-110.

Susha, I., & Grönlund, Å. (2012). eParticipation research: systematizing the field. *Government Information Quarterly, 29*(3), 373-382.

The Data Seal of Approval Board. (2013). Implementation of the data seal of approval. Retrieved April 22, 2015, from https://assessment.datasealofapproval.org/assessment_47/seal/html/.

The European Parliament and the Council of the European Union. (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Retrieved July 15, 2015, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF.

Thompson, J. D. (1967). *Organizations in action. Social science bases of administrative theory*. New York: McGraw-Hill.

Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Research commentary-digital infrastructures: the missing IS research agenda. *Information Systems Research, 21*(4), 748-759.

Trauth, E. M., & Jessup, L. M. (2000). Understanding computer-mediated discussions: positivist and interpretive analyses of group support system use. *MIS Quarterly, 24*(1), 43-79.

Uhlir, P. F., & Schröder, P. (2007). Open data for global science. *Data Science Journal, 6*, 36-53.

Vaishnavi, V. K., & Kuechler Jr, W. (2008). *Design science research methods and patterns. Innovating information and communication technology*. Boca Raton: Auerbach Publications, Taylor & Francis Group.

Vaishnavi, V. K., & Kuechler, W. (2004). Design science research in information systems. Retrieved April 22, 2015, from http://desrist.org/desrist/.

Van den Brink, L., Janssen, P., Quak, W., & Stoter, J. E. (2014). Linking spatial data: automated conversion of geo-information models and GML data to RDF. *International Journal of Spatial Data Infrastructures Research, 9*, 59-85.

Van Loenen, B. (2006). *Developing geographic information infrastructures. The role of information policies*. Delft, the Netherlands: Delft University Press.

Van Loenen, B., & Grothe, M. (2014). INSPIRE empowers re-use of Public Sector Information. *International Journal of Spatial Data Infrastructures Research, 9*, 86-106.

Van de Ven, A. H., Delbecq, A. L., & Koenig, R. (1976). Determinants of coordination modes within organizations. *American Sociological Review, 41*, 322-328.

Van Voorhis, C. W., & Morgan, B. L. (2001). Statistical rules of thumb: what we don't want to forget about sample sizes. *Psi Chi Journal of undergraduate research, 6*(4), 139-141.

Vardaki, M., & Papageorgiou, H. (2007). *Statistical data and metadata quality assessment*. New York: IGI Global.

Vardaki, M., Papageorgiou, H., & Pentaris, F. (2009). A statistical metadata model for clinical trials' data management. *Computer methods and programs in biomedicine, 9*(5), 129-145.

Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: toward a standard for the social sciences. *International Journal of Digital Curation, 3*(1), 107-113.

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: an open data perspective. *Government Information Quarterly, 31*(2), 278–290.

Venable, J. R., Pries-Heje, J., Bunker, D., & Russo, N. L. (2010). *Creation, transfer and diffusion of innovation in organizations and society: information*

systems design science research for human benefit. Paper presented at the International Working Conference, Perth, Australia.

Venkatesh, V., Thong, J. Y. L., Chan, F. K. Y., Hu, P. J.-H., & Brown, S. A. (2011). Extending the two-stage information systems continuance model: incorporating UTAUT predictors and the role of context. *Information Systems Journal, 21*(6), 527–555.

Verschuren, P., & Hartog, R. (2005). Evaluation in design-oriented research. *Quality & Quantity, 39*(6), 733-762.

Visitatiecommissie WODC. (2014). Het WODC als baken in feiten-gedreven Justitieel en Veiligheidsbeleid. Naar een nieuwe balans tussen wetenschap, beleid en uitvoering voor de toekomst? Retrieved April 22, 2015, from file:///H:/My%20Documents/Case%20studies/WODC/Documents%20studied/visitatie-wodc-2014_tcm44-547410.pdf.

Volpi, V., Ingrosso, A., Pazzola, M., Opromolla, A., & Medaglia, C. (2014). Roma crash map: an open data visualization tool for the municipalities of Rome. In S. Yamamoto (Ed.), *Human Interface and the Management of Information. Information and Knowledge in Applications and Services* (pp. 284-295): Springer International Publishing.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research, 3*(1), 36-59.

Walsham, G. (1995). Interpretive case studies in IS research: nature and method. *European Journal of Information Systems, 4*(2), 74-81.

Walsham, G. (2001). The emergence of interpretivism in IS research. *Information Systems Research, 6*(4), 376-394.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1973). *Unobtrusive measures. Nonreactive research in the social sciences* Chicago: Rand McNally & Company.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly, 26*(2), xiii-xxiii.

Wetenschappelijk Onderzoek en Documentatie Centrum. (2013). Zelfevaluatie Wetenschappelijk Onderzoek- en Documentatiecentrum. Periode 2006 tot en met 2012. Retrieved April 22, 2015, from file:///H:/My%20Documents/Case%20studies/WODC/Documents%20studied/zelfevaluatie-wodc-2006-2012_tcm44-550429.pdf.

Wetenschappelijk Onderzoek en Documentatie Centrum. (2014). Organisatie. Retrieved November 27, 2014, from http://www.wodc.nl/organisatie/.

Whitmore, A. (2014). Using open government data to predict war: a case study of data and systems challenges. *Government Information Quarterly, 31*(4), 622–630.

Wittenberg, M. (2009). Een Verjaardagsfeestje met Vergezichten. Retrieved November 16, 2012, from http://dans.knaw.nl/content/categorieen/symposia/symposia-archief/een-verjaardagsfeestje-met-vergezichten.

World Wide Web Consortium. (2014). Data Catalog Vocabulary (DCAT). Retrieved February 2, 2015, from http://www.w3.org/TR/vocab-dcat/.

Yannoukakou, A., & Araka, I. (2014). Access to government information: right to information and open government data synergy. *Procedia - Social and Behavioral Sciences, 147*, 332-340.

Yildiz, M. (2007). E-government research: reviewing the literature, limitations, and ways forward. *Government Information Quarterly, 24*(3), 646-665.

Yin, R. K. (2003). *Case study research. Design and methods*. Thoasand Oaks: SAGE publications.

Yu, H., & Robinson, D. G. (2012). The new ambiguity of "Open Government". *UCLA Law Review Discourse, 59*, 178-208.

Zhang, J., Dawes, S., & Sarkis, J. (2005). Exploring stakeholders' expectations of the benefits and barriers of e-government knowledge sharing. *Journal of Enterprise Information Management, 18*(5), 548-567.

Zhu, X., & Gauch, S. (2000). *Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web*. Paper presented at the 23rd Annual International ACM SIGIR conference on research and development in information retrieval, Athens, Greece.

Zuiderwijk, A., Gascó, M., Parycek, P., & Janssen, M. (2014). Special issue on transparency and open data policies: guest editors' introduction. *Journal of Theoretical and Applied Electronic Commerce Research, 9*(3), I-IX.

Zuiderwijk, A., Helbig, N., Gil-García, J. R., & Janssen, M. (2014). Guest Editors' Introduction. Innovation Through Open Data: A Review of the State-of-the-Art and an Emerging Research Agenda. *Journal of Theoretical and Applied Electronic Commerce Research, 9*(2), I-XIII.

Zuiderwijk, A., & Janssen, M. (2013a). *A coordination theory perspective to improve the use of open data in policy-making*. Paper presented at the 12th Conference on Electronic Government, Koblenz, Germany.

Zuiderwijk, A., & Janssen, M. (2013b). *Re-engineering the open data publishing process at a Dutch government organization (research in progress)*. Paper presented at the 10th Scandinavian Workshop on E-Government, Oslo, Norway.

Zuiderwijk, A., & Janssen, M. (2014a). Barriers and development directions for the publication and usage of open data: a socio-technical view. In M. Gascó-Hernández (Ed.), *Open Government. Opportunities and Challenges for Public Governance* (pp. 115-135). New York: Springer.

Zuiderwijk, A., & Janssen, M. (2014b). *The negative effects of open government data - investigating the dark side of open data*. Paper presented at the Proceedings of the 15th Annual International Conference on Digital Government Research, Aguascalientes, Mexico.

Zuiderwijk, A., & Janssen, M. (2015). *Participation and data quality in open data use: open data infrastructures evaluated*. Paper presented at the 15th European Conference on eGovernment, Portsmouth, United Kingdom.

Zuiderwijk, A., Janssen, M., Choenni, S., & Meijer, R. (2014). Design principles for improving the process of publishing open data *Transforming Government: People, Process and Policy, 8*(2), 185 - 204.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Sheikh Alibaks, R. (2012). Socio-technical impediments of open data. *Electronic Journal of eGovernment, 10*(2), 156 - 172.

Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: essential elements of open data ecosystems *Information Polity, 19*(1-2), 17–33.

Zuiderwijk, A., Janssen, M., & Jeffery, K. (2013). *Towards an e-infrastructure to support the provision and use of open data*. Paper presented at the Conference for e-Democracy and Open Government, Krems an der Donau, Austria.

Zuiderwijk, A., Janssen, M., Meijer, R., Choenni, S., Charalabidis, Y., & Jeffery, K. (2012). *Issues and guiding principles for opening governmental judicial research data*. Paper presented at the 11th Conference on Electronic Government, Kristiansand, Norway.

Zuiderwijk, A., Janssen, M., & Parnia, A. (2013). *The complementarity of open data infrastructures: an analysis of functionalities*. Paper presented at the 14th Annual International Conference on Digital Government Research, Quebec, Canada.

Zuiderwijk, A., Janssen, M., Poulis, K., & Vandekaa, G. (2015). *Open data for competitive advantage: insights from open data use by companies*. Paper presented at the 16th Annual International Conference on Digital Government Research, Phoenix, Arizona, U.S.A.

Zuiderwijk, A., Janssen, M., & Susha, I. (forthcoming). Improving the speed and ease of open data use through metadata, participation mechanisms and quality indicators. *Journal of Organizational Computing and Electronic Commerce*.

Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012a). *The necessity of metadata for open linked data and its contribution to policy analyses.* Paper presented at the Conference on E-Democracy and Open Government, Krems, Austria.

Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012b). The potential of metadata for linked open data and its value for users and publishers. *Journal of e-Democracy and Open Government, 4*(2), 222-244.

Zuiderwijk, A., Klievink, B., Janssen, M., Tan, Y. H., Charalabidis, Y., & Argyzoudis, E. (2013). *Workshop on open information infrastructures enabling innovations*. Paper presented at the 12th IFIP Electronic Government Conference, Koblenz, Germany.

Summary

Problem statement

The focus of this research is on the operational use of structured research Open Government Data (OGD) from the domains of social sciences and humanities by researchers outside the government through OGD infrastructures. Governmental organisations create, collect and pay for large amounts of data. These data are increasingly pro-actively published on the internet. The disclosed governmental data are referred to as Open Government Data (OGD). OGD have many potential benefits that are often not realised. Governments and scholars traditionally focus on the publication of OGD, whereas the actual use of the data resulting in benefits is often neglected. Moreover, OGD use activities are often not coordinated, and tools for using OGD are fragmented. The objective of this study is *to develop an infrastructure that enhances the coordination of OGD use*. An OGD infrastructure can be defined as *a shared, (quasi-)public, evolving system, consisting of a collection of interconnected social and technical elements*. Outside the scope of this study are the OGD providers and the policy makers, and a premise is that improved OGD use will support governmental policy making.

Research methods and contributions

A design science research approach is used to attain the research objective. This approach is relevant, since we aimed to develop an artefact (i.e. an OGD infrastructure) that did not yet exist, and since we aimed to contribute to scientific and practical developments for the design of OGD infrastructures. The following five research questions were answered in our study.

1. *Which factors influence OGD use?* Many factors affect OGD use, however, there was no overview of these factors. A literature review was carried out to answer the first research question. We sought for influencing factors regarding five types of OGD use, namely searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. Various factors were identified and they were integrated in fourteen clusters. This study is among the first to provide a comprehensive overview of the factors, including

the barriers, that need to be taken into account when one wants to improve OGD use.

2. *What are the functional requirements for an infrastructure that enhances the coordination of OGD use?* We sought for functional requirements in each of the fourteen clusters that were identified through the first research question. Functional requirements were derived from two explorative case studies concerning open judicial data use and open social data use. Twenty-eight functional requirements were identified in the cases and used for the design of the OGD infrastructure. We contributed to the existing literature by offering a comprehensive overview of functional user requirements for improving OGD use based on practical case studies.
3. *Which functional elements make up an infrastructure that enhances the coordination of OGD use?* From an enhanced literature review we identified three functional elements that potentially meet all the functional requirements from the case studies, namely metadata, interaction mechanisms and data quality indicators. Based on the assumption that metadata, interaction mechanisms and data quality indicators can enhance the coordination of OGD use, three design propositions were developed:
    - Proposition 1: metadata positively influence the ease and speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis;
    - Proposition 2: interaction mechanisms positively influence the ease and speed of interaction about OGD;
    - Proposition 3: data quality indicators positively influence the ease and speed of OGD quality analysis.

    Building on the design propositions, we developed design principles, which provided more specific directions for the design of the OGD infrastructure. Kernel theories about coordination, metadata, interaction and data quality aided the development of the design principles. Building on the design principles, the design of the OGD infrastructure was described. The OGD infrastructure design incorporated the system design, the coordination patterns and the function design. This study is among the first to describe the design of an OGD infrastructure, including the functional elements it encompasses.

4. *What does the developed OGD infrastructure look like?* A working version of the designed OGD infrastructure, a prototype, was created. The prototype allowed for refining and detailing the functional user requirements, as well as for measuring the effects of the designed OGD infrastructure. The prototype was tested and was improved through many iterative phases. This research contributes to the literature and to practice by showing what the designed OGD infrastructure including its three functional elements looks like and how it can be developed.
5. *What are the effects of the developed infrastructure on the coordination of OGD use?* Three quasi-experiments were conducted which incorporated measurements through surveys, observations and time measures. Students and professional open data users completed OGD use scenarios with the developed OGD infrastructure prototype and evaluated it. This study contributed to the literature and to practice by providing insight in the strengths and weaknesses of the developed OGD infrastructure. The insights obtained through the evaluations can be used by practitioners to further enhance the coordination of OGD use through infrastructures. Moreover, policy makers may use the infrastructure to derive useful information from OGD use, and to consider this in the development of governmental policy making.

Design of the OGD infrastructure

This study revealed that combining metadata, interaction mechanisms and data quality indicators in one OGD infrastructure is an essential condition for managing OGD use dependencies. Whereas kernel theories about coordination, metadata, interaction and data quality are often studied separately, this research showed that these four kernel theories need to be combined and integrated for the development of OGD infrastructures. This study builds on existing metadata studies and confirms that different types of metadata (discovery, contextual and detailed metadata) need to be integrated to enhance the coordination of OGD use by researchers. This study builds on the coordination literature and shows that the coordination of OGD use does not merely require a focus on processes, but additionally requires a technical perspective including the integration of tools, a social perspective including interaction between researchers, OGD providers and

policy makers, and the interaction between the social and technical perspective. Moreover, this study contributes to practice by providing functional infrastructure elements that can be implemented in the design of existing or new OGD infrastructures. The coordination patterns explain how the functional elements can be used in and applied to practice, and the function design shows which functions can be implemented by developers of existing and future OGD infrastructures.

Evaluation of the OGD infrastructure

Scenarios, surveys, observations and time measures were used to evaluate the ease and speed of OGD use (as indicators of coordination) on the developed infrastructure. The evaluation participants completed scenarios that prescribed them to use various tools, to interact with other OGD users and to use tools that allowed for interaction with OGD providers and policy makers. This means that they used OGD in a way that corresponds to our definition of coordination.

The surveys and the observations showed that on average the students and the professional open data users of the treatment group found it significantly easier to conduct scenario tasks related to searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis than the students of the control group. Moreover, the treatment groups needed significantly less time to complete the scenario tasks than the control group. The evaluations indicated that the developed OGD infrastructure enhanced the coordination of OGD use.Through the evaluations we also identified various areas for the improvement of the OGD infrastructure and for the quasi-experimental design.

Conclusions and implications

The objective of this study is *to develop an infrastructure that enhances the coordination of OGD use*. Whereas governments and scholars traditionally focus on the supply of OGD, this study described a demand-driven OGD infrastructure that addressed the needs of OGD users. This study focused on how OGD use by researchers can be improved by enhancing coordination. A premise of this study is that improved OGD use will support governmental policy making.

This dissertation contributes both to science and to practice. The scientific contributions of this thesis lay in the identification of factors and barriers influencing OGD use, the identification of functional requirements for an infrastructure that enhances the coordination of OGD use, the definition of a combination of functional elements that make up an infrastructure that enhances the coordination of OGD use, the description of how the OGD infrastructure can be developed, and the overview of the positive and negative effects of the developed infrastructure on the coordination of OGD use. This dissertation indicates that the key functional elements of the infrastructure, i.e. metadata, interaction mechanisms and data quality indicators, together enhance the coordination of OGD use. These functional infrastructure elements together improve searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis. Practitioners can use the described infrastructure design to enhance the coordination of OGD use of existing and future OGD infrastructures.

Research limitations of this study concern taking an interpretivistic and an open data proponent perspective (which might be biased), not considering non-functional requirements for the OGD infrastructure, the generalisation of the findings from the selected cases, the evaluation of a prototype instead of a completely designed OGD infrastructure, and the generalisation of the findings from the quasi-experiments. Five recommendations for an emerging open data research agenda were identified, namely: 1) balance the benefits and the risks of OGD publication, 2) examine the evolving design of the OGD infrastructure, 3) study to which extent the identified infrastructure requirements also apply to other types of data, other types of data use, on other governmental levels and in other countries and cultures, 4) further improve OGD use utilising the developed OGD infrastructure, and 5) evaluate the OGD infrastructure.

Samenvatting (summary in Dutch)

Probleemstelling

Dit onderzoek richt zich op het operationele gebruik van gestructureerde onderzoeks Open Overheidsdata (OOD) in de domeinen van de sociale wetenschappen en geesteswetenschappen door onderzoekers buiten de overheid door middel van OOD-infrastructuren. Overheidsorganisaties creëren, verzamelen en betalen voor grote hoeveelheden data. Deze data worden in toenemende mate proactief op internet beschikbaar gesteld. Naar de ontsloten overheidsdata wordt verwezen met de term Open Overheidsdata (OOD). OOD hebben vele potentiele voordelen, die vaak niet gerealiseerd worden. Overheden en wetenschappelijk onderzoekers richten zich van oudsher op de publicatie van OOD, terwijl vaak geen acht wordt geslagen op het gebruik van de data dat in de voordelen moet resulteren. Daarnaast zijn OOD-gebruiksactiviteiten vaak nauwelijks gecoördineerd, en zijn hulpmiddelen voor het gebruik van OOD gefragmenteerd. Het doel van dit onderzoek is *om een infrastructuur te ontwikkelen die de coördinatie van het gebruik van OOD vergroot.* Een OOD-infrastructuur kan worden gedefinieerd als *een gedeeld (quasi-)publiek, zich ontwikkelend systeem, bestaande uit een verzameling van onderling verbonden sociale en technische elementen*. De OOD-aanbieders en beleidsmakers vallen buiten de reikwijdte van deze studie, en het is een aanname dat verbeterd OOD-gebruik zal leiden tot betere beleidsontwikkeling door overheidsorganisaties.

Onderzoeksmethoden en contributies

Een *design science* benadering is gebruikt om het onderzoeksdoel te realiseren. Deze benadering is relevant, omdat ons doel is om een artefact (d.w.z. een OOD-infrastructuur) te ontwikkelen dat nog niet bestaat, en omdat we willen bijdragen aan wetenschappelijke en praktische ontwikkelingen voor het ontwerp van OOD-infrastructuren. De volgende vijf onderzoeksvragen zijn in dit onderzoek beantwoord.

1. *Welke factoren beïnvloeden OOD-gebruik?* Vele factoren beïnvloeden OOD-gebruik, echter, er was geen overzicht van deze factoren beschikbaar. Een

literatuuronderzoek is uitgevoerd om de eerste onderzoeksvraag te beantwoorden. We zochten naar beïnvloedende factoren voor elk van de vijf typen OOD-gebruik, namelijk zoeken naar en vinden van OOD, OOD-analyse, OOD-visualisatie, interactie over OOD en OOD-kwaliteitsanalyse. Verschillende factoren werden geïdentificeerd en geïntegreerd in veertien clusters. Dit onderzoek behoort tot de eerste onderzoeken die een uitgebreid overzicht geven van de factoren, inclusief de barrières, die in aanmerking moeten worden genomen wanneer men het gebruik van OOD wil verbeteren.

2. *Wat zijn de functionele eisen voor een infrastructuur die de coördinatie van het gebruik van OOD vergroot?* We zochten naar functionele eisen in elk van de veertien clusters die met de eerste onderzoeksvraag werden geïdentificeerd. Functionele eisen werden afgeleid uit twee exploratieve casus analyses betreffende het gebruik van gerechtelijke en sociale data. 28 functionele eisen werden geïdentificeerd in de cases en gebruikt voor het ontwerp van de OOD-infrastructuur. We hebben bijgedragen aan de bestaande literatuur door een uitgebreid overzicht te bieden van gebruikerseisen voor het verbeteren van OOD-gebruik gebaseerd op praktische casusonderzoeken.
3. *Welke functionele elementen vormen een infrastructuur die de coördinatie van het gebruik van OOD vergroot?* Met een tweede literatuuronderzoek hebben we drie functionele elementen geïdentificeerd die potentieel voldoen aan alle functionele eisen van de casusonderzoeken, namelijk metadata, interactiemechanismen en data kwaliteitsindicatoren. Op basis van de assumptie dat metadata, interactiemechanismen en data kwaliteitsindicatoren de coordinatie van OOD-gebruik kunnen vergroten hebben we de volgende ontwerpproposities opgesteld:
    - Propositie 1: metadata hebben een positieve invloed op het gemak en de snelheid van het zoeken naar en vinden van OOD, OOD-analyse, OOD-visualisatie, interactie over OOD en OOD-kwaliteitsanalyse;
    - Propositie 2: interactiemechanismen hebben een positieve invloed op het gemak en de snelheid van interactie over OOD;
    - Propositie 3: data kwaliteitsindicatoren hebben een positieve invloed op het gemak en de snelheid van OOD-kwaliteitsanalyse.

Voortbouwend op de ontwerpproposities hebben we ontwerpprincipes ontwikkeld, welke meer specifiek richting geven aan het ontwerp van de OOD-infrastructuur. Kerntheorieën over coördinatie, metadata, interactie en datakwaliteit hielpen bij het ontwikkelen van de ontwerpprincipes. Voortbordurend op de ontwerpprincipes is het ontwerp van de OOD-infrastructuur beschreven. Het OOD-infrastructuurontwerp omvatte het systeemontwerp, de coördinatiepatronen en het functioneel ontwerp. Dit onderzoek is een van de eerste onderzoeken die het ontwerp van een OOD-infrastructuur beschrijft, inclusief de functionele elementen die het omvat.

4. *Hoe ziet de ontwikkelde OOD-infrastructuur eruit?* Een werkende versie van de ontwikkelde OOD-infrastructuur, een prototype, is gecreëerd. Het prototype maakte het mogelijk om de functionele gebruikerseisen verder te verfijnen en te detailleren, en om de effecten van de ontwikkelde OOD-infrastructuur te meten. Het prototype is getest en verbeterd door middel van vele iteraties. Dit onderzoek draagt bij aan de literatuur en aan de praktijk door te laten zien hoe de ontworpen OOD-infrastructuur inclusief zijn drie functionele kernelementen eruit ziet en ontwikkeld kan worden.
5. *Wat zijn de effecten van de ontwikkelde infrastructuur op de coördinatie van OOD-gebruik?* Drie quasi-experimenten zijn uitgevoerd welke bestonden uit metingen door middel van enquêtes, observaties en tijdsmetingen. Studenten en professionele OOD-gebruikers hebben OOD-gebruiksscenario's uitgevoerd met het prototype van de ontwikkelde OOD-infrastructuur, en hebben het prototype geëvalueerd. Dit onderzoek draagt bij aan de literatuur en aan de praktijk door inzicht te geven in de sterke en de zwakke kanten van de ontwikkelde OOD-infrastructuur. De inzichten verkregen met de evaluaties kunnen in de praktijk gebruikt worden om de coördinatie van OOD-gebruik door middel van infrastructuren verder te verbeteren. Daarnaast kunnen beleidsmakers de infrastructuur gebruiken om nuttige informatie af te leiden uit OOD-gebruik en om bij te dragen aan het ontwikkelen van overheidsbeleid.

Ontwerp van de OOD-infrastructuur

Dit onderzoek liet zien dat het combineren van metadata, interactiemechanismen en data kwaliteitsindicatoren in één infrastructuur een essentiële voorwaarde is voor het beheersen van de afhankelijkheden van OOD-gebruik. Waar kerntheorieën over coördinatie, metadata, interactie en datakwaliteit vaak apart van elkaar bestudeerd werden heeft dit onderzoek laten zien dat deze vier kerntheorieën gecombineerd en geïntegreerd moeten worden voor de ontwikkeling van OOD-infrastructuren. Dit onderzoek bouwt voort op bestaand metadata-onderzoek en bevestigt dat verschillende typen metadata (ontdekkings-, contextuele en gedetailleerde metadata) geïntegreerd moeten worden om de coördinatie van het gebruik van OOD door onderzoekers te verbeteren. Dit onderzoek bouwt voort op de coördinatieliteratuur en laat zien dat de coördinatie van OOD-gebruik niet alleen een focus op processen vereist, maar dat daarnaast een technisch perspectief inclusief de integratie van hulpmiddelen, een sociaal perspectief inclusief de interactie tussen onderzoekers, OOD-aanbieders en beleidsmakers, en de interactie tussen het technische en sociale perspectief nodig zijn. Bovendien draagt dit proefschrift bij aan de praktijk door functionele infrastructuurelementen aan te bieden die geïmplementeerd kunnen worden in bestaande of toekomstige OOD-infrastructuren. De coördinatiepatronen laten zien hoe de functionele elementen gebruikt en toegepast kunnen worden in de praktijk, en het functionele ontwerp laat zien welke functies ontwikkelaars van OOD-infrastructuren kunnen implementeren.

Evaluatie van de OOD-infrastructuur

Scenario's, enquêtes, observaties en tijdsmetingen zijn gebruikt om het gemak en de snelheid van OOD-gebruik (als indicatoren voor coördinatie) op de ontwikkelde infrastructuur te evalueren. De evaluatiedeelnemers voltooiden scenarios die hen voorschreven om verschillende hulpmiddelen te gebruiken, om te interacteren met andere OOD-gebruikers en om hulpmiddelen te gebruiken die het mogelijk maakten om te interacteren met OOD-aanbieders en met beleidsmakers. Dit betekent dat zij OOD gebruikten op een manier die overeenkomt met onze definitie van coördinatie.

De enquêtes en de observaties lieten zien dat de studenten en de professionele open data gebruikers uit de behandelgroep het gemiddeld significant gemakkelijker vonden om de scenariotaken uit te voeren met betrekking tot het zoeken naar en vinden van OOD, OOD-analyse, OOD-visualisatie, interactie over OOD en OOD-kwaliteitsanalyse dan de studenten in de controlegroep. Daarnaast hadden de behandelgroepen significant minder tijd nodig om de scenariotaken uit te voeren dan de controlegroep. De evaluaties indiceerden dat de ontwikkelde OOD-infrastructuur de coördinatie van OOD-gebruik vergrootte. De evaluaties lieten verschillende gebieden voor verbetering van de OOD-infrastructuur zien en voor de quasi-experimentele opzet.

Conclusies en implicaties

Het doel van dit onderzoek is *om een infrastructuur te ontwikkelen die de coördinatie van het gebruik van OOD verbetert.* Hoewel overheden en wetenschappelijk onderzoekers zich vanouds richtten op het aanbod van OOD heeft dit onderzoek een vraaggedreven OOD-infrastructuur beschreven die de behoeften van OOD-gebruikers adresseert. Dit onderzoek richtte zich op hoe het gebruik van OOD door onderzoekers verbeterd kan worden door het vergroten van coördinatie van OOD-gebruik. Een vooronderstelling van deze studie is dat verbeterd OOD-gebruik het proces waarin overheidsbeleid wordt ontwikkeld kan ondersteunen.

Dit proefschrift biedt zowel wetenschappelijke als praktische bijdragen. De wetenschappelijke bijdragen betreffen de identificatie van factoren en barrières die OOD-gebruik beïnvloeden, de identificatie van functionele eisen voor een infrastructuur die de coördinatie van OOD-gebruik verbeteren, de beschrijving van hoe de OOD-infrastructuur ontwikkeld kan worden, en een overzicht van de positieve en negatieve effecten van de ontwikkelde infrastructuur op de coördinatie van OOD-gebruik. Deze dissertatie heeft laten zien dat de functionele elementen van de infrastructuur, dat wil zeggen metadata, interactiemechanismen en data kwaliteitsindicatoren, gezamenlijk de coördinatie van OOD-gebruik kunnen ondersteunen. Deze infrastructuurelementen hebben samen tot doel om vijf typen OOD-gebruik te verbeteren, namelijk het zoeken naar en vinden van OOD, OOD-analyse, OOD-visualisatie, interactie over OOD en OOD-kwaliteitsanalyse. De

bevindingen van dit onderzoek kunnen in de praktijk worden gebruikt door de coördinatie van OOD-gebruik van bestaande en toekomstige OOD-infrastructuren te verbeteren.

Beperkingen van dit onderzoek betreffen het nemen van een interpretivistisch perspectief van een open data voorstander (welke vooringenomen kan zijn), het niet in aanmerking nemen van niet-functionele eisen voor de OOD-infrastructuur, de generalisatie van de bevindingen van de geselecteerde cases, de evaluatie van een prototype in plaats van een compleet ontwikkelde OOD-infrastructuur, en de beperkingen betreffende de generalisatie van de bevindingen van de quasi-experimenten. De volgende vijf aanbevelingen zijn geformuleerd voor een zich ontwikkelende open data onderzoeksagenda: 1) balanceer de voordelen en de risico's van OOD-publicatie, 2) onderzoek het zich ontwikkelende ontwerp van de OOD-infrastructuur, 3) bestudeer in welke mate de geïdentificeerde eisen voor de infrastructuur ook van toepassing zijn voor andere typen data, op ander type OOD-gebruik, op andere overheidsniveaus en in andere landen en culturen, 4) verbeter OOD-gebruik door verder gebruik van de OOD-infrastructuur, en 5) evalueer de OOD-infrastructuur.

Appendices

Appendix A: Factors influencing OGD use derived from the literature

	Clusters	Factors which influence OGD use	Source
Searching for and finding OGD	Data fragmentation	Locating existing OGD is complex and accompanied with high costs	Ding et al. (2012)
		It is not clear how open data can be found after they have been disclosed	Cowan and McGarry (2014)
		Datasets are fragmented as they are offered on many different open data infrastructures	Conradie and Choenni (2014), De Vocht et al. (2014)
		Data are offered at many different infrastructures, and can sometimes be hard to find	Braunschweig et al. (2012a), Conradie and Choenni (2014)
	Terminology heterogeneity	Considerable different terminology is used to describe datasets; there is a lack of common data definitions	Zhang et al. (2005), Yannoukakou and Araka (2014)
		OGD use different vocabularies	Zhang et al. (2005), Yannoukakou and Araka (2014)
		Each discipline has its own terminologies which leads to heterogeneity	(Reichman et al., 2011)
	Data search support	Many open data infrastructures provide only limited search options (e.g. no advanced search)	Petychakis et al. (2014)
		Often there is no support for searching for OGD in multiple languages	Petychakis et al. (2014)
		A lack of guidance in data service discovery leads to higher thresholds for using the service	Chen and Paik (2013)
	Information overload	The availability of an overwhelming amount of information may result in an information overload	Ho and Tang (2001)
		The number of open governmental datasets provided to the public is increasing	Kulk and Van Loenen (2012), Magalhaes et al. (2014)

Table 3-8: Overview of factors which influence OGD use, derived from the literature.

	Clusters	Factors which influence OGD use	Source
Searching for and finding OGD	Information overload	The number of open datasets is almost unlimited, while people have a limited ability to curate, search, analyse and visualise data	Cowan and McGarry (2014)
		The more open datasets are available, the more difficult it is to make effective use of them	Magalhaes et al. (2013)
OGD analysis	Data context	Traditional open data infrastructures provide data without any contextual information	Alexopoulos, Spiliotopoulou, et al. (2013)
		The ease of finding and understanding a dataset depends on the availability of data about the dataset and the contextual information	Dawes and Helbig (2010)
		Extensive knowledge of the data context is beyond the reach of a large part of the population	Foulonneau et al. (2014)
		Data about the data' (i.e. metadata) is important for the reuse of open data	Braunschweig et al. (2012b)
	Data interpretation support	The fear of drawing false conclusions from open data use is commonly heard	Conradie and Choenni (2014)
		When information collected for one purpose is used for a different purpose, there is potential for misuse, misunderstanding, and misinterpretation	Dawes et al. (2004), Conradie and Choenni (2014)
		Open data might be misinterpreted (either intentionally or unintentionally)	Kucera and Chlapek (2014)
		Data may be completely inappropriate for some uses (e.g. uses that have different temporal, security, or granularity requirements)	Dawes (2010)
		Knowledge about how to use OGD may be limited to a small community	Martin (2014)
		Lack of insight in information that can be easily understood by the general public	Novais et al. (2013)
		It is of critical importance that open data are available in a user-oriented way while this can be problematic for certain open data domains (focused on subsidiary legislation)	McGibbney and Kumar (2013)
	Data heterogeneity	Datasets are released in numerous different formats	Jeffery et al. (2014), Yannoukakou and Araka (2014)
		The semantics of the data may be ambiguous	Conradie and Choenni (2014)
		Heterogeneous data formats are used for the disclosure of open data	Mora Segura et al. (2014)

Table 3-8 (continued): Overview of factors which influence OGD use, derived from the literature.

	Clusters	Factors which influence OGD use	Source
OGD analysis	Data heterogeneity	The use of open data through applications requires the interpretation and combination of heterogeneous data from a variety of sources.	Mora Segura et al. (2014)
	Data analysis support	Analysing data sources requires using different tools	Braunschweig et al. (2012a)
		There is a lack of tools to generate information that can be easily understood by the general public	Novais et al. (2013)
		Services for processing, analysing and integrating open data are needed	Immonen, Palviainen, and Ovaska (2014)
		Most traditional open data infrastructures provide only basic data download and upload functionalities	Alexopoulos, Spiliotopoulou, et al. (2013), Charalabidis et al. (2014)
OGD visualisation	Data visualisation support	Visualisation tools are useful for open data use	De Vocht et al. (2014)
		Visualisation tools are needed for open data use	Shadbolt et al. (2012)
		In general, data visualisations can be used to make information more visible, to tell stories and to simplify, clarify and analyse data	De Vocht et al. (2014), Stowers (2013)
		Maps help in making sense of data	O'Hara (2012), Alani et al. (2008), Dimou et al. (2014)
		OGD sites barely provide visual or interactive possibilities to their users	Liu et al. (2014)
		Only a part of the existing OGD portals provides visualisation features	Sayogo et al. (2014)
Interaction about OGD	Lack of interaction	Better access to information is not enough for active participation	Alani et al. (2008)
		Conversations about released data are lacking	Lee and Kwak (2012)
		Many OGD providers do not know who their external users are	Archer et al. (2013)
		Further work on relationships between open government data supply and use would be valuable	Davies (2010)
	Interaction support	Open data can be used for collaboration amongst citizens and between citizens and the state	Parycek and Sachs (2010)
		The type of participatory tool might affect the extent to which users can participate and collaborate	Sayogo et al. (2014)
		The delivery of open data is characterized by a lack of opportunity for public participation	Lee and Kwak (2012), Whitmore (2014)

Table 3-8 (continued): Overview of factors which influence OGD use, derived from the literature.

	Clusters	Factors which influence OGD use	Source
Interaction about OGD	Interaction support	Participation can be stimulated through social media and interactive communications (blogging, micro blogging, tagging, photo and video sharing)	Veljković et al. (2014)
		Social media can be used to stimulate interaction	Mora Segura et al. (2014)
		Social media technologies allow for access to and interaction with government operations, programs and data	Bertot et al. (2012)
		Existing social media can be used to engage people in open data	Garbett et al. (2011)
		Feedback can be used to correct, enhance and improve data sources	Dawes and Helbig (2010)
		If feedback mechanisms are provided, feedback is typically through informal communications as part of institutional collaborations, comments on blogs and replies to Tweets	Archer et al. (2013)
		Most governmental agencies do not offer feedback mechanisms for open data	Archer et al. (2013), Alexopoulos, Spiliotopoulou, et al. (2013)
		Feedback may be used for updating resources	Lee and Kwak (2012)
		Most traditional open data infrastructures do not allow for improving published data (e.g. through cleaning and processing)	Alexopoulos, Spiliotopoulou, et al. (2013)
		In a case on open parcel data it was found that data almost always flows from the data source to the data requester, and barely the other way around. There is a lack of feedback between data users and data providers	Dawes and Helbig (2010)
		Successful open data adoption should integrate 1) the usefulness of data, 2) IT use and 3) Web 2.0 applications governed by a policy analysis model	Puron-Cid et al. (2012)
OGD quality analysis	Dependence on the quality of OGD	Open data success depends strongly on the quality of released datasets	Behkamal et al. (2014)
		Data quality plays an essential role in the use of government portals	Detlor et al. (2013)
		OGD reuse requires that potential data users can trust that datasets which they want to use are of sufficient quality	O'Hara (2012)
		To be able to assess the quality of datasets in general, data users need to have information about the nature of the data	(Dawes & Helbig, 2010)

Table 3-8 (continued): Overview of factors which influence OGD use, derived from the literature.

	Clusters	Factors which influence OGD use	Source
OGD quality analysis	Poor data quality	Data users may have unrealistic assumptions about the quality of government data (e.g. the common beliefs that information is objective, neutral, and readily available)	Radin (2006), Dawes (2010)
		Data quality remains a major issue for OGD	Karr (2008), Whitmore (2014)
		Much OGD has poor quality (e.g. bad data formats, infrequent data releases, lack of granularity, and inconsistency in naming or choice of identifiers)	Kuk and Davies (2011)
		The quality of OGD may be low and open data users may be concerned about the quality of the data	Martin (2014)
	Quality variation and changes	In general, the quality of data that flows from one information system to multiple others can quickly degrade over time without control of the processes and information input	Batini et al. (2009)
		The quality of data on the Web varies	Auer et al. (2013)
		The quality of open data varies, e.g. per country and per data provider	Petychakis et al. (2014)
		The quality of open data can easily be affected because of the reuse of the data	Oviedo et al. (2013)

Table 3-8 (continued): Overview of factors which influence OGD use, derived from the literature.

Appendix B: Documents studied for the case studies

Both cases

Data Archiving and Networked Services (2013). About DANS. Retrieved August 16, 2013, from http://dans.knaw.nl/en/content/about-dans

Data Archiving and Networked Services, Centraal Bureau voor de Statistiek, Huygens Instituut, Internationaal Instituut voor Sociale Geschiedenis, Koninklijke Bibliotheek en de Vereniging voor Geschiedenis en Informatica (2008). Data uit beleidsonderzoek voor rijk: verplicht archiveren. *E-data & Research, 3*(1), p. 1.

Dublin Core Metadata Initiative. (2010). Dublin Core Metadata Element Set, Version 1.1. Retrieved December 2, 2011, from http://dublincore.org/documents/dces/

Grootveld, M., & Egmond, J. v. (2011). Data Reviews, peer-reviewed research data. *DANS studies in digital archiving.* Retrieved January 6, 2015, from http://www.dans.knaw.nl/sites/default/files/file/publicaties/DANS_SDA_5_Data_Reviews_peer_reviewed_research_data_NL_DEF.pdf

Staat der Nederlanden, & Ministerie van Onderwijs, Cultuur en Wetenschap. *Raamovereenkomst ARVODI inzake beleidsgericht onderzoek*.

Staat der Nederlanden, & Ministerie van Onderwijs, Cultuur en Wetenschap. *Nadere overeenkomst met kenmerk [...] inzake beleidsgericht onderzoek*.

The Data Seal of Approval Board. (2013). Implementation of the Data Seal of Approval. Retrieved January 6, 2015, from https://assessment.datasealofapproval.org/assessment_47/seal/html/

Van der Graaf, M. (2013). De bekendheid van DANS en zijn diensten in 2013 en 2011 onder onderzoekers, promovendi en researchmaster studenten. Retrieved January 6, 2015, from http://www.dans.knaw.nl/sites/default/files/file/De%20bekendheid%20van%20DANS%20en%20zijn%20diensten%202013%20DEF.pdf

Van der Schaaf, J. (2013). *Toegang tot restricted access data in het DANS Data-archief*. DANS.

Wittenberg, M. (2009). Een verjaardagsfeestje met vergezichten. Retrieved November 16, 2012, from http://dans.knaw.nl/content/categorieen/symposia/symposia-archief/een-verjaardagsfeestje-met-vergezichten

Case study 1: Judicial data use

Klein Haarhuis, C.M., & Hagen, L.L.C. (2009). *Toetsen en verbinden. Over het WODC in de tijd en de relatie tussen onderzoek en beleid in het bijzonder.* Den Haag: Boom Juridische Uitgevers.

Staatscourant. (2011). *Organisatieregeling Ministerie van Veiligheid en Justitie* (Nr. 22848).

Wetenschappelijk Onderzoek en Documentatie Centrum (2008). *Circulaire Informatiebeveiliging en gegevensbeheerplan WODC*.

Wetenschappelijk Onderzoek en Documentatie Centrum (2013). *Circulaire Informatiebeveiliging en gegevensbeheerplan WODC*.

Wetenschappelijk Onderzoek en Documentatie Centrum. (2013). Organisatie. Retrieved August 16, 2013, from http://www.wodc.nl/organisatie/

Winter, H. B., Jong, P. O. d., Sibma, A., Visser, F. W., Herweijer, M., Klingenberg, A. M., & Prakken, H. (2008). *Wat niet weet, wat niet deert. Een evaluatieonderzoek naar de werking van de Wet bescherming persoonsgegevens in de praktijk*. Groningen: Pro Facto, Rijksuniversiteit Groningen, De Jong Beleidsadvies, WODC.

Wetenschappelijk Onderzoek en Documentatie Centrum (2007). *DANS als alternatief voor het WODC Onderzoeksdata-archief.*

Wetenschappelijk Onderzoek en Documentatie Centrum (2011). *Memo Resultaten Archivering Onderzoeksbestanden bij DANS.*

Wetenschappelijk Onderzoek en Documentatie Centrum (2014). *WODC Data Archivering en Open Data. Rapportage van de werkgroep Data Archivering en Open Data (concept).*

Case study 2: Social data use

Fisher, K., Tucker, J., Altintas, E., Bennett, M., Jahandar, A., & Jun, J. (2013, 15 July 2013). Technical Details of Time Use Studies. Retrieved December 4, 2013, from http://www.timeuse.org/information/studies/

Overheid.nl - Ministry of Culture, Recreation and Societal Services (2012). *Besluit houdende instelling van een Sociaal en Cultureel Planbureau*, nr. 4760.

Overheid.nl - Ministry of General Affairs (2013). Regeling van de minister-president, Minister van Algemene Zaken, houdende de vaststelling van de Aanwijzingen voor de Planbureaus, nr. 120.

Sociaal en Cultureel Planbureau. (2000). *Jaarverslag informatievoorziening en automatisering 1999*.

Sociaal en Cultureel Planbureau. (2001). *Jaarverslag automatisering, informatievoorziening en statistische advisering 2000*.

Sociaal en Cultureel Planbureau. (2002). *Jaarverslag 2001 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau. (2003). *Jaarverslag 2002 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau. (2004). *Jaarverslag 2003 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau (2005). *Jaarverslag 2004 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau (2006). *Jaarverslag 2005 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau (2007). *Jaarverslag 2006 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau (2008). *Jaarverslag 2007 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau (2009). *Jaarverslag 2008 automatisering, informatievoorziening en statistische advisering*.

Sociaal en Cultureel Planbureau (2010). *Jaarverslag I&A 2009*.

Sociaal en Cultureel Planbureau (2011). *Jaarverslag I&A 2010*.

Sociaal en Cultureel Planbureau (2012). *Jaarverslag I&A 2011*.

Sociaal en Cultureel Planbureau (2013a). *Bestanden bij het SCP: regels en afspraken.*

Sociaal en Cultureel Planbureau (2013b). *Bestanden bij SCP: Praktische kanten*.

Sociaal en Cultureel Planbureau. (2013c). *Data.* Retrieved November 23, 2013, from http://www.scp.nl/Onderzoek/Tijdsbesteding/Achtergronden/Data
Sociaal en Cultureel Planbureau (2013d). *Data en de Wet bescherming persoonsgegevens (Wbp).*
Sociaal en Cultureel Planbureau (2013e). *Jaarverslag I&A 2012*.
Sociaal en Cultureel Planbureau (2013f). *Meta-informatie: wat is MISS?*
Sociaal en Cultureel Planbureau (2013g). *Van datagroep naar datamanagement*.
Sociaal en Cultureel Planbureau. (2013h). *Wat doet het SCP?* Retrieved November 23, 2013, from http://www.scp.nl/Organisatie/Wat_is_het_SCP/Wat_doet_het_SCP
Sociaal en Cultureel Planbureau. (2013i). *Wat is het SCP?* Retrieved November 23, 2013, from http://www.scp.nl/Organisatie/Wat_is_het_SCP
Sociaal en Cultureel Planbureau. (2013j). *Wat maakt het SCP?* Retrieved November 23, 2013, from http://www.scp.nl/Organisatie/Wat_is_het_SCP/Wat_maakt_het_SCP

Appendix C: First survey (pre-test)

Dear participant,
The aim of this survey is to gather information for the improvement of open data infrastructures. You will be asked to fill out another survey after performing the scenarios. This survey consists of 19 questions. Completing the survey will take about 8-12 minutes.
The asterisk (*) behind a question indicates that this question is obligatory.
Thank you very much in advance for participating in this survey. We appreciate your time and input.

Before you start, what is your participant code? *

_____

1: Demographics

Question 1: What is your gender? *

O Male

O Female

Question 2: What is your age? *

_____

Question 3: What is your nationality?

O Dutch

O Chinese

O Iranian

O Greek

O Indonesian

O Other: _____

Question 4: Which role describes your daily occupation best?

O Student or citizen

O Researcher

O Civil servant and/or policy-maker

O Developer or entrepreneur

O Other: _____

Question 5: In daily life, how often have you been involved in publishing open data? *

Being involved in publishing open data refers to making datasets collected/ created/ produced by the government (federal, provincial, municipal) openly available on the internet.

O Never → Skip to question 9.

O Yearly or a few times per year → Skip to question 6.

O Monthly or a few times per month → Skip to question 6.

O Weekly or a few times per week → Skip to question 6.

O Daily or multiple times per day → Skip to question 6.

Question 6: Since when have you been involved in publishing open data? *

O 0-1 years ago

O 1-2 years ago

O 2-5 years ago

O 5-10 years ago

O 10-20 years ago

O More than 20 years ago

Question 7: To which extent do you have experience with publishing open data? *

	1	2	3	4	5	6	7	8	9	10	
No experience	O	O	O	O	O	O	O	O	O	O	Experienced

Question 8: What are your reasons for publishing open data? *

☐ To publish data as part of my job

☐ For data linking (combining and integrating different datasets)

☐ For news / media reporting (journalism)

☐ As a justification of my work

☐ For curiosity

☐ Other: _____

Question 9: In daily life, how often have you been involved in using open data?*

Being involved in using open data refers to studying, downloading, checking, analysing, visualising, investigating, and linking open data provided by the government or using open data in any other way.

O Never → Skip to question 13.

O Yearly or a few times per year → Skip to question 10.

O Monthly or a few times per month → Skip to question 10.

O Weekly or a few times per week → Skip to question 10.

O Daily or multiple times per day → Skip to question 10.

Question 10: Since when have you been involved in using open data? *

O 0-1years ago

O 1-2 years ago

O 2-5 years ago

O 5-10 years ago

O 10-20 years ago

O More than 20 years ago

Question 11: To which extent do you have experience with using open data? *

	1	2	3	4	5	6	7	8	9	10	
No experience	O	O	O	O	O	O	O	O	O	O	Experienced

Question 12: What are your reasons for using open data?

☐ For my study

☐ To write academic publications

☐ To perform a statistical analysis

☐ To perform policy research

☐ To perform investigations (non−scientific and non−policy)

☐ For political and policy-making decisions

☐ For data linking (combining and integrating different datasets)

☐ For news / media reporting (journalism)

☐ For daily operation in work

☐ For curiosity and/or recreation

☐ Other: _____

2: Open data metadata

The following questions concern the infrastructure support for certain open data activities that make use of metadata. To be able to answer these questions, think about the open data infrastructures that you are familiar with.

Question 13: To which extent do you agree with the following statements? At least one of the open data infrastructures that I know enables me to... *

	Strongly disagree	Disagree	Slightly disagree	Neither disagree, nor agree	Slightly agree	Agree	Strongly agree
...easily see where and how to search for data.	O	O	O	O	O	O	O
...use various options to search for data (e.g. key words, categorisations, filters, translations).	O	O	O	O	O	O	O
...clearly understand the search results.	O	O	O	O	O	O	O
...find a dataset that I want to use.	O	O	O	O	O	O	O
...understand what the dataset that I found is about.	O	O	O	O	O	O	O
...analyse the datasets that I found.	O	O	O	O	O	O	O
...view datasets without downloading them.	O	O	O	O	O	O	O
...draw conclusions based on the data that I found.	O	O	O	O	O	O	O
...visualise data in a table.	O	O	O	O	O	O	O
...visualise data in a chart.	O	O	O	O	O	O	O
...visualise data on a map.	O	O	O	O	O	O	O

3: Open data interaction and data quality analysis

The following questions concern the infrastructure support for certain open data activities that make use of interaction mechanisms and data quality indicators. To be able to answer these questions, think about the open data infrastructures that you are familiar with.

Question 14: To which extent do you agree with the following statements? At least one of the open data infrastructures that I know enables me to… *

	Strongly disagree	Disagree	Slightly disagree	Neither disagree, nor agree	Slightly agree	Agree	Strongly agree
...discuss what can be learned from data use by leaving a discussion post.	O	O	O	O	O	O	O
...share and discuss on social media what can be learned from data use.	O	O	O	O	O	O	O
...discuss what can be learned from data use by looking at previous uses of the data (e.g. visualisations, publications and applications).	O	O	O	O	O	O	O
...discuss what can be learned from data use by publishing experiences and articles about this on the infrastructure.	O	O	O	O	O	O	O
...discuss what can be learned from the data use on a Wiki or forum.	O	O	O	O	O	O	O
...view quality ratings of the dataset.	O	O	O	O	O	O	O
...rate different quality aspects of the dataset.	O	O	O	O	O	O	O
...view discussions about the quality of the dataset.	O	O	O	O	O	O	O
...discuss the data quality.	O	O	O	O	O	O	O

4: The DANS/ENGAGE open data infrastructure

The following questions concern the DANS/ENGAGE open data infrastructure. You may not yet be familiar with this infrastructure. In that case, please indicate how you expect DANS/ENGAGE to perform on the various aspects. You do not need to investigate DANS/ENGAGE before answering the questions. There are no wrong answers.

Question 15: To which extent do you agree with the following statements? I expect that... *

	Strongly disagree	Disagree	Slightly disagree	Neither disagree, nor agree	Slightly agree	Agree	Strongly agree
...using DANS/ENGAGE would enable me to use open data more quickly.	O	O	O	O	O	O	O
...using DANS/ENGAGE would make it easier to use open data.	O	O	O	O	O	O	O
...using DANS/ENGAGE would enhance my effectiveness in using open data.	O	O	O	O	O	O	O
...using DANS/ENGAGE would be easy for me.	O	O	O	O	O	O	O
...learning to use DANS/ENGAGE would be easy for me.	O	O	O	O	O	O	O
...it would be easy for me to become skilful at using DANS/ENGAGE.	O	O	O	O	O	O	O
...people who influence my behaviour would think that I should use DANS/ENGAGE.	O	O	O	O	O	O	O
...people who are important to me would think that I should use DANS/ENGAGE.	O	O	O	O	O	O	O
...people who are in my social circle would think that I should use DANS/ENGAGE.	O	O	O	O	O	O	O
...I would have the resources necessary to use DANS/ENGAGE.	O	O	O	O	O	O	O

...I would have the knowledge necessary to use DANS/ENGAGE.	O	O	O	O	O	O	O
...a specific person (or group) would be available for assistance with using DANS/ENGAGE.	O	O	O	O	O	O	O
... DANS/ENGAGE would provide open data in my best interest.	O	O	O	O	O	O	O
... DANS/ENGAGE would provide access to sincere and genuine open data.	O	O	O	O	O	O	O
... DANS/ENGAGE would perform its role of providing open data very well.	O	O	O	O	O	O	O

Question 16: All things considered, to which extent do you expect that using DANS/ENGAGE would be a bad or good idea? *

	1	2	3	4	5	6	7	8	9	10	
Bad idea	O	O	O	O	O	O	O	O	O	O	Good idea

Question 17: All things considered, to which extent do you expect that using DANS/ENGAGE would be a foolish or wise move? *

	1	2	3	4	5	6	7	8	9	10	
Foolish move	O	O	O	O	O	O	O	O	O	O	Wise move

Question 18: All things considered, to which extent do you expect that using DANS/ENGAGE would be a negative or positive step? *

	1	2	3	4	5	6	7	8	9	10	
Negative step	O	O	O	O	O	O	O	O	O	O	Positive step

5: Suggestions and comments

Question 19. Do you have any other comments and/or suggestions? If so, could you please write them down here?

Thank you

Thank you very much for participating in this survey. Please submit your answers. Thereafter you can go to the next part of this experiment by conducting scenarios.

Appendix D: Second survey (scenario survey) and scenario instructions

Scenarios for ENGAGE/DANS group
You have just completed the first survey. The next part of this session concerns conducting scenarios. There are 18 scenario tasks in total, divided in the five following sections.

1. Searching for and finding open data
2. Analysing open data
3. Visualising open data (tables, charts, maps)
4. Providing feedback and discussing what can be learned from the use of open data
5. Discussing the quality of open data

Each section is represented in a different colour. In addition there is a sixth section with four general questions at the end.

Thank you very much in advance for conducting these scenarios. We appreciate your time and input.

Question 1: What is your participant code?

……..

Question 2: What is the time at this moment?
We ask for the time to measure how long it takes to conduct certain tasks.

……..

Please go to the next page.

Scenario tasks - Searching for and finding open data (1/5)

Question 3. To which extent do you find it difficult or easy to complete the following scenario tasks related to searching for and finding open data?
Please first complete the tasks shown in the column on the left below. Then answer the question about how difficult/easy this was. If any of the tasks below is not possible in your opinion, then select 'very difficult'.

Mark only one oval per row.

	Very difficult	Difficult	Neither difficult, nor easy	Easy	Very easy
Task 1. Start the Mozilla Firefox browser and go to www.engagedata.eu/ http://www.dans.knaw.nl/en.	○	○	○	○	○
Task 2. Search for the dataset "Dutch parliamentary election study" for the years 2002-2003, published by the Data Archiving and Networked Services in The Netherlands.	○	○	○	○	○
Task 3. Examine to which extent the search results are useful and understandable. Sort or filter the data if necessary.	○	○	○	○	○
Task 4. Click on the search result entitled "Dutch Parliamentary Election Study 2002 2003 - DPES 2002 2003" (published by DANS).	○	○	○	○	○

Question 4: Please rank the tasks according to their level of difficulty.
Fill in the numbers of the tasks. Choose from task 1-4 and choose each task only once. Write only one number on each line.

The easiest task was task ……..
The slightly more difficult task was task ……..
The more difficult task was task ……..
The most difficult task was task ……..

Or: □ All tasks were equally easy/difficult

Question 5: To which extent were you able to complete tasks 1-4?
Check all that apply.

□ I was able to complete task 1
□ I was able to complete task 2
□ I was able to complete task 3
□ I was able to complete task 4
□ I was not able to complete any of these tasks

Question 6: What is the time at this moment?
We ask for the time to measure how long it takes to conduct certain tasks.
……..

Scenarios tasks - Analysing open data (2/5)

➔ Go to http://www.engagedata.eu/dataset/17024/ (this is an extension of the original dataset that you just selected).

Question 7 (Task 5): You have selected a dataset. Now have a look at the different types of metadata of the dataset. Subsequently briefly describe here what the dataset is about.

……..

……

Question 8 (Task 6): View the dataset without downloading it.

Question 9 (Task 7): Write down two conclusions based on the dataset that you just analysed.

……..

……

Question 10: To which extent did you find it difficult or easy to complete the scenario tasks related to analysing open data?

If any of the tasks below was not possible in your opinion, then select 'very difficult'.

	Very difficult	Difficult	Neither difficult, nor easy	Easy	Very easy
Task 5. To which extent was it difficult/easy to understand what this dataset was about?	○	○	○	○	○
Task6.To which extent was it difficult/easy to view the dataset without downloading it?	○	○	○	○	○
Task 7. To which extent was it difficult/easy to write down two conclusions based on the dataset that you just analysed?	○	○	○	○	○

Question 11: Please rank the tasks according to their level of difficulty.

Fill in the numbers of the tasks. Choose from task 5-7, and choose each task only once. Write only one number on each line.

The easiest task was task ……..

The more difficult task was ……..

The most difficult task was task ……..

Or: □ All tasks were equally easy/difficult

Question 12: To which extent were you able to complete tasks 5-7?

Check all that apply.

□ I was able to complete task 5

□ I was able to complete task 6

□ I was able to complete task 7

□ I was not able to complete any of these tasks

Question 13: What is the time at this moment?

We ask for the time to measure how long it takes to conduct certain tasks.

……..

Scenario tasks - Visualising open data (3/5)
To be able to conduct the following tasks, you need to either sign in using one of your social media accounts or register on the Engage platform. Please do so and then go back to the same dataset as you analysed before (http://www.engagedata.eu/dataset/17024/).

Question 14: To which extent do you find it difficult or easy to complete the following scenario tasks related to visualizing open data?
Please first complete the tasks shown in the column on the left below. Then answer the question about how difficult/easy this was. If any of the tasks below is not possible in your opinion, then select 'very difficult'.

	Very difficult	Difficult	Neither difficult, nor easy	Easy	Very easy
Task 8. Visualise the dataset in a table.	○	○	○	○	○
Task 9. Visualise the dataset in a chart.	○	○	○	○	○

➔ Go to http://www.engagedata.eu/dataset/14559/

	Very difficult	Difficult	Neither difficult, nor easy	Easy	Very easy
Task 10. Visualise the dataset on a map.	○	○	○	○	○

Question 15: Please rank the tasks according to their level of difficulty.
Fill in the numbers of the tasks. Choose from task 8-10 and choose each task only once. Write only one number on each line.

The easiest task was task ……..
The more difficult task was task ……..
The most difficult task was task ……..

Or: □ All tasks were equally easy/difficult

Question 16: To which extent were you able to complete tasks 8-10?
Check all that apply.

□ I was able to complete task 8
□ I was able to complete task 9
□ I was able to complete task 10
□ I was not able to complete any of these tasks

Question 17: What is the time at this moment?
We ask for the time to measure how long it takes to conduct certain tasks.
……..

Scenario tasks - Providing feedback and discussing what can be learned from the use of open data (4/5)
Go to http://www.engagedata.eu/dataset/17024/.

Question 18: To which extent do you find it difficult or easy to complete the following scenario tasks related to giving feedback and discussing open data?
Please first complete the tasks shown in the column on the left below. Then answer the question about how difficult/easy this was. If any of the tasks below is not possible in your opinion, then select 'very difficult'.

	Very difficult	Difficult	Neither difficult, nor easy	Easy	Very easy
Task 11. Participate in a discussion about the dataset that you just used by providing feedback or discussing the dataset (e.g. post the conclusions that you derived from the use of the data in task 7).	○	○	○	○	○
Task 12. View whether the dataset can be discussed by sharing it via social media (e.g. Twitter, Facebook, LinkedIn).	○	○	○	○	○
Task 13. View whether the dataset can be connected to results of reuse of this dataset (e.g. publications, visualisations, applications).	○	○	○	○	○
Task 14. View whether there is a forum or Wiki in this open data infrastructure where general data use can be discussed.	○	○	○	○	○

Question 19: Please rank the tasks according to their level of difficulty.
Fill in the numbers of the tasks. Choose from task 11-14 and choose each task only once. Write only one number on each line.

The easiest task was task ……..
The slightly more difficult task was task ……..
The more difficult task was task ……..
The most difficult task was task ……..
Or: □ All tasks were equally easy/difficult

Question 20: To which extent were you able to complete tasks 11-14?
Check all that apply.
□ I was able to complete task 11
□ I was able to complete task 12
□ I was able to complete task 13
□ I was able to complete task 14
□ I was not able to complete any of these tasks

Question 21: What is the time at this moment?
We ask for the time to measure how long it takes to conduct certain tasks.
……..

Scenario tasks - Discussing the quality of open data (5/5)
Go to http://www.engagedata.eu/dataset/17024/.

Question 22: To which extent do you find it difficult or easy to complete the following scenario tasks related to discussing the quality of open data?
Please first complete the tasks shown in the column on the left below. Then answer the question about how difficult/easy this was. If any of the tasks below is not possible in your opinion, then select 'very difficult'.

	Very difficult	Difficult	Neither difficult, nor easy	Easy	Very easy
Task 15. View quality ratings of the dataset.	○	○	○	○	○
Task 16. Rate the quality of the dataset.	○	○	○	○	○
Task 17. View whether the infrastructure allows for discussing the quality of the dataset by leaving a message.	○	○	○	○	○
Task 18. Discuss the quality of the dataset by leaving a message.	○	○	○	○	○

Question 23: Please rank the tasks according to their level of difficulty.
Fill in the numbers of the tasks. Choose from task 15-18 and choose each task only once. Write only one number on each line.

The easiest task was task ……..
The slightly more difficult task was task ……..
The more difficult task was task ……..
The most difficult task was task ……..

Or: □ All tasks were equally easy/difficult

Question 24: To which extent were you able to complete tasks 15-18?
Check all that apply.

□ I was able to complete task 15
□ I was able to complete task 16
□ I was able to complete task 17
□ I was able to complete task 18
□ I was not able to complete any of these tasks

Question 25: What is the time at this moment?
We ask for the time to measure how long it takes to conduct certain tasks.

……..

General questions

Question 26: Did the user interface influence the difficulty/ease to conduct the scenario tasks?

- ○ **No**
- ○ **Yes, positively/negatively (strike through the incorrect answer), namely by**

……

……

Question 27: Did the number of datasets provided by the open data infrastructure influence the difficulty/ease to conduct the scenario tasks?

- ○ **No**
- ○ **Yes, positively/negatively (strike through the incorrect answer), namely by**

……

……

Question 28: Did the programmes available (e.g. the programmes for analysing and visualising data) influence the difficulty/ease to conduct the scenario tasks?

- ○ **No**
- ○ **Yes, positively/negatively (strike through the incorrect answer), namely by**

……

……

Question 29: Did any other factors influence the difficulty/ease to conduct the scenario tasks?

- ○ **No**
- ○ **Yes, positively/negatively (strike through the incorrect answer), namely by**

……

……

Thank you!
Thank you for completing the scenario tasks. Please hand in your answers at the end of this quasi-experiment. You can now first proceed with the third survey.

Appendix E: Observers' instructions

First of all many thanks for your help with observing in this quasi-experiment. Your help is much appreciated.

This quasi-experiment is part of my PhD-research. In the three quasi-experiments that I planned to conduct in March and April 2014, students and professional open data users will evaluate parts of the open data infrastructure that I have been working on as part of my PhD-research. You have been asked to observe in one or more of these quasi-experiments. See the table below for an overview of the planned quasi-experiments.

Description	Date	Time
Quasi-experiment	March 3, 2014	10.30-12.30
Quasi-experiment	March 5, 2014	13.30-15.30
Quasi-experiment (as part of a workshop)	April 23, 2014	14.55-16.05

The aim of the quasi-experiments is to find out whether the open data infrastructure that I have been working on performs better than another existing open data infrastructure. To attain this objective, I use three types of measurements: surveys (including pre-test and post-test), time measurements and observations (see Figure A-1). The time measurements will be used to find out how long it takes to conduct the scenarios and whether this is different for the treatment and for the control group. With the observations I want to measure whether the ease of using the developed infrastructure is different from the ease of use of the control infrastructure.

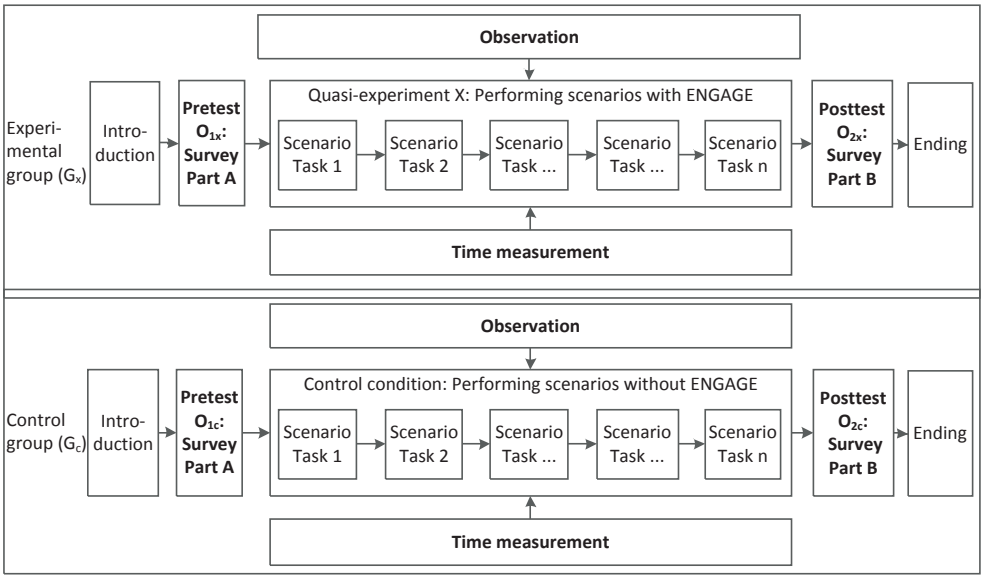


Figure A-1: Structure of the quasi-experiments.

In this document the instructions for the observations are provided. All quasi-experiments are organised as shown in Figure A-2.

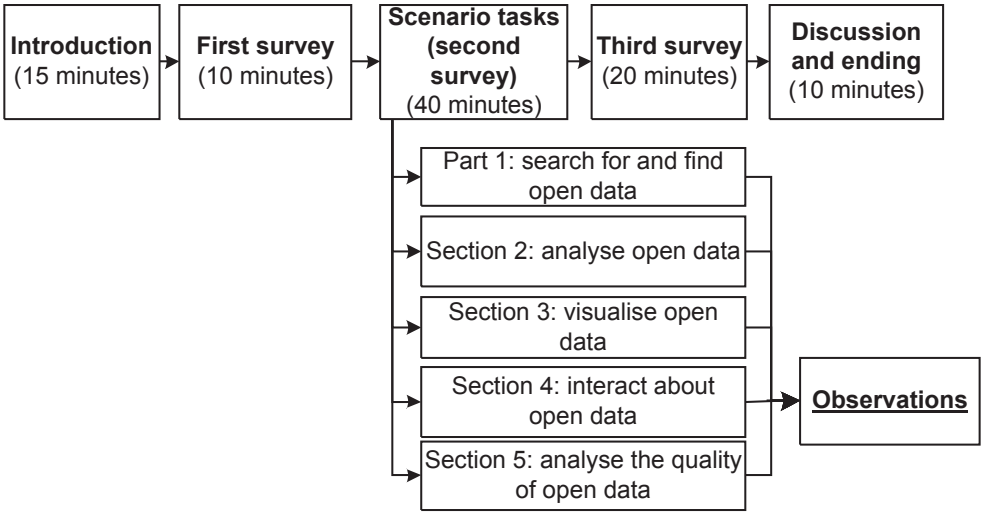


Figure A-2: Organisation of the quasi-experiments.

Observations only need to be done with regard to the five sections of the scenarios. There will be no observations during the introduction, the first and the third survey and the discussion and ending.

I have created an observation form which should guide you through the observation questions. Each observer will receive this observer form in advance of the quasi-experiments. Please have a look at the questions in the observer form before the observations take place.

The table below shows how the observer form is structured.

<table>
<tr><td>Part 1: Observations per scenario task</td><td colspan="3">In this section you will be asked about your observations concerning the scenario tasks that the participants conduct in the five different sections presented above. This section is divided into three columns.</td></tr>
<tr><td></td><td>Column 1: Sections in the quasi-experiment. This column shows how the quasi-experiment has been organised and in which section you are working. The colours in this column correspond with the colours used in the scenario instructions that the participants use to conduct their tasks.</td><td>Column 2: Time. In this column you will be asked to write down the time, so that it can be measured how long it approximately took for the participants to complete the different parts of the scenarios. You can look at the colours used in the scenario instructions that the participants received to see on which section they are working.</td><td>Column 3: Observation per scenario step. This column asks for your observations with regard to the ease or difficulty of conducting the scenario tasks, guided by particular questions. There are open and closed questions.</td></tr>
<tr><td>Part 2: Observations of the scenario tasks in general</td><td colspan="3">In this section you will be asked a number of open and closed questions about your observations concerning the scenario tasks in general.</td></tr>
<tr><td>Part 3: Observations of collaboration and questioning</td><td colspan="3">In this section you will be asked a number of open questions about your observations concerning the way that the observed participants collaborated and asked questions.</td></tr>
</table>

Please bring a watch or phone, so that you can check approximately how long it takes for the participants to complete the scenario tasks.

Each observer will be assigned to the observation of a number of participants. I have created a map to show where each observer should stand and which participants should be observed by this observer. You will be provided with this map.

The participants know that they will be observed (they are told so in the introduction of this experiment). Feel free to look at their computer screens and on their paper to see what they are doing and on which part of the scenarios they are

working. Especially have a look at the colours of the section in which they are working, as this is important for some of the questions. Yet, try not to disturb the participants.

If any of the participants asks you a question, please ask me to come to answer the question. I will not observe the participants myself (as my observations may be biased), but of course I will be in the room.

Please do not talk to the students and do not ask them questions. Only observe them.

If a student asks you a question or if there is any other behaviour that you notice, please write down the 'participant code of this student' (you can observe this, because each student has his code just in front of him/her. The participant code is also taped to the table with tape.

Do not hesitate to ask questions if the observer form and/or the instructions are not clear.

Thank you very much for your help; it is much appreciated!

Anneke Zuiderwijk

Appendix F: Semi-structured observer survey

Date of observation: ……
Name of the observer:……
Group observed: DANS/ENGAGE group
Number of participants observed: …… participants

Part 1: Observations per scenario task

Section 1: Searching for and finding OGD

1a: How much time did it approximately take for the observed participants to conduct the tasks in this section of the scenarios?…… minutes

1b. To which extent do you think that the participants found it difficult or easy to search for and find open data as part of the scenarios?

Very difficult	Difficult	Not difficult, not easy	Easy	Very easy
O	O	O	O	O

1c. Please explain your answer. Why do you think that they found it easy/difficult to search for and find open data?
……
……

Section 2: OGD analysis

2a: How much time did it approximately take for the observed participants to conduct the tasks in this section of the scenarios?…… minutes

2b. To which extent do you think that the participants found it difficult or easy to analyse open data as part of the scenarios?

Very difficult	Difficult	Not difficult, not easy	Easy	Very easy
O	O	O	O	O

2c. Please explain your answer. Why do you think that they found it easy/difficult to analyse open data?
……
……

Section 3: OGD visualisation

3a: How much time did it approximately take for the observed participants to conduct the tasks in this section of the scenarios?…… minutes

3b. To which extent do you think that the participants found it difficult or easy to visualise open data as part of the scenarios?

Very difficult	Difficult	Not difficult, not easy	Easy	Very easy
O	O	O	O	O

3c. Please explain your answer. Why do you think that they found it easy/difficult to visualise open data?

……
……

Section 4: Interaction about OGD

4a: How much time did it approximately take for the observed participants to conduct the tasks in this section of the scenarios?……. minutes

4b. To which extent do you think that the participants found it difficult or easy to give feedback and discuss what can be learned from the use of open data as part of the scenarios?

Very difficult	Difficult	Not difficult, not easy	Easy	Very easy
O	O	O	O	O

4c. Please explain your answer. Why do you think that they found it easy/difficult to give feedback and discuss what can be learned from the use of open data?

……
……

Section 5: OGD quality analysis

5a: How much time did it approximately take for the observed participants to conduct the tasks in this section of the scenarios? ……. minutes

5b. To which extent do you think that the participants found it difficult or easy to discuss the quality of open data as part of the scenarios?

Very difficult	Difficult	Not difficult, not easy	Easy	Very easy
O	O	O	O	O

5c. Please explain your answer. Why do you think that they found it easy/difficult to discuss the quality of open data?

……
……

Part 2: Observations of the scenarios in general
Please answer the questions below.

6a. To which extent do you think that the participants found it difficult or easy to conduct the scenarios in general?

Very difficult	Difficult	Not difficult, not easy	Easy	Very easy
O	O	O	O	O

6b. Please explain you answer. Why do you think that they found it easy/difficult to conduct the scenarios in general?
……
……

7. To which extent do you think that the user interface influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

8. To which extent do you think that the facilitators (*facilitator names*) influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

9. To which extent do you think that the observer(s) influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

10. To which extent do you think that other participants influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

11. To which extent do you think that the participant's gender influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

12. To which extent do you think that the participant's age influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

13. To which extent do you think that the participant's nationality influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

14. To which extent do you think that the participant's experience with publishing and using open data influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

15. To which extent do you think that the setting (e.g. the room, the lights, noise) influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……..
……..

16. To which extent do you think that the organisation of this experiment (e.g. the instructions, the surveys, the task descriptions) influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

17. To which extent do you think that the any other aspects influenced the way that the observed participants conducted the scenarios? Please explain you answer.
……
……

Part 3: Observations of collaboration and questioning

18. How many participants worked individually all the time while conducting the scenarios without discussing with anyone else? (e.g. their neighbour)
……. participants

19. How many participants discussed at least once with (one of) their neighbour(s) while conducting the scenarios? ……. participants

20. How many participants intensively discussed more than once with (one of) their neighbour(s) while conducting the scenarios? ……. participants

21. How many participants asked questions to the facilitator of the quasi-experiment while conducting the scenarios? ……. participants

22. Did any of the participants complain about the difficulty of conducting the scenarios?
O No
O Yes, namely …… participants. The main complaints were:
……
……

23. Do you have any other remarks?
……
……

Thank you very much!

Appendix G: Third survey (post-test)

Dear participant,

You have just conducted the scenarios. As a final part of this session, we would like to ask you to fill out a third survey. The aim of this survey is to gather information for the improvement of open data infrastructures. This survey consists of 26 questions. Completing the survey will take about 15-20 minutes of your time.

The asterisk (*) behind a question indicates that this question is obligatory.

Thank you very much in advance for participating in this survey. We appreciate your time and input.

Before you start, what is your participant code? *

_____

1: Session and the scenarios
The following questions concern this practical session (quasi-experiment) on open data use as a whole and certain parts in particular.

Question 1: To which extent do you agree with the following statements? *

	Strongly disagree	Disagree	Slightly disagree	Neither disagree, nor agree	Slightly agree	Agree	Strongly agree
The practical session on open data use was well-organised.	O	O	O	O	O	O	O
The session was well-structured (clear sequence).	O	O	O	O	O	O	O
It was clear to me what my role was in the session.	O	O	O	O	O	O	O
The instructions made it possible for me to fulfil my tasks well.	O	O	O	O	O	O	O
I received sufficient information to participate in the session.	O	O	O	O	O	O	O
The facilitator of this session influenced my behaviour during the	O	O	O	O	O	O	O

session.							
The facilitator of this session had a neutral attitude.	O	O	O	O	O	O	O
The scenarios reflected the use of open data in a realistic way.	O	O	O	O	O	O	O
The scenarios reflected the use of metadata in a realistic way.	O	O	O	O	O	O	O
The scenarios reflected feedback and discussion in a realistic way.	O	O	O	O	O	O	O
I learned something by participating in the session.	O	O	O	O	O	O	O
In general, participating in the session was a valuable experience.	O	O	O	O	O	O	O

2: Open data metadata
The following questions concern the infrastructure support for certain open data activities that make use of metadata.

Question 2: To which extent do you agree with the following statements? DANS/ENGAGE enabled me to... *

	Strongly disagree	Disagree	Slightly disagree	Neither disagree, nor agree	Slightly agree	Agree	Strongly agree
...easily see where and how to search for data.	O	O	O	O	O	O	O
...use various options to search for data (e.g. key words, categorizations, filters, translations).	O	O	O	O	O	O	O
...clearly understand the search results.	O	O	O	O	O	O	O
...find a dataset that I want to use.	O	O	O	O	O	O	O
...understand what the dataset that I found was about.	O	O	O	O	O	O	O
...analyse the dataset that I found.	O	O	O	O	O	O	O

...view a dataset without downloading it.	O	O	O	O	O	O	O
...draw conclusions based on the data that I found.	O	O	O	O	O	O	O
...visualise the data in a table.	O	O	O	O	O	O	O
...visualise the data in a chart.	O	O	O	O	O	O	O
...visualise the data on a map.	O	O	O	O	O	O	O

3: Open data interaction and data quality analysis
The following questions concern the infrastructure support for certain open data activities that make use of interaction mechanisms and data quality indicators.

Question 3: To which extent do you agree with the following statements? DANS/ENGAGE enabled me to... *

	Strongly disagree	Disagree	Slightly disagree	Neither disagree, nor agree	Slightly agree	Agree	Strongly agree
...discuss what can be learned from data use by leaving a discussion post.	O	O	O	O	O	O	O
...share and discuss on social media what can be learned from data use.	O	O	O	O	O	O	O
...discuss what can be learned from data use by looking at previous uses of the data (e.g. visualisations, publications and applications).	O	O	O	O	O	O	O
...discuss what can be learned from data use by Publishing experiences and articles about this on the infrastructure.	O	O	O	O	O	O	O
...discuss what can be learned from the data use on a Wiki or forum.	O	O	O	O	O	O	O
...view quality ratings of the dataset.	O	O	O	O	O	O	O
...rate different quality aspects of the dataset.	O	O	O	O	O	O	O

...view discussions about the quality of the dataset.	O	O	O	O	O	O	O
...discuss the data quality.	O	O	O	O	O	O	O

4: The DANS/ENGAGE open data infrastructure
The following questions concern the DANS/ENGAGE open data infrastructure.

Question 4: To which extent do you agree with the following statements? *

	Strongly disagree	Disagree	Slightly disagree	Neither disagree, nor agree	Slightly agree	Agree	Strongly agree
Using DANS/ENGAGE enables me to use open data more quickly.	O	O	O	O	O	O	O
Using DANS/ENGAGE makes it easier to use open data.	O	O	O	O	O	O	O
Using DANS/ENGAGE enhances my effectiveness in using open data.	O	O	O	O	O	O	O
I find it easy to use DANS/ENGAGE.	O	O	O	O	O	O	O
Learning to use DANS/ENGAGE is easy for me.	O	O	O	O	O	O	O
It is easy for me to become skilful at using DANS/ENGAGE.	O	O	O	O	O	O	O
People who influence my behaviour think that I should use DANS/ENGAGE.	O	O	O	O	O	O	O
People who are important to me think that I should use DANS/ENGAGE.	O	O	O	O	O	O	O
People who are in my social circle think that I should use DANS/ENGAGE.	O	O	O	O	O	O	O
I have the resources necessary to use DANS/ENGAGE.	O	O	O	O	O	O	O
I have the knowledge	O	O	O	O	O	O	O

necessary to use DANS/ENGAGE.							
A specific person (or group) is available for assistance with using DANS/ENGAGE.	O	O	O	O	O	O	O
DANS/ENGAGE provides open data in my best interest.	O	O	O	O	O	O	O
DANS/ENGAGE provides access to sincere and genuine open data.	O	O	O	O	O	O	O
DANS/ENGAGE performs its role of providing open data very well.	O	O	O	O	O	O	O
I intend to continue using DANS/ENGAGE.	O	O	O	O	O	O	O
I plan to continue using DANS/ENGAGE.	O	O	O	O	O	O	O
I will continue using DANS/ENGAGE.	O	O	O	O	O	O	O

Question 5: All things considered, to which extent would using DANS/ENGAGE be a bad or good idea? *

	1	2	3	4	5	6	7	8	9	10	
Bad idea	O	O	O	O	O	O	O	O	O	O	Good idea

Question 6: All things considered, to which extent would using DANS/ENGAGE be a foolish or wise move? *

	1	2	3	4	5	6	7	8	9	10	
Foolish move	O	O	O	O	O	O	O	O	O	O	Wise move

Question 7: All things considered, to which extent would using DANS/ENGAGE be a negative or positive step? *

	1	2	3	4	5	6	7	8	9	10	
Negative step	O	O	O	O	O	O	O	O	O	O	Positive step

Question 8: Compared to my initial expectations, the ability of DANS/ENGAGE to enable me to use open data more quickly was: *

	1	2	3	4	5	6	7	8	9	10	
Much worse than expected	O	O	O	O	O	O	O	O	O	O	Much better than expected

Question 9: Compared to my initial expectations, the ability of DANS/ENGAGE to make it easier to use open data was:

	1	2	3	4	5	6	7	8	9	10	
Much worse than expected	O	O	O	O	O	O	O	O	O	O	Much better than expected

Question 10: Compared to my initial expectations, the ability of DANS/ENGAGE to enhance my effectiveness in using open data was: *

	1	2	3	4	5	6	7	8	9	10	
Much worse than expected	O	O	O	O	O	O	O	O	O	O	Much better than expected

Question 11: Compared to my initial expectations, using DANS/ENGAGE to use open data was: *

	1	2	3	4	5	6	7	8	9	10	
Much more difficult than expected	O	O	O	O	O	O	O	O	O	O	Much easier than expected

Question 12: Compared to my initial expectations, learning to use DANS/ENGAGE to use open data was: *

	1	2	3	4	5	6	7	8	9	10	
Much more difficult than expected	O	O	O	O	O	O	O	O	O	O	Much easier than expected

Question 13: Compared to my initial expectations, the degree to which people who influence my behaviour think that I should use DANS/ENGAGE to use open data was: *

	1	2	3	4	5	6	7	8	9	10	
Much higher than expected	O	O	O	O	O	O	O	O	O	O	Much lower than expected

Question 14: Compared to my initial expectations, the degree to which people who are important to me think that I should use DANS/ENGAGE to use open data was : *

	1	2	3	4	5	6	7	8	9	10	
Much higher than expected	O	O	O	O	O	O	O	O	O	O	Much lower than expected

Question 15: Compared to my initial expectations, the degree to which people who are in my social circle think that I should use DANS/ENGAGE was: *

	1	2	3	4	5	6	7	8	9	10	
Much higher than expected	O	O	O	O	O	O	O	O	O	O	Much lower than expected

Question 16: Compared to my initial expectations, the resources necessary to use DANS/ENGAGE were: *

	1	2	3	4	5	6	7	8	9	10	
Much less than expected	O	O	O	O	O	O	O	O	O	O	Much more than expected

Question 17: Compared to my initial expectations, my knowledge of using DANS/ENGAGE was: *

	1	2	3	4	5	6	7	8	9	10	
Much less than expected	O	O	O	O	O	O	O	O	O	O	Much more than expected

Question 18: Compared to my initial expectations, availability of assistance (from a specific person or a group) for using DANS/ENGAGE was: *

	1	2	3	4	5	6	7	8	9	10	
Much worse than expected	O	O	O	O	O	O	O	O	O	O	Much better than expected

Question 19: Compared to my initial expectations, the degree to which DANS/ENGAGE provides open data in my best interest was: *

	1	2	3	4	5	6	7	8	9	10	
Much lower than expected	O	O	O	O	O	O	O	O	O	O	Much higher than expected

Question 20: Compared to my initial expectations, open data provided by DANS/ENGAGE were: *

	1	2	3	4	5	6	7	8	9	10	
Much less sincere and genuine than expected	O	O	O	O	O	O	O	O	O	O	Much more sincere and genuine than expected

Question 21: Compared to my initial expectations, the ability of DANS/ENGAGE to perform its role of providing open data was: *

	1	2	3	4	5	6	7	8	9	10	
Much worse than expected	O	O	O	O	O	O	O	O	O	O	Much better than expected

Question 22: Becoming skilful at using DANS/ENGAGE to use open data was: *

	1	2	3	4	5	6	7	8	9	10	
Much more difficult than expected	O	O	O	O	O	O	O	O	O	O	Much easier than expected

Question 23: To which extent are you displeased or pleased with your use of DANS/ENGAGE? I am…… with my use of DANS/ENGAGE: *

	1	2	3	4	5	6	7	8	9	10	
Extremely displeased	O	O	O	O	O	O	O	O	O	O	Extremely pleased

Question 24: To which extent are you frustrated or contended with your use of DANS/ENGAGE? I am…… with my use of DANS/ENGAGE: *

	1	2	3	4	5	6	7	8	9	10	
Extremely frustrated	O	O	O	O	O	O	O	O	O	O	Extremely content

Question 25: To which extent are you dissatisfied or satisfied with your use of DANS/ENGAGE? I am…… with my use of DANS/ENGAGE: *

	1	2	3	4	5	6	7	8	9	10	
Extremely dissatisfied	O	O	O	O	O	O	O	O	O	O	Extremely satisfied

5: Suggestions and comments

Question 26. Do you have any other comments and/or suggestions? If so, could you please write them down here?

……
……

Please write down your e-mail address if you wish to receive the results of the study.

……
……

Thank you very much for participating in this survey.

Appendix H: Publications by the author

Journal articles

- Zuiderwijk, A., Janssen, M., & Dwivedi, Y. (forthcoming). Acceptance and use predictors of open data technologies: Drawing upon the Unified Theory of Acceptance and Use of Technology. *Government Information Quarterly.*
- Zuiderwijk, A., Janssen, M., & Susha, I. (forthcoming). Improving the speed and ease of open data use through metadata, interaction mechanisms and quality indicators. *Journal of Organizational Computing and Electronic Commerce*.
- Susha, I., Zuiderwijk, A., Janssen, M., Grönlund, Å. (2015). Benchmarks for evaluating the progress of open data adoption: usage, limitations, and lessons learned. *Social Science Computer Review, 33*(5), 613-630.
- Zuiderwijk, A. & Janssen, M. (2015). Towards decision support for disclosing data: closed or open data? *Information Polity, 20*(2, 3), 103-117.
- Nugroho, R.P., Zuiderwijk, A., Janssen, M., & de Jong, M. (2015). A comparison of national open data policies: lessons learned. *Transforming Government: People, Process and Policy, 9*(3), 286-308.
- Kaasenbrood, M., Zuiderwijk, A., Janssen, M., de Jong, M., Bharosa, N. (2015). Exploring the factors influencing the adoption of open government data by private organisations. *International Journal of Public Administration in the Digital Age, 2*(2), 75-92.
- Janssen, M., & Zuiderwijk, A. (2014). Infomediary business models for connecting open data providers and users. *Social Science Computer Review, 32*(5), 694-711.
- Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: essential elements of open data ecosystems. *Information Polity, 19*(1-2), 17-33.
- Zuiderwijk, A., Janssen, M., Choenni, S., & Meijer, R. (2014). Design principles for improving the process of publishing open data. *Transforming Government: People, Process and Policy, 8*(2), 185 - 204.
- Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: a comparison framework. *Government Information Quarterly, 31*(1), 17-29.
- Zuiderwijk, A., Jeffery, K. & Janssen, M. (2012). The potential of metadata for linked open data and its value for users and publishers. *eJournal of eDemocracy and Open Government, 4*(2), 222-244.
- Zuiderwijk, A., Janssen, M., Choenni, S, Meijer, R. & Sheikh Alibaks, R. (2012). Socio-technical impediments of open data. *Electronic Journal of e-Government, 10*(2), 156-172.

- 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.

Special issue guest editor introduction papers

- Zuiderwijk, A., Gascó, M., Parycek, P., & Janssen, M. (2014). Special issue on transparency and open data policies: guest editors' introduction. *Journal of Theoretical and Applied Electronic Commerce Research, 9*(3), I-IX.
- Zuiderwijk, A., Helbig, N., Gil-García, J., & Janssen, M. (2014). Guest editors' introduction. Innovation through open data: a review of the state-of-the-art and an emerging research agenda. *Journal of Theoretical and Applied Electronic Commerce Research, 9*(2), I-XIII.

Conference papers

- Janssen, M., Matheus, R., & Zuiderwijk, A. (2015). *Big and Open Linked Data (BOLD) to create smart cities and citizens: insights from smart energy and mobility cases.* Paper presented at the 14th IFIP Electronic Government Conference 2015, Thessaloniki, Greece.
- Zuiderwijk, A. & Janssen, M. (2015). *Participation and data quality in open data use: open data infrastructures evaluated*. Paper presented at the 15th European Conference on eGovernment, Portsmouth, United Kingdom.
- Zuiderwijk, A., Janssen, M., Poulis, K., & Vandekaa, G. (2015). *Open data for competitive advantage: insights from open data use by companies.* Presented at the 16th Annual International Conference on Digital Government Research, Phoenix, Arizona, U.S.A.
- Zuiderwijk, A., Susha, I. Charalabidis, Y., Parycek, P., & Janssen, M. (2015). *Open data disclosure and use: critical factors from a case study.* Paper presented at the International Conference for E-Democracy and Open Government, Krems and der Donau, Austria. **Nominated for 'Best paper award'.**
- Alexopoulos, C., Zuiderwijk, A., Charalabidis, Y., Loukis, E., & Janssen, M. (2014). *Designing a second generation of open data platforms: integrating open data and social media.* Paper presented at the 13th Conference on Electronic Government, Dublin, Ireland.
- Klievink, B., Zuiderwijk, A., & Janssen, M. (2014). Interconnecting governments, businesses and citizens: comparing digital infrastructures. Paper presented at the 13th Conference on Electronic Government, Dublin, Ireland.
- Zuiderwijk, A., & Janssen, M. (2014). The negative effects of open government data: investigating the dark side of open data. Paper

presented at the 15th Annual International Conference on Digital Government Research, Aguascalientes, Mexico. **Winner of the 'Best paper award' in the category of Best Policy/Management Paper.**

- Zuiderwijk, A., Loukis, E., Alexopoulos, C., Janssen, M., & Jeffery, K. (2014). *Elements for the development of an open data marketplace*. Paper presented at the International Conference for E-Democracy and Open Government, Krems an der Donau, Austria.
- Zuiderwijk, A., & Janssen, M. (2013). *A coordination theory perspective to improve the use of open data in policy-making.* Paper presented at the 12th Conference on Electronic Government, Koblenz, Germany.
- Alexopoulos, C., Loukis, E., Charalabidis, Y., & Zuiderwijk, A. (2013). *An evaluation framework for traditional and advanced open public data e-infrastructures.* Paper presented at the 13th European Conference on e-Government, Como, Italy.
- Zuiderwijk, A., Janssen, M., & Parnia, A. (2013). *The complementarity of open data infrastructures: an analysis of functionalities.* Paper presented at the 14th Annual International Conference on Digital Government Research, Quebec, Canada.
- Zuiderwijk, A., Janssen, M., & Jeffery, K. (2013). *Towards an e-infrastructure to support the provision and use of open data.* Paper presented at the International Conference for E-Democracy and Open Government, Krems an der Donau, Austria.
- Zuiderwijk, A., Janssen, M., Meijer, R., Choenni, R., Charalabidis, Y., & Jeffery, K. (2012). *Issues and guiding principles for opening governmental judicial research data.* Paper presented at the 11th Conference on E-Government, Kristiansand, Norway. **Winner of Outstanding Paper Award in the category of Most Promising Practical Concept.**
- Zuiderwijk, A., Janssen, M., & Choenni, S. (2012). *Open data policies: impediments and challenges.* Paper presented at the 12th European Conference on eGovernment, Barcelona, Spain.
- Zuiderwijk, A., & Janssen, M. (2012). *A comparison of open data policies and their implementation in two Dutch ministries.* Paper presented at the 13th Annual International Conference on Digital Government Research, College Park, Maryland, U.S.A.
- Van den Braak, S., Choenni, S., Meijer, R., & Zuiderwijk, A. (2012). *Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector.* Paper presented at the 13th Annual International Conference on Digital Government Research, College Park, Maryland, U.S.A.
- Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012). *The necessity of metadata for linked open data and its contribution to policy analyses.* Paper presented at the International Conference for E-Democracy and Open Government, Krems an der Donau, Austria.

Book chapter

- Zuiderwijk, A., & Janssen, M. (2014). Barriers and development directions for the publication and usage of open data: a socio-technical view. In M. Gascó-Hernández (Ed.), *Open Government. Opportunities and Challenges for Public Governance* (pp. 115-135). New York: Springer.

Workshops and workshop presentations

- Ojo, A., Mergel, I., Janssen, M., & Zuiderwijk, A. (2015). *Workshop: open data to solve societal issues.* Workshop organised at the 16th Annual International Conference on Digital Government Research, Phoenix, Arizona, U.S.A.
- Susha, I., Zuiderwijk, A., Parycek, P., Janssen, M., & Charalabidis, Y. (2015). *Workshop on context-specific critical success factors for open data publication and use.* Workshop organised at the International Conference for E-Democracy and Open Government, Krems an der Donau, Austria.
- Alexopoulos, C., Zuiderwijk, A., Charalabidis, Y., & Loukis, E. (2014). *Closing the open public data feedback loop: the ENGAGE platform.* Position paper presented as part of the Samos Workshop: Uses of Open Data Within Government for Innovation and Efficiency Forward-Looking at the 5th Samos Summit on ICT-enabled Governance, Pythagorion, Greece.
- Zuiderwijk, A. (2014). *Research methods for measuring open government data transparency*. Workshop presentation as part of the ESF Exploratory Workshop on Government Transparency, Lausanne, Switzerland.
- Zuiderwijk, A., Klievink, B., Janssen, M., Tan, Y-H, Charalabidis, Y. & Argyzoudis, E. (2013). *Workshop on open information infrastructures enabling innovations*. Workshop organised at the 12th Conference on Electronic Government, Koblenz, Germany.
- Zuiderwijk, A., & Janssen, M. (2013). *OGD Metadata standards: the ENGAGE metadata architecture.* Presentation as part of the Workshop on Open Gov Data Standardisation at the Informatik2013 Conference, Koblenz, Germany.
- Zuiderwijk, A., Jeffery, K., & Mouzakitis, S. (2013). *e-Infrastructures for open data.* Workshop organised at the International Conference for E-Democracy and Open Government, Krems an der Donau, Austria.
- Zuiderwijk, A., & Janssen, M. (2013). *Re-engineering the open data publishing process at a Dutch government organization.* Workshop paper presented at the 10th Scandinavian Workshop on E-Government, Oslo, Norway.
- Zuiderwijk, A., Janssen, M., & Charalabidis, Y. (2012). *A workshop about using open public sector data: the ENGAGE project*. Workshop organised

at the 11th Conference on Electronic Government and Electronic Participation, Kristiansand, Norway.

- Zuiderwijk, A., Alexopoulos, C., & Janssen, M. (2012). *Open data requirements*. Workshop organised at the 3rd Samos Summit on Open Data an Interoperability for Governance, Industry and Society, Pythagorion, Greece.
- Zuiderwijk, A., Janssen, M., van den Braak, S., & Charalabidis, Y. (2012). *Workshop: linking open data - challenges and solutions*. Workshop organised at the 13th Annual International Conference on Digital Government Research, College Park, Maryland, U.S.A.
- Zuiderwijk, A., Janssen, M., Jeffery, K. & Charalabidis, Y. (2012). *Open linked public sector data for citizen engagement. A workshop about the benefits and restrictions of open linked public sector data and the role of metadata in citizen engagement*. Workshop organised at the International Conference for E-Democracy and Open Government, Krems an der Donau, Austria.
- Janssen, M. & Zuiderwijk, A. (2012). *Open data and transformational government.* Workshop paper presented at the Transforming Government Workshop 2012, London, United Kingdom.

Keynote speech

- Zuiderwijk, A. (2014, July 17). *Business innovation through open data.* Keynote Speech at the 1st Conference on Future Environment and Innovation, Malang, Indonesia.

Curriculum Vitae

Anneke Zuiderwijk – van Eijk was born in Leidschendam in The Netherlands on July 4, 1988. In 2006, Anneke started studying Criminology at Leiden University. As part of two internships, Anneke conducted criminological research at Statistics Netherlands and at the Curium-LUMC Academic Centre for Child and Youth Psychiatry. In 2010 Anneke completed her Masters in Criminology and started working at the Research and Documentation Centre (Wetenschappelijk Onderzoek en Documentatie Centrum; WODC) of the Dutch Ministry of Security and Justice. She worked on a study concerning case processing times in the criminal justice chain, and the considerable amounts of data that were created and collected by the WODC aroused her interest in governmental research data. In November 2011 Anneke started her PhD research at the Faculty of Technology, Policy and Management at Delft University of Technology.

During her PhD research, Anneke has published 30 peer-reviewed journal and conference articles. She published in journals such as Government Information Quarterly (GIQ), Information Systems Management (ISM), Social Science Computer Review (SSCR), the Journal of Organizational Computing and Electronic Commerce (JOCEC), Information Polity, and Transforming Government: People, Process and Policy (TGPPP). Together with her co-authors, Anneke received best paper awards at the Conference on E-Government in 2012 in the category of Most Promising Practical Concept, and at the Annual International Conference on Digital Government Research in 2014 in the category of Best Policy/Management Paper. Anneke guest-edited two special issues for the Journal of Theoretical and Applied Electronic Commerce Research (JTAER) on open data innovation and open data policies and transparency and served as a track chair at the Conference on E-Democracy and Open Government (CeDEM).

Anneke has been involved in the FP7 EU funded ENGAGE project which aimed to develop an infrastructure for diverse public sector information resources to support scientific collaboration and research. She supervised 12 Masters and Bachelors students, taught various lectures at Delft University of Technology, and obtained her University Teaching Qualification. In 2014, Anneke gave an invited keynote presentation at the International Conference on Future Environment and Innovation. After her PhD, Anneke continues her teaching and research activities in the area of open data at Delft University of Technology. She will be involved in the VRE4EIC project (A Europe-wide interoperable Virtual Research Environment to Empower multidisciplinary research communities and accelerate Innovation and Collaboration). For more information about Anneke's research and teaching activities please visit http://tinyurl.com/AnnekeZuiderwijk.

Open Data Infrastructures

The design of an infrastructure to enhance the coordination of open data use

Governments and researchers traditionally focus on the publication of Open Government Data (OGD), whereas the actual use of the data is often neglected. Open data initiatives are often criticized for not realising the promoted benefits, yet only the use of OGD can result in these benefits. OGD use requires several actors, activities and tools; however, these are fragmented and depending on each other. The OGD infrastructure presented in this dissertation aims to enhance the coordination of OGD use. Core components are an advanced and interoperable three-tier metadata model to find, analyse, visualise, interact about and assess OGD, interaction mechanisms to stimulate interaction between OGD users, OGD providers and governmental policy makers, and data quality indicators to assess the data's fitness for use.

This study is among the first to describe the design of an OGD infrastructure. This dissertation contributes to science by providing a comprehensive overview of barriers and functional requirements for OGD use from the perspective of the OGD user, by defining functional building blocks for the design of the OGD infrastructure, and by developing and evaluating a prototype of the OGD infrastructure. Furthermore, this study is the first to apply coordination theory in the field of OGD and shows that coordination of OGD use does not merely require a focus on processes, but additionally requires a technical perspective including the integration of tools, a social perspective including interaction between involved actors, and the interaction between the social and technical perspective. Moreover, while OGD infrastructures traditionally mainly provide discovery metadata, this study confirms several recent studies that different types of metadata (discovery, contextual and detailed metadata) need to be combined to improve OGD use. Finally, whereas kernel theories concerning coordination, metadata, interaction and data quality are often studied separately, this study reveals that it is essential for the development of OGD infrastructures to combine these four kernel theories.

Keywords: open data, open government data, use, infrastructures, coordination, metadata, interaction, data quality

Anneke Zuiderwijk has been working at Delft University of Technology as a researcher since November 2011. For more information about Anneke's research and teaching activities please visit http://tinyurl.com/AnnekeZuiderwijk.

ISBN: 978-94-6295-351-2